



**UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS**



Alan Elias Santana Nascimento

**ESTIMATIVA POR APRENDIZAGEM DE MÁQUINA PARA O VALOR DE MERCADO DE
JOGADORES DE FUTEBOL**

São Cristóvão - SE

2023

Alan Elias Santana Nascimento

**ESTIMATIVA POR APRENDIZAGEM DE MÁQUINA PARA O VALOR DE MERCADO DE
JOGADORES DE FUTEBOL**

Trabalho de Conclusão de Curso apresentado ao
Departamento de Estatística e Ciências Atuariais
da Universidade Federal de Sergipe, como parte
dos requisitos para obtenção do grau de Bacharel
em Estatística.

Orientador: Prof. Dr. Cleber Martins Xavier

São Cristóvão - SE

2023

Alan Elias Santana Nascimento

**ESTIMATIVA POR APRENDIZAGEM DE MÁQUINA PARA O VALOR DE MERCADO DE
JOGADORES DE FUTEBOL**

Trabalho de Conclusão de Curso apresentado ao
Departamento de Estatística e Ciências Atuariais
da Universidade Federal de Sergipe, como parte
dos requisitos para obtenção do grau de Bacharel
em Estatística.

Aprovado em DATA DE APROVAÇÃO.

Prof. Dr. Cleber Martins Xavier
Orientador

**Prof. Dr. Sadraque Eneas de Figueriredo
Lucena**
Convidado 1

Prof. Dr. José Rodrigo Santos Silva
Convidado 2

São Cristóvão - SE
2023

Este trabalho é dedicado à minha mãe, Evalda Souza de Santana (in memoriam).

Agradecimentos

Foram árduos 5 anos, desde o dia 05/06/2017, a realização de um sonho que não era não só meu, mas também o sonho da minha mãe Evalda, falecida em 2019, que compartilhava do mesmo sentimento que sempre me apoiou e nunca mediu esforços para que este dia chegasse, a ela, do fundo do coração, meu muito obrigado e minha eterna gratidão, hoje com o sentimento de dever cumprido por nós e para nós, sempre.

Agradeço também em especial aos meus familiares, meu pai José Alberto e meus irmãos Abel Lyncon, João Pedro e José Alberto Filho. A minha avó Jovelivia e ao meu avô Elias, o qual tenho a honra de carregar seu nome. Por fim, a todos os parentes, amigos e irmãos de vida que são especiais e também caminharam junto comigo se fizeram presente durante esse processo.

Aos elos que criei dentro desse período de curso, de maneira especial aos amigos: Anderson, Glecio Lucas, Higor e João Guilherme. Onde compartilhei todas as conquistas e evolução dentro da universidade, e que com essa parceria desde o primeiro período de curso com o aperto em Cálculo I, e não somente, até a conclusão do curso, como também ao resto da vida, tive a satisfação de aprender e de construir bons momentos que levarei para a vida, inclusive no trabalho enquanto membros diretores da empresa júnior e também aos demais amigos da ATUE Jr.

Ao departamento de Estatística, no qual fui muito presente e especialmente ao meu orientador Cleber Xavier, pela paciência ensinamentos durante as disciplinas e a orientação e também aos professores Sadraque Lucena, Rodrigo, Luiz Henrique e Carlos Raphael os quais tive maior contato e a oportunidade de aprender, errar, sanar dúvidas e extrair o máximo de conhecimento durante a jornada acadêmica.

Ao professor Kleber Fernandes, o agradecimento por ter confiado no meu potencial e me projetado para experiências como estagiário no Governo do Estado de Sergipe, meu pontapé na carreira profissional onde fui muito feliz e pude construir elos profissionais e de fraternidade. E por fim aos amigos de trabalho da SIDI e PROPLAN que corroboraram com meu crescimento pessoal e profissional no dia a dia com incentivo, muito café e boas risadas.

Aos demais amigos e familiares não citados, mas que contribuíram com uma parcela significativa nesse meu ciclo da vida, minha eterna gratidão.

*“Obsessão ganha do talento todas as vezes.
Não podem te matar se você está morto.”
(Arremessando Alto)*

Resumo

O objetivo do presente estudo é identificar e analisar os fatores que influenciam a precificação de valor de mercado para jogadores de futebol por meio de método de aprendizagem de máquina baseado em Florestas Aleatórias (*Random Forest*, na terminologia em inglês). Para alcançar este objetivo foi necessário utilizar técnicas de raspagem de dados com o software livre R, além de técnicas estatísticas descritivas para compreender o perfil dos atletas com as informações coletadas e o método de regressão baseado em árvore para prever os valores de mercado desses atletas. Com os resultados obtidos, foi possível observar a presença da maioria de atletas jovens com idade abaixo de 25 anos. Além disso, verificar especificidades em relação a aspectos físicos de atletas, subdivididos por suas zonas de atuação em campo e a percepção da tendência física em cada posição, como é o caso dos Goleiros. A realização do algoritmo *Random Forest* demonstrou ser um modelo com boa capacidade preditiva para os valores de transferência. Como resultado, identificaram-se as variáveis mais relevantes para o modelo, sendo o salário dos atletas e atributos de habilidade, como drible e defesa.

Palavras-chave: Floresta Aleatória. FIFA22. Valor de mercado.

Abstract

The objective of the present study is to identify and analyze the factors contribute to the pricing of soccer players' market value using a machine learning method based on Random Forests. For this purpose, it was necessary to use a data scraping technique with the free software R. Additionally, descriptive statistics were used to understand the athletes' profile of with the collected information, and the tree-based regression method to predict these athletes' market values. The results showed that the a majority of the were young, below 25 years of age. In addition, specificities regarding the physical aspects of athletes, subdivided by their zones of action on the field, were observed, as well as the perception of the physical tendencies for each position, such as in the case of Goalkeepers. The implementation of the Random Forest algorithm demonstrated a model of good predictive capacity for transfer values. Therefore, the variables with the greatest relevance for the model were identified, including the athletes' salary and skill attributes, such as dribbling and defense.

Keywords: Random Forest. FIFA22. Market value.

Resumen

El objetivo de este estudio es identificar y analizar los factores que intervienen en la fijación de precios del valor de mercado de los jugadores de fútbol a través de un método de aprendizaje automático basado en Random Forests. Para ello fue necesario utilizar una técnica de raspado de datos con el software libre R, además de técnicas de estadística descriptiva para conocer el perfil de los atletas con la información recolectada y el método de regresión basado en árboles para predecir los valores de mercado de estos. Atletas. Con los resultados obtenidos, fue posible observar la presencia de una mayoría de atletas jóvenes menores de 25 años, además, verificar especificidades sobre los aspectos físicos de los atletas subdivididos por sus zonas de acción en el campo y la percepción de la Tendencia física para cada posición, como en el caso de los Porteros. La realización del algoritmo Random Forest, demostró en un modelo de buena capacidad predictiva para los valores de transferencia, y como resultado, se identificaron las variables con mayor relevancia para el modelo, siendo el salario de los atletas, y atributos de habilidad como regate y defensa.

Palabras clave: Árbol de decisión. FIFA22. Valor comercial.

Lista de ilustrações

Figura 1 – Demonstração do processo de reamostragem bootstrap	22
Figura 2 – Demonstração do algoritmo de Árvore de Decisão	24
Figura 3 – Histograma de distribuição do tempo de vida decorrido dos jogadores do FIFA22	28
Figura 4 – Boxplot de características físicas estruturais por zona de posição dos jogadores do FIFA22	29
Figura 5 – Correlação das variáveis coletadas na raspagem de dados do site SoFifa . . .	30
Figura 6 – Gráfico de densidade para os atributos de habilidades específicas dos atletas do FIFA22 por posição	31
Figura 7 – Predição do valor de mercado de jogadores de futebol masculino do Fifa 22 pelo modelo de floresta aleatória	33
Figura 8 – Variáveis influentes no modelo de regressão Random Forest	33

Lista de quadros

Quadro 1 – Variáveis selecionadas do WebScraping do site Sofifa	19
---	----

Lista de tabelas

Tabela 1 – Classificação Geral do atleta Casemiro	20
Tabela 2 – Modelos Random Forest com menores RMSE construídos pelo HyperGrid	32

Sumário

1	INTRODUÇÃO	13
2	OBJETIVOS	14
2.1	Geral	14
2.2	Específicos	14
3	REVISÃO LITERÁRIA	15
3.1	História do Futebol	15
3.2	Mercado de transferência	15
3.3	Aprendizagem de máquina aplicada ao Futebol	17
4	METODOLOGIA	19
4.1	<i>Web Scraping</i> (Raspagem de dados)	19
4.2	Cálculo do <i>Overall</i> (FIFA22)	20
4.3	Machine Learning	21
4.4	<i>Decision Tree</i>	23
4.5	<i>Random Forests</i>	25
4.5.1	Hiperparâmetros	26
4.5.2	<i>Grid Search</i>	27
4.5.3	Software e Pacotes	27
5	RESULTADOS	28
5.1	Análise Descritiva	28
5.2	Aplicação do modelo <i>Random Forests</i>	31
6	CONCLUSÃO	34
	REFERÊNCIAS	35
	APÊNDICE A – CÓDIGO DA RASPAGEM, LIMPEZA DOS DADOS E APLICAÇÃO DO MODELO	38

1 INTRODUÇÃO

O futebol mundial, nos dias atuais, é um grande negócio que envolve agentes diretos, como clubes e federações, e indiretos, como indústrias de equipamentos e produtos esportivos, além da mídia. A caracterização da cadeia de produção da indústria de futebol, com seus vários agentes interdependentes e interesses muitas vezes conflitantes, representa uma das principais condicionantes dessa indústria. (LEONCINI; SILVA, 2005).

De acordo com Coelho (2015), na segunda metade da década de 2000, surgem no futebol os fundos de investimentos. Consequentemente, isso impacta nas parcerias de clubes nos contratos de atletas, com a pretensão de vender atletas mais jovens para obter um retorno imediato sobre o investimento. Entretanto, esse efeito contribui também para a manutenção e suporte de uma parte da elite dos craques brasileiros em clubes de primeiro escalão nacional.

Tonini e Giglio (2019) analisaram o sistema de transferência da FIFA para jogadores de futebol. Assim, os movimentos migratórios de atletas de grande renome no cenário esportivo, que envolvem altas cifras, apresentam-se de maneira mais objetiva como um novo modelo de negociação. Esse fato começa a despertar o interesse de empresários que vislumbram no futebol uma oportunidade de transformação em uma máquina de geração de investimento a curto prazo.

Para compreender a precificação do valor de mercado de jogadores, o método de extração de dados *web scraping* é aplicado para informações de atleta de futebol masculino do FIFA 22, do site Sofifa. Com isso, é possível analisar o perfil dos atletas e seus respectivos valores de mercado, utilizando-se de algoritmos de *machine learning*, especificamente modelos baseados em florestas aleatórias para regressão, com a intenção de observar o comportamento das covariáveis, avaliar o impacto das mesmas na explicação do modelo e identificar os movimentos da precificação de mercado.

A estrutura do trabalho está subdividido em seis seções, iniciando pelo Capítulo 1, de caráter introdutório que contém as considerações iniciais sobre o tema a ser abordado. O Capítulo 2 descreve os objetivos a traçados para o desenvolvimento deste estudo. O Capítulo 3 busca embasar e organizar fundamentação teórica, demonstrando características da economia no futebol. Em seguida, o Capítulo 4 é dedicado à descrição dos materiais e métodos necessários para ser possível chegar nos resultados, exemplificando sobre a base de dados, *Machine Learning* e *Random Forests*. O Capítulo 5 apresenta os resultados obtidos alinhados com a discussões sobre os mesmos. Para finalizar, o Capítulo 6 demonstra as conclusões obtidas conforme os objetivos apresentados.

2 OBJETIVOS

2.1 Geral

Estimar o valor de mercado de jogadores de futebol, por meio de modelo de regressão baseado em árvore.

2.2 Específicos

- Coletar dados dos jogadores de futebol masculino do FIFA22 via *web scraping*;
- Realizar análise descritiva dos dados coletados para observar e resumir o comportamento e as características dos jogadores;
- Aplicar o modelo de regressão baseado em árvore *Random Forest* e realizar o treinamento do mesmo utilizando as variáveis selecionadas;
- Avaliar a acurácia do modelo e avaliar os resultados apresentados;
- Identificar as variáveis influentes no modelo e avaliar sua capacidade preditiva para a estimativa do valor de mercado dos jogadores de futebol.

3 REVISÃO LITERÁRIA

3.1 História do Futebol

O futebol é uma modalidade esportiva que originou-se no Reino Unido, nadurante a segunda metade do século XIX. É um esporte praticado em um campo com dimensões e regras específicas, com 11 jogadores de cada equipe em disputa, sendo que apenas um jogador de cada lado poder tocar a bola com as mãos, desde que dentro da área delimitada (SANTOS, 2020).

A Federação Internacional de Futebol Associação (FIFA) foi criada em 21 de maio de 1904, em Paris. O estatuto de 1912 faz referência a uma decisão tomada em junho de 1908 e trazia em sua composição que um atleta não poderia jogar por duas associações nacionais na mesma temporada sem que houvesse um mútuo consenso entre as associações envolvidas. Este foi o primeiro apontamento da FIFA sobre a circulação de jogadores tendo como foco as seleções nacionais. Com isso, a preocupação amplia-se a partir de dois aspectos que produzem marcas importantes na entidade: as disputas em torno do amadorismo e do profissionalismo, e a presença da transferência dos atletas entre os clubes de futebol (TONINI; GIGLIO, 2019).

Souza e Angelo (2005) definem e dividem o mercado de jogadores profissionais de futebol em dois períodos distintos, antes e depois de 1996. Até este ano, o método de transferência de atletas era o passe, modelo de negociação em os jogadores somente podiam ser negociados de um clube para outro com o consentimento mútuo entre os dois clubes. A negociação consistia basicamente em um valor a ser pago pela equipe compradora à detentora dos direitos sobre o profissional. No mercado europeu, além da existência do passe, havia a regra que um clube só poderia contar com até três jogadores de outra nacionalidade em uma mesma partida, o que limitava a contratação de atletas de outros países pelas equipes.

Em 15 de dezembro de 1995, a Lei Bosman revolucionou o futebol na União Europeia, ao eliminar as transferências de jogadores sem contrato com suas equipes que desejavam mudar de clube dentro e entre países da União Europeia. Além disso, a lei tornou ilegais as cotas sobre o número de estrangeiros que jogam em um clube. Como resultado, os órgãos dirigentes de vários países europeus aboliram imediatamente as cotas de jogadores estrangeiros da União Europeia para partidas de clubes disputadas em suas ligas domésticas (BINDER; FINDLAY, 2012).

3.2 Mercado de transferência

Uma questão fundamental para a contabilidade de um clube de futebol, objeto de pesquisa científica, são os direitos de desempenho dos jogadores de futebol. Eles são contabilizados como ativos intangíveis no balanço patrimonial. Embora os futebolistas tenham contratos diferentes, na maioria deles existe uma cláusula de pagamento para a renúncia ao contrato. Isso significa que se

algum clube quiser contratar tal jogador de futebol, deverá pagar ao atual empregador a cota descrita na cláusula. Outro ponto de vista é que os direitos de desempenho dos jogadores são uma questão comercial e seu valor de mercado é mais de duas vezes maior que o valor dos ativos intangíveis no balanço (MAJEWSKI; University of Szczecin, 2016).

Os clubes de futebol estão envolvidos em três atividades principais. A primeira atividade é a geração de receitas com a venda de ingressos, transmissão, serviços de alimentação e patrocínios. A segunda atividade é a negociação (aquisição e alienação) de contratos de jogadores. A terceira atividade é desenvolver e nutrir talentos internos (AMIR; LIVNE, 2005).

O objetivo principal de qualquer mercado de transferências deve ser duplo: facilitar e organizar a aquisição e troca de jogadores pelos clubes para permitir a reconstituição de equipes com o objetivo de aumentar a força de jogo e melhorar o desempenho da equipe; para facilitar o movimento de jogadores entre clubes em busca de melhores oportunidades, maiores ganhos ou maior satisfação no trabalho. Na prática, a maioria dos mercados de trabalho esportivos em esportes coletivos profissionais estão sujeitos a controles restritivos. Esses controles geralmente envolvem a incapacidade dos jogadores de transferir ou serem liberados para um novo clube sem a permissão de seus clubes atuais e a necessidade de todas as transferências estarem em conformidade com os regulamentos do órgão regulador relevante. (CARMICHAEL; THOMAS, 1993)

Como em qualquer outro setor da economia, os clubes de futebol podem perder os jovens que desenvolveram para outros clubes, o que poderia ser preocupante se o objetivo fosse mantê-los e integrá-los na equipe principal. Por outro lado, os clubes podem investir em academias de base para gerar lucro com a venda desses jogadores no mercado de transferências, se o foco for a maximização do lucro no curto prazo. Trazer jogadores de fontes externas traz o risco inerente de não se adequarem à nova equipe e sistema, o que pode decorrer de suas características pessoais e profissionais. A escala de um clube e os recursos financeiros disponíveis também são determinantes importantes nessas decisões; clubes maiores têm mais opções disponíveis a esse respeito, enquanto clubes menores não têm capacidade financeira para competir pelos melhores talentos no mercado de transferências, o que pode ditar sua decisão de se concentrar no canal de desenvolvimento juvenil (RADOMAN; VOIA, 2022).

Proni (1998) define a associação entre marketing esportivo e o modelo de negócios do futebol, no qual as grandes equipes buscam consolidar e viabilizar a participação em torneios internacionais, fator financeiro importante para aumentar a geração de receitas e atrair patrocinadores com maior investimento. Ele também destaca que as ligas nacionais com economias mais fortes fortalecem seu posicionamento financeiro para ampliar sua participação no mercado do “produto futebol”, que possui a concentração de atletas mais valiosos e espetáculos com maior qualidade técnica, a fim de garantir as maiores e melhores cotas de transmissão e preferência do produto pelo público.

Os valores de mercado dos jogadores são estimativas das taxas de transferência que

provavelmente serão pagas por eles. Assim, as taxas de transferência são uma aproximação do valor de mercado que pode ser interpretada como uma quantia aproximada de dinheiro que é atraente o suficiente para que o jogador e o clube de futebol aceitem a oferta. Os jogadores são vitais em um clube de futebol profissional por causa de seu talento e seu valor de mercado. O valor de mercado de um jogador reflete o valor financeiro que o clube empregador atribui ao desempenho do jogador. Portanto, há um obstáculo ao tentar avaliar os valores de mercado dos jogadores de futebol, pois é extremamente subjetivo (ARRUL; SUBRAMANIAN; MAFAS, 2022).

Determinar as compensações para um jogador de futebol é uma tarefa complexa na economia do esporte, afetada por uma combinação de inúmeros aspectos. Na era pré-informativa, essa tarefa era realizada usando principalmente análises qualitativas, também porque as estatísticas e os dados dos jogadores de futebol não eram coletados de forma consistente, tornando mais difícil comparar o desempenho e as habilidades dos jogadores de futebol de maneira quantitativa (FRICK, 2006).

Uma transferência bem-sucedida é aquela que afeta altos resultados em todos os aspectos do jogador. O sucesso da transferência pode depender da qualidade geral da equipe alvo. Por exemplo, é muito difícil marcar gols com frequência quando os companheiros de equipe são tecnicamente mais fracos e não dão suporte devidamente ao jogador. Posteriormente, compara-se a avaliação do jogador com a avaliação média da equipe. Se a avaliação da pessoa for maior ou igual ao valor médio, a transferência é considerada bem-sucedida. A última definição de transferência bem-sucedida é novamente baseada na avaliação do jogador médio da equipe, mas introduz uma modificação: a média é calculada com apenas 18 pontuações (onze iniciais mais sete substituições). Isso ocorre, de um modo geral, que outros jogadores não aparecem na escalação (ĆWIKLINSKI; GIELCZYK; CHORAŚ, 2021).

3.3 Aprendizagem de máquina aplicada ao Futebol

Nos últimos anos, um modelo de valor esperado do gol (xG) foi desenvolvido para avaliar o desempenho ofensivo de jogadores e times. O modelo xG atribui um valor entre 0 e 1 (com 1 sendo o máximo e representando um determinado gol) para cada tentativa, com base na quantidade e qualidade, como o tipo de assistência, ângulo de chute e distância do gol, e se foi um chute de cabeça (HEROLD et al., 2019).

Embora vários desses modelos tenham se mostrado valiosos na previsão de resultados de arremates a gol e observação de jogadores com altas taxas de conversão, eles não reconheceram ou capturaram o posicionamento do oponente, e portanto, falharam em fornecer contexto que treinadores e analistas possam aplicar à partida. Em geral, os estudos conduzidos para prever o sucesso individual ou da equipe usaram amostras maiores para construir seu modelo. No entanto, eles usaram principalmente modelos de regressão bastante simples para construir modelos de classificação ou previsão (HEROLD et al., 2019).

A pesquisa em análise de futebol com técnicas de *Machine Learning* é limitada. A principal razão para isso é a falta de um conjunto de dados em grande escala para os jogadores, o que é um problema, pois reunir informações tão detalhadas sobre os jogadores pode ser caro, tornando os dados sensoriais limitados a equipes com alto poder aquisitivo. Na análise de futebol, videogames como FIFA e Football Manager são considerados como outra fonte de dados. Desde 2014, pesquisadores e clubes utilizam os videogames como fontes alternativas de dados (AL-ASADI; TASDEMIR, 2022).

Para fornecer esse nível de autenticidade, especialistas foram contratados para coletar informações sobre o estilo de jogo de cada jogador e selecionar esses dados para que cada jogador possa ser representado com precisão como uma combinação de mais de 33 recursos. Esses recursos incluem habilidades físicas (velocidade, potência, aceleração, etc.) e técnicas (driblar, cabecear, precisão, etc.) representadas como um número inteiro de 1 a 100. Esses recursos são atualizados constantemente à medida que a forma dos jogadores muda ao longo da temporada. É interessante ressaltar que esses dados já estão sendo usados para resolver outros problemas, os olheiros da Premier League usam videogames para procurar jogadores jovens e promissores com habilidades específicas. E não faz muito tempo que um jogador de videogame foi nomeado gerente geral de um time de futebol real (SHIN; GASPARYAN, 2014).

De acordo com a análise dos dados, o valor de mercado do jogador de futebol é afetado por inúmeros fatores que não estão diretamente relacionados ao desempenho ou habilidades. Por exemplo, a reputação internacional de um jogador é a segunda característica mais importante na determinação do valor de um jogador, depois do potencial do jogador, de acordo com todos os métodos usados pela experiência. Portanto, concluímos que o valor de mercado dos jogadores depende em grande parte de seu poder de atração de público (AL-ASADI; TASDEMIR, 2022)

4 METODOLOGIA

4.1 Web Scraping (Raspagem de dados)

Para desenvolver o presente estudo e aplicar as técnicas necessárias, foi necessário utilizar a da técnica de *web scraping*. Como fonte da coleta desses dados foi utilizado o site [Sofifa](#) que reúne exclusivamente, informações do jogo *FIFA22* produzido e licenciado pela *Eletronic Arts*.

[Mitchell \(2018\)](#) em seu livro descreve a *web scraping* como uma prática de extração de dados que não envolve a interação de um ser humano usando um navegador de internet. De modo geral, ela é feita escrevendo um código de programação automatizado que consulta um servidor *web*, requisita dados na linguagem de marcação HTML e de outros arquivos e linguagens que compõem as páginas *web*.

Com o banco de dados do *FIFA22* tendo sua última atualização em 18/08/2022, o processo de extração das informações obteve detalhes sobre 19.320 jogadores de futebol masculino com 23 variáveis sobre as características individuais de cada jogador, detalhado no Quadro 1.

Quadro 1 – Variáveis selecionadas do WebScraping do site Sofifa

Variável	Descrição
sofifa_id	Código de identificação do site
name	Nome do atleta
team	Clube do atleta
position	Posição em campo por terço do campo
foot	Pé dominante
best_position	Posição preferida do atleta
height	Altura
weight	Peso
age	Idade
ova	Overall (pontuação do jogo)
pot	Potencial máximo a alcançar
gk_handling	Habilidade de goleiro
weak_foot	Habilidade com pé não dominante
international_reputation	Reputação internacional do jogador
pac	Pace
sho	Chute
pas	Passe
dri	Drible
def	Defesa
phy	Físico
wage	Salário
value	Valor de mercado

Fonte: Elaborado pelo autor

4.2 Cálculo do *Overall* (FIFA22)

A habilidade do jogador no *FIFA22* não é mensurada apenas a partir dos atributos de pace, chute, passe, drible, defesa e físico. De acordo com [Shin e Gasparyan \(2014\)](#) no Capítulo 3, ao todo, somam-se mais de 33 estatísticas para serem transformadas no valor do *Overall*, que é a classificação geral do atleta no jogo. Entretanto, para as diferentes posições, há diferentes mensurações nessa classificação. Ou seja, para atletas que atuam no último terço do campo, há peso maior em habilidades como precisão no chute, posicionamento no ataque e passe, que obtém um peso maior comparado a habilidades de defesa, que prioritariamente não é uma função da posição.

O cálculo é uma ponderação dada por habilidades da posição pré-definidas, que somadas chegam a 1, e multiplicadas pela classificação do atleta naquela determinada habilidade, gerando assim o coeficiente do *Overall* para cada habilidade. As somas dessas multiplicações retornam a classificação geral do atleta pelos atributos.

Além das estatísticas, o reconhecimento internacional do atleta, que é classificado de 1 a 5, também é um peso na hora de definir a classificação geral. Desta maneira, permite o balanceamento entre atletas de ligas menos populares com boas características físicas como velocidade e força, que não sejam melhor classificados que atletas mais lentos que atuam em alto nível nas ligas de alto rendimento. A reputação internacional não modifica nem representa a capacidade técnica do atleta. Trata-se apenas de uma métrica para convergir a classificação geral do jogadores mais comuns entre a comunidade que consome o jogo.

Tabela 1 – Classificação Geral do atleta Casemiro

Estatística	Coeficiente	Pontuação	C x P
Passe Curto	0,13	84	10,92
Interceptações	0,12	87	10,44
Passe Longo	0,11	84	9,24
Marcação	0,10	85	8,50
Desarme em pé	0,10	88	8,80
Reações	0,09	87	7,83
Controle de bola	0,09	79	7,11
Reações	0,08	87	6,96
Visão	0,06	80	4,80
Força	0,06	90	5,40
Agressão	0,05	91	4,55
Total	1	942	84,55

Fonte: Elaborado pelo autor

A Tabela 1 apresenta as estatísticas do meio-campista defensivo brasileiro Carlos Henrique Casemiro, que é avaliado no jogo *FIFA22* com *Overall* igual a 89, com reputação internacional igual a 4.

Após o cálculo dos atributos, onde a multiplicação entre o coeficiente e a pontuação em cada atributo é feita a soma dessas multiplicações resultando no valor final da classificação geral do volante Casemiro de 84,55, que arredondado para cima retorna um *Overall* igual à 85. No entanto o valor final do *Overall* do atleta é 89 devido a outros ajustes realizados pelo jogo.

Os valores da reputação internacional para balanceamento dos atletas é adicionado posteriormente, já que Casemiro atuava na liga espanhola, uma das cinco principais ligas europeias, também conhecidas como *Big Five*, que incluem a liga espanhola, inglesa, francesa, italiana e alemã. Portanto a classificação geral do jogador após o cálculo de coeficiente e pontuação e a soma total dos pontos, além da reputação e os balanceamentos adicionais realizados, obtém-se a pontuação final do jogador no jogo.

4.3 Machine Learning

Machine Learning (Aprendizado de Máquina) é uma área de estudo interessada no desenvolvimento de algoritmos computacionais para transformar dados em ações inteligentes. Utilizando métodos estatísticos como base, os algoritmos de aprendizado de máquina são divididos em categorias de acordo com sua finalidade. As duas principais categorias são: aprendizado supervisionado, que constrói modelos preditivos, e o não supervisionado, que constrói modelos descritivos (LANTZ, 2015).

Modelos de aprendizado de máquina são capazes de aprender com dados sem a ajuda explícita de um humano. Essa é a principal diferença entre os modelos de aprendizado de máquina e os modelos de algoritmos clássicos. Algoritmos clássicos são ensinados a encontrar a melhor resposta em um complexo sistema e o algoritmo, em seguida, procura essas melhores soluções e muitas vezes funciona mais rápido e mais eficiente do que um humano (OZDEMIR, 2016).

Em síntese, para aplicação do método supervisionado, o modelo necessita de informações prévias para assimilar e compreender o comportamento e, assim, poder prever a partir da variável dependente. Enquanto que no método não-supervisionado não há referências prévias para que o modelo adote como base, ou seja, o modelo aprende a analisar sozinho por algumas técnicas, como a de agrupamento.

A performance de um algoritmo de *machine learning* é avaliada separando os dados: uma parte para treinamento e outra para teste. As recomendações típicas para dividir os dados em divisões de treinamento-teste é 60% e 40%, 70% e 30% ou 80% e 20%. Em geral, essas são diretrizes apropriadas a serem seguidas. No entanto, é importante observar que definir uma proporção muito alta para a base de treinamento pode levar a um desempenho preditivo insuficiente. (*Underfitting*), enquanto ajustar um modelo muito bem aos dados de treinamento, pode resultar em uma falta de generalização para novos dados (*Overfitting*). Além disso, uma grande quantidade de observações em banco de dados de testes não nos permite obter uma boa avaliação dos parâmetros do modelo (BOEHMKE; GREENWELL, 2020).

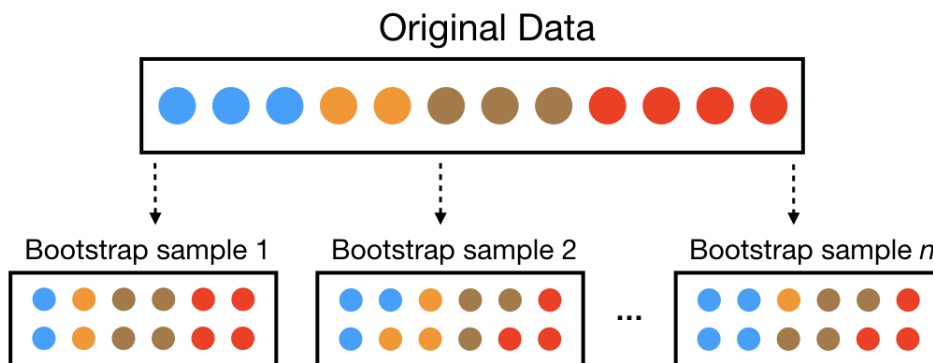
Izbicki e Santos (2020) definem como boa prática a escolha das amostras para compor os dados de treinamento e os dados de validação, por meio de processos aleatórios. O método mais comumente usado para gerar números aleatórios é adotado para escolher as amostras pertencentes aos dados de treinamento, e os dados de validação.

Os métodos de reamostragem fornecem uma abordagem alternativa, permitindo ajustar repetidamente um modelo de interesse a partes dos dados de treinamento e testar seu desempenho em outras partes. Os dois métodos de reamostragem mais usados incluem validação cruzada k-fold e *bootstrapping* (BOEHMKE; GREENWELL, 2020).

A validação cruzada k-fold, divide os dados aleatoriamente em k lotes disjuntos com aproximadamente o mesmo tamanho, utilizando-se de um processo aleatório. Sejam $L_1, \dots, L_k \subset \{1, \dots, n\}$ os índices associados a cada um dos lotes. A ideia do método é criar k estimadores da função de regressão, $\hat{g}_{-1}, \dots, \hat{g}_{-k}$, em que $\hat{g}_{-j}$ é criado usando todas as observações do banco menos aquelas do lote L_j (IZBICKI; SANTOS, 2020).

Uma amostra *bootstrap* é uma amostra aleatória dos dados obtidos com reposição definida por Efron e Tibshirani (1986). Isso significa que, após um ponto de dados ser selecionado para inclusão no subconjunto, ele é colocado de volta no conjunto de dados original e pode ser selecionado para inclusão em outra amostra *bootstrap*. Uma amostra *bootstrap* tem o mesmo tamanho do conjunto de dados original a partir do qual foi construída. A Figura 1 fornece um esquema de amostragem *bootstrap* onde cada amostra *bootstrap* contém 12 observações exatamente como no conjunto de dados original. Além disso, a amostragem *bootstrap* contém aproximadamente a mesma distribuição de valores (representados por cores) que o conjunto de dados original (BOEHMKE; GREENWELL, 2020).

Figura 1 – Demonstração do processo de reamostragem bootstrap



Fonte: (BOEHMKE; GREENWELL, 2020)

O processo de ajuste automático de um modelo consiste na seleção do melhor modelo dentre os candidatos disponíveis. Para modelos de classificação, as estatísticas utilizadas para avaliação incluem acurácia, kappa e área sob a curva ROC (AUC), enquanto modelos de regressão são avaliados através de métricas como R^2 ou RMSE. Adicionalmente, medidas sensíveis ao custo, como sensibilidade e especificidade, também podem ser empregadas durante a avaliação (LANTZ, 2015).

Neste presente estudo, a métrica utilizada para avaliar o desempenho do modelo foi a Raiz do Erro Quadrático Médio (RMSE), definida como a raiz quadrada da média dos quadrados dos desvios entre os valores observados e previstos. Quanto menor for o valor do RMSE, melhor será o desempenho do modelo de regressão. Ela é definida como

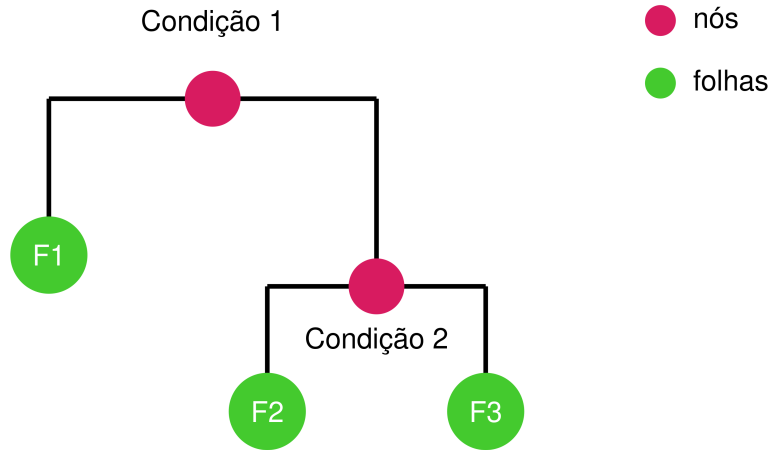
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (4.1)$$

4.4 Decision Tree

Os modelos baseados em árvore são uma classe de algoritmos não paramétricos que funcionam particionando o conjunto de dados em várias regiões menores e não sobrepostas cujos valores de resposta são semelhantes, usando um conjunto de regras de divisão. As previsões são obtidas ajustando um modelo mais simples como uma constante (por exemplo, o valor médio da resposta), em cada região. Esses métodos de divisão e conquista podem produzir regras simples que são fáceis de interpretar e visualizar com diagramas de árvore (BOEHMKE; GREENWELL, 2020).

Na Figura 2 é possível notar como é construída a árvore de decisão. Inicialmente, o topo da árvore é composto pelo nó raiz, que representa todo o conjunto de dados. Em seguida, a árvore é dividida em ramos representados pelas setas do diagrama. Se a condição imposta for satisfeita, o algoritmo continua percorrendo a árvore pela esquerda. Caso contrário, percorre pela direita, resultando nos nós folhas onde é impresso o resultado final.

Figura 2 – Demonstração do algoritmo de Árvore de Decisão



Fonte: (IZBICKI; SANTOS, 2020)

Entretanto, ao analisar o gráfico da árvore de decisão, (BOEHMKE, 2018) conclui que o algoritmo utiliza um particionamento recursivo binário. A regra dependem do nó anterior, e a intenção é encontrar o melhor nó para Y_i a fim de continuar particionando em duas regiões (R_1, R_2). O modelo prevê Y com uma constante c_m para região R_m , demonstrado na Equação 4.2:

$$\hat{f}(X) = \sum_{m=1}^n c_m I(X_1, X_2) \in R_m. \quad (4.2)$$

Para minimizar o erro entre o valor real e o valor previsto, (BOEHMKE, 2018) também destaca que o modelo explora cada valor distinto de cada variável de entrada para encontrar o preditor e o valor de divisão que particiona os dados em duas regiões, como é explicitado na Equação 4.3:

$$\text{minimize} \left\{ SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2 \right\}. \quad (4.3)$$

Para otimizar o desempenho preditivo em dados não visto, é necessário encontrar um equilíbrio entre a profundidade e a complexidade da árvore. Encontrar esse equilíbrio, normalmente é construída uma árvore grande, como descrito na seção anterior, em seguida é feita uma poda para encontrar a subárvore ideal. Para encontrar a subárvore ótima, utiliza-se um parâmetro de complexidade de custo (α), que penaliza a função objetivo com base no número de nós terminais da árvore (T) (BOEHMKE, 2018). Matematicamente, temos

$$\text{minimize}\{SSE + \alpha|T|\}. \quad (4.4)$$

Ou seja, em algum momento α assumirá um valor que definirá o limiar para podar a fim de obter a menor r possível com o menor erro de penalização. Porém, para valores mais baixos de α , mais complexa se torna a construção do modelo. Por esta razão, à medida que a árvore cresce, a redução do SSE deve ser maior que α para que a poda seja realizada.

4.5 Random Forests

Em 2001, Breiman descreveu o método *Random Forests* em uma combinação de *bagging* e ramificação aleatória, descrevendo como um modelo com alta precisão e robustez. Breiman (2001) destaca também que as florestas aleatórias tem a capacidade de trabalhar com múltiplos atributos e relações não-lineares, o que os torna uma ferramenta produtiva para classificação e regressão.

Florestas aleatórias são compostas por um conjunto de árvores de decisão, em que cada árvore é construída com base em um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta. O erro de generalização para florestas converge até um limite à medida que o número de árvores na floresta aumenta. O erro de generalização de uma floresta de classificadores de árvores depende da força das árvores individuais na floresta e da correlação entre elas (BREIMAN, 2001).

De acordo com Lantz (2015), as florestas aleatórias combinam versatilidade e poder em uma única abordagem de aprendizado de máquina. Como o conjunto usa apenas uma parte pequena e aleatória do conjunto completo de recursos, as florestas aleatórias são capazes de lidar com conjuntos de dados extremamente grandes.

Lantz (2015) completa afirmando que em comparação com outros métodos baseados em conjuntos, as florestas aleatórias são altamente competitivas e tendem a ser mais fáceis de usar, além de serem menos propensas a superajuste. O método é demonstrado no Algoritmo 1, como definido por (IZBICKI; SANTOS, 2020).

Algoritmo 1: Aplicação do Random Forest para modelos de regressão

1. Com $b = 1, 2, \dots, B$, aplica-se:
 2. Criar B amostras bootstrap de tamanho N a partir dos dados originais;
 - a) Cada nó só pode escolher uma entre todas as covariáveis
 3. Obtém as árvores $g^b(x)$
 4. Novas entradas no modelo em regressão define-se: $g(x) = \frac{1}{B} \sum_{b=1}^B g^b(x)$
-

Liaw e Wiener (2002) também definem que uma estimativa da taxa de erro pode ser obtida, com base nos dados de treinamento em cada iteração de *bootstrap*, prevendo os dados que não foram selecionados na amostra *bootstrap* (definido por Breiman (2001) como “out-of-bag” ou OOB).

4.5.1 Hiperparâmetros

Um dos aspectos mais desafiadores, mas também potencialmente importantes, da otimização de um modelo é escolher os valores para os hiperparâmetros. Em teoria, o objetivo é escolher a melhor combinação possível, embora seja improvável encontrar o máximo global. Melhores valores para os hiperparâmetros podem melhorar a precisão de um modelo (WILEY, 2016).

O desempenho e a performance dos modelos *Random Forests* dependem de seus hiperparâmetros, os quais são ajustados e balanceados durante o processo de treinamento do modelo. Dentre os diversos hiperparâmetros disponíveis na metodologia do modelo, alguns dos abordados por Boehmke e Greenwell (2020) serão explicados, como o *mtry*, *ntrees*, *node size*, *sample size* e *splitting rule*.

O **mtry** é um dos mais importantes hiperparâmetros e, em síntese, trata-se da definição do número de covariáveis que serão levadas em consideração ao definir cada nó. Em problemas de regressão, o valor recomendado para *mtry* é $m_{try} = \frac{p}{3}$, onde p é o número de variáveis preditoras. Valores mais baixos para o *mtry* tendem a resultar em baixa diversificação das árvores e sensibilidade menor a ruídos, enquanto valores mais altos levam a ramificações com maior diversidade e mais propensas a problemas com *overfitting*. Com isso, o *mtry* é um hiperparâmetro crucial para controlar a variação das árvores e a performance do modelo.

O hiperparâmetro **ntrees** define o número de árvores a serem criadas na floresta aleatória. Quanto maior o valor de *ntrees* mais robusto será o modelo, porém, aumentará a probabilidade de *overfitting*. Portanto, esse hiperparâmetro controla a robustez do modelo e é sensível ao tamanho da amostra.

Node Size, dentro do modelo de floresta aleatória tem como função controlar o tamanho mínimo dos ramos finais, ou seja, o *node size* define o número mínimo de amostras necessárias para que um nó seja dividido. Se o *node size* for muito pequeno, obtém-se árvores com grande profundidade e ramos pequenos, caindo em problemas de *overfitting*. Por outro lado, se o *node size* for muito grande, pode-se obter árvores curtas e com poucas ramificações, se tornando propensa a *underfitting*. Portanto, infere-se que o *node size* tem a função de controlar a profundidade das árvores e equilibrar a capacidade de generalização do modelo com a capacidade de ajuste aos dados de treinamento.

Para controlar o tamanho da amostra de cada árvore a ser treinada, existe um hiperparâmetro conhecido como **sample size**, ou seja, o número de amostras a serem escolhidas de

maneira aleatória da base de treino para cada árvore. Intuitivamente, há problema com *overfitting* se o *sample size* for igual ao número total de dados de treinamento, pois isso resultará em um ajuste excessivo dos dados, em que resultará numa árvore igual a árvore de decisão. Por outro lado, se o valor for menor que o tamanho base de treino, o modelo terá menos capacidade de memorização e menor tendência a *overfitting*.

Splitting Rule, ou regra de divisão, é um hiperparâmetro responsável por definir a maneira como a amostra de treinamento será dividida em modelos de árvore de decisão. Em modelos de regressão, a regra de divisão é responsável por identificar a melhor maneira de dividir os dados com base em suas características para gerar previsões da variável preditora dentro da amostra de teste. Pode-se utilizar de métricas como a média, mediana ou outras para definir o critério de divisão em cada nó. A escolha deste hiperparâmetro afeta na capacidade da floresta aleatória de generalizar novas entradas no modelo.

4.5.2 Grid Search

Em uma pesquisa de grade, diversos valores para hiperparâmetros são especificados e todas as combinações possíveis são tentadas. A pesquisa de grade é excelente quando há poucos valores para alguns parâmetros. No entanto, embora essa seja uma forma abrangente de avaliar diferentes valores de parâmetros, quando há muitos valores para alguns ou muitos parâmetros, rapidamente se torna inviável (WILEY, 2016).

Se não houver interações entre os parâmetros e o desempenho do modelo, ou pelo menos as interações forem pequenas em relação aos efeitos principais, as pesquisas em grade serão uma abordagem ineficiente porque muitos valores de parâmetros serão repetidos, fazendo com que apenas um pequeno conjunto de valores seja amostrado, mesmo que muitas combinações sejam tentadas. Em R é possível utilizar a função `expand.grid()` para criar todas as combinações possíveis de variáveis (WILEY, 2016).

4.5.3 Software e Pacotes

Durante todo o estudo foi utilizado o software estatístico R, que é livre e de código aberto, em sua versão 4.2.0, juntamente com o ambiente de desenvolvimento RStudio na versão 2022.07.2-576. Na etapa de *web scraping* foi utilizado o pacote `rvest`, além do pacote base e do `dplyr` para manipulação e limpeza dos dados, o `ggplot2` para construção dos gráficos e o pacote `ranger`, dos autores Wright e Ziegler (2017), para uso de funções como a `expand.grid()` e demais aplicações do modelo de floresta aleatória.

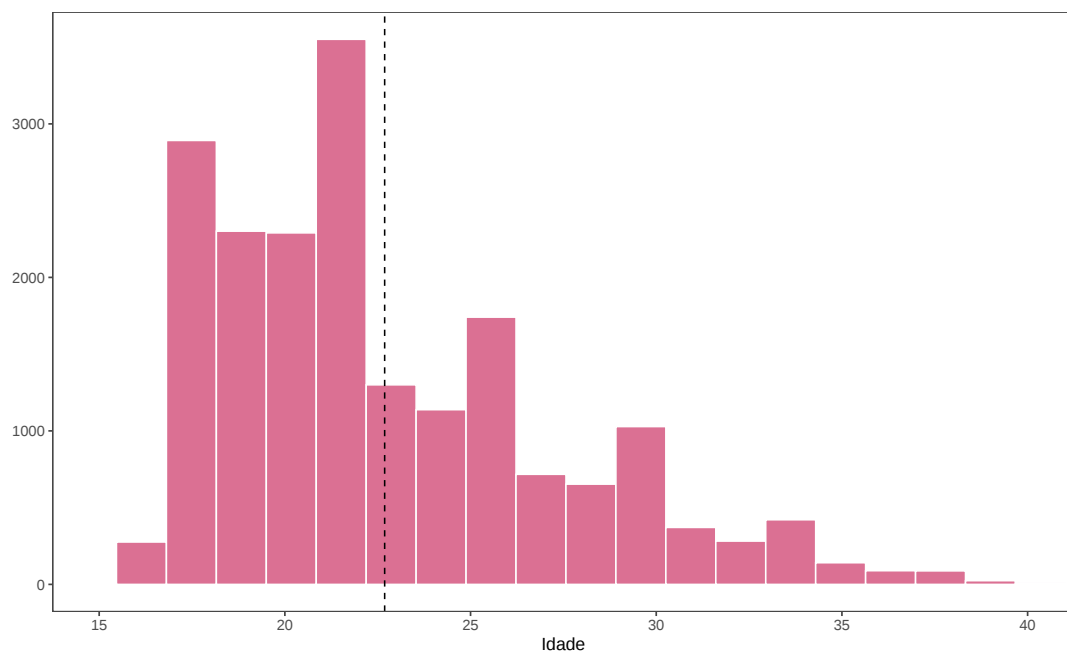
5 RESULTADOS

5.1 Análise Descritiva

De maneira analítica, na Figura 3, é destacada uma densa concentração nas idade mais jovens, o que é comum devido ao tempo de carreira de um atleta de futebol, geralmente dura de 15 a 20 anos. Conforme demonstrado pela linha tracejada, a média de idade de 22,7 anos reflete a maior concentração de jogadores entre 16 a 23 anos.

Carli et al. (2009) destacam a vantagem da idade entre os jogadores de futebol masculino, pois é preciso atingir a capacidade máxima de velocidade e agilidade nos movimentos, além de ter um excelente domínio espacial e temporal para realizar divididas com contato físico entre os jogadores ao longo de toda a partida, disputando espaço dentro das normas do jogo. Eles concluem dizendo que essas vantagens são transitórias e são influenciadas pela idade cronológica e desenvolvimento biológico, portanto, não necessariamente continuarão após o fim da adolescência.

Figura 3 – Histograma de distribuição do tempo de vida decorrido dos jogadores do FIFA22



Fonte: Elaborado pelo autor

O peso e a altura são características que têm importância em algumas posições no futebol, porém não necessariamente na qualidade do jogador. Na Figura 4, os goleiros são os atletas em sua maioria mais altos e mais pesados. Isso pode ser explicado pela necessidade de uma estatura

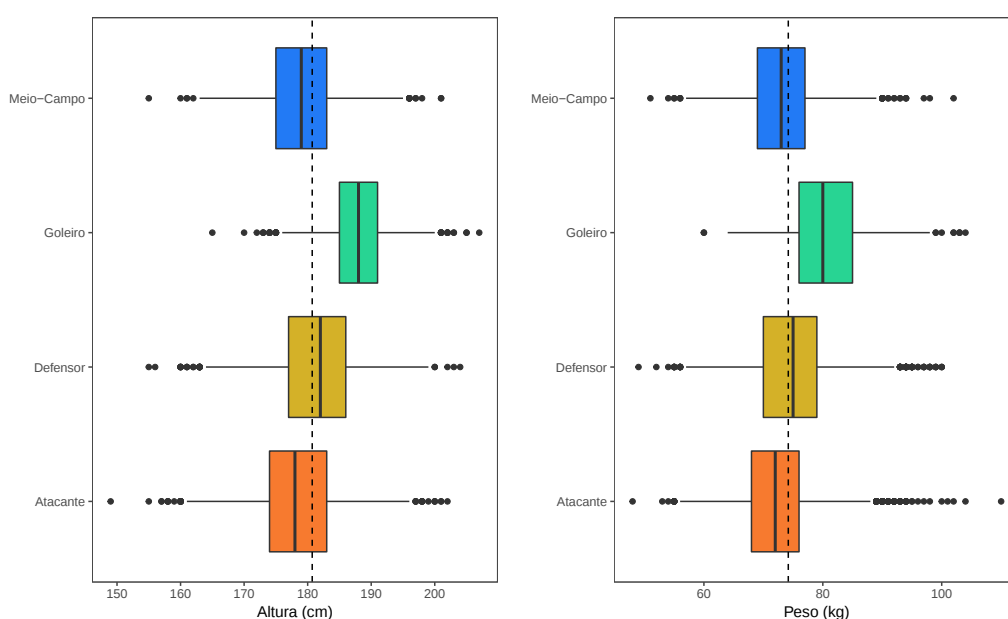
maior em jogadores nessa posição para proteger as dimensões da trave, que são de 2,44m de altura por 7,32m de largura, dimensões das arestas que formam a trave.

Entretanto, é importante ressaltar que quanto mais alto e mais pesado o goleiro for, maior a possibilidade dele se tornar um pouco lento e ter menos reações rápidas. Em seguida, temos os defensores, que também necessitam de estatura e peso elevados para disputar bolas aéreas e ter força em combates terrestres, como é o caso dos zagueiros centrais. Os alas ou laterais não necessariamente seguem esse perfil, já que existem basicamente dois tipos: defensivo e ofensivo, como dissertam [Guimarães et al. \(2015\)](#) no estudo sobre as posições no futebol e suas especificidades.

De acordo com o estudo de [Guimarães et al. \(2015\)](#), os alas ou laterais atuam de maneiras distintas dentro de campo e de acordo com suas características fisiológicas. Os laterais que avançam mais para o campo de ataque tendem a ser mais leves, já que percorrem toda extensão do campo em comprimento e o excesso de massa muscular dificulta na performance e na resistência. Já os laterais defensivos tendem a ter um perfil oposto, pois priorizam a segurança defensiva e, portanto, tendem a ter mais peso e massa muscular para ajudar na proteção da defesa.

Os meio-campistas e atacantes apresentam comportamentos semelhantes, mas existem também atletas com características mais específicas, variando de posição para posição e estilo de jogo de cada um. Essa diversidade é discutida por [Behravan e Razavi \(2021\)](#) na Seção 5 do seu estudo, no qual utilizam de métodos de agrupamento para identificar jogadores e realizar comparações dentre seus respectivos grupos.

Figura 4 – Boxplot de características físicas estruturais por zona de posição dos jogadores do FIFA22

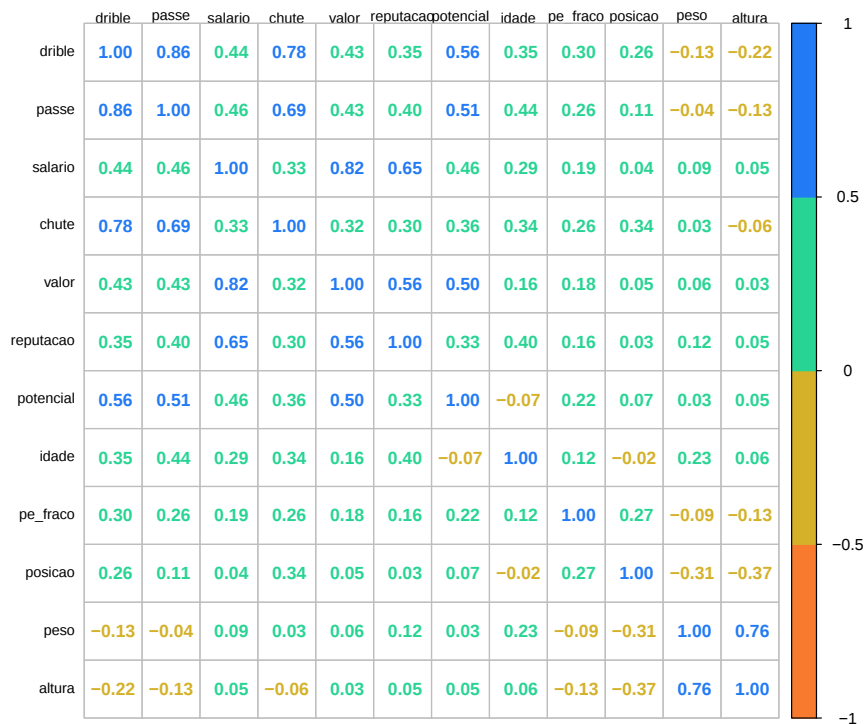


Fonte: Elaborado pelo autor

Na Figura 5 é possível identificar a correlação entre as variáveis consideradas no estudo. Destaque-se, a forte correlação diretamente proporcional entre as variáveis drible e passe (0,86), drible e chute (0,78), valor de mercado e salário (0,82), peso e altura (0,76) e salário e reputação internacional (0,65).

Em outras palavras, compreende-se que, à medida que a reputação do atleta tem uma pontuação maior, o salário tende a aumentar. Isso pode explicar outro fator econômico que o futebol tem vendido enquanto produto: o marketing e publicidade com atletas mais reconhecidos no cenário internacional tendem a valorizá-los e contribuir com o aumento de receitas do clube, como em vendas de camisas e demais ações que geram retorno econômico vinculados à imagem do atleta.

Figura 5 – Correlação das variáveis coletadas na raspagem de dados do site SoFifa



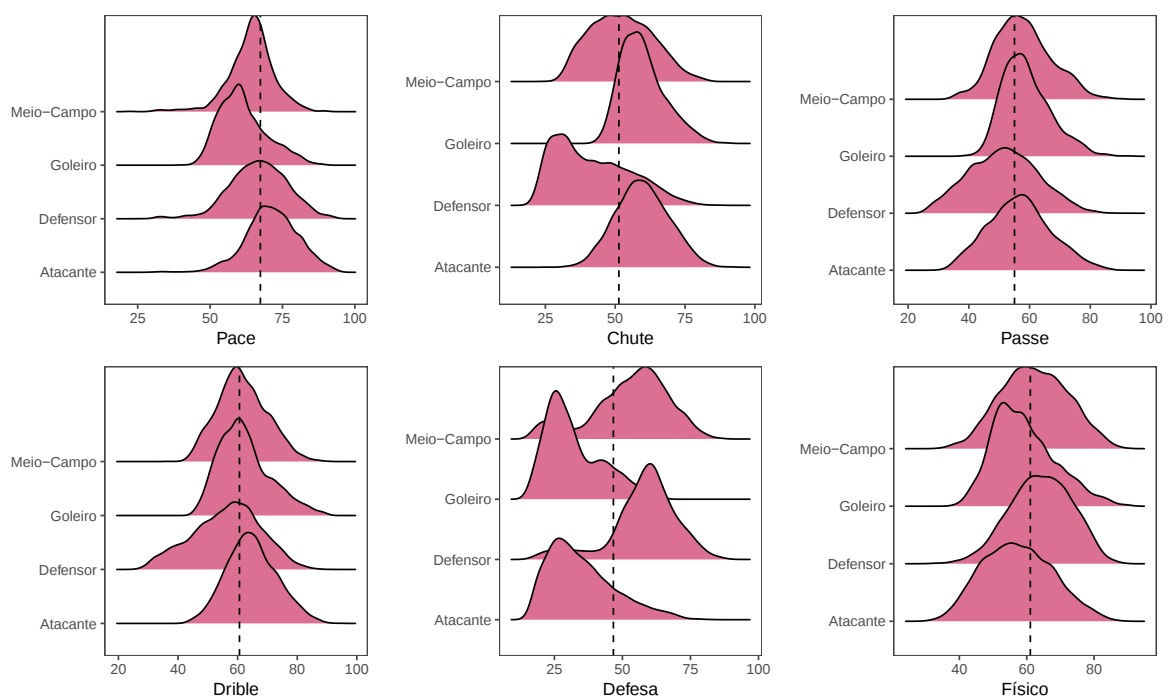
Fonte: Elaborado pelo autor

Na Figura 6, o gráfico de densidade por posição apresenta as estatísticas de cada grupo, diferenciando-as por seus aspectos e características específicas. Inicialmente, pode-se observar o *pace*, que mostra uma densidade maior acima da média nos atacantes e uma densidade abaixo da média para os Goleiros e Meio-campistas. Isso pode ser explicado pela atuação dos atacantes nas faixas laterais do campo e pela tendência atual do futebol em que eles recuam em situação de ataque adversário para apoio no combate defensivo, o que exigem, um ritmo médio acima do comum.

As habilidades Chute e Defesa apresentam um comportamento diferente dos demais que seguem um determinado padrão acerca da média. No gráfico da habilidade Chute, é

perceptível que os defensores, por terem características de posição que não exigem tanto esta habilidade apresentam uma densidade mais baixa nos dados, diferentemente dos atacantes, que são responsáveis por definir as jogadas e apresentam valores acima da média. No gráfico da habilidade Defesa, ocorre o oposto, os defensores sobressaem em relação aos atacantes devido à maior especificidade da posição no combate homem a homem. Como afirma [Behravan e Razavi \(2021\)](#), o físico do atleta é um fator importante para a valorização dos defensores no mercado, além de características como idade e altura.

Figura 6 – Gráfico de densidade para os atributos de habilidades específicas dos atletas do FIFA22 por posição



Fonte: Elaborado pelo autor

5.2 Aplicação do modelo *Random Forests*

Dentre todo o conjunto de dados, apresentado no Capítulo 4, especificamente no Quadro 1, foram dispostas para o modelo a escolha das covariáveis: `age`, `international_reputation`, `dri`, `def`, `sho`, `pas`, `phy` e `wage`. Por conveniência, a variável independente `value` e a covariável `wage`, sofreram transformações logarítmicas apenas para fins de padronização e melhor visualização das variáveis dentro do contexto do modelo.

Com a intenção de obter o modelo que melhor desempenha predições na base de dados, utilizando o método de pesquisa de grade previamente detalhado no Capítulo 4, foram listados, em ordem decrescente, os dez melhores modelos com menor erro quadrático médio (RMSE) e maior

percentual de ganho dispostos na Tabela 2, para seleção do modelo com melhor desempenho a ser aplicado.

Tabela 2 – Modelos Random Forest com menores RMSE construídos pelo HyperGrid

	mtry	min.node.size	replace	sample.fraction	RMSE	perc_gain
1	6	5	TRUE	0.63	0.3670	2.1831
2	6	5	TRUE	0.80	0.3672	2.1504
3	6	8	TRUE	0.80	0.3672	2.1484
4	6	10	TRUE	0.80	0.3672	2.1317
5	6	7	TRUE	0.80	0.3673	2.1066
6	6	1	TRUE	0.80	0.3675	2.0489
7	5	8	TRUE	0.80	0.3678	1.9708
8	6	1	TRUE	0.63	0.3682	1.8698
9	6	7	TRUE	0.63	0.3682	1.8655
10	6	10	TRUE	0.63	0.3684	1.8297

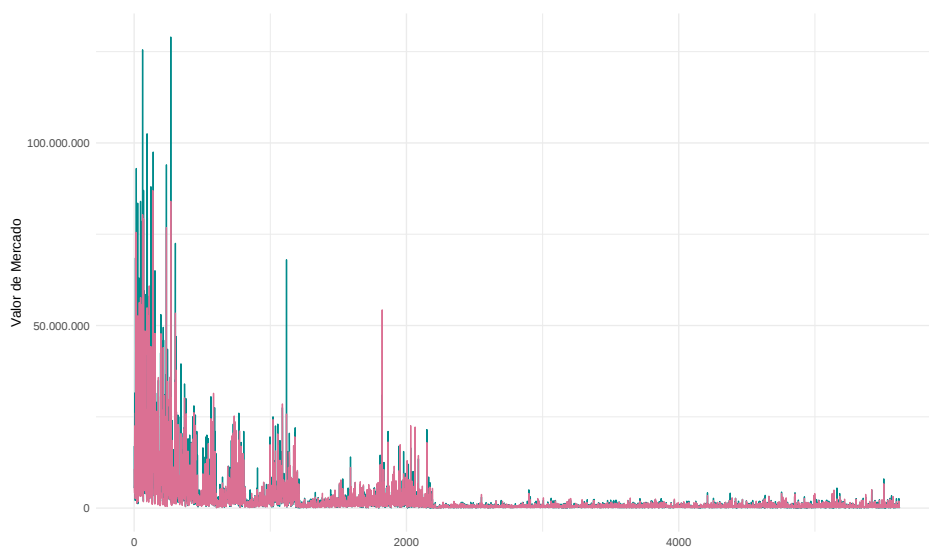
Fonte: Elaborado pelo autor

O modelo 1, que apresenta melhores métricas para ser aplicado, implica em um valor para definição da variação e performance do modelo, além da sensibilidade a ruído, com um valor para *mtry* igual a 6. O ajuste *sample size* define o tamanho da amostra a ser treinada pelo método de amostragem com reposição, baseado na proporção de 63% da amostra.

Como métrica de avaliação do modelo, o valor do RMSE em escala logarítmica foi igual a 0,3670, obtendo melhor desempenho ao inserir novas entradas de dados, como na nossa base de dados de teste.

Após a definição e posterior a construção do modelo, na Figura 7 é realizada a comparação entre os valores observados e os valores preditos pelo método de florestas aleatórias, aplicado na base de dados de teste. O comportamento dos valores preditos em relação aos valores reais, demonstra um bom desempenho. Destaca-se a desenvoltura do modelo em acompanhar os valores preditos bem próximos dos valores reais. Contudo, nos valores considerados *outliers*, o modelo apresenta fragilidade em acompanhar o padrão presente dentro da base de dados, com valores de mercado fora da curva de alguns jogadores.

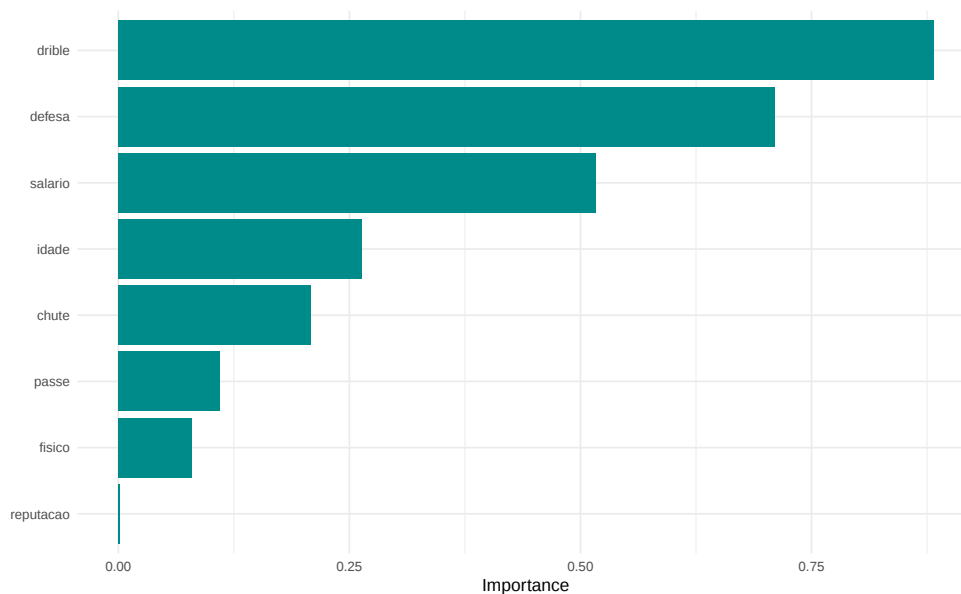
Figura 7 – Predição do valor de mercado de jogadores de futebol masculino do Fifa 22 pelo modelo de floresta aleatória



Fonte: Elaborado pelo autor

Na Figura 8, é possível ver a importância da variável *drible* que apresenta total influência no modelo, seguido dos atributos de *defesa* com influência acima de 70% e o *salário* (em euros) com pouco mais de 50% de relevância para o desenvolvimento das previsões do modelo. Contudo, a *idade* e demais atributos técnicos presentes no modelo atuam de modo mais coadjuvante.

Figura 8 – Variáveis influentes no modelo de regressão Random Forest



Fonte: Elaborado pelo autor

6 CONCLUSÃO

Durante o presente estudo, foi possível compreender a importância da estatística e da ciência de dados a partir de algoritmos de *Machine Learning* para tomada de decisões em situações de negócios no mercado de transferências do futebol. Além disso, não somente existe a oportunidade de aplicação nessa temática, como também possíveis estudos na área de desempenho de atletas ao longo da temporada ou de previsões para analisar o crescimento de receita a partir de vendas de produtos e bilheteria.

Com a análise descritiva dos dados, foi possível identificar as tendências dos distintos perfis de jogadores, bem como compreender as correlações existentes entre as variáveis. A examinação dos dados também permitiu a diferenciação em comparações com atributos e especificidades das diferentes posições dentro de campo, incluindo a análise de remunerações e valores de mercado estimados para jogadores com habilidades excepcionais e grande poder de decisão, além das consequências de jogar em grandes clubes e elevar o status internacional.

O método e aplicação do algoritmo de aprendizado *Random Forest*, obteve êxito significativo pois conseguiu compreender, desenvolver e ajustar um bom modelo para o conjunto de dados extraídos pela *web scraping* do site Sofifa. Dessa maneira, foi possível realizar previsões que acompanharam os valores reais e inferir análises satisfatória sobre o comportamento do modelo e a inserção ou retirada das variáveis com grau de relevância dispostas a compor o modelo.

A forte correlação vista na análise descritiva entre as variáveis de salário e valor de mercado foi reforçada na aplicação do modelo, sua relevância foi notada ao analisar a Figura 8. Destacam-se também as habilidades técnicas, como as características de drible e atributos de defesa, que figuraram entre as variáveis de importância para o modelo. Esse fator pode ser explicado pela evolução ao longo dos anos no futebol, com o domínio de teorias táticas e técnicas alinhadas à ciência e tecnologias.

Os resultados contribuem para o entendimento e funcionamento do mercado de transferência no futebol mundial, tendo em vista que o valor de mercado estipulado nem sempre é o exercido de fato nas transações. Porém são um excelente parâmetro para projeção e planejamento econômico dos clubes que buscam obter sucesso nas grandes competições disputadas ao longo da temporada. É importante que os clubes estruturem elencos do mais alto nível do ponto de vista técnico e exerçam cada vez mais políticas que contribuam não só para o êxito dentro de campo, como também na execução e gestão financeira.

REFERÊNCIAS

- AL-ASADI, M. A.; TASDEMIR, S. Predict the value of football players using fifa video game data and machine learning techniques. *IEEE Access*, IEEE, v. 10, p. 22631–22645, 2022. Disponível em: <<https://ieeexplore.ieee.org/abstract/document/9721908>>. Citado na página 18.
- AMIR, E.; LIVNE, G. Accounting, valuation and duration of football player contracts. *Journal of Business Finance & Accounting*, 2005. Citado na página 16.
- ARRUL, V. S.; SUBRAMANIAN, P.; MAFAS, R. Predicting the football players' market value using neural network model: A data-driven approach. In: *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. [S.l.: s.n.], 2022. Citado na página 17.
- BEHRAVAN, I.; RAZAVI, S. M. A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, Springer, v. 25, n. 3, p. 2499–2511, 2021. Citado 2 vezes nas páginas 29 e 31.
- BINDER, J. J.; FINDLAY, M. The effects of the bosman ruling on national and club teams in europe. *J. Sports Econom.*, SAGE Publications, v. 13, n. 2, p. 107–129, abr. 2012. Citado na página 15.
- BOEHMKE, B. *UC Business Analytics R Programming Guide*. 2018. Disponível em: <https://uc-r.github.io/regression_trees>. Citado na página 24.
- BOEHMKE, B.; GREENWELL, B. M. *Hands-on machine learning with R*. London, England: CRC Press, 2020. (Chapman & Hall/CRC The R Series). Citado 4 vezes nas páginas 21, 22, 23 e 26.
- BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 25 e 26.
- CARLI, G. et al. Relative age effect in soccer players. *Revista Brasileira de Ciência e Movimento*, v. 17, p. 25–31, 01 2009. Citado na página 28.
- CARMICHAEL, F.; THOMAS, D. Bargaining in the transfer market: Theory and evidence. *Applied Economics*, v. 25, p. 1467–76, 12 1993. Citado na página 16.
- COELHO, P. V. *Bola fora: A historia do êxodo do futebol brasileiro*. [S.l.]: Panda Books, 2015. Citado na página 13.
- ĆWIKLINSKI, B.; GIEŁCZYK, A.; CHORAŚ, M. Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, MDPI, v. 23, n. 1, p. 90, 2021. Citado na página 17.
- EFRON, B.; TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, v. 1, p. 54–75, 1986. Citado na página 22.
- FRICK, B. *Salary determination and the pay-performance relationship in professional soccer: evidence from Germany*. [S.l.]: Ediciones de la Universidad de Oviedo, 2006. Citado na página 17.

GUIMARÃES, M. et al. As posições no futebol e suas especificidades. *Revista Brasileira de Futebol (The Brazilian Journal of Soccer Science)*, v. 7, n. 2, p. 71–83, 2015. Citado na página 29.

HEROLD, M. et al. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, SAGE Publications Sage UK: London, England, v. 14, n. 6, p. 798–817, 2019. Citado na página 17.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4. Citado 3 vezes nas páginas 22, 24 e 25.

LANTZ, B. *Machine Learning with R*. 2. ed. Birmingham, England: Packt Publishing, 2015. Citado 3 vezes nas páginas 21, 23 e 25.

LEONCINI, M. P.; SILVA, M. T. d. Entendendo o futebol como um negócio: um estudo exploratório. *Gestão & Produção*, SciELO Brasil, v. 12, p. 11–23, 2005. Citado na página 13.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002. Citado na página 26.

MAJEWSKI, S.; University of Szczecin. Identification of factors determining market value of the most valuable football players. *J. Manag. Bus. Adm. Cent. Eur.*, Kozminski University, v. 24, n. 3, p. 91–104, set. 2016. Citado na página 16.

MITCHELL, R. *Web Scraping with Python, 2e*. Sebastopol, CA: O'Reilly Media, 2018. Citado na página 19.

OZDEMIR, S. *Principles of Data Science*. Birmingham, England: Packt Publishing, 2016. Citado na página 21.

PRONI, M. W. Esporte-espetáculo e futebol-empresa. *Versão preliminar da tese de douto*, 1998. Citado na página 16.

RADOMAN, M.; VOIA, M. C. Internal promotion and the bosman ruling: Evidence from the english premier league. *Labour (Rome)*, Wiley, v. 36, n. 4, p. 445–470, dez. 2022. Citado na página 16.

SANTOS, J. M. C. M. Futebol e história. *O futebol nas ciências humanas no Brasil*. Campinas: Editora da Unicamp, p. 139–151, 2020. Citado na página 15.

SHIN, J. gon; GASPARYAN, R. A novel way to soccer match prediction. In: . [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 18 e 20.

SOUZA, F. A. P. de; ANGELO, C. F. de. O fim do passe e seu impacto sobre o desequilíbrio competitivo entre as equipes de futebol. *Revista de Administração-RAUSP*, Universidade de São Paulo, v. 40, n. 3, p. 280–288, 2005. Citado na página 15.

TONINI, M. D.; GIGLIO, S. S. A transferência de jogadores no sistema FIFA e a migração de brasileiros para a europa (1920-1970). *Estud. Hist.*, FapUNIFESP (SciELO), v. 32, n. 68, p. 609–632, dez. 2019. Citado 2 vezes nas páginas 13 e 15.

WILEY, J. R *Deep Learning Essentials: Build Automatic Classification and Prediction Models Using Unsupervised Learning*. [S.l.]: Packt Publishing, 2016. (Community Experience Distilled). ISBN 9781785280580. Citado 2 vezes nas páginas 26 e 27.

WRIGHT, M. N.; ZIEGLER, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, v. 77, n. 1, p. 1–17, 2017. Citado na página 27.

APÊNDICE A – Código da raspagem, limpeza dos dados e aplicação do modelo

```

1
2 pacman::p_load(rvest, dplyr, stringr, lubridate, tidyr, janitor)
3
4 page <- seq(0,19260,60)
5 link <- paste0("https://sofifa.com/players?showCol%5B0%5D=pi&showCol
    %5B1%5D=ae&showCol%5B2%5D=hi&showCol%5B3%5D=wi&showCol%5B4%5D=pf&
    showCol%5B5%5D=oa&showCol%5B6%5D=pt&showCol%5B7%5D=bp&showCol%5B8
    %5D=gu&showCol%5B9%5D=vl&showCol%5B10%5D=wg&showCol%5B11%5D=gh&
    showCol%5B12%5D=wk&showCol%5B13%5D=ir&showCol%5B14%5D=pac&showCol
    %5B15%5D=sho&showCol%5B16%5D=pas&showCol%5B17%5D=dri&showCol%5B18
    %5D=def&showCol%5B19%5D=phy&offset=",page)
6
7 # xpath <- //*[@id="body"]/div[1]/div/div[2]/div/table
8 read_html(iconv(page_source[[1]], to = "UTF-8"), encoding = "utf8")
9 get_table <- function(url){
10   url %>%
11     read_html(encoding="utf-8") %>%
12     html_nodes(xpath = '//*[@id="body"]/div[1]/div/div[2]/div/table')
13     %>%
14     html_table(header = T)
15 }
16
17 # Applying the created function to URLs
18 scraping <- sapply(link, get_table)
19
20 # Joining scraping a data frame
21 data <- do.call(rbind, scraping)
22 data <- as.data.frame(data)
23
24 ## Cleaning ----
25 sofifa <- data
26 sofifa <- clean_names(sofifa)
27 sofifa[1] <- NULL
28 names(sofifa)[5] <- "team"
29 names(sofifa)[6] <- "sofifa_id"
30 names(sofifa)[10] <- "best_position"
31 names(sofifa)[15] <- "weak_foot"

```

```

32 names(sofifa)[16] <- "international_reputation"
33
34 sofifa <- sofifa %>% relocate(sofifa_id, .before = "name")
35
36
37
38 # Position
39 sofifa <- sofifa %>%
40   mutate(position = case_when(best_position == "GK" ~ "GK",
41                               str_detect(name, "CB") ~ "CB",
42                               str_detect(name, "B") ~ "FB",
43                               str_detect(name, "AM") ~ "ATT_MID",
44                               str_detect(name, "RW") | str_detect(
45                                 name, "LW")
46                               | str_detect(name, "RM")
47                               | str_detect(name, "LM") ~ "WING",
48                               str_detect(name, "M") ~ "MID",
49                               str_detect(name, "ST") | str_detect(
50                                 name, "CF") ~ "ST"))
51
52 sofifa$position <- recode(sofifa$position,
53                           "GK" = "Goleiro",
54                           "CB" = "Defensor",
55                           "FB" = "Defensor",
56                           "MID" = "Meio-Campo",
57                           "ATT_MID" = "Atacante",
58                           "WING" = "Atacante",
59                           "ST" = "Atacante") #Pt-br
60
61 sofifa$team_position <- recode(sofifa$position,
62                                "Goleiro" = 1,
63                                "Defensor" = 2,
64                                "Meio-Campo" = 3,
65                                "Atacante" = 4) #Pt-br
66
67 # Name
68
69 #name_prefix <- str_extract(sofifa$name, "[A-Z]\\.+") %>% replace_na
70   ("")
71
72 #name_sufix <- str_extract(sofifa$name, "[A-Z][a-z]+")

```

```

71 remove_symbol <- function(x) iconv(x, to = "ASCII//TRANSLIT")
72 sofifa$name <- remove_symbol(sofifa$name)
73 name_extract <- str_extract(sofifa$name, "[^-[[:digit:]]+[a-z]+")
74 sofifa$name <- name_extract
75
76 # Team name
77 team_split <- str_split(sofifa$team, "~")
78 team_split <- sapply(team_split, "[[" , 1)
79 team_split <- str_split(team_split, "\n")
80 team_split <- sapply(team_split, "[[" , 1)
81 sofifa$team <- team_split
82
83 # Extract numbers
84 sofifa$height <- str_extract(sofifa$height, "[0-9]+") #cm
85 sofifa$weight <- str_extract(sofifa$weight, "[0-9]+") #kg
86
87
88 # reorder variables
89 sofifa <- sofifa %>% relocate(team, .after = "name")
90 sofifa <- sofifa %>% relocate(position, .after = "team")
91 sofifa <- sofifa %>% relocate(foot, .after = "position")
92 sofifa <- sofifa %>% relocate(best_position, .after = "foot")
93 sofifa <- sofifa %>% relocate(height, .after = "best_position")
94 sofifa <- sofifa %>% relocate(weight, .after = "height")
95 sofifa <- sofifa %>% relocate(team_position, .after = "international_
    reputation")
96
97
98 sofifa <- sofifa %>% relocate(wage, .after = "phy")
99 sofifa <- sofifa %>% relocate(value, .after = "wage")
100
101 # Market value
102 sofifa$value <- str_extract(sofifa$value, "[0-9\\.[A-Z]]+")
103 sofifa$Unity <- str_detect(sofifa$value, "M")
104 sofifa$Unity <- if_else(sofifa$Unity == "FALSE", 1000, 1000000)
105 sofifa$value <- as.numeric(str_extract(sofifa$value, "[0-9\\.]+"))
106 sofifa$value <- sofifa$value * sofifa$Unity
107
108 # wage
109 sofifa$wage <- str_extract(sofifa$wage, "[0-9\\.[A-Z]]+")
110 sofifa$UnityW <- str_detect(sofifa$wage, "K")
111 sofifa$UnityW <- if_else(sofifa$UnityW == "FALSE", 1, 1000)

```

```

112 sofifa$wage <- as.numeric(str_extract(sofifa$wage, "[0-9]+"))
113 sofifa$wage <- sofifa$wage * sofifa$UnityW
114
115 sofifa[25:26] <- NULL
116 rm(team_split, name_extract)
117
118 # export data
119 write.table(format(sofifa, scientific=F), 'sofifa.csv',
120             sep = ";", dec = ",",
121             row.names = F, quote = F,
122             fileEncoding = "UTF-8")
123
124
125
126 ## APPLING RANDOM FOREST MODEL
127
128 options(scipen = 999)
129 pacman::p_load(ggplot2, scales, tidyverse, ggtext, showtext, rsample,
130               quantmod,
131               randomForest, rpart, rpart.plot, ranger, vip, stringr,
132               MASS)
133
134 # load database from github ----
135
136 fifa <- read.csv(
137   "https://raw.githubusercontent.com/alan-elias/transfer-football/
138     updates/r-code/sofifa.csv",
139   sep = ";", dec = ",")
140 fifa <- na.omit(fifa)
141 fifa$name <- str_trim(fifa$name)
142 fifa$dup <- duplicated(fifa)
143 fifa <- fifa |> filter(dup == FALSE)
144 fifa$dup <- NULL
145
146 fifa22 <- fifa |> filter(value>0 & wage>0)
147
148 # 'value' transformation ----
149
150 ## Log
151 fifa22$wage <- log(fifa22$wage)
152 fifa22$value <- log(fifa22$value)

```

```

151
152 ## boxcox
153 c <- boxcox(lm(fifa22$wage~1))
154 lambda <- c$x[which.max(b$y)]
155 fifa22$b_wage <- (fifa22$wage^lambda - 1)/lambda
156
157 b <- boxcox(lm(fifa22$value~1))
158 lambda <- b$x[which.max(b$y)]
159 fifa22$b_value <- (fifa22$value^lambda - 1)/lambda
160
161 fifa22$b_wage <- log(fifa22$wage)
162 fifa22$b_value <- log(fifa22$value)
163
164 # database model selection ----
165
166 fifa_model <- fifa %>%
167   select(name, position, age, height, weight, weak_foot,
168           international_reputation, ova, pot, growth,
169           pac, sho, pas, dri, def, phy, gk_handling, wage, value)
170
171 ##### Best base
172
173 fifa_model <- fifa22 %>%
174   select(sofifa_id, age, height, weight, weak_foot, international_
175           reputation, ova, sho, pas, dri, def, phy, wage, value)
176
177 fifa_model <- fifa22 %>% dplyr::select(age, height, weight, weak_foot
178   , international_reputation, pac, sho, pas, dri, def, phy, wage,
179   value)
180
181 fifa_model <- fifa22 %>%
182   dplyr::select(age, international_reputation, sho, pas, dri, def, phy
183   , wage, value)
184
185 fifa_model <- na.omit(fifa_model)
186
187 # sample ----
188
189 set.seed(123)
190 extraction <- rsample::initial_split(fifa_model, prop = .7, strata =
191   "value")
192 train <- rsample::training(extraction)
193 test <- rsample::testing(extraction)

```

```
187
188 # Decision Tree ----
189
190 set.seed(123)
191 ad1 <- rpart(
192   formula = value ~.,
193   data = train,
194   method = "anova")
195
196 options(scipen = 999)
197 rpart.plot(ad1)
198 plotcp(ad1)
199
200 ## Sem penalidade
201 set.seed(123)
202 adsp = rpart(
203   formula = value ~.,
204   data = train,
205   method = "anova",
206   control = list(cp = 0, xval = 10)
207 )
208
209 rpart.plot(adsp)
210 plotcp(adsp)
211
212 ## SSE
213 adsp$cptable
214
215 ## cross valid
216
217 set.seed(123)
218 (adcv = train(
219   value ~.,
220   data = train,
221   method = "rpart",
222   trControl = trainControl("cv",5),
223   tuneLength = 12)
224 )
225
226 ggplot(adcv)+theme_minimal()
227
228 # Random Forest ----
```

```

229
230 n_features <- length(setdiff(names(train), "value"))
231
232 mrf1 <- ranger(
233   value ~ .,
234   data = train,
235   mtry = floor(n_features / 3),
236   respect.unordered.factors = "order",
237   seed = 123
238 )
239
240 (default_rmse <- sqrt(mrf1$prediction.error))
241
242 ## Hypergrid construction
243
244 hyper_grid <- expand.grid(
245   mtry = floor(n_features * c(.2, .5, .7, .8)),
246   min.node.size = c(1, 5, 7, 8, 10),
247   replace = c(TRUE, FALSE),
248   sample.fraction = c(.5, .63, .8),
249   rmse = NA
250 )
251
252 ## Gridsearch
253
254 for(i in seq_len(nrow(hyper_grid))){
255   fit <- ranger(
256     formula = value ~.,
257     data = train,
258     num.trees = n_features * 10,
259     mtry = hyper_grid$mtry[i],
260     min.node.size = hyper_grid$min.node.size[i],
261     replace = hyper_grid$replace[i],
262     sample.fraction = hyper_grid$sample.fraction[i],
263     verbose = FALSE,
264     seed = 123,
265     respect.unordered.factors = 'order',
266   )
267   hyper_grid$rmse[i] <- sqrt(fit$prediction.error)
268 }
269
270

```

```

271 models <- hyper_grid %>%
272   arrange(rmse) %>%
273   mutate(perc_gain = (default_rmse - rmse)/default_rmse * 100) %>%
274   head(10)
275
276 xtable::xtable(models, label="modelos_RF", digits=4)
277
278 mrf2 <- ranger(
279   value ~ .,
280   data = train,
281   num.trees = n_features * 10,
282   mtry = 6,
283   min.node.size = 5,
284   replace = T,
285   sample.fraction = 0.63,
286   importance = "permutation",
287   respect.unordered.factors = "order",
288   seed = 123
289 )
290
291 (default_rmse <- sqrt(mrf2$prediction.error))
292
293 ## Predict
294
295 p1_rf <- predict(mrf2 , train)
296 sqrt(mean((train$value-p1_rf$predictions)^2))
297
298 p2_rf <- predict(mrf2, test)
299 sqrt(mean((test$value-p2_rf$predictions)^2))
300
301
302 data.gg2 <- as.data.frame(cbind(x = 1:length(test$value),
303                                predicts = exp(p2_rf$predictions),
304                                test = exp(test$value)))
305
306 ggplot(data=data.gg2 , aes(x=x,y=test)) +
307   geom_line(col="darkcyan") +
308   ylab("Valor de Mercado")+xlab("")+
309   geom_line(aes(x=x,y=predicts),col="palevioletred")+
310   scale_y_continuous(labels = comma_format(big.mark = ".", decimal.
311     mark = ","))+
311   theme_minimal()

```

```
312  
313 # df <- cbind(fifa_model, mrf2$predictions)  
314  
315 vip::vip(mrf2, aesthetics = list(fill = "darkcyan"), num_features = n  
    _features) + theme_minimal(base_size = 13)
```