U.PORTO

2º ciclo de estudos
Mestrado em Informática médica

Análise de dados biométricos de jogadores de futebol para previsão de performance.

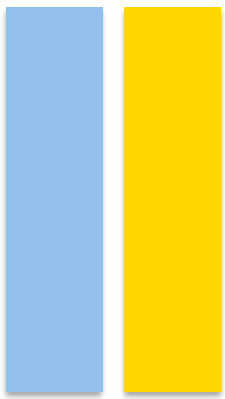
Vanessa Margarida Fonseca Moreira

M

2020

ORIENTADOR: Doutora Inês Dutra

COORIENTADOR: Doutor Pedro Brandão



Agradecimentos

No momento em que se fecha mais um capítulo torna-se difícil agradecer a todas as pessoas que contribuíram, de alguma forma, para a minha vida académica.

Primeiramente quero agradecer aos meus pais, irmãos e avós, que estejam onde estiverem, contribuíram com todo o carinho, apoio, paciência, incentivo e força. Não mediram esforços para que eu chegasse até esta etapa da minha vida. Valeu a pena todo o esforço, todo o sofrimento e todas as renúncias, hoje estamos a colher, juntos, os frutos do nosso empenho. Esta vitória é minha, mas também é vossa!

Dedico especial agradecimento á Professora Doutora Inês Dutra e ao Professor Doutor Pedro Brandão, por toda a paciência, disponibilidade, explicação e apoio na orientação, assim como, toda a ajuda na transmissão dos seus conhecimentos e incentivo que tornaram possível a conclusão desta dissertação.

A todos os professores da Faculdade de Ciências da Universidade do Porto e da Faculdade de Medicina da Universidade de Porto, que foram importantes no meu percurso académico e no desenvolvimento desta dissertação e a todos aqueles que, de alguma forma, deram um pouco de si e dos seus conhecimentos para que a conclusão deste projeto fosse possível.

Sumário

A presente dissertação tem como objetivo a análise de dados biométricos dos jogadores assim como prever a criação de um modelo capaz de prever os atributos dos jogadores e perceber como variam ao longo do tempo.

Desta forma, utilizou-se para análise uma base de dados do *Kaggle* e fez-se a análise descritiva das tabelas e percebeu-se as interações entre elas assim como a correlação entre os diferentes atributos dos jogadores.

Em seguida, gerou-se um novo *dataset* que continha todas as informações relevantes para estudo de evolução ao longo do tempo como as datas de avaliação dos jogadores, o *id* dos jogadores e os atributos em análise que foram analisados na ferramenta *Python.*

Posteriormente os jogadores foram divididos em grupos consoante a posição que ocupam em campo como guarda-redes, defesas, médios e avançados.

Fez-se uma modelagem onde se criou um modelo que analisa-se as características e o comportamento dos jogadores e elaborou-se o pré-processamento dos dados onde se criou um novo *dataset* que continha as datas de avaliação, os jogadores e os atributos mais relevantes de acordo com cada grupo, ou seja, de acordo com cada posição.

Por último, fez-se um *forecasting* para previsão do comportamento do jogador e analisou-se consoante o resultado se o jogador deve ou não ser substituído.

Concluiu-se que existe correlação entre variáveis, observou-se o comportamento dos jogadores e analisou-se os que mais se diferenciavam obtendo resultados de jogadores acidentados tanto positivos como negativos. Na previsão dos jogadores o modelo previu algumas avaliações observadas

Palavras- chave: Futebol, *Machine Learning, Python*, Atributos, VAR e *Forecasting.*

Abstract

The present dissertation has as objective the analysis of biometric data of the players as well as predicting the creation of a model capable of predicting the attributes of the players and understanding how they vary over time.

In this way, a *kaggle* database was used for analysis and the descriptive analysis of the tables was made and the interactions between them were perceived as well as the correlation between the different attributes of the players.

Then, a new dataset was generated that contained all the relevant information for studying the evolution over time, such as the players 'evaluation dates, the players' id and the attributes under analysis that were analyzed in the *Python* tool .

Subsequently the players were divided into groups according to the position they occupy on the field as goalkeepers, defenders, midfielders and forwards.

A modeling was done where a model was created that analyzes the characteristics and behavior of the players and the pre-processing of the data was elaborated, where a new dataset was created that contained the evaluation dates, the players and the players. most relevant attributes according to each group, that is, according to each position.

Finally, a *forecast* model was made to predict the player's behavior and analyzed according to the result whether the player should be replaced or not.

It was concluded that there is a correlation between variables, the behavior of the players was observed and the ones that most differed were analyzed, obtaining results from both positive and negative injured players.

In the prediction of the players the model predicted some observed evaluations.

Keywords: Football, *Machine Learning, Python,* Attributes, VAR and *Forecasting*.

Preâmbulo

Durante o meu percurso académico e através dos conhecimentos adquiridos no primeiro ano do mestrado em informática médica que me auxiliaram no desenvolvimento do projeto surgiu-me o interesse de unir a informática médica ao desporto de alta competição mais concretamente o futebol profissional.

De forma a tornar este projeto fazível entrei em contacto com o Professor Doutor Pedro Brandão onde pudemos analisar quais os projetos possíveis e interessantes para uma abordagem em tema de dissertação. Em seguida, juntamente com a Professora Doutora Inês Dutra decidimos utilizar uma base de dados sobre futebol com o objetivo de implementação de mecanismos de *data mining* e extração de conhecimentos dados e como modo de previsão.

A base de dados utlizada está disponível no Kaggle (https://www.kaggle.com/datasets) referente ao futebol europeu.

De modo a ser possível a realização deste tema foi definido o título da dissertação e a respetiva pergunta de investigação que orientou a formulação dos objetivos.

Considero o tema interessante pelo que, em conjunto com os meus orientadores, decidimos que o tema da minha dissertação seria a análise de dados biométricos de jogadores de futebol para previsão de performance.

Durante a elaboração da presente dissertação tive a possibilidade de adquirir competências em áreas relevantes ao *data mining,* uso das ferramenta *Orange* e *Python* e métodos de *machine learning.*

Com esta dissertação pretendo contribuir de forma significativa na utilização de mecanismos de previsão temporal no futebol profissional de forma a melhorar a avaliação dos jogadores a nível individual e institucional.

Índice

Acrónimos

FIFA – *Fédération Internationale de Futeball Association*

EA– *Electronic Arts*

FUT – FIFA *Ultimate Team*

ML – *Machine Learning*

IA – Inteligência Artificial

VAR – *Verbo Autoregression*

CSV – *Comma-Separated Values*

SVM – *Support Vector Machine*

EPL – *English Premier League*

LASSO – *Least Absolute Selection and Shrinkage Operator*

MARS – *Multivariate Adaptive Regression Splines*

RFE – *Recursive Feature Elimination*

PCA – *Principal Component Analysis*

RNA – Redes Neurais Artificiais

MLP – *Multilayer Perceptron*

MAE – Erro Absoluto Médio

Índice de figura

Índice de tabelas

1 Introdução

O exercício físico propicia a saúde, o bem-estar físico e psicológico e prevenção/melhoria de patologias comuns na sociedade. Todos os desportistas de alta competição apresentam uma elevada predisposição para risco de lesões, nomeadamente no Futebol.

Resultados de estudos feitos demonstram que o treino com pesos (treino de força), para além de aumentar a performance dos jogadores de futebol e de outros atletas profissionais, também diminui de forma significativa o risco de sofrer lesões. Assim, quanto maior a capacidade muscular do atleta, menor será a probabilidade de sofrer lesões durante a época. Para obtenção de melhores resultados é necessário que não haja diferenças corporais a nível de peso, especialmente nos membros inferiores onde existe uma elevada predominância de lesões.

O treino de força quando bem programado e estruturado para cada jogador individualmente, e que não provoque uma sobrecarga de treino excessiva, pode efetivamente proporcionar benefícios na performance do atleta. Assim sendo, torna-se pertinente estudar e compreender as vantagens deste tipo de treino de forma a incorpora-lo nos jogadores de alta competição proporcionando-lhes resultados positivos e mais eficazes. (Integrado et al. 2011; Cunha et al. 2016; Murphy et al. 2003; Ekstrand. 2006).

Os atributos são os grandes responsáveis por determinar a qualidade dos jogadores e conhece-los bem pode ser uma grande vantagem para a equipa na hora de formar o 11 inicial. Os jogadores que sofrem alguma lesão acabam por ter um custo muito elevado para a instituição e a sua ausência pode levar a uma má classificação do clube e/ou a saída antecipada da competição (Woods *et al.*, 2002).

Uma evidência relacionada com o futebol e as lesões é que existe um número desproporcional entre o inicio da época (pré-época) e o fim da época (Woods *et al.*, 2004). Com isto, torna-se importante fazer uma análise biométrica dos jogadores e prever o seu comportamento durante os jogos de forma a perceber se o jogador escolhido seria ou não o indicado para jogar com o intuito de zelar pelo condicionamento físico do jogador que quando forçado e pressionado em jogo pode sofrer alguma lesão ou agravar uma já existente.

Para a implementação de mecanismos de *data mining* e análise de dados biométricos para previsão da performance do jogador usou-se a base de dados de futebol

europeu do *Kaggle*[1] que foi analisada no *Orange* e *Python.*

O presente trabalho tem como objetivo a análise estatística e descritiva para cada equipa e jogadores e prever como é que estas variáveis variam ao longo do tempo ajudando a perceber o que poderia ser alterado anteriormente de forma a obter um melhor resultado no final do jogo e consequentemente no final da época.

Uma das grandes finalidades é mostrar como criar um modelo preditivo que seja capaz de prever o quão bom um jogador de futebol é com base nas suas estatísticas de jogo.

Este trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada uma abordagem teórica onde se abordam conceitos fundamentais do futebol como a sua terminologia, organização em campo, substituições, posição em campo e medidas de desempenho dos jogadores. No Capítulo 3 abordam-se técnicas e conceitos relacionados com o tema em estudo e uma revisão dos trabalhos anteriores. O Capítulo 4, apresenta modelagem dos jogadores como a descrição da base de dados, metodologia utilizada assim como as métricas de avaliação e método de validação que ajudou a alcançar os resultados obtidos e descritos no Capítulo 5. As conclusões finais e perspetivas futuras estão descritas no Capítulo 6.

[1] https://www.kaggle.com/

2 Terminologia e Conceitos fundamentais

Nos últimos anos, temos vivenciado novas descobertas e avanços técnicos e científicos muito importantes para a crescente área do saber, seja a nível social ou desportivo.

A evolução técnica e científica e os grandes desenvolvimentos e potencialidades têm vindo a mudar drasticamente a prática do desporto, nomeadamente o futebol. Este tem vindo a ganhar cada vez mais visibilidade pelo que, a qualidade de treino dos atletas tem influenciado, consequentemente, a performance e desempenho do jogador individual e da equipa.

As grandes mudanças que têm surgido no futebol atualmente passam pela crescente evolução tática, padrões de treinos diferentes, a interdisciplinaridade, cooperatividade e vontade de evoluir.

2.1 **Futebol**

2.1.1 Terminologia do Futebol

O futebol é considerado o desporto mais popular e praticado do mundo que transcende barreiras econômicas, culturais e sociais (Metzl & Micheli, 1998; Shephard, 1999).

Causado pelo seu crescimento exponencial o futebol tornou-se no desporto praticado por centenas de milhões de jogadores de todas as idades com diferentes níveis de especialização e é jogado em mais de 200 países no mundo (Vendite & Moraes, 2006).

O futebol é governado internacionalmente pela *Fédération Internationale de Futeball Association* (FIFA), com sede em Zurique, na Suíça. Entre os torneiros de futebol, o mais conhecido e que mais chama à atenção de todos os espectadores do mundo é O Campeonato do Mundo de Futebol masculino e feminino realizado a cada quatro anos (Matuda & Tagnin, 2014). A competição internacional teve a primeira edição em 1930 e a 21ª em 2018, com exceção da 1942 e 1946, devido à segunda guerra mundial.

Desporto conhecido mundialmente como competição, manifestação cultural e mercado económico e que devido à grande popularidade acarreta, consequentemente, uma grande implicação a nível financeiro, principalmente quando se fala de futebol

profissional.

O futebol é um desporto de natureza complexa, no qual a junção de diferentes técnicas, táticas, psicologias, biomecânica e a fisiologia dos jogadores produz um melhor desempenho uma vez que falamos de um desporto de alta intensidade e que é caracterizado por ações intermitentes. Um jogo de futebol completo exerce um grande esforço físico dos jogadores uma vez que estão sempre em constante movimento e tem a duração de cerca de 90 minutos dividido em duas metades de 45 minutos (Giménez *et al.*, 2020).

2.1.2 Organização em campo

Jogado de forma coletiva onde a individualidade e disparidade de acontecimentos manifestam-se a partir da disputa entre duas equipas adversárias compostas por um máximo de onze jogadores cada e onde nenhum jogo pode começar ou continuar se uma das equipas tiver menos de sete jogadores, os melhores resultados só são possíveis de obter se a equipa jogar cooperativamente.

A organização e formação da equipa tem um único objetivo, marcar golos e evitar que a equipa adversária marque para assim atingir o objetivo final, a vitória (Noda *et al.*, 1998; Santos, 2016).

Os jogadores que formam o onze inicial de cada equipa são chamados de titulares, estes geralmente são os jogadores mais usados pelo clube ao longo da época e escolhidos em primeira instância para o confronto de cada jogo.

O confronto entre duas equipas num jogo de futebol é criado através de regras pré-implementadas pela FIFA (IFAB, 2018) e também pela criatividade do atleta e/ou equipa em campo que é capaz de ganhar oportunidades para obter uma maior vantagem sobre o adversário, driblando-os e fintando-os. É a partir do equilíbrio das regras e da vontade dos jogadores que a organização da equipa em jogo é criada.

A dificuldade está em equilibrar essas regras a fim de se construir uma organização capaz de vencer a organização da equipa adversária (Santos, 2016).

No futebol cada equipa é composta por onze jogadores titulares, um guarda-redes e dez jogadores de campo que o preenchem como defesas, médios e avançados.

A disposição dos atletas em campo depende do esquema tático utilizado pelo coordenador técnico sendo que cada um dos onze jogadores tem uma posição particular em campo durante cada jogo que detalha o papel principal do jogador e a sua área de

atuação (Silva, 2008).

2.1.3 Substituições

Dependendo da competição o número de suplentes difere, estão disponíveis onze suplentes em jogos de seleções e doze suplentes no caso do campeonato do mundo, dado que existe um terceiro guarda-redes. Quando se trata de jogos de equipas existem, geralmente, sete suplentes (Board, 2017; Santos, 2016).

O número de substituições em jogo, até um máximo de cinco, que podem ser usadas em qualquer jogo de uma competição oficial será determinada pela FIFA, pela confederação ou pela federação nacional de futebol. Exceto em competições de futebol masculino e feminino que envolvam as equipas principais dos clubes da mais alta divisão ou seleções nacionais A, em que o máximo é de três substituições (IFAB, 2018).

Um jogador pode ser substituído por motivos de segurança, lesão, infrações, sanções e por opção do treinador do clube (IFAB, 2018).

2.1.4 Posições no Futebol

A natureza fluida do jogo moderno sugere que as posições no futebol representados na Figura 1 não sejam definidas de forma rígida, ou seja, os jogadores mais versáteis podem, inclusive, mudar de posição durante o jogo, no entanto a maioria dos jogadores profissionais vai jogar num número limitado de posições ao longo da sua carreira, uma vez que cada posição em campo exige habilidades e atributos diferentes que diferem de posição para posição (Di Salvo *et al.*, 2007).

Toda a discussão que se segue sobre as posições dos jogadores está baseada na informação contida nos sítios web de Caetano, R. (2020), Menezes, B. (2020) Tampsell, C. (2019).

O guarda-redes é a posição mais defensiva no futebol, é o único que pode tocar na bola com as mãos e agarrá-la, desde que esteja dentro dos limites da área de grande penalidade. A sua grande função em campo é evitar que a equipa adversária marque golo.

Os guarda-redes mantêm-se na área de grande penalidade durante a maior parte do jogo, o que lhes oferece uma maior visão do campo. Conhecidos pelas reações e reflexos rápidos, capacidade de antecipar acontecimentos de forma a mergulhar ou travar

a equipa adversária de marcar golo, têm impulsos e decisões rápidas, e grande agilidade e flexibilidade (Guimarães *et al.*, 2014; Alves, 2017).

Os defesas posicionam-se na linha imediatamente a seguir ao guarda-redes e têm três sub-posições, o defesa central, o defesa esquerdo e o defesa direito.

Os defesas jogam atrás dos médios e permanecem na metade do campo que contém a baliza que estão a defender, a da sua própria equipa. Têm como função dar apoio para a equipa e parar jogadores, especialmente os atacantes levando a bola para fora da área de grande penalidade evitando o golo. Os defesas precisam de ser capazes de interpretar o jogo e ter um grande poder de concentração.

Os defesas centrais ocupam a região da grande área defensiva. São normalmente jogadores altos, fortes e com bom salto, o que lhes permite o cabeceamento de bolas. Costumam ter habilidades como a antecipação, impulsão, velocidade e marcação que ajudam à função do jogador em campo, permitindo o bloqueio e impedindo a aproximação dos adversários da grande área (Guimarães *et al.*, 2014; Alves, 2017).

Os laterais esquerdo e direito são defesas que auxiliam os centrais, cobrem a parte lateral do campo e fazem a comunicação entre os guarda-redes e os médios. Estes são jogadores que apresentam elevada resistência, velocidade e aceleração auxiliando a sua função de apoio ao ataque pelas áreas laterais fazendo cruzamentos e finalizações. Também impedem o avanço dos adversários pelos lados do campo.

Os médios são os jogadores que ocupam o meio do campo, estão entre os defesas e os avançados, e têm como principal função em campo manter a posse de bola, fazer sair a bola da defesa para o ataque e de recuperar a bola quando perdida para a equipa adversária. Eles normalmente iniciam o ataque contra a outra equipa funcionando como distribuidor de jogo e podem também auxiliar a sua equipa com uma linha extra de defesa anulando as jogadas ofensivas quando esta se encontra sob ataque e segurar ou defender bolas paradas. Podem também dividir-se em esquerdo e direito. Apresentam como principais características a sua criatividade, técnica, finta, resistência, velocidade, aceleração, cruzamentos e remates longos e curtos. Também chamados de laterais, estes fazem a ligação entre os defesas laterais e os avançados promovendo o desenvolvimento do jogo em direção ao ataque (Guimarães *et al.*, 2014; Alves, 2017)

Os avançados são os atacantes da equipa, e movimentam-se de forma a abrir e finalizar o jogo com a marcação do golo. As principais capacidades dos atacantes são a finalização, cabeceamento, técnica, velocidade, agilidade e finta da defesa da equipa adversária (Guimarães *et al.*, 2014; Alves, 2017)

Figura 1- Sistema de jogo 4-3-3 (baseado em Teodoro & Veronez, 2013)

2.1.5 Atributos por posição

De forma a calcular e classificar os jogadores em campo usam-se os atributos que permitem simular a precisão (% de sucesso) dos jogadores.

A simulação é a experimentação de um modelo capaz de imitar a realidade possibilitando o trabalho em condições que se assemelhem à realidade. O objetivo de simular acontecimentos facilita a sua verificação e facilmente se corrige falhas e erros ainda durante a simulação criando uma experiência de realidade mais correta e vantajosa.

A percentagem de sucesso é calculada pelo número de sucessos a dividir pelo número de tentativas feitas.

A junção dos atributos formam o jogador dentro de jogo. Quando se forma uma equipa é necessário ter em atenção todos os atributos de maneira a construir uma equipa mais forte e conseguir um bom resultado final do jogo e consequentemente no final da época.

Os atributos são definidos de acordo com a posição do jogador em campo. Por exemplo, um defesa apresenta valores de atributos diferentes dos avançados podendo ser maiores ou menores dependendo do atributo em questão. É nesta fase que temos de ter atenção escolhendo um jogador que apresente valores altos nos atributos mais importantes para a posição que ocupa em campo. Esta escolha leva a equipa a se orientar e ficar o mais equilibrada possível.

Todas estas decisões anteriores ao jogo e à formação da equipa é que podem levá-la à vitória (Caetano, R.,2020; Menezes, B.,2020; Tampsell, C.,2019).

Tabela 1- Atributos dos jogadores de Futebol por posição em campo (Caetano, R.,2020;Menezes, B.,2020; Tampsell, C.,2019).

Guarda-redes	Laterais	Defesas	Médios	Avançados
Mergulho Defesa Reflexo Habilidade Posicionamento Alcance e precisão de reposição de boal em campo	Velocidade Passe alto Habilidade defensiva Força	Cabeceamento Desarme Salto	Desarme Força Resistência Habilidade de distribuição do jogo Passes curtos Drible Controle de bola.	Finalização Equilíbrio Velocidade

A Tabela 1, mostra os atributos mais relevantes dos jogadores de acordo com a posição que ocupam em campo.

Para um guarda-redes são necessários bons números nos atributos de mergulho, defesa, reflexo, habilidade, posicionamento e *kicking* que é usado para medir o alcance e a precisão da reposição da bola em jogo.

Para os laterais é fundamental atributos como a velocidade, passe alto, habilidade defensiva e força uma vez que estes jogadores atacam e defendem (Caetano, R.,2020; Menezes, B.,2020; Tampsell, C.,2019).

Os defesas centrais são a última linha de defesa e precisam de suportar a pressão dos avançados. Por esse motivo, atributos como o cabeceamento e o desarme, são atributos chave para essa posição.

Os médios devem distribuir o jogo e ter bom desempenho no desarme, por isso atributos como força, resistência, habilidade, passes curtos, drible e controle de bola são essenciais ao seu bom desempenho. Os avançados são os que tentam finalizar o jogo escapando ao ataque dos defesas e apresentam atributos como a finalização, equilíbrio e velocidade (Caetano, R.,2020;Menezes, B.,2020; Tampsell, C.,2019).

Os atributos permitem determinar a qualidade dos jogadores e conhecê-los bem pode ser uma grande vantagem para a equipa na hora de formar o onze inicial.

2.1.6 Aumento dos atributos

De forma a melhorar a performance dos jogadores profissionais é necessário ter

em conta os atributos de cada um dependendo da posição que ocupa em campo.

De acordo com a *Electronic Arts* (EA) a FIFA *Chemistry* (química) no FIFA *Ultimate Team* (FUT) é muito importante e as suas interações estão expressas na Figura 2.

EA explica como é que os atributos podem ser afetados e como é calculado o FUT.

A classificação individual dos jogadores na FUT pode ser melhorada pela química. A classificação global de *Chemistry* tem dois tipos de *Chemistry* em FUT, e existem também os *Chemistry Styles* que influencia o seu funcionamento

A junção perfeita de *Individual Chemistry*, *Team Chemistry* e *Chemistry Styles* ajuda a que os atributos sejam aumentados por cerca de 10 pontos cada, perfazendo um total de 90 pontos de atributos por jogador.

A grande questão é sobre *Chemistry,* como o que é que este afeta e como é que o *Individual Chemistry*, *Team Chemistry* e *Chemistry Styles* podem ser combinados de forma a que o aumento de FUT seja notável.

Os atributos podem ser aumentados de forma individual (*Individual Chemistry*) com uma classificação de 10 pontos por jogador, por equipa (*Team Chemistry*) onde o valor pode atingir os 100 pontos e o aumento do *Overall Chemistry* que é a junção dos dois anteriores. O facto de os atributos dos jogadores subirem e descerem são definidos pelo valor escondido de *Overall Chemistry,* um valor alto irá aumentar os atributos dos jogadores e um valor baixo irá causar o oposto. O aumento ou diminuição dos atributos chave pode alterar positiva ou negativamente a equipa.

De forma a entender melhor analisou-se os cálculos de *Chemistry* no FIFA *Ultimate Team* (FUT). A *Chemistry* é aplicada a cada jogador no início de cada jogo e é determinada pela *Individual* e *Team Chemistry.*

Os atributos que a *Chemistry* altera são definidos pelo Estilo de Química (*Chemistry Style*) e são uma forma de melhorar as estatísticas dos atributos de forma individual.

Quando se forma uma equipa pretendemos que um jogador apresente valores altos de *Individual* e *Team Chemistry.* No entanto quando o atleta entra em campo os seus valores podem subir ou descer. Contudo, o aumento do *Individual Chemistry* e *Team Chemistry* não são afetados de forma igual.

O aumento de *Individual Chemistry* do:

- Jogador do onze inicial diz respeito à cerca de 75% do aumento dos atributos de forma individual e os restantes 25% fazem parte do aumento de *Team Chemistry.*

- Jogador que entra como suplente têm uma de *Individual Chemistry* de jogador estática de 5. O que significa que o a sua *Individual Chemistry* é de 25% e de *Team Chemistry* de 75% *Individual Chemistry* de jogador estática de 5.

Quer dizer que os jogadores suplentes apresentam uma maior vantagem em relação à equipa com Química alta, mas que a sua Química individual de jogador não.

Considera-se mais importante que um jogador individual tenha um aumento de 10 pontos em *Individual Chemistry* do que o facto da equipa *Team Chemistry* possuir um total de 100 pontos. O ideal seria o aumento do valor de *Overall Chemistry* que é o que tem mais influência no valor dos atributos uma vez que abrange tanto o jogador individual como a equipa.

A Química nos atributos do jogador é calculada pela seguinte fórmula de acordo com as informações do sítio da web da *Electronic Arts do* FIFA 21 reveladas a 28 de Setembro de 2020 e atualizadas a 9 de outubro de 2020.

(*Team Chemistry* x 0,25) + ((*Individual Chemistry* x 10) x 0,75)

- Se o valor de *Overall Chemistry* for superior a 50 os atributos que sejam afetados pelo seu *Chemistry Styles* irão aumentar para o seu valor máximo.

- Se a melhoria do valor aumentasse os atributos acima de 99, então a sua melhoria tem um limite de valor máximo de 99.

- Se for igual a 50 o valor dos atributos não sofre alterações.

- Se for igual ou inferior a 49 vão diminuir até ao valor mínimo de 1.

De forma a compreender melhor o *Overall Chemistry* vamos mostrar o cálculo em dois exemplos distintos.

Exemplo 1: Para um jogador que faz parte do onze inicial com 100 de *Team*

Chemistry e 10 de *Individual Chemistry.*

(*Team Chemistry* x 0,25) + ((*Individual Chemistry* x 10) x0,75) =
= (100 x 0,25) + ((10 x 10) x 0,75) =
= 25 + 75 =
= 100

Posteriormente faz-se a fórmula de *Individual Chemistry* e calcula-se, o resultado desta equação refere-se às alterações dos atributos individuais.

100 – 50 = 50
50/50 = 1
1 x (valor máximo de melhoria) = alteração do atributo

Exemplo 2- Para um jogador suplente com 100 de *Team Chemistry* e 5 de *Individual Chemistry* estática.

(*Team Chemistry* x 0,25) + ((*Individual Chemistry* x 10) x0,75) =
= (100 x 0,25) + ((5 x 10) x 0,75) =
= 25 + 37,5 =
= 62,5

Fórmula de *Individual Chemistry* e calcula-se.

62,5 – 50 = 12,5
12,5/50 = 0,25
0,25 x (valor máximo de melhoria) = alteração do atributo

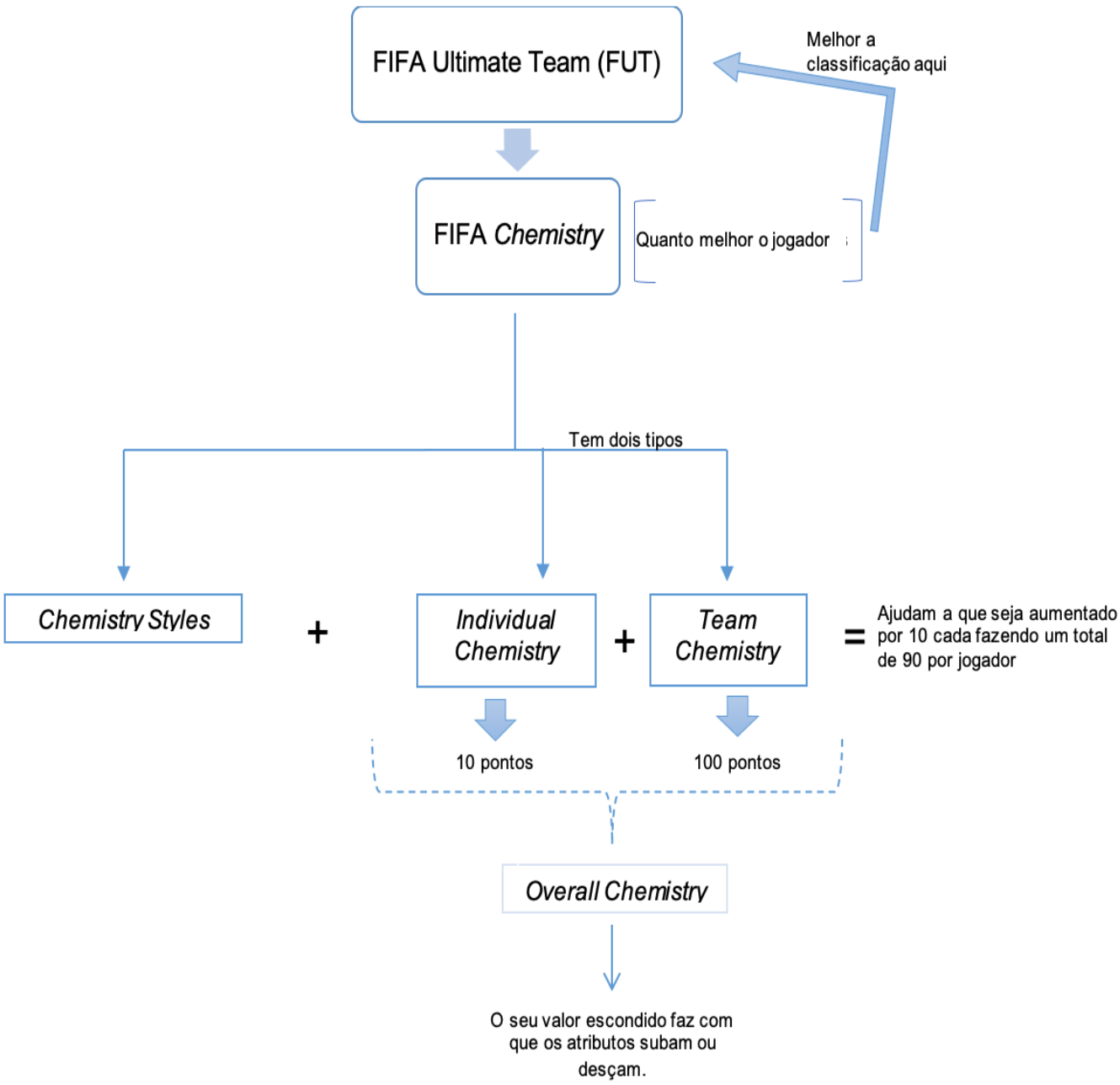


Figura 2- Resumo das interações da FIFA Ultimate Team

2.2 Métodos estatísticos

A evolução tecnológica tem vindo ano após ano a alterar hábitos do nosso dia a dia, e também atingiu o mundo do desporto. O futebol é um dos desportos mais populares do mundo. Assim sendo, a previsão de resultados e perceção do jogo tem ganho um grande interesse por parte do público e treinadores (Mehrez & Hu, 1995; Owramipur, Eskandarian & Mozneb, 2013).

No entanto, existem inúmeros fatores que influenciam os resultados como a aptidão física (atributos), lesões, fadiga, golos e cartões que todos juntos alteram a previsão do resultado final de um jogo (Min *et al.*, 2007). A análise e previsão do perfil dos jogadores também é um dos grandes fatores de interesse para atingir o sucesso de uma equipa e clube.

Com isto, surgem cada vez mais ferramentas técnicas e métricas capazes de recolher e armazenar grandes quantidades de dados que ajudam nas tomadas de decisão dentro e fora de campo. Estas informações são bastante relevantes para as equipas de futebol profissional uma vez que ajudam na obtenção de uma melhor performance física do atleta e uma melhor prestação no jogo levando a equipa a vencer.

Existem diversos métodos estatísticos desenvolvidos para aperfeiçoar a rapidez e sofisticação dos sistemas de recolha de dados. No entanto ainda não se avançou de igual forma na correta interpretação da informação que têm como objetivo o auxílio aos treinadores e *staff* na melhoria e aperfeiçoamento dos métodos de treino que em parte tem relação com as limitações na análise estatística de dados.

Os dados por si só não têm muito valor se não existirem mecanismos que consigam recolher a essa informação. Com isto, surge a abordagem de problemas de apoio à decisão que envolve a estatística e técnicas para a extração do conhecimentos de grande volume de dados (Janosik, 2005). Estas abordagens passam pelas análises univariada, bivariada e multivariada e pela modelagem dos dados procurando encontrar padrões que expliquem o comportamento de cada grupo de jogadores e a previsão de desempenho.

A análise univariada é um ramo da estatística que utiliza diversos métodos para descrever conjuntos de dados. Esta inclui todos os métodos de estatística descritiva organizando-os e sumariando-os e permite a análise e descrição de cada variável individualmente. A análise univariada geralmente usa medidas de tendência central como a média, mediana, moda e medidas de dispersão como o desvio de padrão, valor mínimo e máximo, variância e obliquidade (Babbie, 2010).

A análise bivariada permite observar o comportamento de duas variáveis na presença uma da outra (Babbie, 2010). Este tipo de análise pode ser útil para testar

hipóteses de associação, correlação e regressão linear simples.

O coeficiente de correlação é uma medida numérica de análise bivariada, e caracteriza-se pela semelhança ou relação entre duas variáveis. Estas variáveis são a interação entre duas hipóteses de um determinado conjunto de dados, normalmente chamado de amostra.

O valor do coeficiente de correlação assume valores entre -1 e 1 e o resultado indica se a correlação é negativa ou positiva, ou seja, -1 significa que a correlação entre as variáveis em estudo é negativa perfeita e 1 é positiva perfeita. Quanto mais o coeficiente de correlação se aproxima de ± 1 mais forte é a correlação entre as variáveis. Se o valor for positivo as variáveis correlacionam entre si, ou seja, quando o valor de uma variável aumenta, o valor da outra também aumenta. Se o valor for negativo as variáveis também se relacionam no entanto de forma inversa, ou seja, quando uma variável aumenta a outra diminui (Mukaka, 2012).

Se o coeficiente de correlação apresentar o valor de 0, indica que as variáveis não se correlacionam, isto é, não dependem uma da outra (Taylor, 1939; Filho & Júnior, 2009). Contudo, os coeficientes de correlação têm erros devido à distorção por *outliers* e quando relacionados e usados de forma incorreta entre as diferentes variáveis o que pode criar interpretações erradas (Boddy & Smith, 2009).

Os métodos multivariados são escolhidos consoante o objetivo do estudo. A análise multivariada é uma análise exploratória de dados que estuda o comportamento de três ou mais variáveis aleatórias e inter-relacionadas simultaneamente. Visto que com um grande número de variáveis torna-se difícil perceber a relação entre os diferentes grupos de variáveis (Vicini, 2005).

Utilizados de forma a simplificar os modelos e métodos estatísticos esta análise encontra as variáveis que sejam redundantes e que tenham correlação alta e as que são complementares. Elimina-a as redundantes e coloca só as complementares e neste caso entram também os métodos de seleção de atributos (features selections).

A aprendizagem de máquina engloba muitas estatísticas multivariadas uma vez que muitas das técnicas mais usadas na análise multivariada como a ordenação, *clustering*, etc., usam algoritmos de aprendizagem não supervisionada.

Para além disto, existem técnicas de aprendizagem supervisionada em *machine learning* fora do domínio de análise multivariada regular por exemplo, se o usuário pretender saber em quais categorias um novo objeto iria estar com base em alguns valores das suas variáveis então é possível que se treine o algoritmo para um conjunto de objetos que já se conhece à priori a classificação e definir o algoritmo que vai ser capaz de classificar o novo objeto. Nitidamente, esta não é uma técnica estatística multivariada e

quando se pensa em aprendizagem de máquina tem que se ter isso em conta uma vez que esta envolve o processo de comunicar o sucesso ou fracasso de uma pesquisa ao sistema. Então é nesta altura que a aprendizagem de máquina se inicia sobrepondo-se á Inteligência Artificial (IA) (Reis, 2001).

Alguns dos métodos de aprendizagem automática mais utilizados são as árvores de decisão, redes neurais, *clustering*, regressão, series temporais, *redes Bayesianas,* Modelo Autorregressivo Vetorial (VAR), e *Multiple Mutual Information, Principal Component Analysis* (PCA), etc.

Na aprendizagem de máquina automática pretende-se verificar se existem padrões que correlacionam as variáveis e aprender esses padrões, aprender algo sobre os dados. Em suma, a aprendizagem de máquina destaca a previsão enquanto que a estatística em geral preocupa-se com a inferência (Härdle & Simar, 2003)

A **Error! Reference source not found.** mostra o exemplo de uma árvore de decisão e de uma rede neural.

As árvores de decisão funcionam como um fluxograma, onde é possível tomar decisões a partir de inúmeras possibilidades de escolha. O nó representa dados ou problemas e cada ramificação possui a solução desse problema.

As redes neurais são nós simples, “neurónios” ou “processadores” interligados de forma a criar uma rede de nós. Este tipo de algoritmo procura solucionar problemas através da simulação do comportamento e das funções de um neurónio.

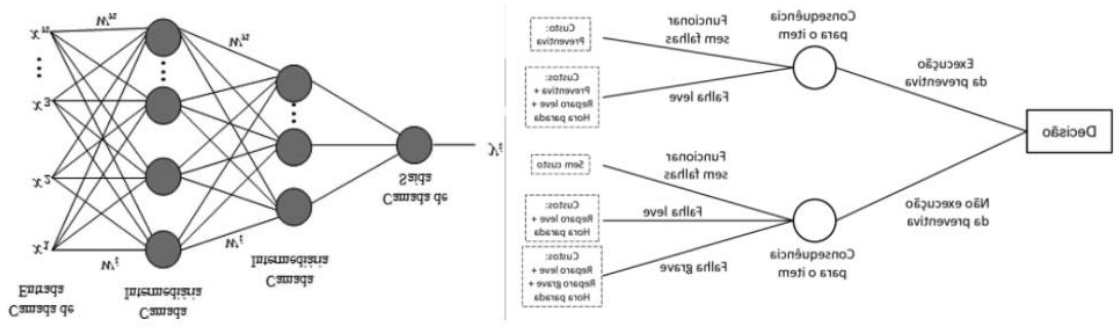


Figura 3- Exemplo de rede neural e árvore de decisão (Picanço et al., 2015; Coutinho, Silva & Delgado, 2016)

O *clustering* ou análise de agrupamento de dados é um modelo que pretende agrupar de forma automática os dados segundo o seu grau de semelhança. O critério de semelhança estabelece a definição do problema e os dados são agrupados conforme a questão, ou seja, se aborda o mesmo problema ou não.

A análise de regressão é um método estatístico que utiliza de duas a mais variáveis quantitativas e é utilizada para previsão de dados numéricos em falta. A regressão é um método estatístico que avalia a tendência de uma distribuição com base nos dados disponíveis. São exemplos a regressão linear, regressão polinomial e simples.

Os modelos de classificação prevêem classes categóricas e os de regressão são capazes de prever valores contínuos.

A análise de séries temporais é das estatísticas mais antigas. Uma série temporal é a sequência de observações/números de uma variável recolhidos em intervalos regulares ao longo do tempo.

Antes de ser aplicado qualquer um destes modelos é necessário fazer uma análise de relevância percecionando se os nossos atributos são mais pertinentes numa análise de regressão ou classificação.

Esta última análise é também aplicada em diversas ferramentas como *Python, Scipy, Matlab,* etc.

Uma das áreas de pesquisa de séries temporais, que tem recebido um especial interesse pelos pesquisadores é a de modelos multiequacionais.

Modelo autorregressivo vetorial (VAR), do inglês *vector autoregression*, é um modelo estatístico normalmente usado em economia e ciências naturais. É um processo aleatório usado para analisar as inter-relações e interdependências entre múltiplas séries temporais a partir de um grupo mínimo de restrições de identificação, ou seja, restrições que permitam reconhecer o componente exógeno das variáveis, este componente exógeno refere-se a uma variável que é determinada fora do modelo e diz respeito às entradas de um modelo, são fixadas no momento em que são introduzidas no modelo (Lu, 2001).

De acordo com a literatura, os modelos VAR foram primeiramente adotados por *Sims* em 1980 com o artigo *Macroeconomics and Reality*. Neste trabalho, o autor analisa modelos de autorregressão já existentes e faz críticas propondo o modelo VAR como alternativa. Segundo *Sims (1986)*, nos modelos anteriores geralmente a informação relevante era perdida devido às poucas restrições impostas, à priori, aos modelos econométricos, o que com o modelo VAR proposto estes requisitos seriam corrigidos. A partir desse momento, estes modelos, enquanto modelos de séries temporais, têm sido utilizados em grande escala na descrição das características estocásticas de series económicas e na realização de previsões (Sims, 1980).

VAR é um modelo de variáveis aleatórias. Estes generalizam o modelo autorregressivo de variável única (univariada), possibilitando séries temporais multivariadas. Deste modo, como o modelo autorregressivo, cada variável possui uma equação que explica a sua evolução ao longo do tempo. Esta equação contém os próprios

valores defasados (passados) da variável, os valores defasados das outras variáveis no modelo e um termo de erro estatístico. Este modelo, não procura grande conhecimento sobre as forças que influenciam uma variável ao contrário dos modelos estruturais com equações simultâneas. Posto isto, pode-se afirmar que os modelos VAR utilizam as vantagens da análise multivariada, atravessando as limitações de modelos univariados. O único conhecimento prévio necessário é uma lista de variáveis que podem supostamente, afetarem-se umas às outras ao longo do tempo.

Apesar das suas limitações, este modelo faz com que a primeira variável não seja afetada por nenhuma das demais variáveis, a segunda seja afetada pela primeira, a terceira pelas duas primeiras , e assim sucessivamente. Posto isto, é possível identificar as interrelações das variáveis.

O modelo VAR representa a evolução de um grupo de k variáveis, denominadas de variáveis endógenas, ao longo do tempo e são modeladas como uma função linear de acordo com os valores anteriores. O período de tempo é numerado $t = 1, \dots, T$.

As variáveis são identificadas por um vetor, y_t com um comprimento k. Este valor pode também ser descrito como uma matriz $k \times 1$. Os componentes do vetor são descritos como $y_{i,t}$, ou seja, a observação no tempo t, com a i que faz referência á variável.

Este modelo é definido pela sua ordem, que se refere ao número de períodos anteriores, ou seja, o VAR de 5ª ordem modela a cada unidade de tempo uma determinada variável com uma combinação linear dos últimos cinco períodos de tempo.

Deste modo, este modelo aborda todas as relações lineares existentes entre as variáveis endógenas e os seus valores passados, permitindo a inclusão de variáveis exógenas (Ramos, 2011).

Assim, o processo VAR() está definido por:

$$y_t = \varphi + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t$$

Onde:

- y_t, é um vetor coluna com k variáveis;
- $A_i, i = 1, \dots p$, são matrizes de coeficientes do tipo $k \times k$;
- φ é um vetor de constantes de dimensão $k \times 1$
- ε_t é um processo de ruído branco vetorial, $\varepsilon_t \sim N(0, \Omega)$, com dimensão k;
- Ω é a matriz de variáveis e covariâncias (definida positiva) (Pascual, 2003)

Para uma compreensão melhor da equação, apresentamos o modelo VAR simples ou também designado VAR(1)

$$y_t = \varphi + A_1 y_{t-1} + \varepsilon_t$$

O que equivale a escrever:

$$\begin{bmatrix} y_{1t} & \varphi_1 \\ y_{2t} & \varphi_2 \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

A estabilidade dos modelos VAR dependem da equação e solução

$$\left|\lambda^p \mathrm{I} - \lambda^{p-1} A_1 - \cdots - \lambda A_{p-1} - A_p\right| = 0$$

O que significa que se todas as raízes estiverem no interior do círculo unitário existe estabilidade, isto é, se $|\lambda| < 1$ o modelo á estável, ou seja, existe uma solução e se $|\lambda| = 1$, o modelo é instável (Jiang *et al.*, 2016)

2.3 **Machine Learning**

A ciência de dados tem vindo a ganhar cada vez mais atenção na sociedade atual e apresenta um mercado socioeconómico muito alto a nível competitivo. A recolha de dados está cada vez mais presente nos dias de hoje (Han, Kamber & Pei, 2014). O elevado crescimento dos altos volumes de dados gera um grande interesse no desenvolvimento de mais ferramentas capazes de suportar grande informação uma vez que se torna humanamente impossível extrair informação dos dados sem uma maior automação do processo. De forma a solucionar esta dificuldade nasce o *data mining* (mineração dos dados em português). Este aborda problemas de apoio à decisão que envolve a estatística e técnicas para a extração do conhecimentos de grande volume de dados (Janosik, 2005).

Data mining como representado na **Error! Reference source not found.** é- como

um processo analítico projetado para explorar grandes quantidades de dados, procurando padrões consistentes, como sequências temporais e regras de associação, de forma a identificar e relacionar variáveis com o objetivo de validá-las aplicando os padrões detetados a novos subconjuntos de dados. Este processo apresenta três etapas: exploração, construção de modelo ou definição do padrão e a validação/verificação.

A premissa do *Data mining* é uma argumentação ativa, ou seja, existe uma pesquisa automática dos dados e métodos. *Data Mining* descobre problemas e oportunidades escondidas na relação entre os dados e analisa e procura anomalias e possíveis relacionamentos, identificando problemas não detetáveis pelo usuário e com a mínima intervenção deste.

Data mining provem de três linhagens: a primeira e mais antiga é a estatística clássica e análise de séries temporais. A estatística é a base da maioria das tecnologias e sem ela não seria possível existir *Data mining*.

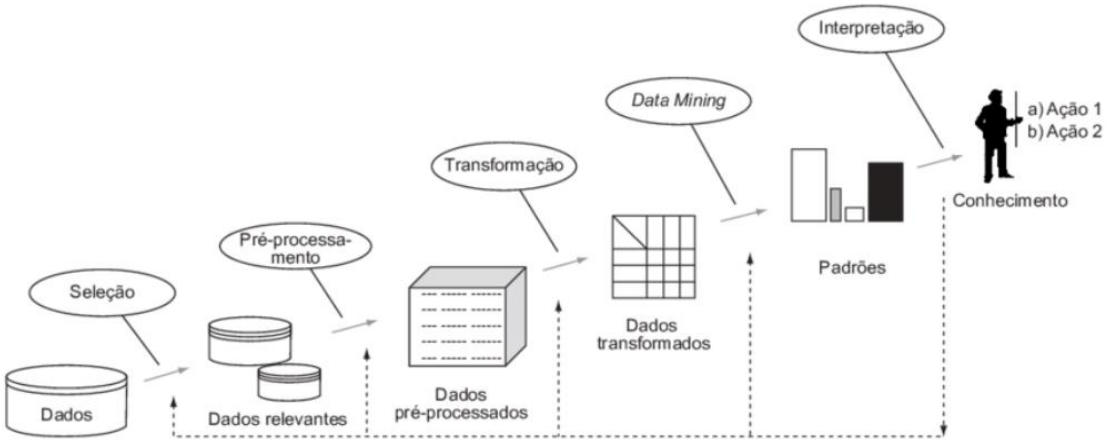


Figura 4- Etapas de Processo de Data Mining (Steiner et al., 2006)

Utilizando a Inteligência Artificial (IA) é possível desenvolver algoritmos computacionais utilizados no *Data Mining.* Estes métodos destinam-se à automação de atividades associadas com o pensamento humano, com a tomada de decisão, aprendizagem, resolução de problemas. Esses algoritmos procuram encontrar soluções humanamente incapazes de serem decifradas devido à limitação de análise de informação.

A maior parte das vezes na técnica de *data mining* não se sabe o que se quer encontrar (aprendizagem não supervisionada), a inteligência do algoritmo deve identificar nos dados relações novas e úteis.

A Inteligência Artificial vem como segunda linhagem, construída a partir de fundamentos de heurística e em oposição à estatística, esta procura prever o pensamento

do homem em relação a problemas estatísticos (Demšar & Zupan, 2013). A IA é uma tecnologia que faz com que o computador pense como algo próximo do raciocínio humano quando executa tarefas. Tal como o ser humano a inteligência artificial analisa dados, encontra padrões e determina conclusões para posteriormente tomar uma decisão.

A terceira linhagem é chamada de aprendizagem automática representada na Figura 5 (*machine learning* em inglês), que é a junção das duas linhagens anteriores. *Machine learning* que pode ser traduzido por aprendizagem de máquina é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. Um dos grandes objetivos do *machine learning é* automatizar a aprendizagem de forma a reconhecer padrões complexos e tomar decisões inteligentes baseadas em dados.

O *Data mining* atinge diversas ramificações importantes. Todas as tecnologias apresentam as suas vantagens e desvantagens uma vez que ainda não é possível responder a todas as necessidades em todas as aplicações (Demšar *et al.*, 2004).

A aprendizagem supervisionada é capaz de tomar decisões precisas quando nos é dado um conjunto de dados cuja resposta de destino já conhecemos, são chamados dados anotados. Os problemas de aprendizagem supervisionados são considerados problemas de regressão e classificação (Gama, 2004). Geralmente os problemas de regressão são quantitativos e os de classificação são qualitativos, ainda que não seja uma regra (James *et al*., 2013).

Num problema de regressão tentamos prever os resultados numa saída contínua, o que significa que estamos a tentar mapear variáveis de entrada para funções contínuas. E num problema de classificação tentamos prever os resultados de saída, o que por outras palavras quer dizer que estamos a tentar mapear variáveis de entrada em grupos diferentes. São utilizados inúmeros algoritmos para criar aprendizagem supervisionada, como as redes neurais, máquinas de vetor de suporte (SVMs) e classificadores *naive bayes* (James, G.*et al.*, 2013)

A aprendizagem não supervisionada, em contrapartida, não possui nenhum tipo de anotação, o objetivo desta aprendizagem é analisar e detetar similaridades e anomalia entre as variáveis e permite tratar problemas com pouca ideia dos resultados. Pode-se derivar a estrutura dos dados e agrupá-los com base em relações entre variáveis. Esta pode também ser usada para diminuir o número de dimensões num conjunto de dados de modo a concentrar apenas os atributos mais úteis, ou para detetar padrões e relações.

Na aprendizagem não supervisionada não existe feedback com base nos resultados de previsão.

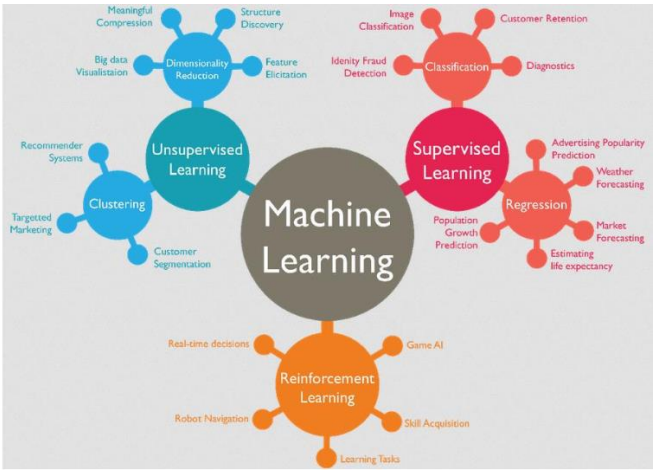


Figura 5- Conceitos básicos de Machine Learning (Clavera, W., 2019)

O presente capítulo da dissertação aborda a terminologia e os conceitos fundamentais do futebol como a organização em campo e as substituições. Estas apresenta elevada importância na hora de construir a melhor equipa e mantê-la forte e com elevada resistência ao longo dos 90 minutos de jogo de forma a enfrentar o adversário e ganhar vantagem obtendo a vitória no jogo.

As posições em campo e os atributos foram analisados e tentou-se perceber a melhor forma de aumento desses atributos dependendo da posição que os jogadores ocupam em campo.

Abordou-se os conceitos básicos de *machine learning* e descreveu-se os métodos estatísticos mais utilizados no futebol como a análise univariada, bivariada e multivariada. O modelo VAR também foi descrito uma vez que vai ser usado na presente dissertação para análise da base de dados em estudo.

3 Estado da arte

Procura-se neste capítulo da dissertação, apresentar, comparar e analisar um grupo de trabalhos onde se utilizaram mecanismos de *data mining* e *machine learning* no futebol profissional.

Para início da dissertação foi realizada uma pesquisa bibliográfica com o intuito de contextualizar o tema de estudo e perceber o que já foi feito. De seguida elaborou-se uma análise detalhada onde a pesquisa foi mais direcionada para durante o jogo de forma a analisar os perfis dos jogadores e a previsão do comportamento destes no futuro.

Para análise dos artigos resultantes das consultas pretende-se uma pesquisa com resultados reais e discussões que forneçam novos resultados.

A pesquisa inicia-se por:

- Observar o título do artigo e se este é relevante para a elaboração do trabalho ou se fornece informações relevantes para o mesmo
- Iniciou-se a leitura do resumo dos artigos de forma a que se compreenda mais sobre o tema em si e sobre os pontos em que o estudo se foca.
- Deu-se especial atenção ao ano em que o artigo foi publicado e o ano em que o estudo do artigo foi elaborado.
- Leu-se atentamente os artigos que ajudam na elaboração deste projeto e retirou-se observações úteis e resultados que pudessem ser proveitosos e benéficos para a elaboração do presente projeto.

Devido à grande evolução do futebol profissional e das suas necessidades torna-se fundamental o desenvolvimento de pesquisas académicas e publicações de forma a compreender e melhorar o perfil dos atletas (Settani Giglio & Spaggiari, 2010; Soares, 2016) e perceber quando é que um treinador vai trocar um jogador dependendo da avaliação e evolução ao longo do tempo.

Segundo Tojo (2015), nos últimos anos o futebol destacou-se pela qualidade de treino que se fez notar na performance individual dos jogadores e enquanto equipa.

O treino de força quando bem programado e estruturado para cada jogador individualmente, e que não provoque uma sobrecarga de treino excessiva, pode efetivamente proporcionar benefícios na performance do atleta, melhorando os seus atributos.

Assim sendo, torna-se pertinente estudar e compreender as vantagens deste tipo de treino de forma a incorporá-lo nos jogadores de alta competição proporcionando-lhes resultados positivos e mais eficazes (Cunha, 2016; Murphy *et al.*, 2003; Ekstrand *et al.*, 2011).

Ribeiro (2009), refere que o "rendimento desportivo pode ser definido como uma manifestação de um determinado desempenho num fenómeno que é o desporto, numa constante tentativa de superação" (Ribeiro, 2009). Neste cenário e no futebol, Braz et al. (2009) afirmam que a versatilidade de movimentos necessária durante o jogo faz com que o atleta se tente superar desenvolvendo as suas habilidades entrando numa competição sempre no seu estado ótimo (Braz *et al.*, 2009)

De forma a melhorar os atributos é fundamental desenvolver uma rotina de exercícios aeróbicos e treino específico e personalizado para os atletas de maneira a aumentar a resistência e as habilidades dos jogadores por um longo período de tempo.

Com o intuito de e consequentemente, apresentar melhores resultados no jogos e assim ganhar mais hipóteses de titularidade e melhores resultados tanto individuais como em equipa no final de cada época (Giménez *et al.*, 2020). Um dos grandes objetivos é a previsão das variáveis que variam ao longo do tempo.

Cientistas e treinadores de futebol ganharam grande interesse em técnicas de *data mining* em especial, *machine learning* que promete beneficiar a previsão da performance do jogador, melhoria do perfil do jogador e resultados dos jogos (Giménez *et al.*, 2020).

Existirem diversos fatores que influenciam o comportamento do atleta em jogo, e o seu perfil irá alterar de acordo com as adversidades que enfrenta, como o local da partida, a qualidade dos jogadores da sua equipa e da equipa adversária, o esquema tático, entre outros. Pelo que é considerado um problema muito interessante para a aplicação de algoritmos de aprendizagem de máquina.

De acordo com Prasetio e Harlili (2016), a previsão de jogos de futebol chama a atenção de adeptos, especialistas e olheiros que procuram adivinhar o resultado antes mesmo do jogo começar (Passos & Kronbauer, 2016; Prasetio & Harlili, 2016).

Este tipo de previsão é feito através de cálculos e variáveis como a equipa ou jogador, aptidão física, lesões, fadiga, cartões e golos que torna essa tarefa difícil, no entanto, interessante, uma vez que o importância de obter um melhor aproveitamento das previsões ganhou bastante relevância. (Passos and Kronbauer, 2016).

Com isto, áreas como a informática e engenharia vêm ajudar neste tema, como a *machine learning*, que consiste em produzir conhecimento a partir de comportamentos, análises de dados, promovendo a extração do máximo de informação para ser utilizada em algoritmos.

O sucesso de uma equipa de futebol passa pela seleção dos jogadores talentosos individualmente e pela formação de uma equipa competente. Esta seleção é feita através das informações disponíveis e das habilidades do atleta.

Na atualidade tem havido uma grande evolução nos trabalhos que abordam as técnicas de previsão, no entanto as diferenças de percentagem são notórias uma vez que cada estudo analisa variáveis de bases de dados diferentes.

Razali et al. (2017) utilizou como método analítico as *redes Bayesianas* para previsão do resultado de jogos de futebol nas épocas 2010/2011, 2011/2012, 2012/2013.

Houve um total de 380 jogos com 20 equipas, em que cada equipa joga entre si duas vezes por ano, uma vez em casa e outra vez fora, como visitante.

Cada jogo foi analisado de forma individual uma vez que cada um apresentou valores diferentes para diferentes fatores que variam com o decorrer do jogo e posteriormente compararam-se os resultados entre as épocas.

Os principais atributos usados para análise de previsão de resultados são os “tiros”, “tiros no alvo”, cantos, faltas cometidas, cartões amarelos e vermelhos, golos da equipa a meio tempo e no fim.

A média de percentagem de previsão de jogos nos três anos foi de 75,09%, valor que os autores consideram bom comparativamente a um estudo realizado por (Joseph, Fenton & Neil, 2006) que era de 59,21%.

A precisão de previsão é de 75,26% para a época 2010/2011, 79,47% para 2011/2012 e 70,53% para 2012/2013 (Razali *et al.*, 2017)

Em outro estudo, realizado por Schneider procura encontrar alternativas para a previsão do resultado de jogos de futebol utilizando métodos de *machine learning*. Este estudo apresenta como objetivo a realização de uma avaliação de desempenho de algoritmos de classificação usados para prever resultados (Schneider, 2018). Lasek; Szlávik; Bhulai (2013) afirmam que é importante que exista uma classificação dos atletas como forma de mostrar a qualidade de cada um de acordo com a performance individual no jogo (Lasek, Szlávik & Bhulai, 2013). Schneider utiliza jogos realizados entre 2000 a 2017 e totalizando 6460 jogos. Para cada jogo disputado, os dados mostram detalhes das equipas como a equipa visitada e a equipa visitante, os resultados dos jogos, o local onde se realizou, o número total de chutes ao alvo e em geral, total de cartões, total de cantos e total e faltas cometidas.

Os dados foram implementados em *Python* e utilizou-se bibliotecas como *Pandas* (para leitura dos ficheiros CSV), *NumPy* (cálculos em geral), *Matplotlib* (gráficos) e *Scikit-learn* (aprendizagem de máquina).

Para avaliação e compararam-se em 21 atributos como o ELO do time visitado e o

do visitante, probabilidade de vitória e empate, a força ofensiva e defensiva, mádia de golos marcados e sofridos, etc.

Optaram por utilizar o sistema ELO é um dos mais usados no futebol como forma de classificação, este sistema é normalmente usado no xadrez, no entanto também é usado para classificação de indivíduos e equipas no futebol. Neste sistema foi criado pelo físico Arpad Emrick Elo, a pontuação de cada jogador é atualizada após cada jogo o que significa que a pontuação atualizada é a atualização da pontuação antiga baseada no resultado e na expetativa antes do jogo, funcionando assim como uma previsão, resultado esperado (Schneider, 2018).

Deste modo, analisaram-se o desempenho dos métodos utilizando validação cruzada e observaram que a exatidão máxima e obteve-se valores maiores de regressão logística de 54,73% $\pm$ 2,06%, a análise discriminante linear de 54,82% $\pm$ 2,01%, *SVM (Support Vector Machine)* com *kernel* linear de 54,35% $\pm$ 2,12% e k-vizinhos 54,34% $\pm$ 1,93%.

No entanto, o classificador *ensemble* constituído por seis classificadores selecionados demonstrou 57% de exatidão na análise de um conjunto de dados de teste.

Em comparação com as restantes *baselines* a classe Vitória para a equipa da casa (mandante) é de 46,56% de exatidão e a segunda *baselines* é assente apenas no sistema ELO da equipa visitada era maior do que o ELO da equipa visitante e, comparando todos os jogos considerou-se que a equipa com maior ELO venceu com 50,24% de exatidão.

Relativamente aos empates, não houve informação suficiente para previsão de instâncias. O melhor classificador foi a Análise Discriminante Quadrática, no entanto dentre os classificadores este foi o que obteve uma menor exatidão com valor de 42,63%$\pm$2,33%.

O estudo concluiu que o fato de o desempenho da exatidão dos classificadores aumentar ao prever resultados de jogos de futebol está diretamente relacionado com a dificuldade de solução de problemas de previsão de empates.

Nabinger realizou estudos de previsão de resultados de jogos do campeonato inglês EPL (English Premier League), aplicando algoritmos de inteligência artificial e *machine learning* (Nabinger, 2018). Ulmer e Hernandez (2014) citam no seu trabalho o estudo realizado por Joseph et al., que utiliza *redes Bayesianas* como previsão. Este artigo tem como objetivo gerar um modelo a partir de uma amostra pequena e várias variáveis de modo a perceber quais das variáveis em estudo são mais ou menos relevantes de maneira a descobrir o vencedor da partida antes desta começar. As variáveis menos relevantes são excluídas pondo o coeficiente a zero, através do encolhimento (*shrinkage*). Os métodos utilizados para encolher os coeficientes de regressão são a regressão Ridge e o LASSO (Least Absolute Selection and Shrinkage Operator) (James et al.;2013).

Os modelos que foram considerados neste artigo são a árvore de decisão onde são usadas várias árvores de maneira a que haja uma maior precisão através de métodos de conjunto como o *bagging* (ensacamento), *random forest* (floresta aleatória) e *boosting, LASSO* e MARS (Multivariate Adaptive Regression Splines).

Deste modo, o objetivo passa por detetar, através de métodos de aprendizagem supervisionada, regressão e classificação quais as variáveis mais importantes para a obtenção de um resultado positivo.

Relativamente à execução de cada um dos métodos para acerto, a floresta aleatória teve o pior desempenho. Conseguiu prever corretamente 43,75% dos resultados, tendo em consideração a variável do primeiro tempo de jogo e sem essa variável teria uma percentagem de acerto de 37,5%. O método LASSO teve de acerto 59,375% com o resultado da primeira parte e sem ele teve 50%. *Boosting* com 56,25% e o melhor modelo com menos variáveis associadas foi o MARS com 60,9375% e 57,8125% de acerto.

Em conclusão, este artigo mostra que as variáveis consideradas mais relevantes são as variáveis defensivas como bloqueios e desarmes e variável ofensiva como a tentativa na direção do golo foram as mais importantes. Nenhum dos modelos aplicados considerou as variáveis de desempenho como significativas estatisticamente.

(Al-Asadi, 2018) procura construir sistemas de apoio a decisões inteligentes de maneira a enfrentar um dos principais desafios no futebol que está relacionado com a escolha do jogador adequado à determinada posição em jogo. Esta atribuição é normalmente efetuada pelo treinador de acordo com as suas observações e experiência, o que leva a uma seleção com grandes vieses. Posto isto e uma vez que não existem fórmulas ou equações usadas para a sua identificação, o objetivo deste trabalho passa por propor um novo sistema inteligente de apoio à decisão com base nos atributos e habilidades dos jogadores de forma a descobrir qual a posição mais adequada para ele.

Foram usadas técnicas de aprendizagem de máquina como regressão logística e linear, rede neuronal, floresta aleatória e k-vizinhos para classificação e problemas de regressão. Foi usado também algoritmo RFE (recursive feature elimination) e algoritmo PCA (principal component analysis) de modo a reduzir a dimensão dos dados.

No presente estudo (Al-Asadi, 2018) usou algoritmos de floresta aleatória com o intuito de encontrar a posição preferencial para cada jogador. Concluiu-se ser mais eficiente na classificação de posição do que os outros algoritmos onde se obteve 88,6% de precisão de previsão para a classificação binária (2 posições) e para precisão preditiva foi de 58,5% para a classificação múltipla (14 posições).

De seguida, avaliou-se o desempenho dos algoritmos usando três técnicas comuns, a retenção (hold-out), validação cruzada (CV) e retenção aleatória repetida. Comparou-se

depois o resultado entre as três técnicas.

Depois de atribuir a cada jogador uma posição em campo, determinou-se o melhor 11 inicial para formação de táticas. Por último, utilizaram-se algoritmos de regressão linear e logística, floresta aleatória e rede neural e o melhor resultado foi obtido pela floresta aleatória onde a precisão de previsão foi de 99,9% usando 17 atributos de desempenho.

Neste artigo (Young, Miller & Talpey, 2015) procuram prever as qualidades físicas dos jogadores de futebol. Têm como objetivo determinar as relações entre as qualidades físicas em estudo, a mudança de velocidade de direção (COD) e o desempenho de agilidade defensiva. O estudo contou com 24 jogadores masculinos que foram avaliados no sprint (10metros), força máxima (3 repetições – meio agachamento máximo), potência na perna (salto de contra movimento), força reativa (salto em queda) e um único teste de velocidade COD e um teste de agilidade defensiva. Fez-se a correlação com a velocidade de mudança de direção e a força reativa (r=-0,645, p=0,001) e aceleração de sprint (r=0,510, p=0,011). Fez-se também a regressão múltipla que indicou que as qualidades físicas quando combinadas obtinham de variância de 56,7% quando associada à velocidade COD (ajustada $R^2 = 0.567$, $p \leq 0.05$).

Os atletas foram divididos em dois grupos de velocidade COD, o mais rápido e o mais lento e comparados com testes t. O grupo mais rápido foi consideravelmente melhor ($p \leq 0.05$) nos testes de aceleração de *sprint* e força máxima.

As correlações entre as qualidades físicas e a agilidade foi (r = -0,101 a 0,123, p>0,05) e apenas 14,2% da variância associada com o desempenho da agilidade (ajustada $R^2 = -0.142$, $p > 0.05$). Quando se comparou a agilidade para os dois grupos existiram grandes diferenças ($p > 0.05$) em todas as qualidades físicas.

Conclui-se neste trabalho que a força reativa e a aceleração do *sprint* têm importância na velocidade COD, no entanto as qualidades físicas não estão associadas ao desempenho da agilidade defensiva.

O artigo de Al-Shboul *et al.*, 2017, pretende auxiliar os gestores e selecionadores de equipas de futebol, ajudando a formar o melhor 11 possível a partir de um leque de jogadores disponíveis. Esta análise e construção inicial da equipa cria uma maior vantagem competitiva e uma maior probabilidade de ganhar os jogos.

O presente trabalho centrou-se em jogos de equipas e no jogador analisando as estatísticas e cálculos a partir de dados anteriores e também analisando os jogos disputados entre oponentes específicos de forma individual (Al-Shboul *et al.*, 2017).

Os grandes objetivos deste artigo foram a seleção da melhor combinação de equipa possível para um adversário específico e prever a probabilidade de ganhar o jogo. Este artigo procura criar uma ferramenta capaz de fazer a gestão da equipa, implementar

análise de dados dos jogadores e classificá-los com base nessa análise.

As duas grandes limitações do presente trabalho foram a falta de dados estatísticos e pouca informação sobre derivação de medidas significativas e estatísticas dos jogos de futebol.

A metodologia aplicada foi o modelo preditivo que é usado para motivar o jogador individualmente relativamente a outros jogadores da mesma equipa. De forma a solucionar este problema utilizou-se a aprendizagem semi-supervisionada com o intuído de descobrir a importância do jogador. Para este fim, usou-se análise de rede neural onde foram atribuídos pesos aos recursos de entrada de todos os jogadores de maneira a obter uma medida de avaliação de cada jogador de acordo com a sua posição, e utilizou-se os resultados do final dos jogos para duas pontuações para equipas individuais. Os resultados das duas equipas são combinados para obter uma vitória ou derrota no final do jogo.

Para a classificação do jogador usou-se a rede neural com 11 jogadores na camada de entrada considerando apenas a questão da equipa e a precisão com validação cruzada de 5 vezes é de 54%, com o conjunto de dados que foi disponibilizando. No entanto o autor acredita que a precisão seria melhor caso houvesse mais dados em análise.

Para a classificação de rede para previsões da equipa usou-se a rede neural com 22 jogadores, 11 jogadores para cada equipa em jogo e a precisão de treino de validação cruzada 5 vezes é de 60%, com uma melhoria comparativamente à anterior de 6 pontos percentuais.

Barron *et al.*, 2018, afirmam que há um elevado interesse no desenvolvimento de perfis de desempenho de fatores técnicos dos jogadores de futebol. O objetivo foi desenvolver um modelo capaz de identificar indicadores-chave de desempenho no futebol profissional que influenciam status da liga dos jogadores externos. Foram usados 966 jogadores em que cada um deles completou os 90 minutos de 1104 jogos no Campeonato da Liga Inglesa de Futebol nos anos 2008/2009 e época de 2009/2010. Foram também usadas 505 variáveis onde se excluíram as que tinham baixa variância e ficou-se com um total de 335 variáveis de desempenho como o número total, precisão, media, medianas e quartis superior e inferior de passes, *tackles*, passes recuperados, folgas e tiros. Foi também usado um sistema para coletar dados de desempenho que fornece 5 variáveis principais nas ações realizadas durante um jogo, evento e hora do evento, o jogador um envolvido e o jogador dois envolvido (se relevante) que mostrou ter bom *inter-observer agreement* (acordo entre as observações) para o número e tipo de eventos, o jogador um envolvido e o jogador dois envolvido $(k > 0{,}9)$.

Existiam três categorias independentemente da posição em campo onde os jogadores foram atribuídos com base em onde estiveram a maior parte do tempo de jogo

durante a época seguinte. A primeira categoria compreende os jogadores que completaram a maior parte dos jogos numa liga inferior (Grupo 0: n=209 e presenças médias de 90 minutos = 10 ± 10), a segunda categoria contém os jogadores que completam a maior parte do tempo de jogo na Liga Inglesa de Futebol (Grupo 1: n=637 e média de 90 minutos de exibição = 18 ± 12) e a terceira categoria inclui os jogadores que progrediram e completaram a maior parte do seu tempo de jogo na *Premier League* Inglesa durante a época seguinte (Grupo 2: n=120 e média de 90 minutos de exibição = 19 ± 12).

De maneira a que o tamanho das amostras do grupo 0 e grupo 1 fossem iguais houve uma seleção interna para conter apenas 209 jogadores de cada grupo. Já o grupo 2, da terceira categoria sofreu uma análise usando Rede Neural Artificial *Stepwise* de maneira a identificar o conjunto de variáveis ideal para previsão do *status* de jogo.

A análise de rede neural artificial não forneceu nenhum modelo capaz de detetar diferenças entre os jogadores do grupo 0 e grupo 1. O melhor modelo produzido para estes dois grupos previu corretamente 67,9% do *status* de jogo dos jogadores do grupo de teste com um erro de 10,8% utilizando a combinação de nove variáveis. O melhor modelo foi para o grupo 1 e 2 que previu corretamente 61,5% com erro de 11,6%. Para o grupo 0 e 2 a percentagem de acerto foi de 78,8% com erro de 8,3%.

Este modelo foi produzido para mostrar que é possível que através de rede neural artificial detetar com precisão quais os jogadores que devem ou não ser promovidos para escalões superiores predizendo a trajetória da sua carreira.

Evwiekpaefe, Bitrus e Ajakaiye afirmam que o sucesso de qualquer equipa de futebol passa pelo desempenho dos seus jogadores. No entanto, torna-se difícil encontrar o melhor jogador dentro de um leque de jogadores.

Posto isto, o objetivo deste artigo é avaliar as habilidades/atributos de desempenho dos jogadores de futebol que ocupem a posição de avançados.

Para a pesquisa foram selecionados 100 jogadores de forma aleatória e divididos em dois grupos, 90 jogadores para treino e 10 para teste que pertencessem a equipas diferentes. Para a sua análise usaram-se Redes Neurais Artificiais (RNA) baseadas num modelo de perceção multicamadas(MLP). Posteriormente e com o auxílio da ferramenta de mineração de dados WEKA contruíram modelos para categorizar os jogadores em bons, médio e abaixo da média (Evwiekpaefe, Bitrus & Ajakaiye, 2020).

Utilizaram o algoritmo J48 que tem como objetivo gerar uma árvore de decisão baseada num conjunto de dados classificando as suas instâncias (Librelotto & Mozzaquatro, 2013).

Os algoritmos de classificação foram avaliados de acordo com os valores de Accuracy, Precision, Recall.

Accuracy (exatidão), faz referência ao resultado da observação prevista corretamente para o conjunto total de dados ou valor real verdadeiro, a melhor á 1,0 e a pior 0,0.

Precision (previsão), é o resultado dos valores de previsões positivas corretas sobre o número total de previsões positivas. Quanto maior o seu valor, melhor é a capacidade preditiva do algoritmo.

Recall, é o número de previsões positivas corretas sobre o número total de casos positivos. Este refere-se à verdadeira taxa positiva. Quanto maior o valor de recuperação, melhor á a capacidade preditiva do algoritmo para a classe positiva.

Um alto valor de *Precision* mostra que o algoritmo retorna resultados mais relevantes do que irrelevantes e um alto valor de *Recall* indica que a maioria dos resultados ajustados pelos algoritmos são relevantes.

Os resultados da validação cruzada 10 vezes indicam que a MLP teve um melhor desempenho do que a J48. MLP apresenta valores de Accuracy de 68%, Precision de 62,3% e Recall de 68% e a J48 tinha Accuracy de 57%, Precision de 52% e Recall de 57%.

A classificação de perceção multicamadas (MLP) previu com precisão 64 dos 87 jogadores têm uma classificação ‘bom’ (73,6%), 2 dos 9 ‘médio’ (22,2%) e 2 dos 4 (50%) ‘abaixo da média’. O classificador J48, previu com precisão 56 dos 78 jogadores apresentam classificação ‘bom’ (71,8%), 1 dos 15 classificação ‘médio’ (6,67%) e 0 dos 7 (0%) ‘abaixo da média’. Concluiu-se que a classificação de perceção de multicamadas superou o classificador J48 (E., Bitrus & Ajakaiye, 2020).

Este artigo (Vroonen *et al.*, 2017) procuram identificar o nível de habilidades de um jogador e como este irá evoluir no futuro. Foram desenvolvidos noutros desportos sistemas de previsão do futuro de um jogador usando sistemas como PECOTA para basebol, MLB e CARMELO para basquete NBA que também analisam e medem a similaridade com base no passado estatístico e características pessoais como peso, idade e altura. O sistema PECOTA compara o perfil do jogador-alvo com os perfis dos jogadores anteriores quando estes estavam no mesmo estágio de desenvolvimento e o CARMELO prevê o desempenho futuro de jogador-alvo com base na evolução de jogadores anteriores.

Neste estudo o sistema proposto foi o APROPOS, um sistema que prevê o futuro potencial de um jogador de futebol onde se projetou o potencial de um jogador-alvo de um banco de dados. De maneira a identificar jogadores com um perfil semelhante quando eles tinham a mesma idade, ou seja, prevê a evolução de determinada habilidade ao longo do tempo. Cada jogador apresenta 24 habilidades diferentes e cada habilidade é classificada numa escala de 0 a 100. Na avaliação do sistema APROPOS o objetivo foi avaliar quatro questões. A primeira sobre o quão bem pode este sistema prever as classificações em um

ano futuro, a segunda sobre a precisão de prever quão longo no futuro este sistema pode classificar, a terceira sobre qual o efeito do número de anos de dados usados para calcular a similaridade entre dois jogadores no desempenho preditivo do sistema APROPOS e por último, a quarta questão pretende responder a como o limite usado para identificar jogadores semelhantes afeta o modelo APROPOS e o seu desempenho de previsão.

Com isto, compararam-se 4 sistemas diferentes:

- *Baseline*, onde seria dada uma idade de previsão de a_2, que encontra todos os jogadores da mesma idade e para cada uma das habilidades prevê a classificação média de todos os jogadores;
- ABS-ABS, usa a métrica de similaridade absoluta e o mecanismo de previsão absoluta;
- ABS-EVO, usa a métrica de similaridade absoluta e mecanismo de previsão evolutiva;
- EVO-EVO, usa a métrica de similaridade evolutiva e pré-mecanismo de dicção.

Foram usados 1000 jogadores das competições inglesas e alemãs e para cada um dos jogadores usou-se dados de 2012 e anteriores para cálculo de semelhança e previsão para o ano de 2013 em diante. Como métrica de erro usou-se o erro absoluto médio (MAE) que é uma média de todos os jogadores e de todas as habilidades.

De forma a dar resposta à primeira questão das quatro acima colocadas previu-se as classificações para 2013 e usaram-se três anos para calcular as semelhanças dos jogadores e o limite para selecionar o melhor jogador é definido como 0,9.

Para a segunda questão, previram-se resultados para cada ano no período de 2013 a 2017 inclusive. Na terceira questão, prevê-se a classificação de 2013 usando o limite de similaridade de 0,9 e o número de anos usado varia para calcular as semelhanças de um a cinco. E a quarta e última pergunta prevê a classificação de 2013 usando três anos para calcular as semelhanças dos jogadores e variar o limite usado para classificar os jogadores de 0,7 a 0,9 em aumentos de 0,05.

Como resultado da pesquisa de Vroonen, cada modelo tem um melhor desempenho do que o modelo de *baseline*. O erro absoluto médio (MAE) da *baseline* é de 12,62%, do modelo APROPOS, e ABS-ABS, tem um erro de 5,45%. O método de previsão

evolutiva parece resultar em previsões mais precisas do que a métrica de previsão absoluta.

Conclui-se também que o período de previsão influência o erro absoluto médio (MAE) uma vez que as previsões mais futuras de cinco anos são duas vezes piores do que as previsões mais próximas, de um ano.

Em conclusão à quarta pergunta o limite usado para identificar jogadores semelhantes afeta o desempenho. No modelo ABS-ABS esta regra tem um efeito forte, quanto que nos outros modelos ABS-EVO e EVO-EVO o efeito não é tão notório.

Tabela 2- Comparativo entre trabalhos relacionados analisados

Autor	Objetivo	Métodos	Resultados
(Razali *et al.*, 2017)	Previsão de resultados de jogos de Futebol vitoria-empate-derrota;	- *Redes Bayesianas* (BNs)	A média de percentagem de previsão dos 3 anos foi de 75,09%. A precisão de previsão é de 75,26% para 2010-2011, 79,47% para 2011-2012 e 70,53% para 2012-2013.
(Schneider, 2018)	Previsão de resultados de jogos de Futebol vitoria-empate-derrota;	- Regressão logística - Análise Discriminante Linear - Sistema ELO	- A exatidão máxima foi de 54,73% ± 2,06% - Análise discriminante linear de 54,82% ± 2,01% - *SVM* com *kernel* linear de 54,35% ± 2,12% - k-vizinhos 54,34% ± 1,93%. - Método *ensemble* demonstrou 57% de exatidão. - Vitória para a equipa da casa (mandante) é de 46,56%. - Sistema ELO da equipa visitada era maior do que o ELO da equipa visitante vencendo com 50,24%. - Análise Discriminante Quadrática foi o melhor método quando falado em empates, no entanto dentre os classificadores este foi o que obteve uma menor exatidão com valor de 42,63%±2,33%.
(Nabinger, 2018)	Previsão de resultados de jogos de Futebol	- *Redes Bayesianas* - Árvore de	Floresta aleatória 43,75% de acerto e na 1ª parte 37,5%. LASSO acertou 59,375% e na 1ª parte 50%.

	vitoria-empate-derrota;	decisão - Ridge - LASSO e MARS - Floresta aleatória - *Bagging e boosting*	*Boosting* 56,25% MARS (melhor modelo) 60,9375% e 57,8125% de acerto na 1ª parte.
(Al-Asadi, 2018)	Escolha do jogador adequado a determinada posição	- Regressão logística e linear - Rede neuronal - Floresta aleatória - k-vizinhos RFE e PCA	Floresta aleatória foi o melhor método com 88,6% de precisão para classificação binária e 58,5% para classificação múltipla. Para precisão de previsão de posição do jogador em campo o melhor método foi de floresta aleatória com 99,9% de acerto.
(Young, Miller and Talpey, 2015)	- Previsão das qualidades físicas dos jogadores de futebol. - Determinar as relações a mudança de velocidade de direção (COD) e o desempenho de agilidade defensiva.	- Correlação - Regressão múltipla - Testes t	- A correlação com a velocidade de mudança de direção e a força reativa (r=-0,645, p=0,001) e aceleração de sprint (r=0,510, p=0,011). - A regressão múltipla que indicou que as qualidades físicas quando combinadas obtinham de variância de 56,7% quando associada à velocidade COD (ajustada R^2 = 0.567, $p \leq 0.05$). - Os atletas foram divididos em dois grupos de velocidade COD, o mais rápido e o mais lento e comparados com testes t. O grupo mais rápido foi consideravelmente melhor ($p \leq 0.05$) nos testes de aceleração de *sprint* e força máxima. - As correlações entre as qualidades físicas e a agilidade foi (r = -0,101 a 0,123, $p > 0,05$) e apenas 14,2% da variância associada com o desempenho da agilidade (ajustada R^2 = -0.142, $p > 0.05$). Quando se comparou a agilidade para os dois grupos existiram grandes diferenças ($p > 0.05$) em todas as qualidades físicas. - Força reativa e a aceleração do *sprint* são importantes para a velocidade COD, mas as qualidades físicas não estão associadas ao desempenho da defesa da agilidade.

(Al-Shboul *et al.*, 2017)	Formação da melhor equipa e previsão da probabilidade de ganhar o jogo	- Aprendizagem semi-supervisionada - Rede Neuronal Artificial	Classificação do jogador é de 54% para a precisão com validação cruzada de 5 vezes Classificação para previsão da equipa utilizou-se a rede neural e a validação cruzada de 5 vezes é de 60%.
(Barron *et al.*, 2018)	Desempenho de fatores técnicos dos jogadores de futebol	- Estatística descritiva - Rede neural artificial *Stepwise*	A rede neural artificial não forneceu nenhum modelo capaz de detetar diferenças entre os jogadores. O melhor modelo previu 67,9% do *status* do jogo com um erro de 10,8%. O melhor modelo previu corretamente 61,5% com erro de 11,6%.
(Evwiekpaefe, Bitrus & Ajakaiye, 2020)	Avaliar atributos de desempenho dos jogadores na posição de avançados.	- Rede neuronal artificial - WEKA - MLP - J48	- MLP apresenta valores de *Accuracy* de 68%, *Precision* de 62,3% e *Recall* de 68%. - J48 tinha *Accuracy* de 57%, *Precision* de 52% e *Recall* de 57%. - A classificação MLP previu com precisão 64 dos 87 jogadores têm uma classificação 'bom' (73,6%), 2 dos 9 'médio' (22,2%) e 2 dos 4 (50%) 'abaixo da média'. - O classificador J48, previu com precisão 56 dos 78 jogadores apresentam classificação 'bom' (71,8%), 1 dos 15 classificação 'médio' (6,67%) e 0 dos 7 (0%) 'abaixo da média'. - A classificação de MLP teve um melhor desempenho do que a J48.
(Vroonen *et al.*, 2017)	Avaliação do nível de habilidades do jogador e sua evolução ao longo do tempo.	- PECOTA - MLB - CARMELO - APROPOS	- Cada modelo tem um melhor desempenho do que o modelo de *baseline*. O erro absoluto médio da *baseline* é de 12,62%, do modelo APROPOS, e ABS-ABS, tem um erro de 5,45%. - O método de previsão evolutiva parece resultar em previsões mais precisas do que a métrica de previsão absoluta. - Conclui-se também que o período de previsão influência o erro absoluto médio (MAE). - Em conclusão à quarta pergunta o limite usado para identificar jogadores semelhantes afeta o desempenho. No modelo ABS-ABS esta

			regra tem um efeito forte, enquanto que nos outros modelos ABS-EVO e EVO-EVO o efeito não é tão notório.

Como se pode observar na Tabela 2, existe uma grande diferença do trabalho proposto na presente dissertação em comparação com a grande maioria das pesquisas e estudos que se têm centrado na previsão de resultados, principalmente e em como modelar a equipa para um melhor resultado no final do jogo, ou seja, concentram-se maioritariamente na parte final do jogo No entanto, a presente dissertação apresenta o seu foco e interesse no durante o jogo, na análise dos perfis dos jogadores e a sua evolução, tentando prever o comportamento dos jogadores e que tipo de jogador serão no futuro, de forma a que se tiver que fazer uma substituição durante um determinado jogo, a escolha seja o mais próxima da ótima quanto possível.

4 Modelação dos jogadores

Procura-se neste capítulo, analisar e descrever os mais relevantes processos inerentes à monitorização dos atletas e análise de dados biométricos.

Pretende-se descrever de forma detalhada as tabelas e as suas variáveis de modo a compreender melhor a base de dados em estudo. Desse modo, analisaram-se as interações entre elas para perceber como tirar proveito das mesmas de maneira a usar essa informação para obtenção de resultados.

Em seguida, abordamos a metodologia em estudo onde se discute de forma detalhada dos passos a elaborar e quais as métricas de avaliação que seriam possíveis utilizar e quais as mais vantajosas e completas.

4.1 **Dados**

Para a realização da presente dissertação utilizou-se a base de dados do *Kaggle* (https://www.kaggle.com/datasets) referente ao futebol europeu.

A presente base de dados foi escolhida de primeira instância pela sua versatilidade, polivalência, boa estruturação e apresentava informação relevante para o trabalho que se queria desenvolver na dissertação. Existiram mais bases de dados, no entanto a informação que continham pareceu pouco relevante e com pouca quantidade de dados o que não ajudaria na polivalência de tarefas que se poderia fazer ficando demasiado regidos pela pouca informação.

O primeiro passo na elaboração da presente dissertação passou por descrever detalhadamente as tabelas e os diferentes atributos procurando justificar a sua inclusão ou exclusão do âmbito desta dissertação. Uma vez que existem atributos que não se alteram ao longo dos anos como o pé preferido por exemplo, então para análise de melhor performance não serão necessárias estas medidas. Os atributos dos jogadores e das equipas são provenientes do videogame FIFA da *Electronic Arts (EA)*, atualizadas semanalmente.

Assim sendo, temos o antecedente e o consequente neste caso, o antecedente será que os *player_attributes* são de um jogo de computador e o consequente é o que eles são da vida real.

No presente trabalho assume-se que aquilo está relacionada com o que a EA

atualizou dos valores dos jogadores e da equipa de acordo com os resultados que vão tendo e do desempenho deles semanalmente. E existindo alguma correlação com realidade, poder-se-ão usar para prever as trocas e substituições, um dos grandes objetivos da dissertação.

Apesar de não estar a trabalhar com dados reais o videogame FIFA da EA tem informação o mais real possível pelo que é também possível transpor isso para o mundo real. Os atributos dos jogadores são definidos para um jogo específico, e quando estão a jogar no jogo FIFA há uma simulação dos atributos de desempenho dos jogadores no jogo o mais próximo possível da realidade, no entanto existe a introdução e atualização de novos dados semanalmente.

Estes atributos são os grandes responsáveis por determinar a qualidade dos jogadores e conhecê-los bem pode ser uma grande vantagem para a equipa na hora de formar o onze inicial.

A base de dados contém sete tabelas interligadas entre si, as sete tabelas são a Tabela *Country, League, Player, Player_attributes, Team e Team_attributes*.

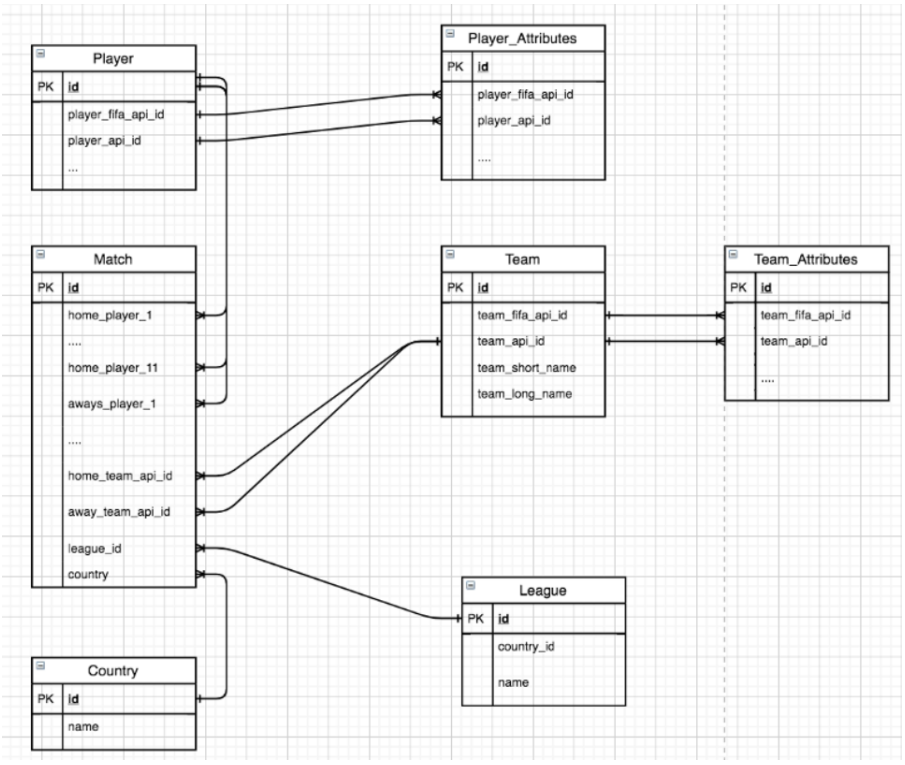


Figura 6- Demostra as tabelas existentes no dataset e a relação entre elas

4.1.1 Tabela *Country* e *League*

Na tabela *Country* e *League* temos a descrição dos 11 países europeus em estudo. Todas as tabelas da base de dados têm uma coluna correspondente ao número de identificação (ID). Nas tabelas *Country* e *League* o número de identificação (ID) referentes às colunas *country_id* e *league_id* apresentam o mesmo valor de identificação e a coluna *Name* faz referência à liga.

Tabela 3-Tabela Country e League da base de dados em estudo

Id	*Country_id*	*Name*
1	1	Belgium Jupiler League
1729	1729	England Premier League
4769	4769	France Ligue 1
7809	7809	Germany 1. Bundesliga
10257	10257	Italy Serie A
13274	13274	Netherlands Eredivisie
15722	15722	Poland Ekstraklasa
17642	17642	Portugal Liga ZON Sagres
19694	19694	Scotland Premier League
21518	21518	Spain LIGA BBVA

4.1.2 Tabela *Player*

Na tabela *Player* temos 11060 linhas que nos indicam o número de jogadores. O *player_name* é o nome do jogador e o *player_api_id* é o número associado ao jogador e o *player_fifa_api_id* é o número associado ao jogador no FIFA. Estes identificam os jogadores em partidas únicas. Ou seja, se um jogador fizesse uma transferência, ele ainda teria esse número. O *player_api_id* e *player_fifa_api_id* não são iguais, mas funcionam da mesma forma, pois cada um se refere a um único jogador. Estas últimas três colunas apresentam valores únicos para cada linha, ou seja, um determinado jogador tem um valor único e singular para o *player_api_id* e *player_fifa_api_id* e isso acontece nas 11060 linhas da tabela.

A coluna *birthday* da Tabela 4 é referente à data de nascimento do jogador e está representada em forma de data e hora.

Tabela 4- Tabela Player da base de dados em estudo

Id	*Player_api_id*	*Player_name*	*Player_fifa_api_id*	*Birthday*	*Height*	*Weight*
1	505942	Aaron Appindangoye	218353	29/02/92 00:00	182.88	187
2	155782	Aaron Cresswell	189615	15/12/89 00:00	170.18	146
3	162549	Aaron Doran	186170	13/05/91 00:00	170.18	163
4	30572	Aaron Galindo	140161	08/05/82 00:00	182.88	198
5	23780	Aaron Hughes	17725	08/11/79 00:00	182.88	154
6	27316	Aaron Hunt	158138	04/09/86 00:00	182.88	161
7	564793	Aaron Kuhl	221280	30/01/96 00:00	172.72	146
8	30895	Aaron Lennon	152747	16/04/87 00:00	165.1	139
9	528212	Aaron Lennox	206592	19/02/93 00:00	190.5	181
10	101042	Aaron Meijers	188621	28/10/87 00:00	175.26	170
11	23889	Aaron Mokoena	47189	25/11/80 00:00	182.88	181
12	231592	Aaron Mooy	194958	15/09/90 00:00	175.26	150
13	163222	Aaron Muirhead	213568	30/08/90 00:00	187.96	168
14	40719	Aaron Niguez	183853	26/04/89 00:00	170.18	143
15	75489	Aaron Ramsey	186561	26/12/90 00:00	177.8	154
16	597948	Aaron Splaine	226014	13/10/96 00:00	172.72	163
17	161644	Aaron Taylor-Sinclair	213569	08/04/91 00:00	182.88	176
18	23499	Aaron Wilbraham	2335	21/10/79 00:00	190.5	159
19	120919	Aatif Chahechouhe	187939	02/07/86 00:00	175.26	150
20	46447	Abasse Ba	156626	12/07/76 00:00	187.96	185
21	167027	Abdelaziz Barrada	192274	19/06/89 00:00	177.8	161
22	245653	Abdelfettah Boukhriss	202425	22/10/86 00:00	185.42	161
23	128456	Abdelhamid El Kaoutari	188145	17/03/90 00:00	180.34	161
24	42664	Abdelkader Ghezzal	178063	05/12/84 00:00	182.88	172
25	425950	Abdellah Zoubir	212934	05/12/91 00:00	180.34	161
26	38423	Abdelmajid Oulmers	52782	12/09/78 00:00	172.72	143
27	3264	Abdelmalek Cherrad	51868	14/01/81 00:00	185.42	165
…						

4.1.3 Tabela Player_Attributes

Toda a discussão que se segue sobre os atributos dos jogadores está baseada na informação contida nos sítios web de FIFAUTeam e Electronic Arts.

Comparando a tabela *Player* com a *Player_attributes,* na tabela *Player* os valores são únicos uma vez que aparecem os jogadores apenas uma vez e as suas informações pessoais como pesa, idade e altura, e na *Player_attributes* apresenta várias vezes o mesmo número, ou seja, apesar de valor único também este aparece várias vezes ao longo da tabela, isto porque um determinado jogador tem várias avaliações ao longo do tempo de acordo com a coluna *date*. Na tabela *Player_attributes* é a data que o jogador tem uma nova avaliação, ou seja, os atributos desta tabela são o desempenho dos jogadores no

jogo. Existem várias datas associadas ao mesmo jogador uma vez que ele joga várias vezes ao longo de cada época.

Para cada jogador iremos ter atributos associados que apresentam um valor importante em jogo. Dependendo do valor, o treinador pode decidir escolher o jogador para entrar em campo.

Como descrito no capítulo 2, o valor que o jogador apresente em determinados atributos que sejam importantes para a posição que ocupa em campo evidencia uma maior ou menor importância.

Overall rating é a avaliação geral do jogador e como descrito acima é uma das colunas mais importantes na avaliação do jogador para aquele jogo uma vez que este valor está associado a uma data, a data do jogo.

Potential do jogador é maior quando maior for a sua experiência.

Preferencial_foot é, como o nome indica, o pé preferido do jogador e cerca de 24.43% dos jogadores prefere o pé esquerdo e 75.57% prefere o direito.

Attacking_work_rate e defensive_work_rate são a taxa de ataque e defesa para cada jogador, este atributo irá depender da posição que o jogador ocupa em jogo, um defesa irá ter um valor de *defensive_work_rate* maior que um avançado e vice-versa.

Volleys é um atributo que mede a precisão e o poder do pontapé aéreo no golo. Afeta a técnica e a precisão dos remates enquanto a bola está no ar. Ou seja, este valor em tabela é orientado de acordo com o jogador de acordo com a posição em campo e de quantos pontapés ele fez num determinado jogo (data).

Crossing, atributo que mede a precisão com que o jogador cruza a bola durante as corridas normais e os lances livres. Se o jogador fica bloqueado e não passa a bola a pontuação de cruzamento tende a baixar.

Finishing, mede o jogador quando este está quase a marcar golo e a precisão dos remates dentro da área de grande penalidade.

Heading_accuracy, mede a precisão de direção do jogador no cabeceamento na bola, e afeta a capacidade do jogador de colocar a cabeça na bola.

Short_passing, mede a precisão e velocidade dos passes rasteiros e de curta distância para cada jogador naquele jogo.

Long_passing, mede a qualidade e precisão que o jogador é capaz de passar a bola pelo alto para outro jogador.

Balance, mede o equilíbrio e coordenação um dos fatores importantes no desenvolvimento de habilidades e no perfil geral de condicionamento físico do jogador.

Dribbling, é a capacidade de um jogador fintar o oponente e manter a posse de bola, tendo-a o mais próxima do pé enquanto executa a jogada.

Curve, mede a capacidade do jogador de curvar a bola ao passar pelo adversário alterando a sua trajetória.

Free_kick_accurancy, mede a precisão de chute livre e quanto maior o valor maior será a precisão.

Ball_control, mede a capacidade e habilidade de um jogador controlar a bola quando a recebe. Quanto maior o valor, menor a probabilidade de a bola saltar para longe do jogador após controlá-la.

Acceleration, mede o tempo que demora um jogador a atingir a sua velocidade máxima. Quanto maior o valor de aceleração, menos tempo demorará o jogador a chegar á sua rapidez.

Speed_sprint, mede a velocidade com que o jogador corre enquanto está na velocidade máxima.

Agility, mede a capacidade de movimentação dos jogadores, o quão ágil o jogador é enquanto se movimenta e o quão rápido e gracioso um jogador é capaz de controlar a bola. Quanto mais alto o valor maior será a rapidez e graciosidade no drible.

Reaction, mede a rapidez com que o jogador reage e responde a uma situação em jogo. É o tempo entre o momento em que vê onde a bola está e o momento em que fica posicionado para a receber.

Strength, é sobre a qualidade ou estado físico do atleta. Este atributo mostra a capacidade do jogador de vencer uma disputa física com o adversário. Quanto maior o valor maior probabilidade o jogador terá no frente a frente com o adversário na hora de ganhar espaço em campo.

Vision, classifica a consciência do jogador sobre a posição dos companheiros de equipa e dos oponentes ao seu redor. Quanto maior este atributo mais amplo será o campo de visão do jogador para localizar os jogadores da equipa e maior será a possibilidade de um passe longo bem-sucedido.

Penalties, mede a precisão dos chutes dentro da área de grande penalidade.

Marking, refere-se à marcação de um jogador a outro, diminuindo os espaços e impedindo de obter um cruzamento ou passe.

Jumping, mede a capacidade e a qualidade do jogador de saltar da superfície para o cabeceamento. Quanto maior o valor maior será o salto.

Intersection, mede a capacidade de ler o jogo, tomar posse da bola e intercetar os passes. Quanto maior o valor maior será a probabilidade de o jogador conseguir evitar o passe impedindo que a bola chegue ao jogador planeado.

Shot_power, avalia o “poder de tiro”, a força máxima e precisão com que o jogador bate na bola para marcar.

Stamina, varia de acordo com a posição do jogador. Considerada pela FIFA a habilidade mais importante do jogo torna-se importante aumentar a resistência dos jogadores, caso contrário, em pouco tempo eles ficam cansados. A *stamina* ajuda também a evitar lesões e funciona como uma barra de fadiga, quanto mais tempo o jogador joga, mais cansado ele irá ficar e as restantes habilidades ficam também comprometidas.

O valor determina o cansaço do jogador durante o jogo e é bastante expresso no final do jogo, isto é, baixos níveis de *stamina* condicionam a boa resistência do jogador e consequentemente da equipa. Avalia o quão cansado fica o jogador quando o jogo se aproxima do intervalo ou do fim do jogo. As equipas que apresentam uma boa resistência têm uma maior vantagem durante as últimas fases do jogo. Se o jogador apresentar valor de stamina baixos os restantes atributos não vão ser relevantes uma vez que já não tem o mesmo resultado, ou seja, que os baixos níveis de resistência irão condicionar as restantes habilidades do atleta o que o condicionará na sua prestação em jogo.

A stamina aumenta ao longo dos anos, e a cada jogo existe um valor de *stamina* para cada jogador, isto significa que quanto maior o valor, mais rápido o jogador fica cansado.

Long_shots, significa “tiros longos”. Enquanto o atributo de finalização mede a precisão apenas dentro da área, este atributo mede a precisão dos chutes de fora da área de grande penalidade.

Agression, mede o nível de agressividade de um jogador e avalia a frequência e a agressão de empurrões, carrinhos, etc.

Positioning, determina o quão bom o jogador é a defender a sua posição e a ocupar boas posições no campo durante um jogo. Quanto maior o valor, maior a probabilidade de o jogador criar espaço suficiente para receber a bola em áreas perigosas. Este atributo liga com a capacidade de o jogador detetar espaços abertos e passar para boas posições que lhe podem dar vantagem no ataque.

Tackle mede a capacidade de um jogador ganhar a posse de bola sem causar um livre, usando a força corporal, sendo o contacto ombro a ombro ou a pé. No jogo, também se traduz em ser capaz de insultar o adversário ao puxar a camisola, agarrar ou outros movimentos não tão justos sem que o árbitro veja.

Standing_tackle é o atributo que mede a habilidade do jogador ganhar a bola ao invés de cometer uma falta. Um valor baixo pode fazer com que o jogador perca tempo em jogo ou que chute a bola para longe.

Sliding são os cortes ou carrinhos. Esta estatística mede a capacidade de o jogador marcar os *tackles* de forma a ganhar a bola ao invés de fazer uma falta.

As restantes colunas são referentes ao guarda-redes, que são o *diving, handling,*

kicking, positioning e *reflexes.*

Diving mede a capacidade de o guarda-redes defender enquanto mergulha no ar.

Handling é usado para medir com clareza a capacidade de apanhar a bola e a segurar

Kicking é um atributo usado para medir o alcance e a precisão da reposição da bola em jogo.

Positioning que para os guarda-redes é diferente, este atributo mede a aptidão de se posicionar corretamente para defender.

Reflexes que mede a agilidade e rapidez com que o guarda-redes reage para defender a bola.

4.1.4 Tabela *Team*

A Tabela 5, faz referência à tabela *Team* que indica- as equipas existentes representadas pelo *id.*

Tabela 5- Estrato da Tabela Team da base de dados

Id	*Team_api_id*	*Team_fifa_api_id*	*Team_long_name*	*Team_short_name*
1	9987	673	KRC Genk	GEN
2	9993	675	Beerschot AC	BAC
3	10000	15005	SV Zulte-Waregem	ZUL
4	9994	2007	Sporting Lokeren	LOK
5	9984	1750	KSV Cercle Brugge	CEB
6	8635	229	RSC Anderlecht	AND
7	9991	674	KAA Gent	GEN
8	9998	1747	RAEC Mons	MON
9	7947		FCV Dender EH	DEN
10	9985	232	Standard de Li√®ge	STL
11	8203	110724	KV Mechelen	MEC
12	8342	231	Club Brugge KV	CLB
13	9999	546	KSV Roeselare	ROS
14	8571	100081	KV Kortrijk	KOR
15	4049		Tubize	TUB
16	9996	111560	Royal Excel Mouscron	MOU
17	10001	681	KVC Westerlo	WES
18	9986	670	Sporting Charleroi	CHA

614	9997	680	Sint-Truidense VV	STT
1034	9989	239	Lierse SK	LIE
1042	6351	2013	KAS Eupen	EUP
1513	1773	100087	Oud-Heverlee Leuven	O-H
2004	8475	110913	Waasland-Beveren	WAA
2476	8573	682	KV Oostende	OOS
2510	274581	111560	Royal Excel Mouscron	MOP
3457	10260	11	Manchester United	MUN
3458	10261	13	Newcastle United	NEW
...				

Na Tabela 5 existem 1458 equipas, a *team_api_id* faz referência ao nome da equipa e o *team_fifa_api_id* é o número de identificação da equipa no FIFA. O nome completo da equipa e a sua sigla correspondente estão identificados na coluna *team_long_name* e *team_short_name.*

4.1.5 Tabela Team_Attributes

A tabela *Team_attributes* como o nome indica está interligada com a tabela *team*, estas apresentam colunas iguais como a coluna *team_fifa_api_id* e *team_api_id*, dadas como forma de identificação.

Apresenta também outras colunas com atributos gerais, referentes à equipa.

Tabela 6- Estrato da tabela Team_attributes

Id	*Team_fifa_api_id*	*Team_api_id*	*Date*	*BuildUpPlaySpeed*	*BuildUpPlaySpeedClass*	*BuildUpPlayDrib-bling*	*BuildUpPlayDrib-blingClass*	*BuildUpPlay-Passing*	*BuildUpPlayPass-ingClass*
1	434	9930	22/02/10 00:00	60	Balanced		Little	50	Mixed
2	434	9930	19/09/14 00:00	52	Balanced	48	Normal	56	Mixed
3	434	9930	10/09/15 00:00	47	Balanced	41	Normal	54	Mixed
4	77	8485	22/02/10 00:00	70	Fast		Little	70	Long
5	77	8485	22/02/11 00:00	47	Balanced		Little	52	Mixed

6	77	8485	22/02/12 00:00	58	Balanced		Little	62	Mixed
7	77	8485	20/09/13 00:00	62	Balanced		Little	45	Mixed
8	77	8485	19/09/14 00:00	58	Balanced	64	Normal	62	Mixed
9	77	8485	10/09/15 00:00	59	Balanced	64	Normal	53	Mixed
10	614	8576	22/02/10 00:00	60	Balanced		Little	40	Mixed
…									

A Tabela 6, é um excerto da tabela *Team_attributes* onde estão demonstradas algumas colunas como a *Date* que indica a data dos jogos. Nesta coluna existem dados do ano 2010 a 2015 para cada valor único de *team_api_id* e *team_fifa_api_id* o que significa que cada equipa tem várias datas diferentes uma vez que corresponde aos jogos realizados nesses anos. As colunas *BuildUpPlaySpeedClass, ChanceCreationPassingClass, ChanceCreationDribblingClass, ChanceCreationCrossingClass, etc*, que estão designadas como *Class são a descrição textual dos atributos.* Todas as colunas apresentam valores de 20-33 (*slow*); 34-66 (*balanced*); 67-80 (*fast*) durante o jogo.

A coluna *BuildUpPlaySpeed* indica o aumento de velocidade de jogo, a coluna *BuildUpPlayDribbling* indica o aumento do drible no jogo.

BuildUpPlayPassing indica o aumento do número de passes no jogo *e BuildUpPlayPositioning* indica o aumento do posicionamento.

A *ChanceCreationPassing* é a oportunidade de criar o passe*, a ChanceCreationCrossing* é a oportunidade de criar cruzamento*, ChanceCreationShooting* é a oportunidade de tiro, isto é, chutar a bola *e ChanceCreationPositioning* é a oportunidade da equipa se posicionar de forma correta no jogo.

A *DefencePressure* é a pressão de defesa. *DefenceAgression* é a agressão de defesa. *DefenceTeamWidth* é a largura da equipa de defesa e a *DefenceDefenderLineClass* é a classe de linha da defesa.

4.1.6 Tabela *Match*

A Tabela 7, faz referência à tabela *Match* onde as linhas representam os jogos e as colunas os jogadores.

A coluna *id* indica os jogadores, representada no número de linhas da coluna houve um total de 25.979 jogos.

Existem 11.3% de valores em falta, que indicam que a tabela *Match* apresenta dados corrompidos, isto é, poderá ter havido falha ao carregar as informações ou extração incompleta por quem disponibilizou o *dataset* o que pode ter gerado valores em falta.

A tabela *Match* inclui o agregado da equipa onde as diferentes colunas indicam os valores da equipa para determinado atributo.

As trocas de jogadores na equipa durante um jogo não são contabilizadas, ou seja, não se sabe quais as substituições que foram feitas nem o jogador que saiu ou entrou em campo, apenas temos informação na tabela *match* dos onze jogadores iniciais.

Tabela 7- Estrato da Tabela Match da base de dados em estudo

Id	*Coun-try_id*	*Leag ue_id*	*Season*	*Stage*	*Date*	*Match_ api_id*	*Home_team_a pi_id*	*Away_team_ api_id*	*Home_team _goal*	*Away_tea m_goal*
1	1	1	2008/2009	1	17/08/08 00:00	492473	9987	9993	1	1
2	1	1	2008/2009	1	16/08/08 00:00	492474	10000	9994	0	0
3	1	1	2008/2009	1	16/08/08 00:00	492475	9984	8635	0	3
4	1	1	2008/2009	1	17/08/08 00:00	492476	9991	9998	5	0
5	1	1	2008/2009	1	16/08/08 00:00	492477	7947	9985	1	3
6	1	1	2008/2009	1	24/09/08 00:00	492478	8203	8342	1	1
7	1	1	2008/2009	1	16/08/08 00:00	492479	9999	8571	2	2
8	1	1	2008/2009	1	16/08/08 00:00	492480	4049	9996	1	2
9	1	1	2008/2009	1	16/08/08 00:00	492481	10001	9986	1	0
10	1	1	2008/2009	10	01/11/08 00:00	492564	8342	8571	4	1
11	1	1	2008/2009	10	31/10/08 00:00	492565	9985	9986	1	2
12	1	1	2008/2009	10	02/11/08 00:00	492566	10000	9991	0	2
13	1	1	2008/2009	10	01/11/08 00:00	492567	9994	9998	0	0
14	1	1	2008/2009	10	01/11/08 00:00	492568	7947	10001	2	2
15	1	1	2008/2009	10	01/11/08 00:00	492569	8203	9999	1	2
16	1	1	2008/2009	10	01/11/08 00:00	492570	9996	9984	0	1
17	1	1	2008/2009	10	01/11/08 00:00	492571	4049	9987	1	3
18	1	1	2008/2009	10	02/11/08 00:00	492572	9993	8635	1	3
19	1	1	2008/2009	11	08/11/08 00:00	492573	8635	9994	2	3
20	1	1	2008/2009	11	08/11/08 00:00	492574	9998	9996	0	0
21	1	1	2008/2009	11	09/11/08 00:00	492575	9986	8342	2	2
22	1	1	2008/2009	11	07/11/08 00:00	492576	9984	10000	2	0
23	1	1	2008/2009	11	08/11/08 00:00	492577	9991	7947	1	1
24	1	1	2008/2009	11	08/11/08 00:00	492578	9999	4049	1	2
25	1	1	2008/2009	11	08/11/08 00:00	492579	8571	8203	0	0
26	1	1	2008/2009	11	08/11/08 00:00	492580	10001	9987	1	0
27	1	1	2008/2009	11	09/11/08 00:00	492581	9993	9985	1	3
…										

Tabela 8- Número de jogos para cada País

Jogos		*Id Country e League*	*Nº de jogos de cada equipa*
1	1728	Id - Bélgica – 1	1728
1729	4768	Id – England - 1729	3040
4769	7808	Id- France - 4769	3040
7809	10 256	Id- Germany - 7809	2448
10 257	13 273	Id- Italy – 10 257	3017

13 274	15 721	Id- Netherlands - 13 274	2448
15 722	17 641	Id- Poland - 15 722	1920
17 642	19 693	Id- Portugal - 17 642	2052
19 694	21 517	Id- Scotland - 19 694	1824
21 518	24 557	Id- Spain - 21 518	3040
24 558	25 979	Id- Switzerland - 24 558	1422

Na Tabela 8, acima descrita, foi elaborada como forma de compreender melhor as informações das tabelas anteriores, esta resume os jogos que existiram de 1 ao 25979.

Tabela 9- Indica a época em que se iniciaram os jogos e quantos jogos houve por época

2008/2009 – 3326	*2012/2013 – 3260*
2009/2010 – 3230	2013/2014 – 3032
2010/2011 – 3260	2014/2015 – 3325
2011/2012 – 3220	2015/2016 – 3326

A coluna *Season* da tabela *Match* indica todas as épocas para cada jogo e de forma a compreender melhor elaborou-se a Tabela 9 de maneira a poder-se analisar o inicio e o fim dos jogos de avaliação que iniciaram na época 2008/2009 e acabam na época 2015/2016.

A coluna *home_team_goal* e *away_team_goal* indica o número de golos no jogo e a *match_api_id* número do jogo no FIFA.

A *home_team_api_id* e *away_team_api_id* significa o número da equipa no FIFA, são as equipas que jogam em casa e as que são visitantes, como as equipas têm valores únicos, as colunas irão apresentar valores iguais ao longo da tabela uma vez que uma equipa tanto é visitante num jogo como é visitada noutro.

O *home_player_1 a 11* e away*_player_1 a 11* corresponde aos jogadores titulares do jogo correspondente á data. A coluna *home_player_X1 a X11* e *home_player_Y1 a Y11, away_player_X1 a X11 e away_player_Y1 a Y11* são as posições em campo para cada jogador, ou seja, as variáveis X e Y indicam a posição do jogador no campo, ou seja, cada jogador apresenta uma coordenada X e Y associada.

O campo está dividido em coordenadas X e Y. As coordenadas X apresentam o valor no intervalo de 1 a 9, e as coordenadas Y o valor varia de 1 a 11, este valor associado é o que distingue a posição exata de um jogador em campo.

Os guarda-redes têm coordenadas (1,1) e os restantes jogadores em campo podem

ter coordenadas X 2 a 9 e Y 2 a 11. Se o segundo jogador em campo da equipa da casa tiver coordenadas (2,3) será homeplayer_X2=2 e homeplayer_Y2=3, pode-se deduzir que o jogador joga do lado esquerdo do campo (X=2) como defesa (Y=3). Este jogador será um defesa esquerdo. As coordenadas Y entre 2 a 5 são defesas, entre 6 a 8 são médios e 9 a 11 são avançados. O X1-Y1 é o *player* 1; X2-Y2 é o *player* 2, etc., tanto para *home* como *away*.

No presente estudo, não se torna necessário saber as coordenadas de cada jogador individualmente uma vez que queremos categorizar os jogadores em guarda-redes, defesas, médios e avançados.

A

Figura 7 abaixo, representa o campo de Futebol dividido por zonas onde a ZDE (zona defensiva esquerda); ZDC (zona defensiva central); ZDD (zona defensiva direita); ZMDE (zona média defensiva esquerda); ZMDC (zona média defensiva central); ZMDD (zona média defensiva direita); ZMOE (zona média ofensiva esquerda); ZMOC (zona média ofensiva central); ZMOD (zona média ofensiva direita); ZOE (zona ofensiva esquerda); ZOC (zona ofensiva central); ZOD (zona ofensiva direita); (Santos et al., 2016)

Figura 7- Campograma da divisão do terreno de jogo em zonas (Santos et al., 2016)

Todas as restantes colunas representadas na tabela Match estão indicadas para um determinado jogo por equipa que joga em casa (visitada) ou que joga fora de casa

(visitante) e também indica o tipo dependendo do atributo.

A coluna *goal* indica o número de golos, a *shoton* especifica cada tentativa de tiro no alvo e a *shotoff* especifica cada tentativa de tiro fora do alvo.

A coluna *foulcommit* indica a falta cometida. O *card* é o tipo de cartão, tipo de falta, A coluna *cross* são os cruzamentos. A coluna *corner* são os cantos. A coluna *possission* é posse de bola.

As medidas de tempo do *dataset* são medidas temporais em que os jogadores são avaliados ao longo dos anos. As datas de avaliação representadas na tabela *Match* são as avaliações dos jogadores feitas pelos olheiros da FIFA antes dos jogos e de forma semanal, são na realidade atualizações dos dados do jogador feitas pela EA no jogo da FIFA.

As datas de avaliação compreendem os anos 2007 a 2016, nos anos 2007 a 2012 as avaliações foram nos meses de Fevereiro e Agosto, apenas duas datas por ano. A partir de 2013 até 2016 as avaliações foram mais contínuas com intervalos de uma semana num total de 172 semanas.

As avaliações semanais podiam ou não pertencer aos mesmos jogadores e equipas, ou seja, uma equipa pode ter reajuste de valores nos atributos no início do mês e outra equipa na semanas seguinte.

A Tabela 10 e Tabela 11 mostram as datas de avaliação por épocas. A Tabela 10 faz referência às datas dos anos 2007 a 2013 com duas avaliações por época e a Tabela 11 dos anos 2013 a 2016 com várias avaliações como explicado acima.

Tabela 10- Datas de avaliação dos jogadores de 2007 a 2013 da tabela Match

Épocas	*Datas de avaliação*	
2007		2007/02/22
2007/2008	2007/08/30	2008/02/22
2008/2009	2008/08/30	2009/02/22
2009/2010	2009/08/30	2010/02/22
2010/2011	2010/08/30	2011/02/22
2011/2012	2011/08/30	2012/02/22
2012/2013	2012/08/31	2013/02/15

Tabela 11- Datas de avaliação dos jogadores desde 2013 a 2016

2013	*2014*	*2015*	*2016*
2013/02/22	2014/01/03	2015/01/02	2016/01/07
2013/03/01	2014/01/10	2015/01/09	2016/01/14
2013/03/08	2014/01/17	2015/01/16	2016/01/21
2013/03/15	2014/01/24	2015/01/23	2016/01/28
2013/03/22	2014/01/31	2015/01/26	2016/02/04
2013/03/28	2014/02/07	2015/01/28	2016/02/11
2013/04/05	2014/02/14	2015/01/30	2016/02/18
2013/04/12	2014/02/21	2015/02/06	2016/02/25
2013/04/19	2014/02/28	2015/02/13	2016/03/03
2013/04/26	2014/03/07	2015/02/20	2016/03/10
2013/05/03	2014/03/14	2015/02/27	2016/03/17
2013/05/10	2014/03/21	2015/03/06	2016/03/24
2013/05/17	2014/03/28	2015/03/13	2016/03/31
2013/05/24	2014/04/04	2015/03/20	2016/04/07
2013/05/31	2014/04/11	2015/03/27	2016/04/14
2013/06/07	2014/04/18	2015/04/10	2016/04/21
...	...	...	...

Os jogos iniciam em 2008, mas as avaliações dos jogadores iniciam-se desde 2007 a 2016.

Assim sendo, existem datas de avaliação antes e depois dos jogos onde é possível fazer a comparação das variáveis. A tabela *player_attributes* tem avaliação temporal, evolui ao longo do tempo. Assim é possível pegar num atributo e avaliá-lo ao longo do tempo para cada jogador.

4.2 **Metodologia**

Na elaboração da presente dissertação pretendeu-se analisar dados biométricos de jogadores de futebol profissionais a partir de uma base de dados pré-concebida e a implementação de mecanismos de *machine learning*.

Desta forma cada equipa seria capaz de melhorar a sua performance em jogo e conseguir uma melhor gestão do painel de forma a obter mais rapidamente o objetivo final, ganhar o jogo e consequentemente ter uma melhor prestação e classificação no final de

cada época.

O presente estudo apresenta os seguintes objetivos específicos demonstrados a Figura 8.

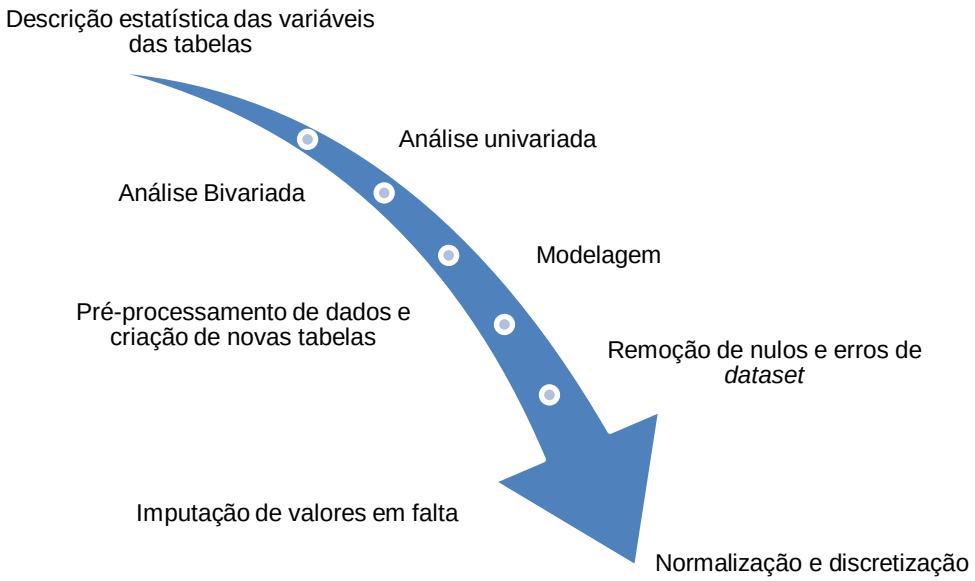


Figura 8- Análise e modelação dos passos da dissertação.

O primeiro objetivo passou pela definição das sete tabelas da base de dados de forma individual explicada no capítulo 4 em 4.1 para melhor compreender as principais características que possuem.

O segundo passo foi a análise univariada onde se fez a descrição estatística das tabelas. Em seguida, fez-se a análise bivariada utilizando a correlação para as diferentes variáveis de forma a compreender a interação entre elas e de que forma se influenciam uma à outra.

Logo depois, fez-se a modelação e pré-processamento dos dados onde se criou um novo *dataset* a partir das tabelas iniciais e logo após removeu-se os nulos e detetou-se os erros do *dataset*.

Posteriormente fez-se imputação de valores *missing* e normalização e discretização dos dados.

Os passos da Figura 8 são descritos de forma detalhada no Capítulo 5.

Tabela 12- Anos em divisão de dados

*2007-02-22 a 2007-08-30 → **2007***	*2012-02-22 a 2012-08-31 → **2012***
*2008-02-22 a 2008-08-30 → **2008***	2013-02-15 a 2013-08-30 → **2013**
*2009-02-22 a 2009-08-30 → **2009***	2014-02-07 a 2014-08-29 → **2014**
*2010-02-22 a 2010-08-30 → **2010***	2015-02-06 a 2015-08-27 → **2015**
*2011-02-22 a 2011-08-30 → **2011***	2016-02-04 a 2016-07-07 → **2016**

Em seguida, iniciou-se a avaliação de cada jogador por ano, de 2007 até 2008, de 2008 a 2009, e assim sucessivamente. Construiu-se um *dataset* em *csv* e abriu-se em *Orange*[2] para análise de gráficos. Porém devido há pouca informação que se conseguiria retirar a partir da tabela no *Orange* e de forma a simplificar todo o processo criaram-se tabelas individuais para casa atributo de avaliação.

As tabelas foram analisadas em *Python* usando ferramentas como o *Jupyter Notebook*[3]*, Pandas*[4]*, Matplotlib*[5] e *NumPy*[6]*.* As tabelas criadas apresentam as datas de avaliação dos jogadores , *player_fifa_api_id* e o atributo em avaliação.

Os dados analisados semanalmente acabam por simplificar e permitem validar mais valores.

O próximo passo passou pela análise por grupos de jogadores com a mesma função/posição em jogo. De forma a organizar esta análise fez-se a modelação onde se criou um modelo que explicasse as características de funcionamento e comportamento dos jogadores. Logo depois, fez-se o pré-processamento dos dados e gerou-se um novo *dataset* que continha as variáveis mais relevantes para estudo de acordo com a posição de cada grupo de jogadores e as datas de avaliação dos jogadores.

Na tabela elaborada removeram-se os nulos e erros de *dataset,* normalizaram-se os dados onde se implementou um conjunto de regras de maneira a organizar o projeto e de forma a reduzir a redundância de dados e aumentar a sua integridade e desempenho.

Em seguida fez-se a discretização que tem como objetivo transformar atributos contínuos em atributos categóricos, este método de discretização reduz e simplifica os dados com o intuito de tornar a aprendizagem mais rápida e com resultados mais compactos.

Analisaram-se os resultados da tabela elaborada anteriormente.

O último passo da metodologia foi a criação do modelo VAR e fez-se um *forecasting*

[2] https://orange.biolab.si/
[3] https://jupyter.org/
[4] https://pandas.pydata.org/
[5] https://matplotlib.org/
[6] https://numpy.org/

com o intuito de prever se o jogador deve ou não ser substituído.

4.3 **Métricas de avaliação**

O futebol oferece muitos desafios fascinantes para a aprendizagem de máquina. Na presente dissertação utilizaram-se ferramentas como *SQLite Studio, Orange Python* e dentro deste utilizaram-se o *Jupyter Notebook*, *Pandas, Matplotlib* e *NumPy* uma vez ser um conjunto de aprendizagem de máquina e *data mining.*

SQLite Studio foi usado para conversão de tabelas da base de dados de *sqlite* em *csv*.

O Orange apresenta várias tarefas, começando no pré-processamento dos dados, amostragem, filtragem, dimensão, discretização, combinações e interações de novos elementos, indução de modelos de classificação e regressão como árvores de classificação, redes *bayesianas*, regressão linear e logística.

Orange é uma ferramenta na qual o usuário insere *widgets* e gera um fluxo de tarefas de análise de dados. Os *widgets* apresentam funcionalidades como a leitura dos dados, tabelas, seleção de recursos, comparação de algoritmos, visualização de dados, etc. A ferramenta permite ao usuário explorar o programa de forma interativa, visual ou em subconjuntos de *widgets*.

Machine learning é maioritariamente utilizada por usuários experientes e novos pesquisadores que pretendam escrever scripts *Python* de forma a prototipar novos algoritmos. Oferece uma programação visual estruturada onde é possível formar diferentes combinações e interações entre todos os elementos em análise.

O *Jupyter Notebook,* foi criada para desenvolver código aberto, padrões abertos e serviços de computação interativa com as mais diferentes linguagens de programação. Na presente dissertação utilizou-se o *Python* como linguagem de programação e dentro deste bibliotecas como *Pandas, Matplotlib* e *NumPy.*

O *Pandas* é uma biblioteca de software usada para manipulação e análise de dados e oferece estruturas e operações para manipular tabelas numéricas e séries temporais.

O Matplotlib é uma biblioteca de software que cria gráficos e visualizações de dados em geral e a sua extensão de matemática *NumPy* que suporta *arrays* e matrizes multidimensionais e possui um enorme conjunto de funções matemáticas para analisar estas estruturas.

4.4 Método de validação dos modelos

Para validação do nosso modelo de classificação usou-se a ferramenta *Python* onde se analisou os jogadores de forma individual através de gráficos de linhas e de histogramas. Utilizou-se o método VAR como forma de obter informações sobre os jogadores de futebol e perceber se o jogador deve ou não ser substituído no próximo período de tempo.

Posteriormente, o *forecasting* para previsão do futuro com base em dados passados e presentes e mais comumente, por análise de tendências.

A análise de dados e descrição das tabelas foi descrita no presente capítulo assim como a metodologia utilizada. Ao longo da metodologia abordou-se como foi feita a análise dos dados que estão descritos de forma mais detalhada nos resultados explicados no capítulo seguinte.

5 Resultados

No presente capítulo pretende-se mostrar e calcular resultados relativos à análise de dados biométricos dos jogadores de futebol de forma a traçar o perfil do jogador para as diferentes variáveis.

Assim sendo, analisou-se de forma mais detalhada e calculada os passos de dissertação referidos na Figura 8.

Iniciou-se com a análise univariada onde se fez a descrição estatística, de seguida fez-se a análise bivariada onde se analisou a correlação entre variáveis.

O terceiro passo passou pela análise por jogador e a análise por grupos de jogadores divididos de acordo com a posição que ocupam em campo.

Por último e como forma de compreender todos os processos anteriores fez-se a aprendizagem onde se respondeu à pergunta sobre o jogador deve-se ou não jogar e quando é que este deve ser substituído.

5.1 Descrição estatística

Na Figura 9 abaixo indicada refere-se ao gráfico de distribuição para a coluna *birthday* que está representada na forma de idades. Utilizaram-se ferramentas *Excel* para transpor os dados de data de nascimento para a idade do jogador a 12 de Dezembro de 2016.

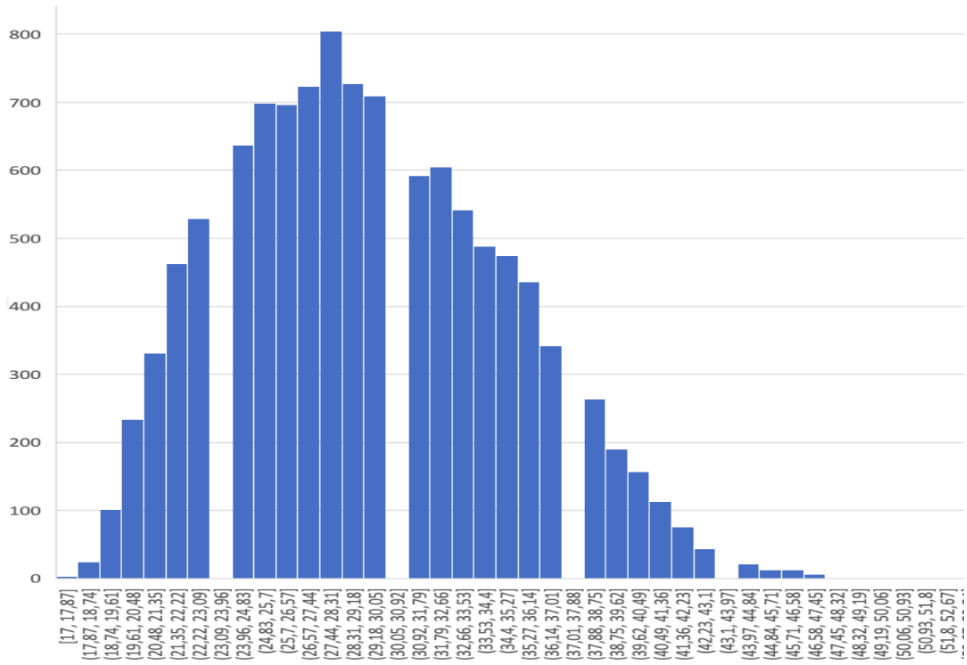
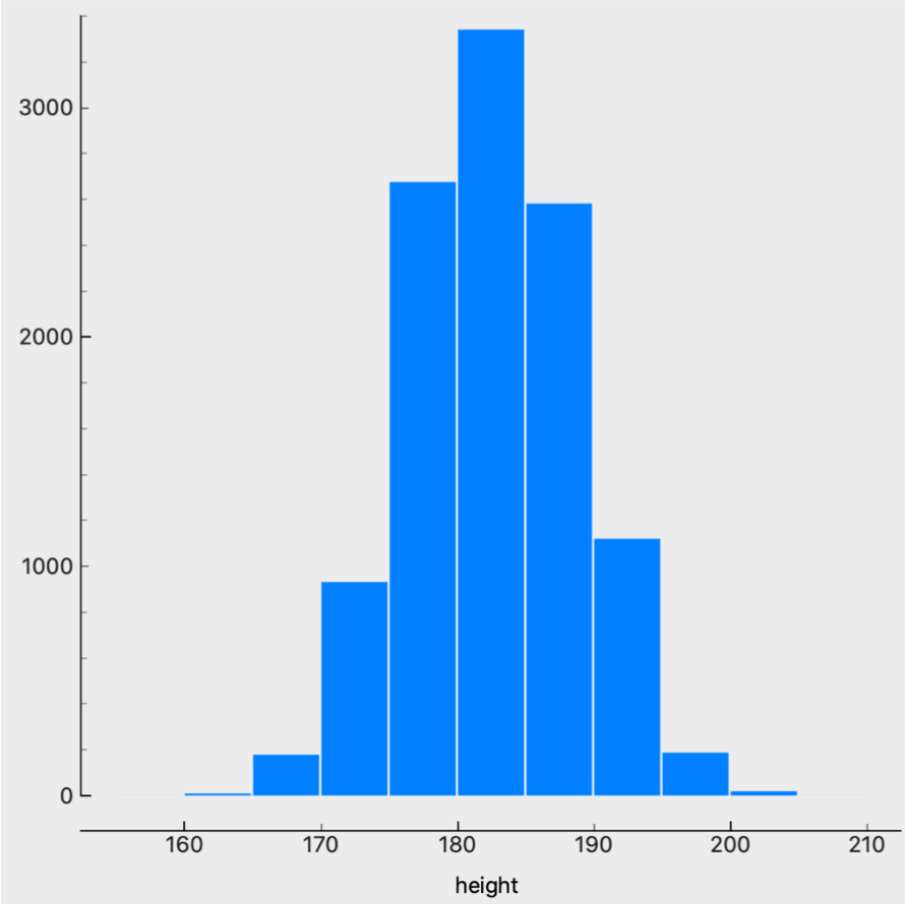


Figura 9- Gráfico de distribuição para a coluna birthday da tabela Player

Os jogos foram iniciados a 2008 e tiveram duração até 2016. Como os jogadores foram alterando ao longo das épocas é de esperar que jogadores mais velhos jogassem em 2008 e não em 2016 e os jogadores mais novos joguem nos anos mais recentes. Existe um elevado aumento do número de jogadores desde os 17 anos de idade até atingir idades entre os 23 e os 29 que são o que compreende um maior número de jogadores e um elevado decréscimo a partir dos 30 anos de idade em diante.

A Figura 10 mostra o gráfico para a altura dos jogadores assim como a altura mínima, máxima e valor médio e a Figura 11 mostra o *Boxplot*.



Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N height		181.867	0.0350195	157.48	208.28	0 (0%)

Figura 10- Gráfico de distribuição para a altura dos jogadores

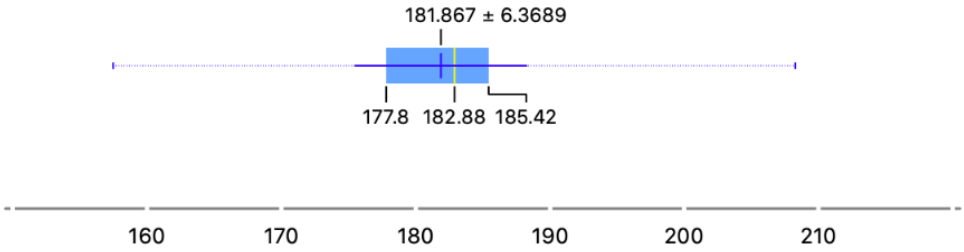


Figura 11- Boxplot da altura dos jogadores

Conclui-se que a altura mínima é de 1,57metros (m) e altura máxima de 2,08m e o valor médio é de 1,82m. O *Boxplot* da altura dos jogadores, mostra o 1º quartil de 177.80; mediana (2ºquartil) é de aproximadamente 182.88; 3ºquartil é 185.42.

A *Figura 12*, abaixo representada, mostra-nos o gráfico de distribuição para o peso dos jogadores assim como o peso mínimo, máximo e valor médio e a *Figura 13* mostra o *Boxplot.*

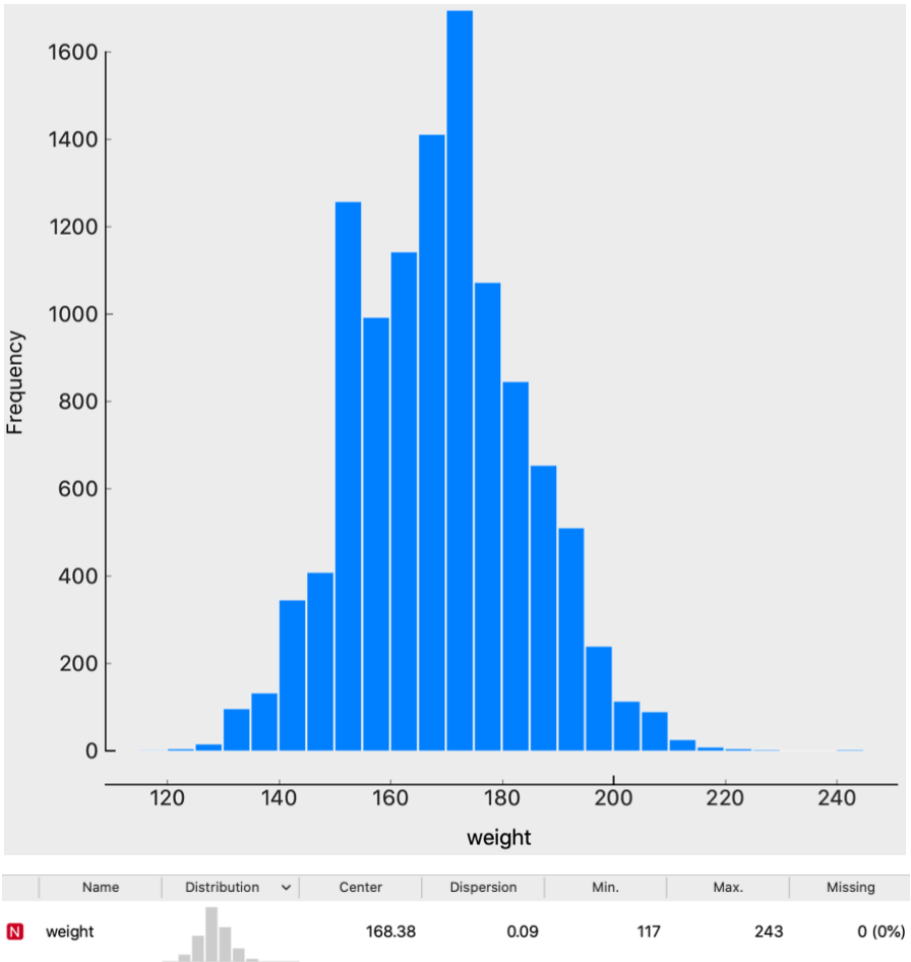


Figura 12- Gráfico da distribuição normal para o peso dos jogadores

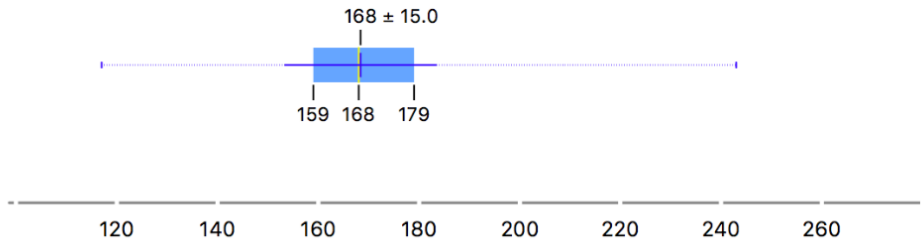


Figura 13- Boxplot do peso dos jogadores de futebol

Conclui-se que o peso mínimo é de 117lb, aproximadamente 53kg, peso máximo de 240lb, cerca de 110kg e valor médio 168lb aproximadamente 76kg. No *Boxplot* da *Figura 13* o 1º quartil é 159; mediana (2ºquartil) é de aproximadamente 168; 3ºquartil é 179.

5.2 **Análise bivariada**

A partir da tabela *player_attibutes* fez-se a correlação para cada variável no programa *Orange*. Esta análise foi feita por ano como mostra a **Error! Reference source not found.** demonstrada em 4.2 na metodologia.

O primeiro passo foi analisar o gráfico do perfil do jogador e gerou-se um *merge* das tabelas *player* e *player_attributes* e posteriormente dividiu-se a tabela *player_attributes* em anos. Tabela 1

Após dividir por anos fez-se a correlação com as diferentes variáveis de avaliação no *Orange* e analisou-se duas variáveis. Os valores mostram não existir correlação (r<0.5).

A correlação entre estas duas variáveis *stamina e agilidade* é positiva fraca. Os valores mínimos e máximo é de 0.27 e 0.45. No entanto, os valores aumentaram ao longo dos anos o que significa que quanto mais agilidade, mais rápido o jogador é no controle de bola e apresenta maior *stamina* (controla o cansaço do jogador durante o jogo) ou seja, ao longo dos anos os jogadores ficam cada vez mais rápidos e ágeis no manuseio da bola e a sua *stamina* ao longo dos anos também aumento o que significa que demora mais até o jogador se sentir cansado no jogo, ou seja, demora mais a que o seu rendimento comece a baixar.

A *stamina com o potential* apresenta uma correlação positiva fraca. Diminui ao longo dos anos com pequenas irregularidades. A *stamina* ao longo dos anos também

aumenta e o *potential* também aumenta, no entanto pouco.

Overall_rating e *Stamina,* conclui-se que o *overall_rating* diminui à medida que a *stamina* aumenta ao longo dos anos à exceção do ano 2014 que teve um pequeno aumento não significativo, com pequenas irregularidades. O *overall_rating* é a avaliação geral do jogador, ou seja, se a *stamina* aumenta, o valor de *overall_rating* também deveria aumentar o que não acontece. Então, conclui-se que não é possível fazer correlação entre estas duas variáveis.

Na *stamina com o sprint_speed* verifica a existência de uma correlação positiva não inferior a 0.5 e não superior a 0.6 ao longo dos anos, ou seja, com pouca variação. A *stamina* aumenta e o *sprint_speed* também aumenta ao longo dos anos.

Para o ano de 2007 → r= 0.52	Para o ano 2012 → 0.50
Para o ano de 2008 → 0.56	Para o ano 2013 → 0.55
Para o ano 2009 → 0.58	Para o ano 2014 → 0.55
Para o ano 2010 → 0.60	Para o ano 2015 → 0.53
Para o ano 2011 → 0.56	Para o ano 2016 → 0.54

A *stamina* com a *aceleração* verifica uma correlação positiva não inferior a 0.49 e não superior a 0.57 ao longo dos anos, ou seja, com pouca variação. À medida que a *stamina* aumenta a aceleração também aumenta. Isto pode-se dever ao ganho de experiência do jogador ao longos dos anos. O jogador apresenta uma maior resistência e demora mais a ficar cansado o que demonstra uma melhoria na sua performance e consequentemente uma melhor aceleração.

Para o ano de 2007 → r= 0.49	Para o ano 2012 → 0.48
Para o ano de 2008 → 0.54	Para o ano 2013 → 0.53
Para o ano 2009 → 0.55	Para o ano 2014 → 0.52
Para o ano 2010 → 0.57	Para o ano 2015 → 0.49
Para o ano 2011 → 0.53	Para o ano 2016 → 0.51

Na *stamina* com a *reaction* ao longo dos anos com exceção dos anos 2014 e 2015 verifica-se que a correlação entre a reação e a *stamina* diminui. Ou seja, quando menor a resistência e maior cansaço, menor será a reação do jogador.

Para o ano de 2007 → r= 0.50	Para o ano 2012 → 0.33
Para o ano de 2008 → 0.53	Para o ano 2013 → 0.31
Para o ano 2009 → 0.51	Para o ano 2014 → 0.39
Para o ano 2010 → 0.50	Para o ano 2015 → 0.33
Para o ano 2011 → 0.40	Para o ano 2016 → 0.31

A *Stamina* com *heading_accuracy* apresenta uma correlação positiva com valores aproximados ao longos dos vários anos com variação entre 0.41 e 0.55. Quanto maior a *stamina* maior a precisão de direção do jogador. O jogador ao se encontrar menos cansado ao longo dos anos e com a evolução de resistência acaba por ter uma melhor performance em jogo e consequentemente uma maior precisão de direção no cabeceamento.

Para o ano de 2007 → r= 0.50	Para o ano 2012 → 0.41
Para o ano de 2008 → 0.54	Para o ano 2013 → 0.47
Para o ano 2009 → 0.53	Para o ano 2014 → 0.48
Para o ano 2010 → 0.55	Para o ano 2015 → 0.43
Para o ano 2011 → 0.51	Para o ano 2016 → 0.48

Stamina com *jumping* apresenta uma correlação positiva fraca. Variando entre 0.21 e 0.31. Quando a *stamina* aumenta, o *jumping* também aumenta, ou seja, quando menor o cansaço do jogador mais resistência ele tem e maior salto ele consegue dar alcançando a bola e ganhando vantagem contra o adversário.

Stamina com *finishing,* apresenta uma correlação positiva fraca. Variando entre 0.30 e 0.38. Quando a *stamina* aumenta, o *finishing* também aumenta, ou seja, quando menor o cansaço do jogador e maior resistência, melhor será as finalizações do jogador.

Stamina e *crossing* apresentam uma correlação positiva com variação no intervalo

0.46 e 0.59 e não apresenta crescimento regular. Quanto maior a *stamina* melhor o jogador cruza a bola em jogo.

Para o ano de 2007 → r= 0.46	Para o ano 2012 → 0.54
Para o ano de 2008 → 0.51	Para o ano 2013 → 0.58
Para o ano 2009 → 0.53	Para o ano 2014 → 0.55
Para o ano 2010 → 0.58	Para o ano 2015 → 0.55
Para o ano 2011 → 0.58	Para o ano 2016 → 0.59

Stamina e passes longos mostra nos primeiros 3 anos a correlação diminuiu ligeiramente. Teve um aumento grande para 2010 e 2011. Em 2012 diminuiu novamente e de 2014 para 2015 volta a aumentar. Correlação positiva baixa o que significa que quanto maior a *stamina* mais resistência o jogador tem e mais e melhores passes longos consegue fazer.

Para o ano de 2007 → r= 0.36	Para o ano 2012 → 0.52
Para o ano de 2008 → 0.35	Para o ano 2013 → 0.54
Para o ano 2009 → 0.33	Para o ano 2014 → 0.49
Para o ano 2010 → 0.48	Para o ano 2015 → 0.56
Para o ano 2011 → 0.58	Para o ano 2016 → 0.56

Stamina e passes curtos, os valores da correlação variam de 0.53 a 0.64 notando-se alguma estabilidade entre os valores no intervalo de 0.60 a 0.64 por isso pode classificar como positiva tendencialmente forte. Ou seja, quanto maior a *stamina* maior e melhor serão os passes curtos do jogador.

Para o ano de 2007 → r= 0.53	Para o ano 2012 → 0.55
Para o ano de 2008 → 0.60	Para o ano 2013 → 0.59
Para o ano 2009 → 0.62	Para o ano 2014 → 0.59

Para o ano 2010 → 0.64	Para o ano 2015 → 0.63
Para o ano 2011 → 0.62	Para o ano 2016 → 0.64

Stamina e *gk_diving*, se o valor da *stamina* aumenta de uma maneira geral a outra variável diminui. Apresenta uma correlação negativa que começa a ser expressiva o que quer dizer que o aumento da *stamina* poderá ter uma influência negativa.

Stamina e *gk_handling*, se o valor da *stamina* aumenta de uma maneira geral a outra variável diminui. Apresenta uma correlação negativa que começa a ser expressiva o que quer dizer que o aumento da *stamina* poderá ter uma influência negativa.

Stamina e gk_kicking, entre os anos 2007 – 2009 a correlação é positiva mais muito fraca. No ano 2010 passou a negativa fraca e a partir de 2011 a correlação manteve-se negativa, mas aumentou o seu valor absoluto variando no intervalo de -0.60 a -0.50 com exceção do ano 2014 em que a correlação diminuiu apesar de negativa. Isto levamos a concluir que nos anos 2011 a 2016 (exceto 2014) o aumento da *stamina* conduz a uma diminuição da outra variável.

Stamina e *gk_positioning***,** correlação negativa, mas mais expressiva a partir de 2011, com exceção do ano 2014. À medida que a *stamina* diminui o posicionamento do guarda-redes aumenta, ou seja, quando menor resistência o jogador tem, melhor será o seu posicionamento em jogo.

Stamina com *gk_reflexes*, correlação negativa, mas mais expressiva a partir de 2011, com exceção do ano 2014. À medida que a *stamina* diminui os reflexos do guarda-redes aumentam, ou seja, quando menor resistência o jogador, melhor serão os reflexos do jogador em jogo.

Balance e *potential* têm correlação positiva fraca. Ainda menos expressiva a partir de 2012. *Balance* aumenta à medida que o *potential* também aumenta.

Balance e stamina, positiva com correlação moderada. Equilíbrio e coordenação com a stamina que determina o cansaço do jogador ao longo tempo. Ou seja, quanto maior a resistência maior o seu balance.

Balance e *overall_rating*, a partir de 2010 e com grande incidência a partir de 2011 conclui-se a inexistência de correlação entre estas duas variáveis.

Balance e *heading_accuracy*, a correlação foi positiva e com alguma expressão de 2007 até 2010 e a partir de 2011 começou a descer tendo atingido valores negativos a partir de 2014. *Balance* é o equilíbrio do jogador e *heading_accuracy* a precisão do jogador ao cabecear a bola, ou seja, se estes apresentam uma correlação positiva significa que à

medida que um aumenta o outro também aumenta.

Potential e *heading_accuracy***,** correlação positiva muito fraca que diminui ao longo dos anos com exceção de 2014. Ou seja, à medida que o *potential* do jogador aumenta a precisão de cabecear a bola também aumenta.

Potential e *overall_rating,* correlação positiva forte e estável. O *potential* de um jogador aumenta de acordo com a sua experiência, quanto maior a experiência maior o *potential* e o *overall_rating* é a avaliação geral do jogador. Significa que quanto maior a experiência do jogador melhor será a sua avaliação geral.

Para o ano de 2007 → r= 0.75	Para o ano 2012 → 0.79
Para o ano de 2008 → 0.79	Para o ano 2013 → 0.79
Para o ano 2009 → 0.80	Para o ano 2014 → 0.77
Para o ano 2010 → 0.80	Para o ano 2015 → 0.78
Para o ano 2011 → 0.79	Para o ano 2016 → 0.77

Potential e agilidade, ao longo dos anos o *potential* diminui e a agilidade aumenta. Correlação positiva fraca, sem grande oscilação.

Overall_rating e *aceleração,* correlação positiva que tende a diminuir ao longo dos anos com exceção de 2014. O *overall_rating* aumenta e a aceleração diminui.

Overall_rating e *sprint_speed,* correlação positiva que tende a diminuir ao longo dos anos com exceção de 2014. O *overall_rating* aumenta e o *sprint_speed* diminui

Overall_rating e agilidade, correlação positiva fraca. O *overall_rating* aumenta e a agilidade diminui.

Overall_rating e reação, correlação positiva forte e estável. Os valores apresentam tendência a aumentar ao longo dos anos. O *overall_rating* significa a avaliação geral do jogador o que significa que a reação aumenta á maneira que a avaliação geral também. Quanto maior a reação do jogador maior a sua avaliação geral.

Para o ano de 2007 → r= 0.66	Para o ano 2012 → 0.76
Para o ano de 2008 → 0.71	Para o ano 2013 → 0.79
Para o ano 2009 → 0.70	Para o ano 2014 → 0.75

Para o ano 2010 → 0.71	Para o ano 2015 → 0.81
Para o ano 2011 → 0.73	Para o ano 2016 → 0.82

Overall_rating e *jumping*, correlação positiva fraca. Significa que a avaliação geral do jogador aumenta da mesma forma que o *jumping* também aumenta. Quanto maior o salto do jogador maior a sua avaliação geral.

Overall_rating e força, correlação positiva fraca. Valores iniciais mais estáveis que tende a diminuir. Significa que a avaliação geral do jogador aumenta da mesma forma que a força também aumenta. Quanto mais força o jogador tem maior a sua avaliação geral.

Overall_rating e volleys, correlação positiva fraca sem grande oscilação. Significa que a avaliação geral do jogador aumenta da mesma forma que os *volleys* também aumenta.

Overall_rating e *finishing*, correlação positiva fraca. Ou seja, a avaliação geral do jogador aumenta da mesma forma que os *finishing* também aumentam.

Overall_rating e *heading_accuracy,* correlação positiva fraca. Valores diminuem ao longo dos anos exceto 2014 e 2015. Ou seja, a avaliação geral do jogador aumenta à medida que a precisão de cabecear a bola também aumenta.

Overall_rating e crossing, correlação positiva fraca. Significa que a avaliação geral do jogador aumenta da mesma forma que os *crossing* do jogador também aumenta.

Overall_rating e *dribbling*, correlação positiva fraca. Ou seja, a avaliação geral do jogador aumenta da mesma forma que o *dribbling* também aumenta.

Stamina com os passes longos, correlação positiva que tende a aumentar a longo dos anos. significa que quando maior a *stamina* a resistência é maior e consequentemente os passes longos são melhores, resultando de uma maior performance do jogador.

Stamina com os passes curtos, correlação positiva. Significa que quando maior a *stamina* a resistência é maior e consequentemente os passes curtos são melhores, resultando de uma maior performance do jogador.

Para o ano de 2007 → r= 0.53	Para o ano 2012 → 0.55
Para o ano de 2008 → 0.60	Para o ano 2013 → 0.59
Para o ano 2009 → 0.62	Para o ano 2014 → 0.59
Para o ano 2010 → 0.64	Para o ano 2015 → 0.63

Para o ano 2011 → 0.62 Para o ano 2016 → 0.64

Em suma, as correlações que mais se relacionam, positivas fortes, são do *potential* com o *overall_rating* e *overall_rating* com a *reação*.

O *overall_rating* tende a diminuir ao longo dos anos quando feita a correlação com variáveis como aceleração, *sprint_speed*, etc.

As variáveis que apresentam correlação positiva fraca, mas que aumenta ao longo dos anos são a *stamina* com a *agilidade*.

As que se matem ao longo dos anos são *a stamina* com o *finishing, o potential* com *agilidade, overall_rating* com a *agilidade, overall_rating* com *volleys, overall_rating* com *finishing, overall_rating* com a *crossing e overall_rating* com a *dribbling;*

A correlação que diminui ao longo dos anos para a *stamina* com *potential, stamina* com *overall_rating, stamina* com *jumping, balance* com *potential, balance* com *overall_rating, potential* com *heading_accuracy, overall_rating* com *aceleração, overall_rating* com *sprint_speed, overall_rating* com *jumping, overall_rating* com *força e overall_rating* e *heading_accuracy*.

A correlação positiva que aumenta ao longo dos anos são a *stamina* com passes longos e os que se mantem mais ou menos igual ao longo dos anos são *a stamina* com *sprint_speed, a stamina* com aceleração*, stamina* com *heading_accuracy, stamina* com *crossing, stamina* com passes curtos e *stamina* com *balance*.

As que diminuem são a *stamina* com a reação e o *balance* com *heading_accuracy* até ficar negativo.

A correlação positiva forte são o *potential* com *overall_rating* que diminui ao longo do tempo e o *overall_rating* com a reação com se mantem.

A correlação negativa *stamina* com *gk_diving, gk_handlig, gk_kicking, gk_positioning, gk_reflexes.*

Em todas as correlações a *stamina* aumenta ao longo dos anos e a outra variável diminui. Significa que o aumento da *stamina* pode influenciar de forma negativa a prestação do jogador em campo. A *stamina* é o cansaço, a resistência, ou seja, á medida que os anos passam a *stamina* aumenta o que quer dizer que o jogador ganha resistência e demora mais a ficar cansado.

5.3 **Análise por jogador**

A análise de cada jogador individualmente iniciou-se a partir da tabela *player_attibutes* com a data de avaliação de cada jogador por ano. Em seguida, construiu-se um *dataset* em *csv* que continha informação sobre as datas de avaliação dos jogadores, o atributo e a identificação de cada jogador e abriu-se em *Orange.* De forma a simplificar e a obter uma maior informação das tabelas em estudo utilizou-se o *Python* e usaram-se ferramentas como o *Jupyter Notebook, Pandas, Matplotlib* e *NumPy* de maneira a obter mais informações sobre os atletas.

Os dados analisados semanalmente acabam por simplificar e permitem validar mais valores do que o uso contínuo uma vez que existe uma discrepância de valores uma vez que os pontos não são nas mesmas alturas.

Os gráficos gerados em *Python* continham as datas de avaliação no eixo do X, e a variável em estudo no eixo dos Y. As linhas indicavam os jogadores onde cada um deles apresentava uma cor de linha diferente e comparou-se os seus atributos com as datas de avaliação.

Pegou-se nos jogadores (*player_fifa_api_id*), uma data e uma variável associada, por exemplo *stamina* e elaborou-se uma nova tabela *csv* onde se estudou a sua evolução ao longo do tempo em *Python* e observou-se o gráfico de linhas como mostra a Figura 15.

No *Python* fez-se uma avaliação do comportamento do jogador, de forma a compreender a diferença de comportamentos de jogador para jogador. O grande objetivo desta análise é identificar jogadores que tenham comportamento "acidentado", ou seja, um comportamento muito diferente dos outros e verificar a existência de jogadores que tenham um comportamento muito diferente dos outros e escolher alguns jogadores que tenham grandes variações.

Assim, seria possível criar um modelo para um jogador com o objetivo de analisar e distinguir um jogador que apresente um comportamento muito diferente dos outros e os que têm um comportamento parecido. Esta análise é importante para os clubes a fim de perceber o comportamento do jogador e decidir se devem ou não investir nele, se vendem ou não, etc.

Esses jogadores são importantes para saber se o treinador na hora de tomar decisões e opta ou não por uma substituição e o que acontece com o jogador no futuro como explicado no capítulo 5.5 em 5.5.3.

Utilizou-se a ferramenta *Python* para perceber o comportamento dos jogadores e verificar se existem diferenças notórias nas avaliações, se existem grupos de jogadores que têm comportamento muito diferente dos outros e comparar datas uma vez que um

jogador que tem um bom resultado de *stamina* por exemplo pode não ter jogado e ter uma boa resistência e apresentar valores de atributos melhores do que os jogadores que jogaram.

Para essa análise pegou-se numa variável como a *stamina* por exemplo e para todos os jogadores e viu-se como é que essa variável se altera ao longo do tempo. Analisou-se e replicou-se o processo para as variáveis que se considerem ser mais relevantes de momento e assim como forma de previsão agrupou-se os jogadores independentemente da equipa e de acordo com todo o histórico anterior e avalisou-se em 5.5.3 de forma exclusiva os jogadores e a sua possibilidade de ser ou não substituídos, avaliando a performance do jogador ou o passado destes.

O primeiro passo, passou por importar para o *python* todas as ferramentas que se usou para a análise estatística das tabelas como o *NumPy, Matplotlib, Pandas* e *SciPy.*

Em seguida na escreveu-se o "caminho" para a pasta onde tem os ficheiros *csv* que vão ser utilizados.

Posteriormente, leu-se o ficheiro *csv* e colocou-se os dados dentro do *dataframe* iniciado anteriormente e visualizou-se os cinco primeiros elementos da tabela em análise que forneceu indicação sobre qual o tipo de data e qual é que são as informações existentes na tabela e que servirão para estudo e análise desta como mostra a Figura 14.

	player_fifa_api_id	22/02/07 00:00	30/08/07 00:00	22/02/08 00:00	30/08/08 00:00	22/02/09 00:00	30/08/09 00:00	22/02/10 00:00	30/08/10 00:00	22/02/11 00:00	...
0	218353	62.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1	189615	64.0	NaN	NaN	64.0	64.0	64.0	63.0	64.0	NaN	...
2	186170	59.0	NaN	NaN	NaN	59.0	59.0	59.0	59.0	59.0	...
3	140161	70.0	70.0	NaN	70.0	72.0	71.0	69.0	NaN	NaN	...
4	17725	76.0	76.0	NaN	76.0	77.0	77.0	77.0	77.0	75.0	...

Figura 14- Extrato das primeiras 5 linhas das tabelas em análise.

Posto isto, o principal interesse foram os valores dos atributos e assim sendo, retirou-se a primeira coluna que faz referência ao *player_fifa_api_id* e criou-se um novo *dataframe* de forma a que seja visualizada e analisada o excerto da tabela em interesse.

Foi gerado o gráfico de linhas para todas as variáveis em estudo a Figura 15 mostra o gráfico para a variável *stamina*. As imagens apresentam nos eixo dos X as datas de avaliação dos jogadores de futebol e no eixo dos Y a variável em estudo. Cada jogador está referenciado com uma cor diferente nos gráficos.

Após gerar o gráfico para cada variável temos a lista (*df)* que é a *player_fifa_api_id*

e criou-se uma nova lista *df2* para cada linha da tabela apenas com as datas de avaliação e os valores dos atributos.

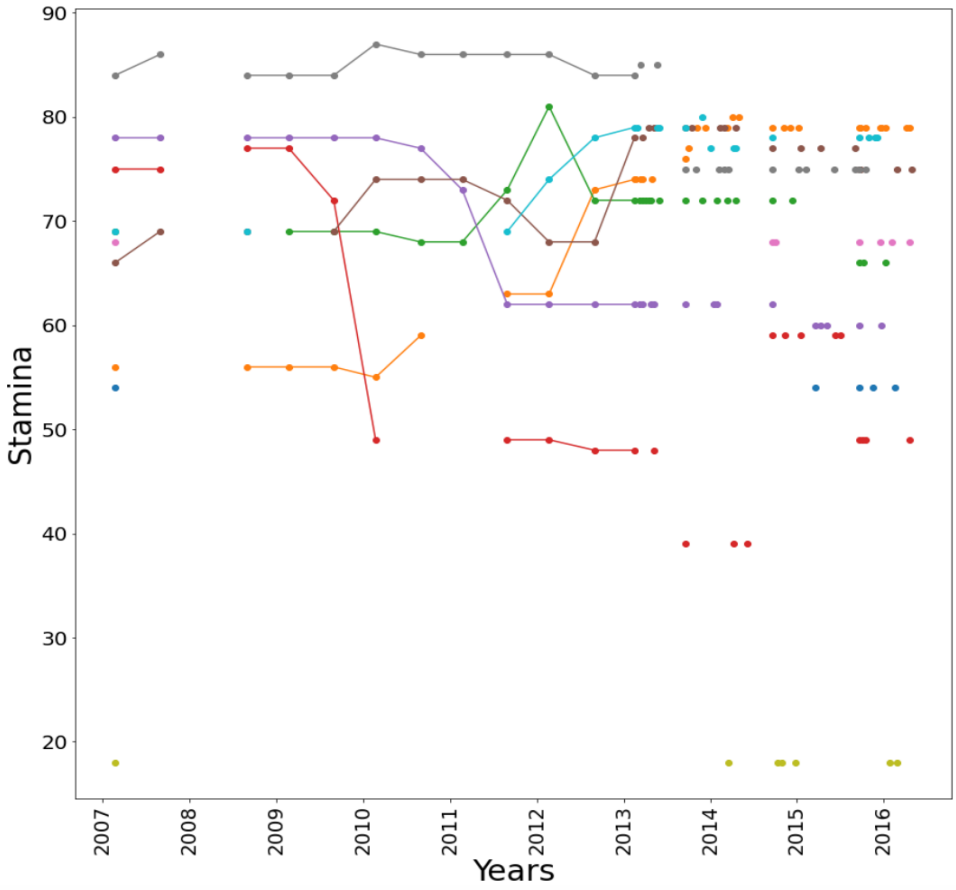


Figura 15- Gráfico de linhas para avaliação do jogadores.

Em *df2* cada linha deu-se o nome de lista *df2.loc*. com os *NaN's* (valores nulos) incluídos. O objetivo é observar a diferença absoluta entre o máximo e o mínimo da linha e se é maior que o critério que se impos na análise. Em seguida, e de maneira a ficar apenas os valores não nulos da tabela fez-se *skipna* e removeu-se todos os *NaN's.*

Como forma de obter resultados das tabelas impôs-se condições de se a primeira condição fosse favorável passaria a executar o segundo *if,* se não fosse favorável passa para a *else*. Ou seja, se não for uma opção, automaticamente seria a outra e se for a primeira opção não o código não executa mais e fica apenas com a primeira opção.

Concluindo, se o tempo do máximo for maior que o tempo do mínimo significa dizer que o jogador subiu de forma e o identificador do jogador inclui-se nos acidentados positivos. A segunda opção (*else*) foi o tempo do máximo ser menos que o tempo do mínimo o que significa dizer que o jogador desceu de forma e será um jogador acidentado negativo.

Para que a função retorne os resultados temos todos os jogadores acidentados e esses são divididos em jogadores positivos ou jogadores negativos.

O resultado[0] da função mostra todos os jogadores acidentados e len (resultado[0]) a quantidade de valores acidentados o resultado[1] mostra os jogadores_positivos e len (resultado[1]) quantidade e os resultado[2] são os jogadores_negativos e len (resultado[2]) quantos jogadores são.

Análise de jogadores e ver quantos pontos os jogadores acidentados subiram e desceram acima do critério, ou seja, o critério é 30 e analisou-se acima do critério para os jogadores acidentados todos, positivos e negativos.

Para elaborar o histograma é necessário ter informações como o valor máximo e mínimo e de forma a obter essa informação auxiliamos da ferramenta *python* e observou-se os valores para cada uma das análises para os jogadores aumentados positivamente e negativamente.

Em seguida, fez-se o mesmo procedimento para os dados em bruto e viu-se o aumento positivamente e negativamente e por ultimo o mesmo processo para jogadores totais onde se viu o histograma e fez-se uma distribuição normal.

As variáveis em avaliação foram a *acceleration, aggression, agility, balance, ball_control, crossing, curve, dribbling, finishing, free_kick_accuracy, heading_accuracy, interceptions, jumping, long_passing, short_passing, long_shots, overall_rating, penalties, positioning, potential, reactions, shot_power, sliding_tackle, sprint_speed, stamina, standing_tackle, strength, vision, volleys, gk_diving, gk_handling, gk_positioning, gk_kicking e gk_reflexes.*

Para a *Acceleration* os jogadores são:

- Resultado [0] - jogadores acidentados - são [180440, 205947, 188770, 194769, 201223, 201818, 188132, 109693, 198117, 142359, 157945, 202605, 150267, 204830, 49026, 20551, 30697, 183361,...]

- Len (resultado[0]) - jogadores acidentados - 667.

- Resultado [1] - jogadores_positivos - são [180440, 205947, 188770, 194769, 201223, 201818, 202605, 204830, 226221, 199763, 186521, 215455, 192521, 190852, 163858, 163713, 211673, 202770,...]

- Len (resultado[1]) - jogadores_positivos - são 169 jogadores.

- Resultado[2] - jogadores_negativos - são [188132, 109693, 198117, 1423

59, 157945, 150267, 49026, 20551, 30697, 183361, 7631, 135466, 18906
8,107280, 6862, 121591, 46815, 200818,...]

- Len (resultado[2])] - jogadores_negativos - são 498.

A análise para jogadores aumentados positivamente o aumento máximo foi 55.0, o aumento mínimo foi 30.0, a média do aumento foi 34.32 e o desvio padrão do aumento foi 4.69

Para a análise de jogadores aumentados negativamente, a diminuição máxima foi de -30.0, a diminuição mínima foi -59.0, a média da diminuição foi -36.87, o desvio padrão da diminuição foi 5.97.

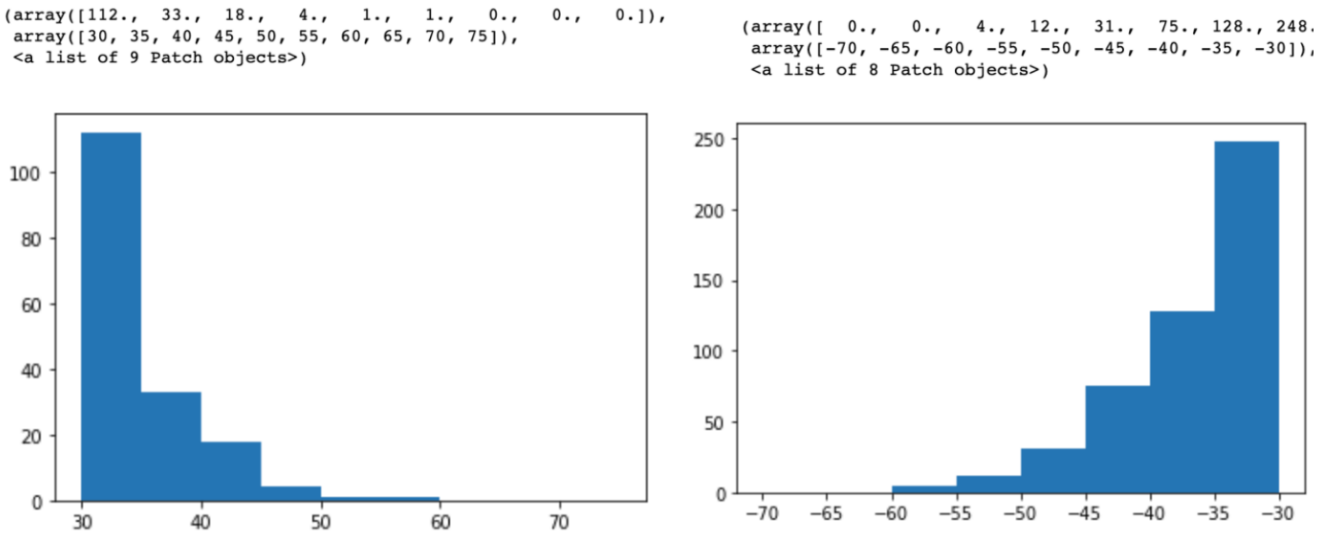


Figura 16- Histograma de Acceleration para os jogadores acidentados positivamente e negativamente.

Os histogramas representados na Figura 16Figura 16Figura 16 indicam o número de jogadores por intervalos definido de 30 a 80 para acidentados positivos e -70 a -30 para acidentados negativos ambos com intervalos de 5 em 5. Entre 30 a 35 há 112 jogadores, 35 a 40 há 33, e assim sucessivamente.

Em seguida, analisou-se os dados em bruto para os jogadores aumentados positivamente e negativamente e gerou-se um novo gráfico.

Os jogadores aumentados positivamente o aumento máximo foi de 55.0, o aumento mínimo foi de 1.0, a média do aumento foi 12.40 e o desvio padrão do aumento foi 7.93.

Os jogadores aumentados negativamente tiveram uma diminuição máxima de -0.0, uma diminuição mínima de -59.0, uma média de diminuição de -10.95 e um desvio padrão

da diminuição de 11.18.

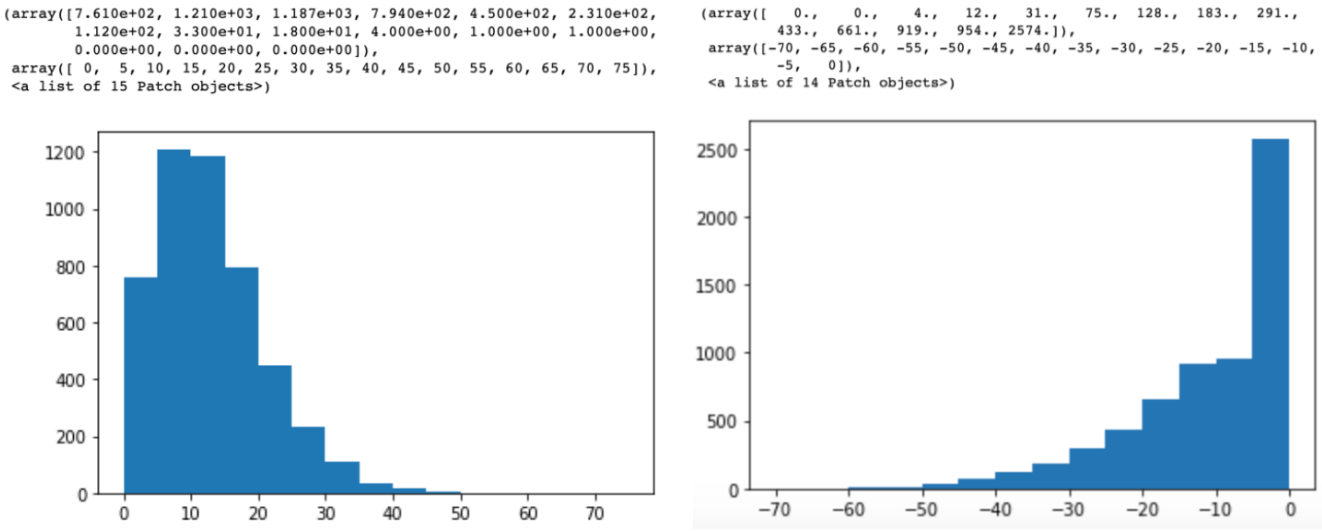


Figura 17- Histograma dos dados brutos para os jogadores aumentados positivamente e negativamente.

Os histogramas representados na Figura 17 indicam que entre 0 a 5 há 761 jogadores acidentados, de 5 a 10 há 1210, e assim sucessivamente. O intervalos definido foi de 0 a 80 para todos os jogadores acidentados com intervalos de 5 em 5.

Por fim, para todos os dados (aumento total) o máximo dos dados foi de 55.0, o mínimo dos dados de -59.0, a média dos dados de -0.82 e o desvio padrão dos dados de 15.23.

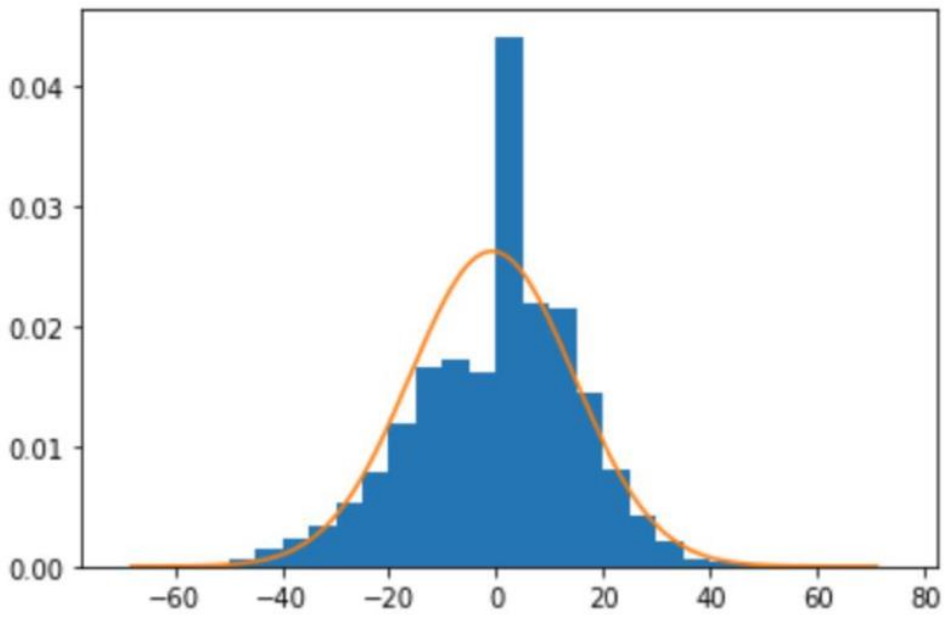


Figura 18- Histograma e distribuição normal com o aumento total dos dados para a Acceleration.

A Figura 18Figura 18 mostra o histograma para o aumento total dos dados para o atributo *acceleration* e a distribuição normal desses dados.

Á exceção do intervalo 0 a 5 que se verifica uma simetria da curva, ou seja, existe quase a mesma quantidade de jogadores positivos e negativos.

Para o atributo *Aggression* o jogadores acidentados são 1212, os jogadores positivos são 950 e os jogadores negativos são 212. A análise dos jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 30.0, a média do aumento foi 38.32 e o desvio padrão do aumento foi 7.37.

Os jogadores aumentados negativamente tiveram uma diminuição máxima de -30.0, uma diminuição mínima de -67.0, a média da diminuição foi -37.25 e o desvio padrão da diminuição foi 6.89.

Analisou-se os dados brutos para jogadores aumentados positivamente e negativamente. Os jogadores aumentados positivamente tiveram um aumento máximo de 70.0, um aumento mínimo de 1.0, uma média do aumento de 15.68 e um desvio padrão do aumento de 12.27.

Os jogadores aumentados negativamente tiveram um aumento máximo de -0.0, uma diminuição mínima de -67.0, uma média da diminuição de -8.08 e um desvio padrão da diminuição de 10.62.

Para o aumento total de todos os dados o máximo dos dados é 70.0, mínimo do s dados é -67.0, a média dos dados é 6.23 e o desvio padrão dos dados é 16.45.

Para o atributo *Agility* o jogadores acidentados são 220, os jogadores positivos s ão 98 e os jogadores negativos são 122.

A análise de jogadores aumentados positivamente o aumento máximo foi 50.0, o aumento mínimo foi 30.0, a média do aumento foi 33.87, o desvio padrão do aumento f oi 4.18.

A análise para jogadores aumentados negativamente a diminuição máxima foi de -30.0, a diminuição mínima foi -53.0, a média da diminuição foi -34.39, o desvio padrão da diminuição foi 4.19.

Para a análise de dados brutos os jogadores aumentados positivamente o aumento máximo foi 50.0, o aumento mínimo foi 1.0, a média do aumento foi 10.40 e o desvio padrão do aumento foi 7.43.

Para os jogadores aumentados negativamente a diminuição máxima foi -0.0, a diminuição mínima foi -53.0, a média da diminuição foi -6.11 e o desvio padrão da diminuição foi 8.12.

Na análise de todos os dados relativos a *Agility* o máximo dos dados é 50.0, o

mínimo dos dados é -53.0, a média dos dados é 0.62 e o desvio padrão dos dados é 11.29.

Para o *balance* os jogadores acidentados são 864, os jogadores positivos são 310 e os jogadores negativos são 554.

Na análise de jogadores aumentados positivamente o aumento máximo foi 66.0, o aumento mínimo foi 30.0, a média do aumento foi 37.05 e o desvio padrão do aumento foi 6.82.

A análise de jogadores aumentados negativamente a diminuição máxima foi -30.0, a diminuição mínima foi -79.0, a média da diminuição foi -38.58 e o desvio padrão da dim inuição foi 7.33.

Os dados brutos para os jogadores aumentados positivamente o aumento máximo foi 66.0, o aumento mínimo foi 1.0, a média do aumento foi 12.69 e o desvio padrão do aumento foi 10.260.

A análise jogadores aumentados negativamente a diminuição máxima foi -0.0, a di minuição mínima foi -79.0, a média da diminuição foi -9.61 e o desvio padrão da diminuiçã o foi 12.09

Para todos os dados o máximo dos dados é 66.0, o mínimo dos dados é -79.0, a média dos dados é -1.32 e o desvio padrão dos dados é 15.72.

O ball_control os jogadores acidentados foram 283, os jogadores positivos foram 2 59 e os jogadores negativos foram 24.

A análise para jogadores aumentados positivamente o aumento máximo foi 63.0, o aumento mínimo foi 30.0, a média do aumento foi 35.87 e o desvio padrão do aumento fo i 5.99.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -56.0, a média de diminuição foi -37.88 e o desvio padrão da diminuição foi 8.52

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 63.0, o aumento mínimo foi 1.0, a média do aumento foi 12.17 e o desvio padrão do aum ento foi 8.42.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -56.0, a média da diminuição foi -6.18 e o desvio padrão da di minuição foi 6.59.

A análise relatava a todos os dados juntos teve um máximo dos dados de 63.0, um mínimo dos dados de -56.0, uma média dos dados de 5.31 e um desvio de padrão do s dados de 11.80.

O crossing tem 806 jogadores acidentados, 711 jogadores positivos e 95 jogado res negativos.

A análise para jogadores aumentados positivamente teve um aumento máximo de 65.0, um aumento mínimo de 30.0, uma média do aumento de 36.77 e um desvio padrão do aumento de 6.42.

A análise para jogadores aumentados negativamente teve uma diminuição máxi ma de -30.0, diminuição mínima de -56.0, uma média da diminuição de -34.68 e um desvi o padrão da diminuição de 4.91.

Os dados brutos para jogadores aumentados positivamente tiveram um aument o máximo de 65.0, um aumento mínimo de 1.0, uma média do aumento de 14.69e um des vio padrão do aumento de 10.72.

Para jogadores aumentados negativamente a diminuição máxima foi -0.0, a dimi nuição mínima foi -56.0, a média da diminuição foi -6.85 e o desvio padrão da diminuição foi 8.24.

Para todos os dados juntos o máximo dos dados é 65.0, o mínimo dos dados é -5 6.0, a média dos dados é 6.22 e o desvio padrão dos dados é 14.39.

Para análise do atribute *curve* os jogadores acidentados 344, os jogadores positiv os foram 329 e os jogadores negativos foram 15.

A análise para jogadores aumentados positivamente, o aumento máximo foi 60.0, o aumento mínimo foi 30.0, a média do aumento foi 35.98 e o desvio padrão do aumento f oi 5.45

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -50.0, a média da diminuição foi -37.27e o desvio padrão da di minuição foi 6.58.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 60.0, o aumento mínimo foi 1.0, a média do aumento foi 10.47 e o desvio padrão do aum ento foi 9.85.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -50.0, a média da diminuição foi -2.59 e o desvio padrão da dimi nuição foi 5.04

Para todos os dados juntos o máximo dos dados é 60.0, o mínimo dos dados é -5 0.0, a média dos dados é 4.48 e o desvio padrão dos dados é 10.32

Para o dribbling a 580 são jogadores acidentados dos quais 518 são positivos e 6 2 são negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 63.0, o aumento mínimo foi 30.0, a média do aumento foi 36.99 e o desvio padrão do aumento fo i 6.36.

A Análise para jogadores aumentados negativamente a diminuição máxima foi -

30.0, a diminuição mínima foi -50.0, a média da diminuição foi -35.06 e o desvio padrão da diminuição foi 5.36.

Os dados brutos para jogadores aumentados positivamente tiveram um aumento máximo de 63.0, um aumento mínimo de 1.0, uma média do aumento de 13.61 e o desvi o padrão do aumento foi 9.76

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -50.0, a média da diminuição foi -7.21 e o desvio padrão da di minuição foi 7.70

Na analise de todos os dados refentes ao atributo *dribbling* o máximo é 63.0 , o mí nimo é -50.0, a média é 5.57 e o desvio padrão é 13.57.

Para *finishing* os jogadores acidentados foram 645, os jogadores positivos foram 4 50 e os jogadores negativos foram 195.

A análise para jogadores aumentados positivamente aumento máximo foi 69.0, au mento mínimo foi 30.0 média do aumento foi 36.54 e o desvio padrão do aumento foi 6.17 .

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -56.0, a média da diminuição foi -35.12 e o desvio padrão da diminuição foi 5.73.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 69.0, o aumento mínimo foi 1.0, a média do aumento foi 13.43 e o desvio padrão do aumento foi 9.98.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -56.0, a média da diminuição foi -8.60 e o desvio padrão da dimi nuição foi 9.29.

Para todos os dados máximo dos dados é 69.0, o mínimo dos dados é -56.0, a mé dia dos dados é 3.35 e o desvio padrão dos dados é 14.63.

O atributo *free_kick_accuracy*, tem 1350 jogadores acidentados, 651 jogadores positivos e 699 jogadores negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 30.0, a média do aumento foi 37.57 e o desvio padrão do aumento foi 7.30.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -83.0, a média da diminuição foi -43.33 e o desvio padrão da diminuição foi 11.52.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 1.0, a média do aumento foi 13.11 e o desvio padrão do aum

ento foi 12.09.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -83.0, a média da diminuição foi -11.27 e o desvio padrão da di minuição foi 15.20.

Para todos os dados juntos o máximo dos dados é 70.0, o mínimo dos dados é -8 3.0, a média dos dados é 1.32 e o desvio padrão dos dados é 18.32.

O atributo *heading_accuracy* tem 480 jogadores acidentados dos quais 403 são p ositivos e 77 negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 30.0, a média do aumento foi 38.12 e o desvio padrão do aumento fo i 8.05.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -66.0, a média da diminuição foi -35.79 e o desvio padrão da diminuição foi 6.54.

Os dados brutos do atributo *heading_accuracy* para jogadores aumentados positiv amente tiveram um aumento máximo de 70.0, um aumento mínimo de 1.0, a média do au mento foi 12.85 e o desvio padrão do aumento foi 10.04.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -66.0, a média da diminuição foi -6.99 e o desvio padrão da dimi nuição foi 8.09.

Para todos os dados o máximo é 70.0, o mínimo é -66.0, a média é 4.27 e o desvi o padrão é 13.49.

Para o atributo *interceptions* os jogadores acidentados são 2135, os jogadores pos itivos são 763 e os jogadores negativos são 1372.

A análise para jogadores aumentados positivamente o aumento máximo foi 71.0, o aumento mínimo foi 30.0, a média do aumento foi 38.13 e o desvio padrão do aumento fo i 7.19.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -78.0, a média da diminuição foi -40.88 e o desvio padrão da di minuição foi 8.64.

Os dados brutos relativo ao atributo *interceptions* o aumento máximo foi 71.0, o au mento mínimo foi 1.0, a médio do aumento foi 15.81 e o desvio padrão do aumento foi 11. 85.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -78.0, a média da diminuição foi -16.71 e o desvio padrão da di minuição foi 16.59.

Para todos os dados juntos máximo é 71.0, o mínimo é -78.0, a média é -0.40 e o desvio padrão é 21.73.

O *jumping* tem 509 acidentados, 284 positivos e 225 negativos. A análise para jog adores aumentados positivamente o aumento máximo foi 65.0, o aumento mínimo foi 30.0 , a média do aumento foi 36.60 e o desvio padrão do aumento foi 6.74.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, diminuição mínima foi -61.0, a média da diminuição foi -37.33 e o desvio padrão da di minuição foi 6.53.

Os dados brutos para análise de jogadores aumentados positivamente o aumento máximo foi 65.0, o aumento mínimo foi 1.0, e média do aumento foi 12.19 e o desvio padr ão do aumento foi 9.41.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -61.0, a média da diminuição foi -6.44e o desvio padrão da dimi nuição foi 9.50.

Para todos os dados juntos o máximo é de 65.0, o mínimo é de -61.0, a média d os dados é 1.83 e o desvio padrão dos dados é 13.24.

Para o atributo *long_passing* os acidentados foram 1135, 607 positivos e 528 nega tivos.

A análise para jogadores aumentados positivamente, aumento máximo foi 65.0, o aumento mínimo foi 30.0, a média do aumento foi 36.53 e o desvio padrão do aumento foi 6.14.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -76.0, a média da diminuição foi -43.95 e o desvio padrão da diminuição foi 9.54.

Os dados brutos para a análise de jogadores aumentados positivamente o aument o máximo foi 65.0, o aumento mínimo foi 1.0, a média do aumento foi 14.21 e o desvio pa drão do aumento foi 10.37.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -76.0, a média da diminuição foi -10.30 e o desvio padrão da d iminuição foi 14.59.

Para todos os dados o máximo é 65.0, o mínimo é -76.0, a média dos dados é 4.7 4 e o desvio padrão dos dados é 17.05.

Para *short_passing* tem 474 jogadores acidentados, 456 jogadores positivos e 18 negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 57.0, o aumento mínimo foi 30.0, a média do aumento foi 35.70 e o desvio padrão do aumento fo

i 5.29.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -52.0, a média da diminuição foi -33.28e o desvio padrão da di minuição foi 5.10.

Para os dados brutos a análise para jogadores aumentados positivamente o aume nto máximo foi 57.0, o aumento mínimo foi 1., a média do aumento foi 13.03 e o desvio pa drão do aumento foi 9.24.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -52.0, a média da diminuição foi -5.78 e o desvio padrão da dimi nuição foi 6.44.

Para todos os dados juntos o máximo é 57.0, o mínimo é -52.0, a média é 6.34 e o desvio padrão é 12.28.

Para *long_shots* 885 são acidentados dos quais 762 positivos e 123 negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 30.0, a média do aumento foi 37.63 e o desvio padrão do aumento fo i 7.29.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -60.0, a média da diminuição foi -36.07 e o desvio padrão da diminuição foi 6.22.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 70.0, o aumento mínimo foi 1.0, a média do aumento foi 14.64 e o desvio padrão do aum ento foi 11.35.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -60.0, a média da diminuição foi -6.71 e o desvio padrão da di minuição foi 8.73.

Para todos os dados juntos o máximo é 70.0, o mínimo é -60.0, a média é 6.23 e o desvio padrão dos dados é 14.73.

Para o *Overall_rating* tem acidentados positivos e não tem jogadores acidentados negativos.

A análise para jogadores aumentados positivamente o aumento máximo foi 42.0, o aumento mínimo foi 30.0, a média do aumento foi 32.70 e o desvio padrão do aumento fo i 2.64.

Para o atributo *penalties* dos 1402, 578 são positivos e 824 negativos. A análise p ara jogadores aumentados positivamente o aumento máximo foi 62.0, o aumento mínimo foi 30.0, a média do aumento foi 37.28 e o desvio padrão do aumento foi 6.58.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30.

0, a diminuição mínima foi -86.0, a média da diminuição foi -41.27 e o desvio padrão da di minuição foi 9.83.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 62.0, o aumento mínimo foi 1.0, a média do aumento foi 13.47 e o desvio padrão do aum ento foi 11.75.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -86.0, a média da diminuição foi -12.99 e o desvio padrão da di minuição foi 14.73.

Para todos os dados o máximo dos dados é 62.0, o mínimo dos dados é -86.0, a média dos dados é -0.32 e o desvio padrão dos dados é 18.82.

Para o atributo *positioning* dos 1891, 663 são positivos e 1228 negativos. A anális e para jogadores aumentados positivamente o aumento máximo foi 72.0, o aumento míni mo foi 30.0, a média do aumento foi 38.04 e o desvio padrão do aumento foi 7.48.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -76.0, a média da diminuição foi -41.41 e o desvio padrão da di minuição foi 8.86.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 72.0, o aumento mínimo foi 1.0, a média do aumento foi 15.36 e o desvio padrão do aum ento foi 11.35.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -76.0, a média da diminuição foi -16.67 e o desvio padrão da di minuição foi 16.13.

Para todos os dados juntos o máximo é de 72.0, o mínimo é de -76.0, a média é -0 .38 e o desvio padrão dos dados é 21.21.

O *potential* tem 26 positivos e não tem jogadores negativos. A análise para joga dores aumentados positivamente o aumento máximo foi 39.0, o aumento mínimo foi 30.0, a média do aumento foi 32.15 e o desvio padrão do aumento foi 2.07.

Para o atributo *reaction* os jogadores acidentados foram 387, dos quais são positiv os 365 e negativos 22. A análise para jogadores aumentados positivamente o aumento m áximo foi 58.0, o aumento mínimo foi 30.0, a média do aumento foi 35.62 e o desvio padr ão do aumento foi 5.47.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -46.0, a média da diminuição foi -33.60 e o desvio padrão da di minuição foi 4.48.

Os dados brutos para análise dos jogadores aumentados positivamente o aument o máximo foi 58.0, aumento mínimo foi 1.0, a média do aumento foi 12.87 e o desvio padr

ão do aumento foi 8.92.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -46.0, a média da diminuição foi -6.12 e o desvio padrão da di minuição foi 6.87.

Para todos os dados juntos o máximo dos dados é 58.0, o mínimo dos dados é -4 6.0, a média dos dados é 5.8 e o desvio padrão dos dados é 12.31.

Para o *shot_power* existem 859 jogadores acidentados, 764 positivos e 95 negativ os. A Análise para jogadores aumentados positivamente o aumento máximo foi 72.0, o au mento mínimo foi 30.0, a média do aumento foi 37.87 e o desvio padrão do aumento foi 7. 42.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -75.0, a média da diminuição foi -37.68 e o desvio padrão da diminuição foi 7.86.

Para os dados brutos da análise de jogadores aumentados positivamente o aume nto máximo foi 72.0, o aumento mínimo foi 1.0, a média do aumento foi 14.34 e o desvio p adrão do aumento foi 11.46.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -75.0, a média da diminuição foi -5.70 e o desvio padrão da dimi nuição foi 8.26.

Para juntar todos os dados o máximo é de 72.0, o mínimo é -75.0, a média é 6.7 e o desvio padrão é 14.21.

Para o atributo *Sliding_tackle* existem 396 jogadores acidentados dos quais 348 s ão positivos e 48 negativos. A análise para jogadores aumentados positivamente o aume nto máximo foi 62.0, o aumento mínimo foi 30.0, a média do aumento foi 38.86 e o desvio padrão do aumento foi 7.75.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -52.0, a média da diminuição foi -35.08 e o desvio padrão da diminuição foi 5.38.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 62.0, o aumento mínimo foi 1.0, a média do aumento foi 11.61 e o desvio padrão do aum ento foi 9.89.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -52.0, a média da diminuição foi -5.07 e o desvio padrão da di minuição foi 6.85.

Para todos juntos o máximo é 62.0, o mínimo é -52.0, a média é 3.64 e o desvio p adrão é 11.96.

Para *sprint_speed* os jogadores acidentados são 700, o positivos 179 e os negativ os 521. A análise para jogadores aumentados positivamente o aumento máximo foi 66.0, o aumento mínimo foi 30.0, a média do aumento foi 35.22 e o desvio padrão do aumento f oi 5.91.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -59.0, a média da diminuição foi -36.80 e o desvio padrão da diminuição foi 6.18.

Os dados brutos para os jogadores aumentados positivamente o aumento máximo foi 66.0, o aumento mínimo foi 1.0, a média do aumento foi 12.62 e o desvio padrão do aumento foi 8.11.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -59.0, a média da diminuição foi -11.17 e o desvio padrão da d iminuição foi 11.26.

Para todos os dados o máximo é 66.0, o mínimo é -59.0, a média é -0.86 e o desvi o padrão é 15.47.

Para a *stamina* tem 1306, 574 dos quais jogadores positivos e 732 jogadores ne gativos. A análise para jogadores aumentados positivamente o aumento máximo foi 71.0, o aumento mínimo foi 30.0, a média do aumento foi 36.72 e o desvio padrão do aumento f oi 6.46.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -68.0, a média da diminuição foi -39.18 e o desvio padrão da diminuição foi 7.42.

Os dados brutos para análise dos jogadores aumentados positivamente, o aument o máximo foi 71.0, o aumento mínimo foi 1.0, a média do aumento foi 15.96 e o desvio pa drão do aumento foi 10.24.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -68.0, a média da diminuição foi -12.62 e o desvio padrão da d iminuição foi 13.44.

Para juntar todos os dados o máximo dos dados é 71.0, o mínimo é -68.0, a média é 1.39 e o desvio padrão dos dados é 18.65.

Para o atributo *standing_tackle* existem 1688 acidentados, 763 positivos e 925 ne gativos. A análise para jogadores aumentados positivamente o aumento máximo foi 68.0, o aumento mínimo foi 1.0, a média do aumento foi 14.95 e o desvio padrão do aumento fo i 11.59.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -54.0, a média da diminuição foi -8.08 e o desvio padrão da di

minuição foi 9.15.

Os dados brutos para a análise dos jogadores aumentados positivamente, o au mento máximo foi 68.0, o aumento mínimo foi 1.0, a média do aumento foi 14.95 e o desv io padrão do aumento foi 11.59.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -54.0, a média da diminuição foi -8.08 e o desvio padrão da dim inuição foi 9.15.

Para todos os dados juntos o máximo é 68.0, o mínimo é -54.0, a média é 5.52 e o desvio padrão é 15.55.

Para o *strength* dos 628, 537 são jogadores positivos e 91 negativos. Análise para jogadores aumentados positivamente o aumento máximo foi 68.0, o aumento mínimo foi 30.0, a média do aumento foi 37.03 e o desvio padrão do aumento foi 6.60.

A análise para jogadores aumentados negativamente a diminuição máxima foi -3 0.0, a diminuição mínima foi -50.0, a média da diminuição foi -34.62 e o desvio padrão da diminuição foi 4.76.

Os dados brutos para a análise dos jogadores aumentados positivamente o aum ento máximo foi 68.0, o aumento mínimo foi 1.0, a média do aumento foi 14.42 e o desvio padrão do aumento foi 9.80.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -50.0, a média da diminuição foi -6.62 e o desvio padrão da di minuição foi 8.10.

Para todos os dados juntos o máximo é 68.0, o mínimo é -50.0, a média é 6.17 e o desvio padrão dos dados é 13.80.

Para o atributo *vision* existem 627 jogadores acidentados, 289 positivos e 338 n egativos. A análise para jogadores aumentados positivamente e o aumento máximo foi de 60.0, o aumento mínimo foi de 30.0, a média do aumento foi 36.76 e o desvio padrão do aumento foi 6.26.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -61.0, a média da diminuição foi -38.18 e o desvio padrão da di minuição foi 7.21.

Os dados brutos para jogadores aumentados positivamente o aumento máximo foi 60.0, o aumento mínimo foi 1.0, a média do aumento foi 11.62 e o desvio padrão do au mento foi 9.80.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -61.0, a média da diminuição foi -8.66 e o desvio padrão da dimi nuição foi 10.66.

Para todos os jogadores juntos o máximo é 60.0, o mínimo é -61.0, a média é 0. 53 e o desvio padrão é 14.41.

Para o atributo *volleys* 219 jogadores, 182 são acidentados positivos e 37 acide ntados negativos. A análise para jogadores aumentados positivamente o aumento máxim o foi 59.0, o aumento mínimo foi 30.0, a média do aumento foi 35.62 e o desvio padrão do aumento foi 5.72.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -49.0, a média da diminuição foi -34.22 e o desvio padrão da di minuição foi 4.11.

Os dados brutos para os jogadores aumentados positivamente têm um aumento m áximo de 59.0, o aumento mínimo de 1.0, a média do aumento de 9.17 e o desvio padrão do aumento foi 8.81.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -49.0, a média da diminuição foi -3.95 e o desvio padrão da di minuição foi 6.20.

Para todos os dados o máximo é 59.0, o mínimo dos dados é -49.0, a média dos d ados é 2.51 e o desvio padrão dos dados é 10.04.

Gk_diving tem 29 jogadores, 26 positivos e 3 negativos. A análise para jogadores aumentados positivamente o aumento máximo foi 72.0, o aumento mínimo foi 30.0, a méd ia do aumento foi 33.15 e o desvio padrão do aumento foi 7.95.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30. 0, a diminuição mínima foi -54.0, a média da diminuição foi -39.67 e o desvio padrão da di minuição foi 10.34.

Os dados brutos para análise de jogadores aumentados positivamente o aumento máximo foi 72.0, o aumento mínimo foi 1.0, a média do aumento foi 4.62 e o desvio padrã o do aumento foi 4.96.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -54.0, a média da diminuição foi -2.19 e o desvio padrão da di minuição foi 3.67.

Para juntar todos os dados o máximo dos dados é 72.0, o mínimo dos dados é -54 .0, a média dos dados é 1.59 e o desvio padrão dos dados é 5.58.

O atributo para *gk_handling* tem 14 acidentados, 13 positivos e 1 negativo. A análi se para jogadores aumentados positivamente o aumento máximo foi 64.0, o aumento mín imo foi 30.0, a média do aumento foi 35.69 e o desvio padrão do aumento foi 8.55

A análise para jogadores aumentados negativamente a diminuição máxima foi -4 7.0, a diminuição mínima foi -47.0, a média da diminuição foi -47.0 e o desvio padrão da d

iminuição foi 0.0.

Para os dados brutos a análise para os jogadores aumentados positivamente foi u m aumento máximo foi 64.0, um aumento mínimo foi 1.0, uma média do aumento foi 6.44 e o desvio padrão do aumento foi 6.93.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -47.0, a média da diminuição foi -7.69 e o desvio padrão da di minuição foi 6.60.

Para todos os dados máximo é 64.0, o mínimo é -47.0, a média é -2.16 e o desvio padrão é 9.64.

Para o atributo *gk_positioning* os jogadores acidentados são 29, os positivos são 2 8 e os negativos apenas 1. A análise para jogadores aumentados positivamente o aument o máximo foi de 56.0, o aumento mínimo foi 30.0, a média do aumento foi 33.46 e o desvi o padrão do aumento foi 5.65.

A análise para jogadores aumentados negativamente a diminuição máxima foi -4 1.0, a diminuição mínima foi -41.0, a média da diminuição foi -41.0 e o desvio padrão da d iminuição foi 0.0.

A análise dos dados brutos para jogadores aumentados positivamente, o aumento máximo foi 56.0, o aumento mínimo foi 1.0, a média do aumento foi 6.66 e o desvio padr ão do aumento foi 7.12.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0 .0, a diminuição mínima foi -41.0, a média da diminuição foi -7.63 e o desvio padrão da di minuição foi 6.58.

Todos os dados o máximo dos dados é 56.0, o mínimo dos dados é -41.0, a média dos dados é -1.92 e o desvio padrão dos dados é 9.76.

Para o *gk_kicking* tem 4874 jogadores, 38 positivos e 4836 negativos. A análise p ara jogadores aumentados positivamente o aumento máximo foi 56.0, o aumento mínimo foi 30.0, a média do aumento foi 36.26 e o desvio padrão do aumento foi 7.38.

A análise para jogadores aumentados negativamente a diminuição máxima foi -30 .0, a diminuição mínima foi -96.0, a média da diminuição foi -52.28 e o desvio padrão da d iminuição foi 11.17.

Os dados brutos na análise para jogadores aumentados positivamente há um aum ento máximo de 56.0, o aumento mínimo foi 1.0, a média do aumento foi 3.77 e o desvio padrão do aumento foi 6.71.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -96.0, a média da diminuição foi -33.08 e o desvio padrão da di minuição foi 25.56.

Para os dados todos o máximo é de 56.0, o mínimo é -96.0, a média é -23.15 e o desvio padrão é 27.51.

Para o *gk_reflexes* tem 38 jogadores acidentados, 33 positivos e 5 negativos. Anál ise para jogadores aumentados positivamente o aumento máximo foi 69.0, o aumento mín imo foi 30.0, a média do aumento foi 34.91 e o desvio padrão do aumento foi 7.65.

Análise para jogadores aumentados negativamente a diminuição máxima foi -33 .0, a diminuição mínima foi -54.0, a média da diminuição foi -40.4 e o desvio padrão da di minuição foi 8.01.

Para análise dos dados embruto os jogadores aumentados positivamente tiveram um aumento máximo de 69.0, o aumento mínimo de 1.0, a média do aumento foi 6.56 e o desvio padrão do aumento foi 7.21.

A análise para jogadores aumentados negativamente a diminuição máxima foi -0.0 , a diminuição mínima foi -54.0, a média da diminuição foi -7.78 e o desvio padrão da dimi nuição foi 6.69.

Para todos os dados o máximo dos dados é 69.0, o mínimo dos dados é -54.0, a média dos dados é -2.22 e o desvio padrão dos dados é 9.82.

Os atributos que mais aumentam são *aggression, free_kick_accuracy, heading_ac curacy, interceptions, long_shots, positioning, shot_power* e *stamina* e os que mais diminu em são *balance, long_passing, free_kick_accuracy, interceptions, positioning* e *shot_pow er.* Os atributos para ao guarda-redes que aumentam são o *gk_diving* e *gk_reflexes* e os que diminuem são o *gk_kicking.*

Contrastando com esta prática dos treinadores as curvas de desempenho temporal dos jogadores não mostram que os atributos todos melhoram ao longo dos anos.

5.4 **Análise por grupos de jogadores com a mesma função em jogo**

Após a análise por jogador de forma individual o próximo passo passou pela análise de todos os jogadores divididos por grupos de acordo com a posição que ocupam em campo como guarda-redes, defesas, médios e avançados.

Após essa divisão fez-se a modelagem dos dados onde se criou um modelo que explicasse as características de funcionamento e comportamento dos jogadores. Logo depois, criou-se o pré-processamento dos dados e gerou-se um novo *dataset* através da ferramenta *Python.*

Os atributos mais relevantes para estudo de acordo com a posição de cada grupo

de jogadores, os *id* dos jogadores (*player_fifa_api_id*) e as datas de avaliações dos jogadores durante cada semana. A nova tabela criada não foi avaliada genericamente, mas sim consoante a posição do jogador em campo divididos quatro em grupos.

A tabela usada para avaliação foi a tabela *Match* e foi criada a partir das colunas dos jogadores, data das avaliações semanais, posição em campo e data de jogo. Nesta tabela avaliaram-se todos os jogos e para todos os jogadores analisaram-se as variáveis mais importantes, as que mais se diferenciam, as que fogem do padrão e as que têm uma maior distância de valores muito expressos positiva ou negativamente para determinada posição.

O principal foco desta análise foi maior em relação à posição do jogador em campo independentemente da equipa em que está e ao analisar o histórico das variáveis percebeu-se a sua evolução ao longo do tempo de maneira a projetar no futuro e auxiliando a decisão do treinador sobre se o jogador deve ou não jogar ou se deve ou não ser substituído.

Na tabela do grupo dos guarda-redes os atributos associados ao jogador foram o *overall_rating, potential, stamina, gk_diving, gk_handling, gk_kicking, gk_positioning e gk_reflexes*.

Em seguida, fez-se a análise descritiva dos dados onde se analisou para cada atributo o valor máximo, mínimo, média, desvio de padrão e quartis. O valor máximo para o atributo *overall_rating* é de 93 e o mínimo de 41, ou seja, o alcance está muito grande, assim como o atributo *potential* o valor máximo é de 93 e o mínimo de 45. No entanto, estes valores são esperados uma vez que na análise bivariada a correlação dos atributos com estas variáveis foram positivas e tendencialmente fortes e na avaliação por jogador também não houveram jogadores acidentados negativos para ambos os atributos. A *stamina* tem um valor máximo de 88 e mínimo de 10 e todos os restantes atributos têm valor mínimos de 1 até 5 e máximos de 92 a 96. Os atributos que apresentam valores muito baixos do mínimo e muito altos de máximo.

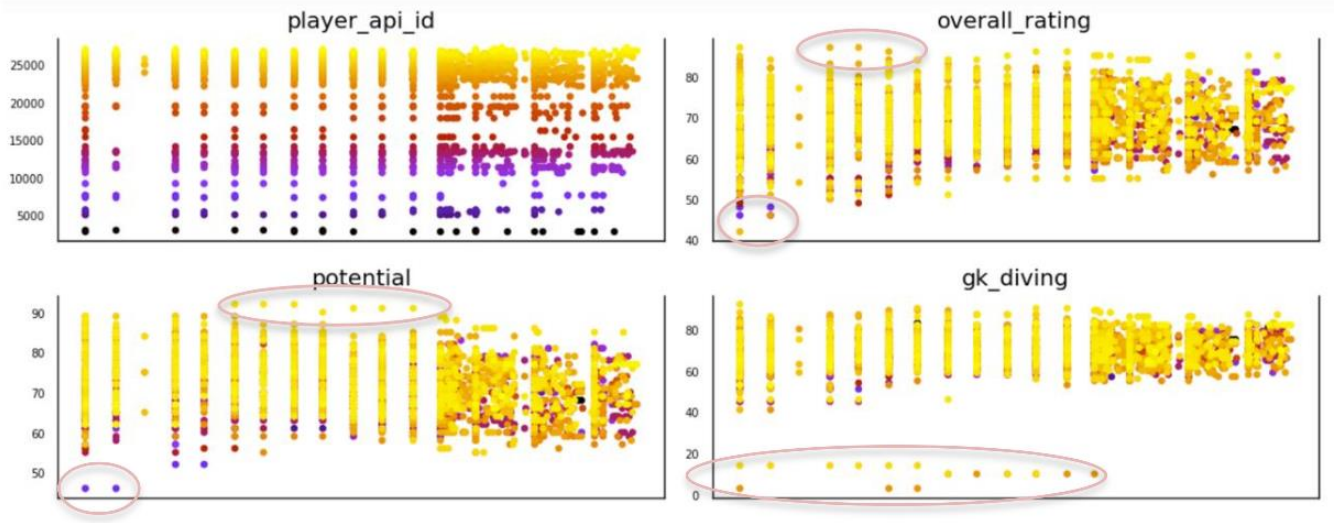


Figura 19-Gráfico de dispersão com os outliers para cada atributo.

Após esta análise gerou-se gráficos para cada uma das variáveis de forma a analisar ao longo do tempo. O primeiro gráfico é as datas de avaliação no eixo dos X e o valor dos atributos para o eixo dos Y. Os gráficos dos atributos mostram o valor das variáveis ao longo do tempo para cada um desses jogadores e apresentam tons de cor diferentes para cada jogador medidos nas datas. Apesar de aparecer ano no eixo do X, não estão por ano mas sim por data do ano de avaliação e há anos que têm mais avaliações e outros que têm menos. Os jogadores aparecem várias vezes nos gráficos dos atributos uma vez que existem vários anos e cada ano apresenta várias datas e os jogadores têm diversas avaliações ao longo dos anos.

Para os gráficos dos atributos analisaram-se os subgrupos de jogadores que fogem do padrão, ou seja, os que têm valores muito altos e os que têm valores muito baixos como expresso na Figura 19 nos círculos vermelhos assinalados. Analisou-se a evolução dos jogadores acidentados ao longo do tempo de maneira a verificar se existem ou não melhorias e se se aproximam dos valores padrão, ou para os valores muito altos, se estes se mantêm.

Os jogadores que obtiveram valor muito altos e muito baixos de acordo com o padrão para cada atributo foram analisados e observou-se o seu historial a partir da base de dados.

Os jogadores em análise foram os mais relevantes para cada posição tendo em conta a Tabela 1 da presenta dissertação.

Os atributos *overall_rating* e *potential* os valores mínimos são sempre acima de 41 e os máximos atingem os 93. A *stamina* tem mínimos de 10 e máximos de 88, existem

jogadores com grandes diferenças de valores repentinas passando de valor por exemplo de 83 para 14 no caso do jogador 27433 de *player_fifa_api_id*, de 88 para 35 o jogador 39584 e 23984. Os atributos *gk* têm valor mínimos de 1 e máximos de 96 no entanto existem jogadores que melhoram a sua avaliação como o jogador 36286 que sobe de 65 para 93 e o jogador 39989 sobe de 75 para 93.já o jogador 30657 apesenta inicialmente valores de 94 e descem ao longo dos anos até 79 assim como o jogador 39989.

Na tabela criada para os defesas os atributos associados ao jogador como o *overall_rating, potential, stamina, defensive_work_rate, sprint_speed, strength, heading_accuracy, balance e jumping.*

O valor máximo para o atributo *overall_rating* é de 92 e o mínimo de 35, ou seja, o alcance está muito grande, assim como o atributo *potential* o valor máximo é de 93 e o mínimo de 42, como esperado. A *stamina*, força e balance têm um valor máximo de 96, o *jumping* tem valor máximo de 95 e mínimo de 20. O atributo *sprint_speed* e *heading_accuracy* têm valores máximo de 96 e 95 e mínimos de 20 e 7, respetivamente.

Os jogadores que apresentam valores de atributos muito baixos do mínimo e muito altos de máximo.

Os atributos *overall_rating* e *potential* os valores mínimos são sempre acima de 35 e os máximos atingem os 92. O jogador 150243 apresenta inicialmente o valor de 53 de *overall_rating* em 2007 e desce para 35 em 2009 e no mesmo ano sobe para 67, o jogador 93457 tem inicialmente valor 37 e sobe para 74. Os jogadores 178603 têm valor de *potential* 69 que sobe para 82 e baixa no mesmo mês para 45, o jogador 68796 têm valor inicial de 74 em 2007 e passado um ano baixa para 54 e posteriormente aos poucos sobe até 75 e em 4 meses de diferença brusca para 47. A *stamina* tem valores mínimos de 22 e máximos de 96. Os jogadores sobem drasticamente de 28 para 80 caso do jogador 193651, de 22 para 93 o jogador 183711. Nos atributos *sprint_speed* e *strength* os valores mínimos são de 20 e máximos de 96 com jogadores de valores 68 a descer para 22. No atributo *heading_accuracy* os valores mínimos são de 11 e máximo de 93. Nos atributos *jumping* e *balance* os mínimos são de 21 e máximos de 96. O jogador 198046 com valor de 85 de *balance* desce para 21.

Na tabela criada para os médios os atributos associados ao jogador como o *overall_rating, potential, stamina, short_passing, dribbling, ball_control, balance, long_passing e strength.*

O valor máximo para o atributo *overall_rating* é de 93 e o mínimo de 33, ou seja, o alcance está muito grande, assim como o atributo *potential* o valor máximo é de 95 e o mínimo de 39, como esperado. Todos os restantes atributos apresentam valores muito altos de máximo entre 96 e 97 e mínimos de 21 para balance e stamina, a força, ball_control

e long_passing a 12, short_passing a 15 e dribbling atinge valores mínimos de 2.

Os jogadores que apresentam valores de atributos muito baixos do mínimo e muito altos de máximo.

Os atributos *overall_rating* e *potential* os valores mínimos são de 33 e os máximos atingem os 95. O jogador 178043 tem valor de *overall_rating* de 33 que sobe até 81. O jogador 178014 tem valor de *potential* de 70 que desce até 39.

Para o atributo *short_passing* o jogador 189658 apresenta valor constante de 57 e em um mês desce drasticamente para 15, o jogador pode provavelmente ter sofrido alguma lesão que justifica a grande diferença de valor no atributo. No atributo balance os jogadores sobem de valores de 22 para 72 e vise versa. No atributo *strength* o jogador 184475 apresenta inicialmente valor de 54 e sobe no mesmo ano para 93 continuando ao longo dos anos até 96.

Na tabela criada para os avançados os atributos associados ao jogador como o *overall_rating, potential, stamina, finishing, shot_power, sprint_speed e balance.*

O valor máximo para o atributo *overall_rating* é de 94 e o mínimo de 33, o atributo *potential* o valor máximo é de 97 e o mínimo de 42. Todos os restantes atributos apresentam valores muito altos de máximo entre 97 e 96 e mínimos de 20 para balance e stamina, o sprint_speed com valor de 15, *shot_power* a 6 e finishing atinge valores mínimos de 2.

Os jogadores que apresentam valores de atributos muito baixos do mínimo e muito altos de máximo. Os atributos *overall_rating* e *potential* os valores mínimos são de 33 e máximos 97. Os valores de *stamina* vão de 20 a 96. O atributo *shot_power* o jogador 177533 inicia a avaliação com valor 11 e o longo dos anos sobe e atinge 83. O atributo *srint_speed* tem o jogador 177485 acidentado que apresenta avaliação inicial de valor 81 e baixa até 2013 para 59 e após uma semana apresenta valor 22, ou seja, os valores expressam que possivelmente o jogador sofreu uma lesão. No atributo *balance* o jogador 189451 inicialmente com 80 desce ao longo dos anos até valor 20.

5.5 **Learning**

O *dataset* usado para a análise dos jogadores por grupos foi também usado para a análise do VAR. Os jogadores escolhidos para análise do VAR foram de acordo com as observações anteriores onde para cada grupo se verificou jogadores acidentados.

No *dataset* gerado não é possível fazer a previsão dentro do próprio jogo, isto é,

durante o jogo, uma vez que não se sabe o que está a acontecer. Temos a avaliação de um grupo de jogadores guarda-redes, defesas, médios e avançados e ao fazer a previsão, o *forecasting*, eles podem ser substituídos ao longo de tempo e a partir de um conjunto de jogadores.

A data de avaliação são medidas que conhecidas antes dos jogos e por isso, são medidas com relação a um jogo em particular, o *match_id* da tabela. Os jogadores que não se encontram na tabela *match* podem não ter jogado devido a uma lesão, a cartões ou simplesmente por opção do treinador. No entanto, estes jogadores foram avaliados na mesma uma vez que podem ser usados para substituição.

Não existe informação se os jogadores foram ou não substituídos e por quem é que foram, uma vez que só temos informação do 11 inicial, por isso não sabemos se eles jogaram, mas sabemos que não jogaram logo no início logo verifica-se se existe alguma correlação com os testes de forma a perceber o porquê de não jogarem inicialmente.

5.5.1 *Forecasting*: um jogador deve ser substituído no próximo período de tempo ?

O modelo de forecasting avalia os jogadores num determinado tempo de jogo e prevê os seus valores de atributos. Nesse tempo aplica-se o classificador e o modelo irá fazer um forecasting para todos os jogadores incluindo os que são suplentes com o intuito de perceber como será o comportamento das variáveis no próximo período de tempo de jogo, cerca de 10/20 minutos. Se o modelo der um alerta informando que um determinado jogador vai ter um decréscimo muito acentuado numa determinada variável, e essa variável precisar de um valor para que o jogador consiga exercer a sua função de forma correta, este modelo de acordo com a coleção de jogadores o classificador imite um alerta para que esse jogador seja substituído apresentando-lhe para isso um conjunto de jogadores com a mesma posição em campo para o substituir, esse é o grande objetivo do classificador. Ou seja, para todos os defesas existe um que se escontra fora da média positiva ou negativa em comparação com o restante grupo de jogadores. Assim sendo, para cada grupo de jogadores traça-se o perfil médio desses jogadores e posteriormente para cada um deles verifica-se a distância entre eles, isto é, verifica-se quais os jogadores que estão fora do padrão, os que têm uma maior distância entre eles.

O classificador mede todos os atributos e projeta essa informação para o futuro dando o parecer ao treinador e auxiliando-o na tomada de decisão.

Fazer um *forecasting* dos jogadores dentro do padrão é mais fácil do que fazer dos jogadores acidentados isto porque parte-se do pressuposto que os jogadores dentro do

padrão não sofrem grandes alterações mas os acidentados sim uma vez que oscilam constantemente de valores e fazer uma previsão torna-se mais complicado devido á sua inconstância.

Analisou-se o jogador 287894 com *lag* 4, *forecast* 40 com 32 instancias e o mesmo jogador com 25 instancias, ao usamos mais dados de treino a previsão é mais fina. Quanto menor for o valor das instancias, menos será erro que é mais baixo no treino para 25 instancias uma vez que são menos projeções com comparação com o de 32 instancias, logo o erro é menor porque o previsto se aproxima mais do real.

A correlação entre as duas instancias for alta o treinador quando quer fazer uma substituição precisa ter em atenção as correlações entre variáveis uma vez que uma depende da outra pelo que o jogador e a equipa teriam mais vantagem se observassem as variáveis como um todo criando interações e não de forma individual porque um atributo baixo pode condicionar um atributo alto. Ou seja, o treinador pode seguir o modelo ou a sua intuição e opção pessoal no entanto assume-se sempre que o treinados faz uma boa substituição.

Correlation matrix of residuals

	gk_diving	gk_handling	gk_positioning	gk_reflexes
gk_diving	1.000000	0.598143	0.672792	0.836413
gk_handling	0.598143	1.000000	0.698717	0.534097
gk_positioning	0.672792	0.698717	1.000000	0.504263
gk_reflexes	0.836413	0.534097	0.504263	1.000000

Correlation matrix of residuals

	gk_diving	gk_handling	gk_positioning	gk_reflexes
gk_diving	1.000000	0.721876	0.909984	0.933297
gk_handling	0.721876	1.000000	0.870049	0.468896
gk_positioning	0.909984	0.870049	1.000000	0.704448
gk_reflexes	0.933297	0.468896	0.704448	1.000000

Figura 20- Correlação para jogador dos atributos em análise.

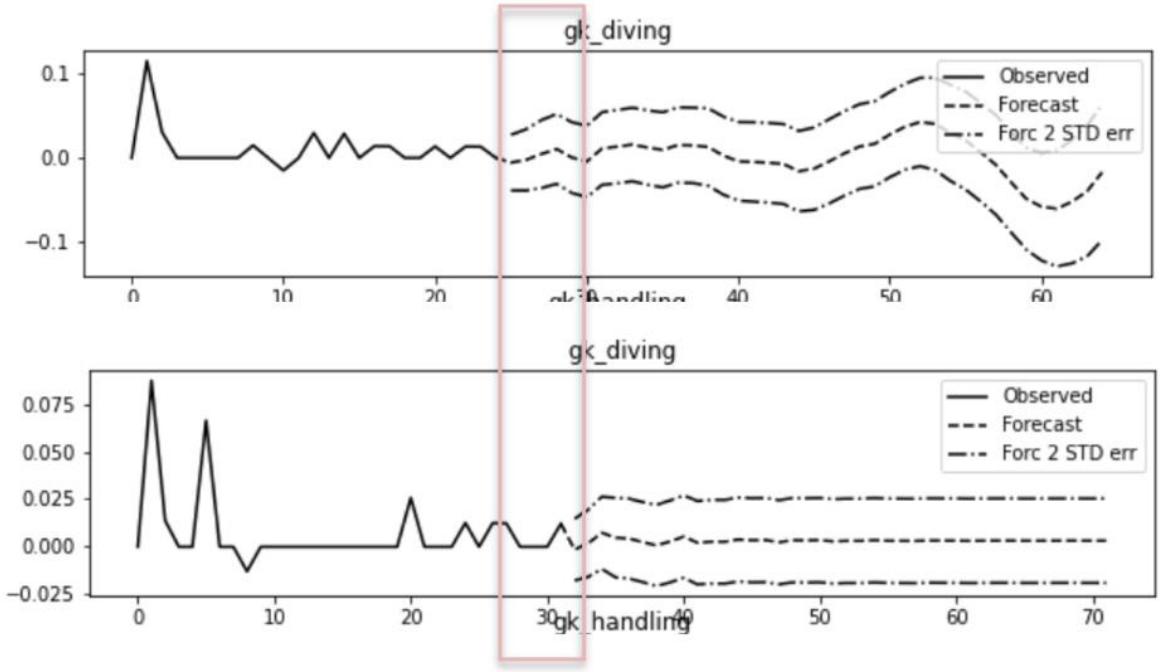


Figura 21- Forecasting para o atributo gk_diving.

O *forecast* do atributo *gk_diving* expresso na Figura 21 apresenta 32 instancias e é aparentemente mais preciso uma vez que tem mais instancias observadas e é mais fácil de prever o futuro apesar de que em ambas as previsões tenderem a ficar constantes. O previsto é quando já se sabe o valor observado mas treina-se para ver se está correto e o projetado é quando não se sabe o valor observado e treina-se para saber.

O retângulo vermelho na figura delimita o observado em cima e o projetado para as instancias de 25 a 32 em baixo e comparando ambos os gráficos estes expressam uma pequena subida e descida como mostra no gráfico em baixo e novamente um subida o que implica dizer que existe alguma conformidade entre o observado e o previsto.

Dado este modelo que se treinou, se um jogador estiver com valor de atributos baixos pode-se prever se o jogador estava abaixo do padrão então é de esperar que o jogador passado uns anos atinga o patamar padrão. No entanto, não se consegue saber em quanto tempo um jogador demora a atingir determinado valor por isso prevê-se um modelo *forecasting.*

Os dados mostram que existem jogadores que partem de um nível muito baixo e conseguem atingir o mesmo valor que os outros que já estão lá antes.

O *forecast* do atributo *gk_reflexes* demonstrado na Figura 22*Figura 22* tem 32 instancias para o gráfico de cima e 25 para o gráfico em baixo e equiparando os dois implica dizer que existe alguma conformidade com o observado e o previsto no entanto o gráfico para 32 instancias tende a ficar sem expressão nos valores projetados e o gráfico de 25

instancias apresenta algumas altos e baixos no jogador o que parece ir mais de encontro com a realidade.

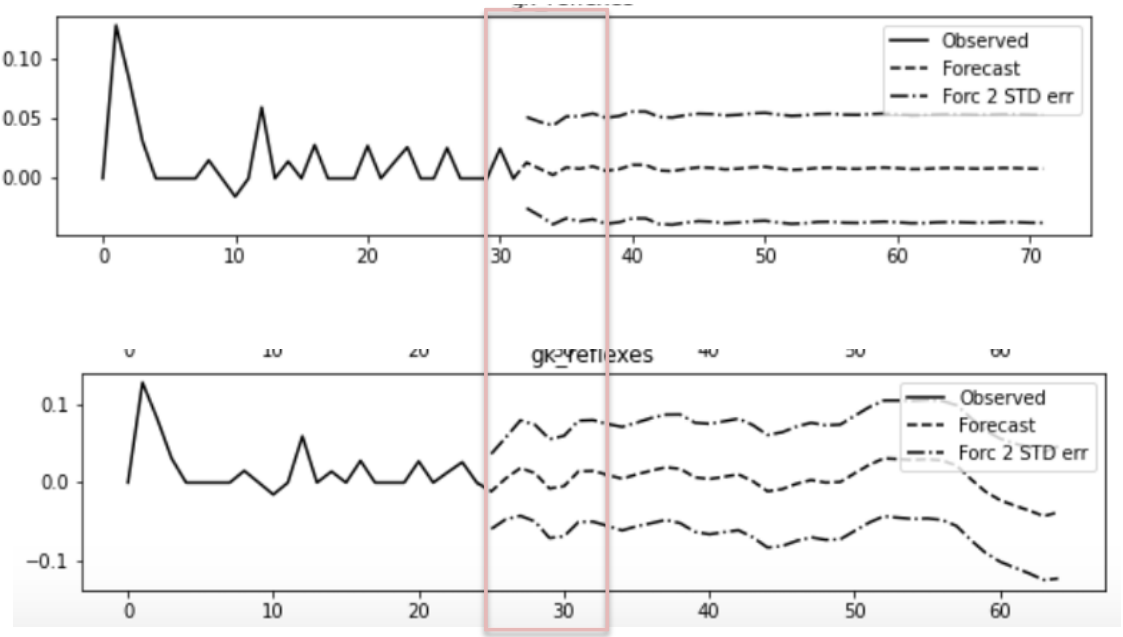


Figura 22. Forecasting para o atributo gk_reflexes.

Analisou-se o jogador 49589 com *lag* 4, *forecast* 40 com 32 instancias e o mesmo jogador com 28 instancias. O erro é menor para 28 instancias do que para 32.

Correlation matrix of residuals

	gk_diving	gk_handling	gk_positioning	gk_reflexes
gk_diving	1.000000	0.364698	0.626233	0.424884
gk_handling	0.364698	1.000000	0.481549	0.288967
gk_positioning	0.626233	0.481549	1.000000	0.455258
gk_reflexes	0.424884	0.288967	0.455258	1.000000

Correlation matrix of residuals

	gk_diving	gk_handling	gk_positioning	gk_reflexes
gk_diving	1.000000	0.424098	0.742286	0.280567
gk_handling	0.424098	1.000000	0.267304	-0.062267
gk_positioning	0.742286	0.267304	1.000000	0.506677
gk_reflexes	0.280567	-0.062267	0.506677	1.000000

Figura 23- Correlação para jogador dos atributos em análise.

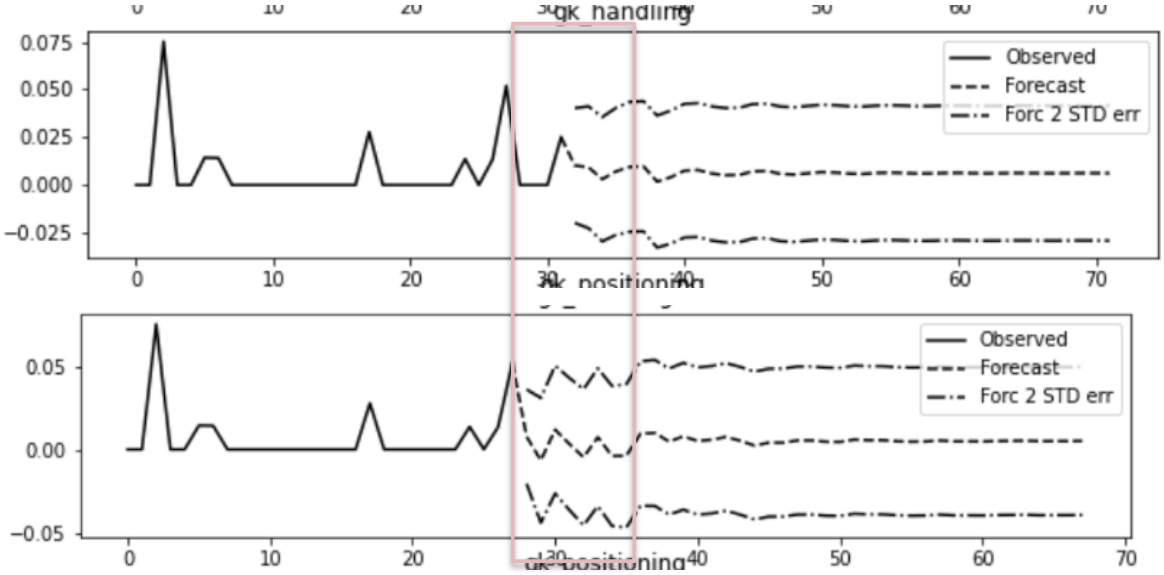


Figura 24- Forecasting para o atributo gk_handling.

O *forecast* do atributo *gk_handling* demonstrado na Figura 24 tem 32 instancias para o gráfico de cima e 28 para o gráfico em baixo e equiparando os dois implica dizer que existe conformidade entre o observado e o previsto. Até ao valor de 30 os dois gráficos apresentam uma descida dos valores e logo em seguida uma subida também demonstrada nos dois gráficos.

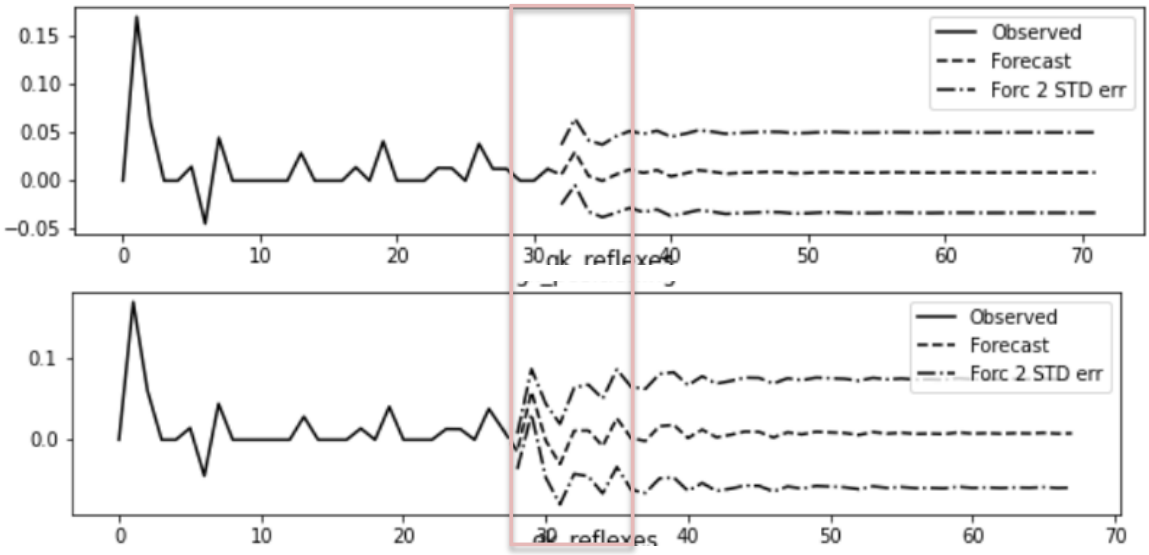


Figura 25- Forecasting para o atributo gk_positioning.

O *forecast* do atributo *gk_positioning* demonstrado na Figura 25*Figura 24* tem 32 instancias para o gráfico de cima e 28 para o gráfico em baixo. Não existe grande conformidade entre o observado e o previsto e só a partir do valor 32 é que há um pico de valores nos dois gráficos.

6 Conclusões e Trabalhos Futuros

O objetivo desta dissertação foi analisar os dados biométricos de jogadores de futebol para previsão de performance durante o jogo de forma a perceber qual o jogador que deveria jogar e não jogou e qual se poderia substituir de acordo com os dados de avaliação dos atributos. De forma a resolver a questão montou-se um modelo onde os coeficientes vão mostrar o peso relativo para cada uma das variáveis e a partir desse peso conclui-se que um jogador deve ser substituído ou este jogador deve substituir outro.

Ao contrário do esperado a análise de previsão no modelo *forecasting* não demostrou grandes oscilações nas projeções do jogador para determinado atributo pelo que uma possível justificação seria o facto de o *lag* ser muito baixo e as medidas não apresentarem valores corretos quando gerado o modelo tendo que baixar o número de instâncias. O modelo gerou relativamente ás medidas anteriores e o facto de estas medidas serem poucas faz com que o modelo se torne constante uma vez que analisa medidas 4 tempos para trás.

O *dataset* em análise não foi separado por *home* e *away,* no entanto existem repetições de jogos em casa e fora pelo que no futuro seria relevante analisar separadamente o jogos em casa e fora de forma a ter uma melhor observação e perceção do comportamento dos jogadores uma vez que o local onde joga pode fazer diferença no atributo visto que o jogador quando joga em casa tem um desempenho e fora tem outro desempenho.

No futuro poder-se-ia utilizar este modelo de previsão para os jogadores mas com o sistema de fornecimento de medidas automáticas através de sensores no equipamento onde o treinador poderia estar a receber a informação em tempo real sobre os atributos do jogador e fazer com que o sistema imite-se um alerta quando o jogador precisar de ser substituído.

7 Referências

Al-Asadi, M. (2018) 'Decision support system for a football team management by using machine learning techniques', *Xinyang Teachers College*, 10(2), pp. 1–15.

Al-Shboul, R. *et al.* (2017) 'Automated Player Selection for Sports Team using Competitive Neural Networks', *International Journal of Advanced Computer Science and Applications*, 8(8), pp. 457–460. doi: 10.14569/ijacsa.2017.080859.

Alves, J. C. F. (2017) 'Avaliação e Controlo da Performance no Futebol: Estudo realizado na equipa de Sub-19 do Gil Vicente Futebol Clube', p. 184. Available at: https://repositorio-aberto.up.pt/bitstream/10216/109997/2/241002.pdf.

Babbie, E. (2010) *The practice of social research*.

Barron, D. *et al.* (2018) 'Artificial neural networks and player recruitment in professional soccer', *PLoS ONE*, 13(10), pp. 1–11. doi: 10.1371/journal.pone.0205818.

Board, I. T. I. F. A. (2017) 'Leis do Jogo 2017/18 Portugal', p. 210.

Boddy, R. and Smith, G. (2009) 'Statistical Methods in Practice: For Scientists and Technologists', *Statistical Methods in Practice: For Scientists and Technologists*, pp. 1–236. doi: 10.1002/9780470749296.

Braz, T. V., Spigolon, L. M. P. and Borin, J. P. (2009) 'Proposta de bateria de testes e classificação de desempenho das capacidades biomotoras em futebolistas', *Revista da Educação Física/UEM*, 20(4). doi: 10.4025/reveducfis.v20i4.7392.

Caetano, R. (22/7/2020) https://www.espn.com.br/esports/artigo/_/id/7304447/barcelona-ou-psg-quem-leva-a-melhor-na-possivel-troca-entre-neymar-e-griezmann-em-fifa-21

Coutinho, E. R., Silva, R. M. and Delgado, A. R. S. (2016) 'Using computational intelligence technique for the meteorological data prediction', *Revista Brasileira de Meteorologia*, 31(1), pp. 24–36. doi: 10.1590/0102-778620140115.

Cunha, P. (2016) 'Teoria e Metodologia de Treino Desportivo', p. 34546.

Demšar, J. *et al.* (2004) 'Orange : From Experimental Machine Learning', *Knowledge Discovery in Databases: PKDD 2004*, pp. 537–539. doi: 10.1007/978-3-540-30116-5_58.

Demšar, J. and Zupan, B. (2013) 'Orange: Data mining fruitful and fun - A historical

perspective', *Informatica (Slovenia)*, 37(1), pp. 55–60.

Ekstrand, J., Hägglund, M. and Waldén, M. (2011) 'Epidemiology of muscle injuries in professional football (soccer).', *The American journal of sports medicine*. United States, 39(6), pp. 1226–1232. doi: 10.1177/0363546510395879.

Electronic Arts (4/5/2020) https://answers.ea.com/t5/General-Discussion/Season-3-Attributes-Explained/td-p/7153524

Electronic Arts (9/10/2020) https://help.ea.com/pt-pt/help/fifa/fifa-ultimate-team-chemistry/

Evwiekpaefe, A., Bitrus, E. and Ajakaiye, F. (2020) 'Selecting Forward Players in a Football Team using Artificial Neural Networks', *International Journal of Computer Applications*, 176(28), pp. 8–13. doi: 10.5120/ijca2020920298.

FIFAUTeam (4/5/2020) https://www.fifauteam.com/fifa-19-attributes-guide/

Filho, D. B. F. and Júnior, J. A. D. S. (2009) 'Desvendando os mistérios do coeficiente de correlação de Pearson (r)', *Revista Política Hoje*, 18(1), pp. 115–146.

Gama, J. (2004) 'Functional trees', *Machine Learning*, 55(3), pp. 219–250. doi: 10.1023/B:MACH.0000027782.67192.13.

Gil Gonçalves dos Santos, P. (2016) 'A Organização Do Treino E Do Jogo no Futebol De Formação', p. 120.

Giménez, J. V. *et al.* (2020) 'Predictive modelling of the physical demands during training and competition in professional soccer players', *Journal of Science and Medicine in Sport*, 23(6), pp. 603–608. doi: 10.1016/j.jsams.2019.12.008.

Guimarães, M. *et al.* (2014) 'As posições no futebol e suas especificidades', *Revista Brasileira de Futebol*, 7(2), pp. 71–83. doi: 10.1590/0100-69912015005010.

Han, J., Kamber, M. and Pei, J. (2014) *Data mining: Data mining concepts and techniques*, *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. doi: 10.1109/ICMIRA.2013.45.

IFAB (2018) 'Leis do Jogo 2018/19'.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*. Available at: https://www.springer.com/gp/book/9781461471370%0Ahttp://www.springer.com/us/book/9781461471370.

Janosik, S. M. (2005) *Anticipating Legal Issues in Higher Education*, *NASPA Journal*. doi: 10.1017/CBO9781107415324.004.

Jiang, H. *et al.* (2016) 'Short-Term Speed Prediction Using Remote Microwave Sensor Data: Machine Learning versus Statistical Model', *Mathematical Problems in Engineering*, 2016. doi: 10.1155/2016/9236156.

Joseph, A., Fenton, N. E. and Neil, M. (2006) 'Predicting football results using Bayesian nets and other machine learning techniques', *Knowledge-Based Systems*, 19(7), pp. 544–553. doi: 10.1016/j.knosys.2006.04.011.

Lasek, J., Szlávik, Z. and Bhulai, S. (2013) 'The predictive power of ranking systems in association football', *International Journal of Applied Pattern Recognition*, 1(1), p. 27. doi: 10.1504/ijapr.2013.052339.

Librelotto, S. R. and Mozzaquatro, P. M. (2013) 'Análise dos algoritmos de mineração J48 e apriori aplicados na detecção de indicadores da qualidade de vida e saúde', *Revista Interdisciplinar de Ensino, Pesquisa e Extensão*, 1(1), pp. 26–37.

Lu, M. (2001) 'Vector autoregression (var) — an approach to dynamic analysis of geographic processes', *Geografiska Annaler: Series B, Human Geography*. Routledge, 83(2), pp. 67–78. doi: 10.1111/j.0435-3684.2001.00095.x.

Matuda, S. and Tagnin, S. (2014) 'A terminologia do futebol: um estudo direcionado pelo corpus', *Letras & Letras*, 30(2), pp. 214–243. doi: 10.14393/ll60-v30n2a2014-11.

Mehrez, A. and Hu, M. Y. (1995) 'Predictors of the outcome of a soccer game - a normative analysis illustrated for the Israeli Soccer League', *ZOR Zeitschrift für Operations Research Mathematical Methods of Operations Research*, 42(3), pp. 361–372. doi: 10.1007/BF01432510.

Menezes, B. (22/7/2020) https://globoesporte.globo.com/e-sportv/fifa/noticia/fifa-20-veja-como-os-atributos-dos-jogadores-afetam-a-gameplay.ghtml

Metzl, J. D. and Micheli, L. J. (1998) 'Youth soccer: An epidemiologic perspective', *Clinics in Sports Medicine*, 17(4), pp. 663–673. doi: 10.1016/S0278-5919(05)70110-1.

Min, B. *et al.* (2007) 'A Compound Framework for Sports Prediction: The Case Study of Football', *Knowledge-Based Systems or Expert Systems with Applications*, 21(February), pp. 551–562. Available at: http://www.sciencedirect.com/science/article/pii/S0950705108000609.

Mukaka, M. M. (2012) 'Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research', *Malawi Medical Journal*, 24(September), pp. 69–71. doi:

10.1016/j.cmpb.2016.01.020.

Murphy, D. F., Connolly, D. A. J. and Beynnon, B. D. (2003) 'Risk factors for lower extremity injury: a review of the literature'.

Nabinger, A. (2018) 'Utilização de algoritmos do tipo machine learning', *Society*, pp. 14–18. Available at: https://movisa.org.mx/images/NoBS_Report.pdf.

Noda, I. *et al.* (1998) 'Soccer server: a tool for research on multiagent systems', *Applied Artificial Intelligence*, 12(2–3), pp. 233–250. doi: 10.1080/088395198117848.

Owramipur, F., Eskandarian, P. and Mozneb, F. S. (2013) 'Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team', *International Journal of Computer Theory and Engineering*, 5(5), pp. 812–815. doi: 10.7763/ijcte.2013.v5.802.

Pascual, A. (2003) 'Assessing European stock market (co)integration', *Economics Letters*, 78, pp. 197–203. doi: 10.1016/S0165-1765(02)00245-8.

Passos, S. and Kronbauer, D. P. (2016) 'Uma Abordagem Estatística em Aprendizagem de Máquina para Previsões em Campeonatos de Futebol', *Univali*, (2016), pp. 2017–2019.

Picanço, A. R. S. *et al.* (2015) 'Um modelo de programação linear inteira para a tomada de decisão de manutenção preventiva', *XLVII Simpósio Brasileiro de Pesquisa Operacional (SBPO 2015)*, (October 2016), pp. 1449–1460.

Prasetio, D. and Harlili (2016) 'Predicting football match results with logistic regression', *4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016*. IEEE, pp. 2–6. doi: 10.1109/ICAICTA.2016.7803111.

Ramos, F. (2011) 'Cointegração, Modelos VAR e BVAR'.

Razali, N. *et al.* (2017) 'Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)', *IOP Conference Series: Materials Science and Engineering*, 226(1). doi: 10.1088/1757-899X/226/1/012099.

Reis, E. (2001) 'Estatística Multivariada Aplicada'.

Ribeiro, A. (2009) 'Forma(s) Desportiva(s) em Futebol.'

Di Salvo, V. *et al.* (2007) 'Performance characteristics according to playing position in elite soccer', *International Journal of Sports Medicine*, 28(3), pp. 222–227. doi: 10.1055/s-2006-924294.

Santos, F. *et al.* (2016) 'Análise do golo em equipas de elite de futebol na época 2013-2014 REDAF', *Revista de Desporto e Actividade Física*, 8(1), pp. 11–22.

Santos, P. (2016) 'a Organização Do Treino E Do Jogo', p. 120.

Schneider, C. (2018) 'Machine learning aplicado na previsão de resultados de partidas de futebol: um estudo de caso para comparação de diferentes classificadores', *Ufrgs*. Available at: https://www.lume.ufrgs.br/handle/10183/179461.

Settani Giglio, S. and Spaggiari, E. (2010) *A produção das ciências humanas sobre futebol no Brasil: um panorama (1990-2009)*.

Shephard, R. J. (1999) 'Biology and medicine of soccer: An update', *Journal of Sports Sciences*, 17(10), pp. 757–786. doi: 10.1080/026404199365498.

Silva, H. L. M. da (2008) 'A organização de jogo de uma equipa de futebol: uma abordagem sistémica para a construção de uma forma de jogar', p. 91.

Sims, C. A. (1980) 'Macroeconomics and Reality', *Econometrica*, 48(1), p. 1. doi: 10.2307/1912017.

Soares, J. (2016) 'Fisiologia do treino do futebolista', (Out).

Steiner, M. T. A. *et al.* (2006) 'Study of a medical problem using KDD, with emphasis on exploratory data analysis', *Gestao e Producao*, 13(2), pp. 325–337. doi: 10.1590/s0104-530x2006000200013.

Tampsell, C. (22/7/2020) https://www.eurogamer.pt/articles/2019-04-29-fifa-19-chemistry-como-aumentar-team-chemistry-individual-chemistry-e-maximizar-chemistry-em-ultimate-team

Tampsell, C. (22/7/2020) https://www.eurogamer.pt/articles/2019-04-29-fifa-19-chemistry-styles-atributos-afectados-e-quais-os-melhores-chemistry-styles-para-cada-posicao

Taylor, J. R. (1939) 'An introduction to error analysis: Mill Valley', p. 349.

Teodoro, P. and Veronez, D. E. S. (2013) 'João Paulo Teodoro De Sousa Veronez Futebol De Campo : As Variações Táticas Ofensivas Dentro Do Futebol De Campo : As Variações Táticas Ofensivas Dentro Do', pp. 1–71.

Vendite, C. and Moraes, A. (2006) 'Estratégia e tática de jogo: Uma análise dos profissionais que atuam no futebol'.

Vroonen, R. *et al.* (2017) 'Predicting the potential of professional soccer players', *CEUR Workshop Proceedings*, 1971, pp. 1–10.

Young, W. B., Miller, I. R. and Talpey, S. W. (2015) ‘Physical qualities predict change-of-direction speed but not defensive agility in australian rules football’, pp. 206–212.

SEDE ADMIN-

faculdade de **medicina**

faculdade dE **ciências**