UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E TRANSPORTES

TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO

APLICAÇÃO DE MACHINE LEARNING PARA IDENTIFICAR VARIÁVEIS DE MAIOR IMPACTO NA VARIAÇÃO DO VALOR DE MERCADO DE JOGADORES DE FUTEBOL

FELIPE ALLEGRETTI

Orientador: MICHEL JOSÉ ANZANELLO

PORTO ALEGRE
AGOSTO/2023

RESUMO

Jogadores de futebol possuem características de desempenho variadas e complexas que influenciam seu potencial de valorização no mercado. A diversidade nas características decorre de fatores como estilo de jogo, peculiaridades dos campeonatos e estratégias de treinamento. Esta multiplicidade torna desafiador prever e avaliar sua variação monetária, exigindo técnicas analíticas robustas. O estudo atual se desenrola em duas fases: inicialmente, a validação de técnicas preditivas supervisionadas - Regressão Linear Múltipla, K-Vizinhos Mais Próximos e Floresta Aleatória - com base nos dados disponíveis; em seguida, o treinamento e análise da técnica mais eficaz focada numa posição específica. A Floresta Aleatória demonstrou ser a abordagem de melhor desempenho médio (R^2=0,87 e MAE=€1.818.917) e foi empregada para discernir a importância de diversas características na previsão da variação de valor de mercado dos meio-campistas. Dentro dos resultados obtidos, o modelo direcionado ao campeonato da Bundesliga foi o mais destacado, com R^2 de 0,89 e MAE de €1.703.015. Constatou-se que a idade do jogador é o fator mais determinante na sua valorização. Adicionalmente, os campeonatos exibem perfis distintos na valorização dos meio-campistas: Bundesliga e La Liga enfatizam ações ofensivas, enquanto Premier League e Serie A valorizam a construção de jogadas.

1. INTRODUÇÃO

O aprendizado de máquina é um campo multidisciplinar que intersecciona as áreas de ciência da computação, engenharia e estatística e tem como premissa fundamental a generalização de um padrão detectável ou a criação de uma regra desconhecida a partir de exemplos fornecidos. Este processo ocorre a partir da geração de um modelo, que é desenvolvido utilizando dados históricos para aprender e usar seu conhecimento na tomada de decisões futuras. Via de regra, ao longo de um período de tempo, as previsões de um modelo se tornam melhores à medida que mais dados são fornecidos (DANGETI, 2017). A partir desse contexto, mostra-se relevante o uso do aprendizado de máquina no cenário esportivo que, segundo Ashley (2020), mesmo sendo um método relativamente novo na ciência do esporte, está fazendo grandes avanços e já impacta muitas áreas nesse campo, desde treinamento pessoal até competições profissionais.

Historicamente, o futebol se encontra atrasado quando comparado a outros esportes mais populares em relação ao uso de dados para análises e tomadas de decisões, sendo considerado o esporte “menos estatístico” em decorrência da escassez de dados disponíveis (KAPLAN, 2010).

Entretanto, o cenário atual apresenta uma perspectiva mais otimista em consequência da consolidação de empresas privadas que coletam dados estatísticos de partidas de futebol voltados a fins específicos, como estimar e prever o valor de mercado de jogadores profissionais (BRANDES E FRANCK, 2012), prever o resultado de partidas com o objetivo de gerar lucro com apostas esportivas (CONSTANTINOU et al., 2013), avaliar o risco de lesão em jogadores (NASSIS et al., 2023), auxiliar olheiros a identificar jovens jogadores com potencial (RICO-GONZÁLEZ et al., 2023) e analisar o desempenho individual em treinos e jogos (PAPPALARDO et al., 2019).

Além de ser um esporte mundialmente popular, o futebol também é um grande negócio que movimenta enormes quantias monetárias. Desde meados da década de 1990, o mercado de trabalho para jogadores de futebol tem sido objeto de diversos estudos empíricos, atraindo um número crescente de economistas de toda Europa a dedicar a sua atenção ao funcionamento e peculiaridades desse campo (FRICK, 2007). Mais especificamente, pesquisadores têm direcionado seus estudos ao valor de mercado associado aos jogadores que, de acordo com Herm et al. (2014), é uma estimativa da quantidade de dinheiro que um clube estaria disposto a pagar para que determinado atleta assine um contrato, independentemente de uma transação real.

Contudo, clubes de futebol ainda encontram dificuldades para identificar jogadores que possam vir a ter uma grande valorização de mercado. Este processo tipicamente se apoia na ação de funcionários do clube - chamados de olheiros - que se deslocam para assistir partidas in loco de potenciais atletas. Tal avaliação, por se valer de premissas subjetivas (WINEMILLER et al., 2020) e que dependem da interpretação humana, acaba se tornando imprecisa. Alternativamente, clubes também vêm utilizando a análise de dados para auxiliar na tomada de decisão, avaliando variáveis atreladas a jogadores que incluem número de gols, passes, dribles, cruzamentos, desarmes, entre outros. Entretanto, com o grande volume de dados e informações, é cada vez mais inviável a identificação de padrões apenas por meio da interpretação humana.

Este artigo tem como principal objetivo desenvolver modelos de previsão de variação do valor de mercado de jogadores de futebol com base em dados estatísticos de partidas utilizando algoritmos de *machine learning* para problemas de regressão em aprendizado supervisionado (Regressão Linear Múltipla, K-Vizinhos Mais Próximos, Árvores de Decisão e Floresta Aleatória) e, a partir

destes modelos gerados, identificar as variáveis que mais afetam tal variação monetária dos jogadores de uma mesma posição em diferentes campeonatos.

O objetivo se mostra relevante pois, do ponto de vista gerencial, um dos fatores responsáveis pelo sucesso de um clube reside na identificação de jogadores promissores que possam contribuir tanto em campo quanto no balanço financeiro (ZHU et al., 2015). Para tal, é de extrema importância que dirigentes de clubes tomem decisões corretas ao investir em jogadores a partir da análise de seu desempenho e o perfil do campeonato de atuação, considerando ainda a dificuldade em prever se o atleta contribuirá de fato ou não.

Este artigo está organizado conforme segue. A seção 2 desenvolve o referencial teórico sobre aprendizado de máquina, aplicação de técnicas preditivas, métricas de desempenho, divisão de dados, seleção de variáveis e indicadores do futebol. A seguir, os dados coletados são apresentados e utilizados para gerar um modelo de previsão na seção 3. Por fim, a partir desta aplicação, os resultados são discutidos na seção 4 e o artigo é finalizado com as conclusões na seção 5.

2. REFERENCIAL TEÓRICO

Este estudo baseia-se em conceitos relacionados ao valor de mercado de jogadores de futebol, bem como importantes variáveis levadas em consideração para a sua estimativa, sistemas de *machine learning*, alguns dos principais algoritmos para resolução de problemas de regressão em aprendizado supervisionado, aplicações destas técnicas para análises específicas no cenário do futebol, e seleção de variáveis na preparação de dados.

2.1. Técnicas Multivariadas

Conforme Faceli et al. (2011), *machine learning* (aprendizado de máquina, em português), é uma área de pesquisa da Inteligência Artificial que visa ao desenvolvimento de programas de computador com a capacidade de aprender aa executar uma determinada tarefa com sua própria experiência. Isto leva à possibilidade de estruturar sistemas capazes de aprender de forma autônoma a partir da leitura e processamento de dados históricos. Trata-se de uma área de pesquisa multidisciplinar que engloba inteligência artificial, probabilidade e estatística, teoria da complexidade computacional, teoria da informação, filosofia, psicologia, neurobiologia, entre outros (CERRI e CARVALHO, 2017).

Segundo os tipos de sistemas de *machine learning* elencados por Géron (2019), este estudo desenvolve-se a partir do treinamento de um modelo que possui um aprendizado classificado como: *Supervisioned, Batch e Model-Based.*

Modelos do tipo de aprendizado supervisionado são treinados a partir de um conjunto de exemplos que contém todos os valores conhecidos para a variável dependente (rótulo do exemplo e valor a ser predito) e as variáveis independentes (também chamadas de *features*) são as características que influenciam a variável dependente. De uma maneira geral, têm como objetivo desenvolver autonomamente um mapeamento entre os valores destas variáveis de um determinado conjunto de dados tal que o classificador resultante possua um poder de generalização suficiente para ser utilizado para predição de dados nunca vistos (KOTSIANTIS, 2007).

Já os sistemas que utilizam a estratégia de Batch Learning (aprendizado por lote, em português), são caracterizados pela incapacidade de aprender incrementalmente, pois devem ser treinados utilizando todos os dados disponíveis. Este tipo de abordagem é recomendado por Huyen (2022) para fins acadêmicos e de pesquisa devido à natureza dos dados - que em sua grande maioria são estáticos e não possuem um fluxo contínuo - e a priorização do rendimento do modelo em detrimento da velocidade de treinamento.

Por sua vez, *model-based learning* diz respeito à forma com que o sistema de *machine learning* realiza a generalização do modelo, ou seja, como realiza predições sobre novos dados nunca vistos a partir da inferência dos dados rotulados utilizados no treinamento. A abordagem de generalização *Model-Based Learning* é caracterizada pela construção de um modelo matemático sobre as variáveis independentes e dependentes dos dados de exemplo e o uso posterior deste modelo para gerar as predições sobre novos dados.

Dentre os algoritmos de *machine learning* para problemas de regressão em aprendizado supervisionado, destacam-se a Regressão Linear Múltipla, K-Vizinhos Mais Próximos, Árvores de Decisão e Floresta Aleatória.

2.1.1. Regressão Linear Múltipla

De acordo com Gazola (2002), atualmente a análise de regressão múltipla é uma das ferramentas ou métodos estatísticos utilizados com maior frequência. É uma metodologia estatística para predizer valores de uma variável resposta (dependente) para uma coleção de valores de variáveis

preditoras (independentes). O modelo de regressão linear múltipla pode ser expresso pela Equação 1.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \ (i = 1, \dots, n) \quad (1)$$

Onde Y_i é a variável dependente (ou explicada), X_{ik} $(k = 1, 2, \dots, p)$ são as variáveis independentes (ou explicativas), β_i são os parâmetros da regressão e ε_i são os erros aleatórios.

As variáveis X e Y têm valores observáveis, diferentemente dos valores de ε, fato que representa o efeito aleatório. Os erros aleatórios representam inúmeros fatores que, em conjunto, podem interferir nas observações da variável dependente Y (CHARNET et al., 1999).

2.1.2. K-Vizinhos Mais Próximos

O algoritmo K-Vizinhos Mais Próximos (KNN) é um classificador baseado na analogia, onde o conjunto de treinamento é formado por vetores de n-dimensões, sendo que cada elemento deste conjunto representa um ponto no espaço n-dimensional. De forma a classificar um elemento o qual não pertence ao conjunto de treinamento, o classificador KNN procura K elementos no conjunto de treinamento, sendo que estes K elementos devem estar próximos do elemento desconhecido. Os K elementos próximos são denominados K-vizinhos mais próximos. Desta forma, verifica-se quais são as classes dos K-vizinhos, sendo que a classe mais frequente é atribuída ao elemento desconhecido (SILVA, 2005).

A métrica mais comum para determinação dos K-vizinhos mais próximos de um ponto (x, y) é a distância euclidiana, definida pela Equação 2.

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (2)$$

O KNN acha então os K vizinhos mais próximos de um dado ponto no espaço, procurando pelo seguinte conjunto de vetores a partir da Equação 3.

$$e = \sum_{i=0}^{K} \sqrt{\sum_{j=0}^{N} \left(x_{i,j} - t_i\right)^2} \quad (3)$$

Onde $x_{i,j}$ são as linhas da matriz x, K é o Número de vizinhos, N é o número de elemento do vetor e t representa a instância que será classificada.

Aplicando-se uma modificação, é possível adaptar este algoritmo para problemas de regressão: ao invés de utilizar a classe mais recorrente nos vizinhos, deve-se calcular a média dos valores destas instâncias, que é dado pela Equação 4.

$$fe(t) = \frac{1}{K}\sum_{i=0}^{K}(v_i) \quad (4)$$

Onde v_i é o vetor de valores dos K vizinhos $(v_0, v_1, \dots, v_K)$.

2.1.3. Árvores de Decisão e Floresta Aleatória

A Árvore de Decisão (AD), segundo Breiman et al. (1984), representa um conjunto de restrições ou condições que são organizadas hierarquicamente e aplicadas sucessivamente de uma raiz a um nó terminal ou folha da árvore. Os principais benefícios de se utilizar esta heurística para realizar decisões de classificação e regressão é que a estrutura de árvore é simples, interpretável, transparente, possui um baixo custo computacional e pode ser representada graficamente. O algoritmo da AD envolve primeiro selecionar vetores de medição de divisão ideais, começando com a divisão do recurso dependente - o nó pai - em partes binárias, onde os nós filhos são "mais puros" que ele. Através deste processo, as ADs pesquisam em todas as divisões candidatas para encontrar a divisão ótima, s*, que maximiza a “pureza” da árvore resultante (conforme definido pela maior diminuição na impureza) através da Equação 5.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (5)$$

Onde t representa o nó, s é a divisão candidata e Δi(s,t) mede a diminuição na impureza. O nó t é dividido por s no nó filho esquerdo (tL) com uma proporção pL e no nó filho direito (tR) com uma proporção pR. i(t) é uma medida de impureza antes da divisão e i(tL) e i(tR) são medidas de impureza após a divisão. Na literatura existem diversos métodos para se calcular a impureza e, dentre eles, o Índice de Gini, que mede i(t) conforme a Equação 6.

$$I_G\left(t_{x(x_i)}\right) = 1 - \sum_{j=1}^{m} f\left(t_{x(x_i)}, j\right)^2 \quad (6)$$

Onde $f\left(t_{x(x_i)}, j\right)$ é a proporção de amostras com o valor x_i pertencentes a folha j como nó t. O critério de divisão da árvore de decisão é baseado na escolha do atributo com o menor índice de impureza Gini (I_G).

Já a Floresta Aleatória (FA) é uma técnica de regressão que combina múltiplos algoritmos AD a fim de prever o valor de uma variável dependente. Quando a FA recebe um vetor de entrada (x), composto pelos valores das diferentes características evidenciais analisadas para uma determinada área de treinamento, FA constrói um número K de árvores de regressão e faz a média dos resultados (BREIMAN, 2001). Depois que tais árvores ($\{T(x)\}_1^K$) são "cultivadas", o preditor de regressão de RF é calculado pela Equação 7.

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} T(x) \quad (7)$$

2.1.4. Métricas de desempenho

As métricas estatísticas de desempenho são uma parte essencial da avaliação de modelos de Machine Learning, pois fornecem medidas objetivas para avaliar a qualidade e a capacidade de generalização dos modelos em relação aos dados de teste.

O Coeficiente de Determinação, também conhecido como R^2 (Equação 8), é uma medida útil para avaliar o ajuste global do modelo, permitindo comparar diferentes modelos e determinar sua capacidade de explicar a variação dos dados e quantificando a proporção da variabilidade dos dados que é explicada pelo modelo. O R^2 varia de 0 a 1, onde 0 indica que o modelo não consegue explicar a variabilidade dos dados, e 1 indica que o modelo explica perfeitamente a variabilidade (MONTGOMERY et al., 2012).

Já a medida estatística do Erro Absoluto Médio (*Mean Absolute Error*, MAE) é utilizada na avaliação de modelos de regressão para medir a magnitude média dos erros de previsão para medir a magnitude média dos erros de previsão. O MAE é calculado como a média das diferenças absolutas entre os valores previstos pelo modelo e os valores reais (Equação 9) e, ao contrário do erro quadrático médio (RMSE), o MAE não pondera os erros de forma quadrática, o que significa que ele é menos sensível a valores discrepantes (*outliers*) (GEORGE, 2021).

Visto que a variável dependente consiste em um valor monetário (variação do valor de mercado, em euros), esta métrica mostra-se interessante em termos de interpretabilidade por fornecer uma medida direta da magnitude média dos erros.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \quad (8)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \quad (9)$$

Onde N é o número de observações, Y_i é o valor observado, $\hat{Y}$ é o valor estimado e $\overline{Y}$ é a média das observações.

2.1.5. Divisão de dados

O processo de divisão dos dados em pelo menos dois subconjuntos independentes, sendo um deles para o treinamento do modelo e outro para o teste do mesmo, é um importante procedimento realizado após o pré-processamento do conjunto de dados pois permite mensurar o desempenho de técnicas preditivas, além de evitar problemas de superajuste (*overfitting*) e garantir a generalização do modelo para dados não vistos (HASTIE et al., 2009). Por meio de pesquisas práticas conduzidas por Rajer-Kanduč et al. (2003) e Faraway (2018), demonstrou-se pelos resultados obtidos que a aplicação de técnicas de divisão dos dados para problemas de *machine learning* tem um significante impacto positivo na performance final de métodos empregados.

O Algoritmo de Kennard-Stone (KENNARD e STONE, 1969) é uma ferramenta utilizada para realizar, de forma sistemática, tal divisão de um conjunto representativo de amostras a partir de uma população maior, baseando-se na maximização da distância Euclidiana entre as amostras selecionadas, permitindo uma representação eficiente e eficaz da população original. Esta técnica tem como característica garantir uma cobertura abrangente do espaço de características, mantendo a maior variabilidade possível de registros nos subconjuntos, minimizando assim a redundância das informações.

2.1.6. Seleção de variáveis

De acordo com Yu e Liu (2004), um importante processo executado antes da aplicação de aprendizado de máquina é a seleção de variáveis (ou *features*), que compreende as etapas de identificação destas variáveis que são consideradas irrelevantes ou redundantes e a sua remoção do conjunto de dados. Esta prática tem em vista facilitar o posterior treinamento dos modelos de aprendizado de máquina por meio da redução da dimensionalidade das *features* com o objetivo de mitigar a interdependência entre elas e, assim, consequentemente, reduzindo o tempo necessário de execução, aumentando a eficiência, aprimorando a acurácia dos resultados, reduzindo a complexidade dos resultados aprendidos (MARKOVITCH e ROSENSTEIN, 2002; YU e LIU, 2004).

Como métrica para mensurar tal interdependência entre as variáveis, é utilizada a Informação Mútua (*Mutual Information*). A Informação Mútua (IM) é um conceito da Teoria Matemática da Informação que foi introduzido por Shannon (1948) no contexto da comunicação digital. De acordo com Cover e Thomas (2006), a IM é uma medida da quantidade de informação que uma variável aleatória tem sobre outra variável. Do ponto de vista do processo de seleção de variáveis, esta técnica se mostra útil pois permite quantificar a dependência mútua entre duas variáveis.

Ela descreve quanta informação duas variáveis aleatórias compartilham entre si, ou seja, a quantidade de incerteza sobre uma variável aleatória dado o conhecimento da outra variável aleatória. A informação mútua para duas variáveis aleatórias é simétrica e sempre não negativa. É igual a zero se e somente se as duas variáveis aleatórias são independentes. Além disso, a informação mútua entre duas variáveis aleatórias contínuas é igual a infinito se houver uma relação funcional entre essas duas variáveis aleatórias. Essas propriedades oferecem a possibilidade de a informação mútua ser usada como medida de dependência (LU, 2011).

A IM entre duas variáveis x_1 e x_2 é dada pela Equação 10.

$$IM(X_1, X_2) = \sum_{x_1 \varepsilon X_1} \sum_{x_2 \varepsilon X_2} p(x_1 x_2) log\left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right) \quad (10)$$

Onde $p(x_1, x_2)$ é a distribuição de probabilidade conjunta de X e Y, $p(x_1)$ e $p(x_2)$ são funções de probabilidade de distribuição marginal de X e Y, respectivamente.

2.2. Futebol

No futebol moderno, jogadores de futebol estão entre os ativos mais importantes de um clube de futebol: eles podem ser fatores-chave para um gerenciamento de sucesso de clubes de futebol ao contribuir com o balanço patrimonial em uma possível venda (ZHU et al., 2015). Conforme Singh e Lamba (2019), o valor de mercado de um jogador é uma pontuação cumulativa que depende de vários fatores como talento, popularidade, habilidade, estilo de jogo, eficiência etc. de um jogador, que também pode ser usado como parâmetro para comparação de vários jogadores. De tal forma, é uma avaliação econômica do jogador que não significa apenas quantificar seu valor em unidade monetária, mas também mensurar seu desempenho, fato que pode transformar o processo de dar uma classificação a um jogador em uma tarefa subjetiva. Os indicadores mais comuns que podem

afetar o valor de mercado dos jogadores se encaixam em duas principais categorias, Características e Desempenho.

2.2.1. Características do Jogador

A idade está entre as características que mais influenciam no valor de mercado (CARMICHAEL E THOMAS, 1993; GYIMESI E KEHL, 2021) pois pode refletir tanto a experiência quanto o potencial de um jogador. Um estudo realizado por Bryson et al. (2012) observou uma relação inversa entre estes dois fatores, apresentando o seu maior valor de mercado ao redor da metade dos 20 anos e o início do seu declínio logo em seguida.

A posição (goleiro, defensor, meio-campista ou atacante) em que o jogador atua também tem um grande peso na sua valorização. Além da possibilidade de comprovar empiricamente esta divergência entre diferentes posições, estudos identificaram que este fenômeno ocorre por dois principais motivos: flexibilidade (FRICK, 2007) e poder de atração de multidões (GARCIA-DEL-BARRIO E PUJOL, 2007; HE et al., 2015). O primeiro diz respeito à versatilidade e de um jogador em campo, por exemplo, os goleiros geralmente são os mais especializados pois atuam somente nesta posição, enquanto meio-campistas podem exercer múltiplas funções, inclusive durante uma única partida. Já o segundo fator é explicado pelo fato de que há posições - no caso, atacantes - que deixam o jogador com uma visibilidade maior em relação às outras simplesmente pelo fato de que o objetivo máximo do futebol é o gol.

Também é identificada a altura do jogador como uma característica de destaque, por que está relacionada com a habilidade de um jogador para cabecear a bola, fator que é importante tanto para posições ofensivas quanto defensivas, pois influencia diretamente na probabilidade de marcar ou evitar gols (FRY et al., 2014).

Alguns estudos foram realizados com o objetivo de inferir se a nacionalidade é uma característica que pode afetar o valor de mercado, o salário, ou a percepção subjetiva da habilidade dos jogadores. Garcia-del-Barrio e Pujol (2007), Frick (2007) e Bryson et al. (2012) identificaram que existe uma relação direta, enquanto Reilly e Witt (1995) e Medcalfe (2008) não encontraram evidências desta hipótese.

Por fim, outro fator característico que pode influenciar na valorização de mercado é a preferência do jogador pela utilização do pé direito, esquerdo ou a ambidestralidade dele. Bryson et al. (2012)

conduziram um extenso estudo no qual inferiu-se que a habilidade de jogar com ambos os pés está relacionada com um maior salário, enquanto Herm et al. (2014) também concluíram que isto pode levar a um aumento da valorização de mercado.

2.2.2. Desempenho do Jogador

O desempenho do jogador durante as partidas é o objeto de estudo mais impactado pelo avanço e desenvolvimento de novas tecnologias tanto para a coleta de dados estatístico quanto a análise destes. Dentre as variáveis que compõem esta categoria, a quantidade de gols marcados - originados por chutes, cabeceios, faltas e pênaltis - por partida é considerada uma unanimidade dentre a literatura, sendo o fator o mais relevante para determinar o valor de mercado baseado no desempenho do jogador (CARMICHAEL E THOMAS, 1993; BRYSON et al., 2012; HE et al., 2015). Outra variável que possui grande impacto no valor de mercado e faixa salarial de jogadores é o número de assistências (FRANCK E NÜESCH, 2012; LUCIFORA E SIMMONS, 2003), pois está diretamente associada à contribuição de marcação de gols dos companheiros de equipe.

Outras variáveis relacionadas ao desempenho dos jogadores em partidas não foram amplamente estudadas por pesquisadores historicamente, muito devido à escassez de dados disponíveis causada pela dificuldade de coleta das estatísticas em partidas e a comercialização destes dados por empresas privadas. Entretanto, outras variáveis relevantes identificadas na literatura são: passes (HERM et al., 2014), interceptações (MEDCALFE, 2008), dribles (HE et al., 2015), faltas cometidas (FRANCK E NÜESCH, 2012) e cartões amarelos e vermelhos (KIEFER, 2014).

2.2.3. Machine learning aplicado no futebol

Dentre as aplicações de *machine learning*, os estudos mais recorrentes na literatura contemplam como objeto de estudo o desenvolvimento de modelos para prever o resultado de uma partida de futebol (vitória do time mandante, vitória do time visitante ou empate) e estimar o valor de mercado de jogadores.

Pappalardo e Cintia (2018) desenvolveram um modelo supervisionado *random forest* treinado com dados de 6396 partidas (de 145 clubes durante 3 temporadas) com o objetivo de identificar o resultado de partidas e campeonatos. Atingindo uma performance de quase 80%, identificou-se que as vitórias e derrotas podem ser explicadas com a performance do time e que a posição final no campeonato disputado também depende da performance técnica. Schneider (2018) realizou um

estudo sobre um conjunto de dados que retrata as partidas de futebol do campeonato Premier League 2000-2017 para comparar o desempenho de diferentes algoritmos (Regressão Logística, Análise Discriminante Linear, *SVM* com *kernel* linear, *K-Nearest Neighbors*), atingindo uma acurácia média de cerca de 54%. Entretanto, ambos estudos apresentaram pouca ou nenhuma previsão de resultado como sendo pertencentes à classe empate.

Devido à escassez de conjuntos de dados completos com estatísticas de partidas de futebol disponibilizados gratuitamente, muitos pesquisadores e entusiastas da área não têm conseguido identificar características ou fazer previsões com acurácia, recorrendo à utilização de dados provenientes do jogo de futebol FIFA para aplicar técnicas preditivas (COTTA et al., 2016).

Um estudo conduzido por Singh e Lamba (2019) utilizou dados das habilidades dos jogadores provenientes do jogo FIFA 18 em conjunto com métricas de popularidade (retiradas da Wikipédia) e fatores identificados em estudos anteriores a fim de elencar quais variáveis possuem mais influência sobre o valor de mercado.

Já Behravan e Razavi (2020) aplicaram uma técnica de agrupamento em dados extraídos do jogo FIFA 20 e treinaram um modelo de regressão híbrida - uma combinação de PSO (otimização por enxame de partículas) e SVR (máquina de vetores de suporte) - com o objetivo de estimar o valor de mercado dos jogadores com base em suas habilidades no jogo. Segundo os autores, foi alcançada uma acurácia de 74%, demonstrando melhores resultados que outras técnicas comparadas no estudo.

3. PROCEDIMENTOS METODOLÓGICOS

De forma geral, a estratégia de implementação adotada por este trabalho consiste em obter dados de partidas de futebol e utilizar diferentes algoritmos de *machine learning* para gerar modelos capazes de prever uma possível valorização ou desvalorização do valor de mercado dos jogadores. Também objetiva-se determinar os fatores (variáveis) que mais influenciam nesta variação.

3.1. Classificação da pesquisa

Segundo Gerhardt e Silveira (2009), uma pesquisa pode ser classificada de acordo com a sua abordagem, sua natureza, seus objetivos e seus procedimentos. Este trabalho caracteriza-se por adotar uma abordagem quantitativa pois enfatiza a objetividade, baseando-se na coleta de dados brutos de partidas de futebol e posterior análise recorrendo a métodos matemáticos e estatísticos

para descrever relações entre variáveis (FONSECA, 2002). Em relação à natureza, adota-se uma pesquisa aplicada pois tem um propósito imediato de embasar tomadas de decisão utilizando dados no que tange futuras contratações e vendas de jogadores de futebol, gerando conhecimentos para aplicação prática (GERHARDT E SILVEIRA, 2009). Quanto aos objetivos, trata-se de uma pesquisa explicativa visto que são desenvolvidos modelos a fim de identificar possíveis variações futuras no valor de mercado de jogadores de futebol a partir de dados de suas partidas, bem como quais são as variáveis que mais influenciam esta variação (GIL, 2002). Por fim, quanto aos procedimentos, a pesquisa classifica-se como experimental pois, de acordo com Gil (2002), ela consiste em determinar um objeto de estudo - valor de mercado dos jogadores de futebol - e selecionar as variáveis que seriam capazes de influenciá-lo - por meio da aplicação de técnicas de *machine learning*.

3.2. Etapas do trabalho

O método proposto se apoia em cinco etapas: coleta dos dados, preparação dos dados, escolha do algoritmo de regressão, treinamento dos modelos e análise dos resultados.

A primeira etapa consiste em gerar um conjunto de dados estruturado que agregue informações acerca (i) dos eventos (jogos) de cada jogador, (ii) de suas características pessoais, e (iii) do histórico do valor de mercado dos jogadores. Informações relativas ao bloco (i) consistem de dados estatísticos de partidas de futebol das temporadas 20/21, 21/22 e 22/23 dos quatro principais campeonatos domésticos europeus: *Premier League* (Inglaterra), *Bundesliga* (Alemanha), *Serie A* (Itália) e *LaLiga* (Espanha). Para isto, será utilizado o serviço de API disponibilizado pela empresa Sportradar, responsável pela coleta, processamento e disponibilização do conjunto de dados, o qual é composto pelos eventos (descritos no Anexo A) executados por cada jogador nas partidas das temporadas analisadas. Tais eventos serão utilizados como as variáveis independentes (*features*) nos subsequentes treinamentos dos modelos. O bloco (ii) é derivado da coleta de informações básicas dos jogadores, ou seja, suas características pessoais e que não estão sujeitas à variação ao longo das temporadas. Estes dados, disponibilizados pela Sportradar, estão descritos no Anexo B e também serão utilizados como variáveis independentes. Já o bloco (iii) contém o histórico do valor de mercado dos jogadores que atuaram nas temporadas indicadas, retirado do *website* Transfermarkt, a maior comunidade *online* destinada para este fim e, inclusive, utilizada por clubes como parâmetro para valores de transferência de jogadores e negociação de salários (HERM et al.,

2014). Este conjunto é composto pelo nome do jogador, seu valor de mercado (que será utilizado como variável dependente), a data em que o valor foi definido (que possui uma frequência média de atualização de 5 meses) e a sua idade nesta mesma data.

A segunda etapa engloba o processo de preparação dos dados para posterior modelagem através dos algoritmos de *machine learning*. Primeiramente, os blocos (i) e (ii) serão unidos a fim de gerar um único conjunto composto por variáveis independentes; tal unificação é simples, visto que os dados são provenientes da mesma plataforma e os jogadores possuem uma chave única de identificação. O banco resultante é então unificado com os dados do bloco (iii). Entretanto, como tratam-se de fontes diferentes, não existe uma correspondência direta entre as chaves únicas dos jogadores. Para tanto, o processo de união utilizará os nomes dos jogadores e, para os casos em que uma relação não pode ser identificada, um mapeamento manual entre as duas plataformas será realizado. Por fim, visando mitigar a multicolinearidade entre variáveis altamente interdependentes, Guyon e Elisseeff (2003) recomendam a combinação de variáveis para formar novos atributos, reduzindo assim os efeitos negativos da correlação e melhorar o desempenho e a interpretabilidade dos modelos de *machine learning*.

A terceira etapa visa determinar o modelo de regressão mais adequado para predição da variação de valor dos jogadores de forma global (ou seja, sem distinções da posição em que o jogador atua). Para isso, os seguintes passos serão aplicados nos conjuntos de dados completos de cada campeonato analisado (*Bundesliga*, *LaLiga*, *Premier League* e *Serie A*): 1) Divisão dos dados em subconjuntos de treino (75%) e teste (25%) utilizando o algoritmo de Kennard-Stone; 2) Cálculo da Informação Mútua de todas as variáveis independentes em relação à variável dependente na porção de treino; 3) Treinamento iterativo dos modelos preditivos (Regressão Linear Múltipla, K-Vizinhos Mais Próximos e Floresta Aleatória), removendo a variável com menor IM em cada iteração e calculando o erro de predição (MAE); esse processo é repetido até que sobre apenas uma variável independente no modelo; 4) Identificação do subconjunto de variáveis responsável pelo menor MAE para cada modelo preditivo; e 5) Cálculo das métricas de desempenho (R^2 e MAE) finais e indicação da melhor técnica de predição e do conjunto recomendado de variáveis a ser usado por tal técnica. A precisão dos modelos selecionados é validada na porção de teste.

A seguir, na quarta etapa, os conjuntos de dados serão desdobrados para representarem apenas uma das posições de interesse (defensor, meio-campista ou atacante). Os dados dos jogadores da

posição escolhida serão então modelados usando uma rotina similar à etapa acima [passos 1) a 5)], porém somente a melhor técnica de predição será avaliada (reduzindo substancialmente o número de modelagens necessárias). A precisão dos modelos selecionados é validada na porção de teste.

Por fim, a quinta última etapa deste estudo consiste na análise dos resultados obtidos a partir do treinamento dos modelos, tendo como objetivo principal a realização de comparações das importâncias das variáveis para uma mesma posição em diferentes campeonatos, a fim de identificar possíveis padrões e características distintas. Essa análise permitirá levantar hipóteses sobre o perfil de cada campeonato, considerando as variáveis que mais influenciam na variação do valor de mercado dos jogadores baseando-se na sua performance nas partidas e suas características intrínsecas.

4. RESULTADOS

4.1. Preparação dos dados

Para se obter um conjunto de dados adequado à modelagem, algumas transformações nos dados foram realizadas. Primeiramente, a variável *country_code*, por possuir uma baixa representatividade por parte de países de menor expressão no futebol, foi transformada em *continent_code*, contendo apenas o código dos continentes associados ao respectivos países. A seguir, foi aplicada a técnica de one-hot encoding (SEGER, 2018) para realizar o tratamento das variáveis categóricas, visto que os modelos preditivos testados são incapazes de processar variáveis categóricas. A técnica one-hot encoding converte variáveis categóricas em uma representação binária, a qual é compreendida pelos modelos de regressão. As variáveis geradas a partir dessa transformação foram: *preferred_foot_both*, *preferred_foot_left*, *preferred_foot_right*, *continent_code_AF*, *continent_code_AS*, *continent_code_EU*, *continent_code_NA*, *continent_code_OC* e *continent_code_SA*.

Na sequência, conforme proposto por Bruce et al. (2020) em problemas de *machine learning*, foi realizada uma análise da correlação entre as variáveis dependentes do conjunto de dados, utilizando o coeficiente de Pearson. A partir dessa análise, observou-se que os maiores valores de correlação estavam relacionados a dois fatores principais:

1. As variáveis com sufixo "*_successful*" e "*_unsuccessful*" apresentaram uma forte correlação com suas respectivas variáveis com sufixo "*_total*" (por exemplo:

passes_successful, *passes_unsuccessful* e *passes_total*), devido à sua relação direta. Essa correlação ocorre devido à natureza das variáveis, que representam estatísticas absolutas para cada jogador/partida.

2. Verificou-se que as variáveis "*minutes_played*" e "*player_starter*" possuíam uma forte relação com várias outras variáveis. Isso ocorre porque os jogadores que participaram de mais jogos e tiveram mais minutos em campo geralmente apresentam um maior número de resultados nas estatísticas, o que contribui para a correlação observada.

Para reduzir os impactos dessas correlações nos modelos preditivos, optou-se por substituir as variáveis com sufixo "*_successful*" e "*_unsuccessful*" por novas variáveis com sufixo "*_pct*", que representam o percentual de ações bem-sucedidas para as respectivas estatísticas. Essa conversão conduziu às seguintes variáveis: *crosses_successful_pct*, *passes_successful_pct*, *long_passes_successful_pct*, *tackles_successful_pct*, *shots_on_target_pct*, *shots_off_target_pct* e *shots_goal_pct*. Essa abordagem visa reduzir a dependência das estatísticas absolutas, fornecendo uma perspectiva relativa das ações bem-sucedidas em relação ao total de ações perpetradas durante o jogo. Além disso, facilita a comparação entre jogadores, pois leva em conta sua eficiência, independentemente do número absoluto de tentativas efetuadas. Adicionalmente, para reduzir o impacto da variável *minutes_played*, todas as estatísticas absolutas foram relativizadas, dividindo-as pela soma dos minutos jogados por cada atleta. Essa abordagem substituiu o número absoluto de ações pelo número de ações executadas por minuto, fornecendo uma perspectiva comparável entre os jogadores, independentemente do tempo em campo.

Em seguida, uma nova análise utilizando o coeficiente de correlação de Pearson reforçou o impacto positivo das transformações aplicadas na redução da multicolinearidade entre as variáveis (Figura 1). Os resultados indicaram uma diminuição significativa das correlações entre as variáveis, o que evidencia que as transformações realizadas foram efetivas na mitigação da dependência linear entre elas. Além de garantir maior robustez nos procedimentos de modelagem, as transformações também visam facilitar a interpretação dos coeficientes dos modelos regressivos.

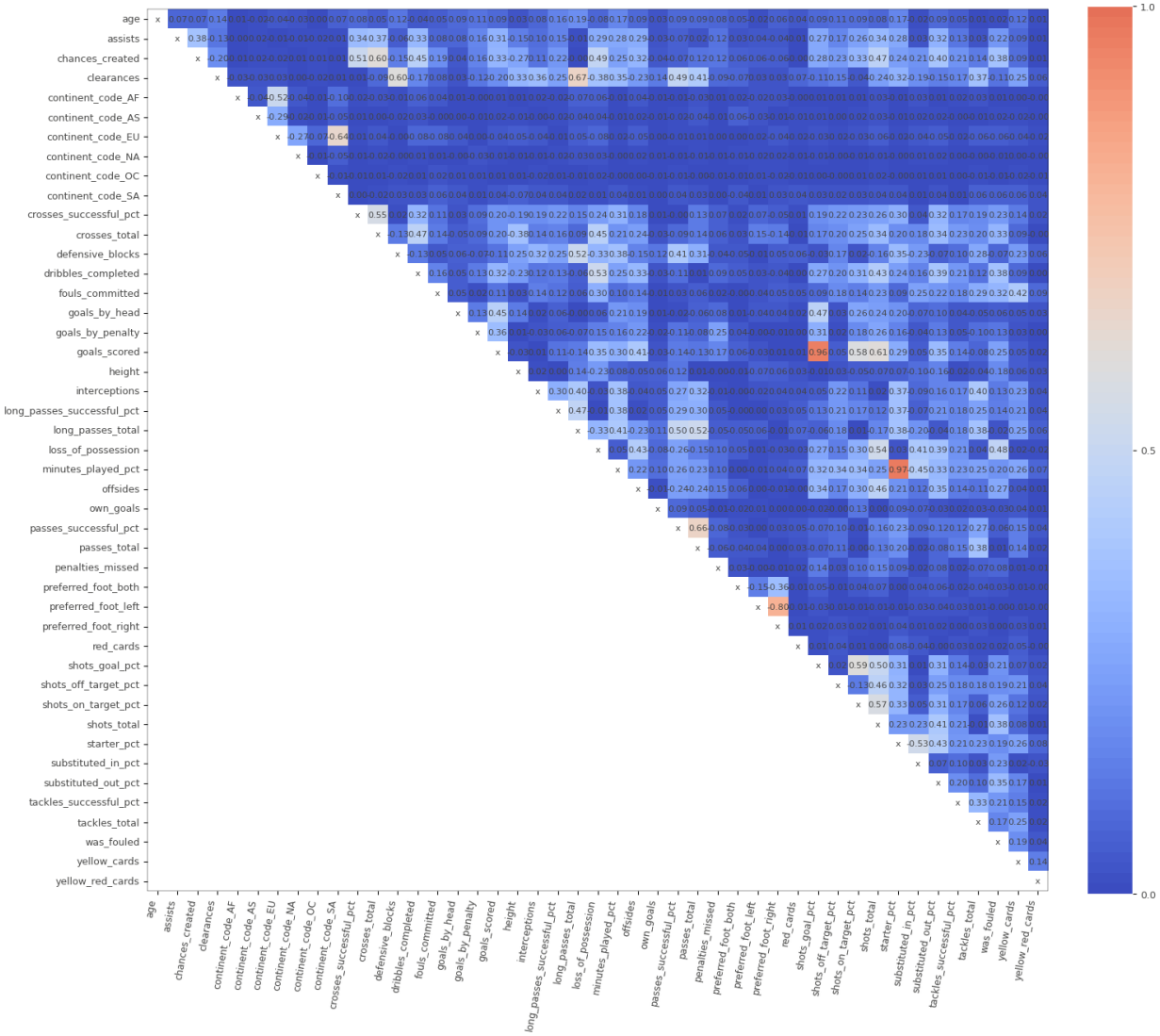


Figura 1 – Correlação entre as variáveis após as transformações

Contudo, destaca-se ainda uma notável correlação entre as variáveis *goals_scored* e *shots_goal_pct*. Essa relação sugere que jogadores que têm um alto percentual de acerto em chutes ao gol tendem, consequentemente, a marcar mais gols, ou seja, a eficiência na finalização pode ser um forte indicador do número total de gols marcados por um jogador. Já a correlação entre *minutes_played_pct* e *starter_pct* também apresenta um alto valor visto que jogadores que frequentemente iniciam as partidas (*starters*) também tendem a ter uma maior minutagem. Paralelamente, observa-se uma alta correlação inversa entre *preferred_foot_left* e *preferred_foot_right*, o que é esperado, pois um jogador que prefere usar o pé esquerdo para jogadas e chutes tende a não preferir o pé direito e vice-versa. É uma relação complementar: a predominância de um exclui, em grande parte, a predominância do outro.

Por fim, é importante abordar o desalinhamento temporal existente entre as datas das partidas e as datas das precificações dos jogadores, uma vez que esses eventos ocorrem em frequências discrepantes (7 e 145 dias, em média, respectivamente). Para resolver tal diferença, optou-se por realizar uma agregação das estatísticas das partidas para cada intervalo de precificação, calculando a média das estatísticas de cada jogador ao longo desse intervalo, considerando-as como variáveis dependentes. Além disso, a variação entre o valor de mercado do período atual e o período seguinte foi considerada como variável independente. Essas transformações permitem uma melhor adequação dos dados e viabilizam a análise e modelagem do desempenho dos jogadores em relação à variação de seus valores de mercado ao longo do tempo.

4.2. Seleção da melhor técnica preditiva

Na sequência, foi realizado um processo para determinar a melhor técnica de regressão por meio de testes no conjunto de dados. Nesse sentido, as técnicas a serem testadas neste estudo foram treinadas de acordo com os passos definidos nos procedimentos metodológicos, utilizando os conjuntos de dados dos quatro diferentes campeonatos e considerando todas as posições extraídas, com o objetivo de identificar a técnica que demonstrasse consistência e oferecesse os melhores resultados nas métricas selecionadas.

Os resultados obtidos, apresentados na Tabela 1, apontam a regressão de Floresta Aleatória como a técnica recomendada para a modelagem do desempenho de uma posição específica nos campeonatos analisados, tendo em vista seu menor MAE e superior R^2. Portanto, a regressão de Floresta Aleatória será utilizada nas próximas etapas metodológicas com o objetivo de modelar e compreender o desempenho de uma posição específica nos campeonatos em questão.

	Bundesliga		La Liga		Premier League		Serie A	
	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
Regressão Linear Múltipla	0,01	1.688.766	0,01	1.450.684	0,04	2.488.450	0,01	1.632.265
K-Vizinhos Mais Próximos	0,29	1.855.784	0,26	1.750.093	0,35	2.656.276	0,34	1.848.028
Floresta Aleatória	0,88	1.661.473	0,86	1.515.817	0,88	2.476.608	0,88	1.621.769

Tabela 1 – Resultado das métricas para as combinações de técnicas preditivas e campeonatos

Ressalta-se que a Floresta Aleatória demonstrou um desempenho superior tanto em termos do MAE quanto do R^2 em comparação aos modelos anteriores nos dados de treinamento. Além disso, reforçou a validade da utilização da Informação Mútua como um procedimento eficaz para a seleção de variáveis, uma vez que a remoção das *features* com menor IM resultou em um MAE

igual ou menor. A Figura 2 ilustra o processo de remoção das variáveis menos relevantes, onde o primeiro gráfico representa os valores do MAE ao longo das iterações e o segundo gráfico exibe os valores de R^2. Em ambos os gráficos, as diferentes linhas simbolizam os diversos campeonatos e o eixo X indica o número de variáveis independentes retidas (começando pela retenção de todas). A partir desta análise, observa-se que a Floresta Aleatória exibiu resultados consistentes em todos os conjuntos de dados analisados, o que sugere sua capacidade de generalização em diferentes contextos de campeonatos. Tal resultado pode ser justificado pela sua flexibilidade em capturar relações não lineares e interações complexas entre as variáveis e lidar com conjuntos de dados de alta dimensionalidade devido à natureza de árvores de decisão independentes.

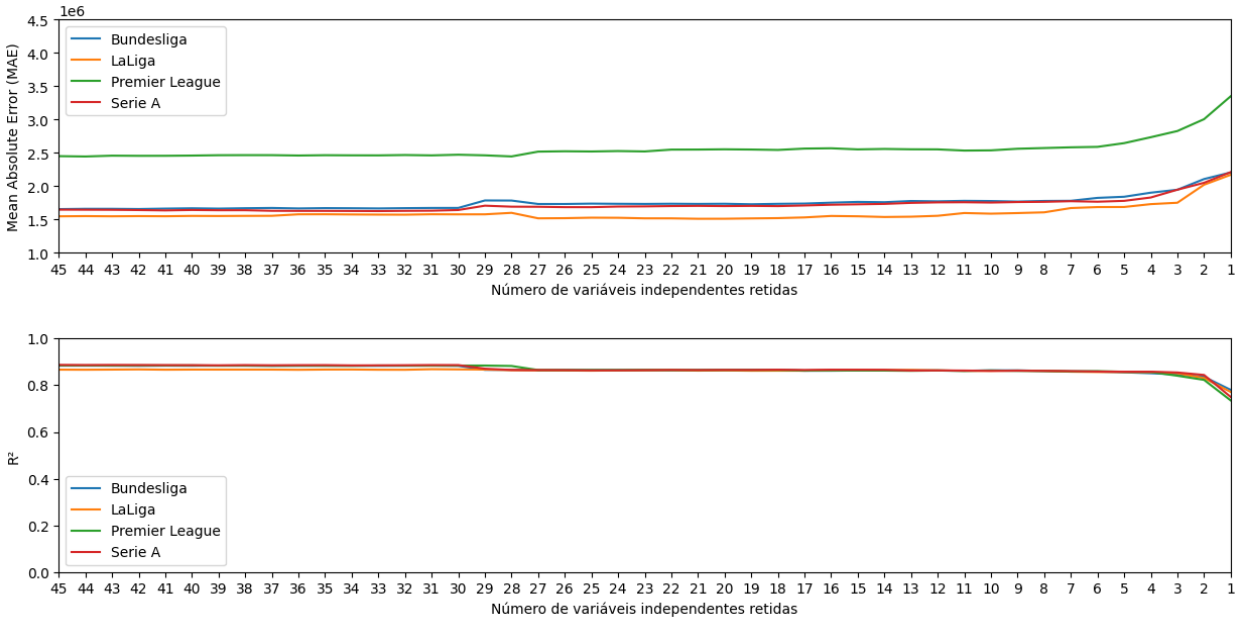


Figura 2 – Resultados da Floresta Aleatória

Os resultados provenientes do treinamento da regressão linear exibiram constância e consistência ao longo das iterações, independentemente do número de variáveis consideradas. O valor do MAE apresentou um comportamento similar nos diferentes conjuntos de dados, com uma exceção notável no campeonato Premier League, onde foi observado um valor relativamente maior nessa métrica em comparação aos outros. No entanto, é importante mencionar que os valores calculados para o R^2 foram próximos de zero em todos os conjuntos, indicando que o modelo não consegue explicar adequadamente a variabilidade dos dados e sugerindo que a regressão linear por si só pode não ser suficiente para capturar os padrões e complexidades presentes nos dados analisados. Uma das razões que pode causar esta incapacidade de generalização é a baixa adaptabilidade desta

regressão a dados não normalmente distribuídos, pois ela parte do pressuposto que os resíduos sigam uma distribuição normal e, caso essa suposição seja violada, a interpretação dos resultados pode ser comprometida.

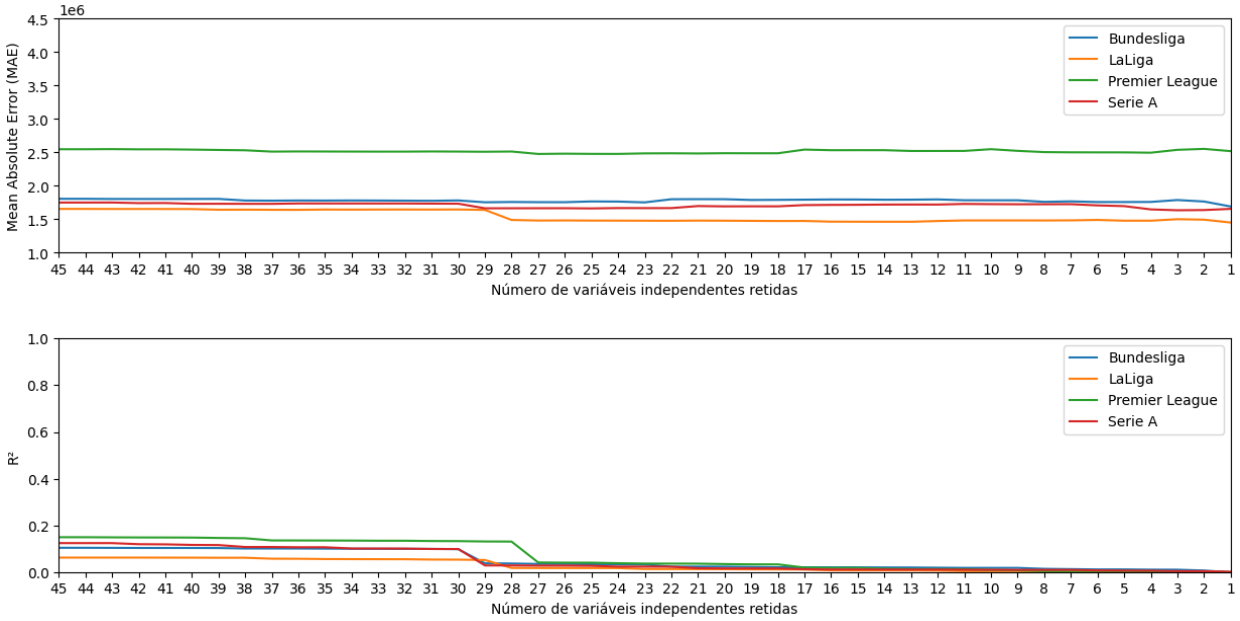


Figura 3 – Resultados da Regressão Linear Múltipla

Já a técnica K-Vizinhos Mais Próximos demonstrou uma maior variação das métricas ao longo do processo de remoção de variáveis, destacando a sua sensibilidade à utilização da Informação Mútua como ferramenta de seleção. Na Figura 3 percebe-se que o MAE obteve resultados maiores que a Regressão Linear Múltipla, enquanto o R^2 obteve um desempenho melhor para todas as iterações. A maior capacidade de generalização deste modelo pode ser explicada pela maior flexibilidade ao não supor uma relação linear entre as variáveis independentes e a variável dependente. Além disso, apresenta uma maior robustez a outliers, pois seu processo de previsão é baseado na vizinhança dos pontos de dados, significando que os valores atípicos têm menos impacto nos resultados do KNN, tornando-o mais adequado para conjuntos de dados com possíveis outliers.

A hiper parametrização é uma etapa fundamental na construção de modelos preditivos, uma vez que a configuração adequada desses parâmetros pode otimizar significativamente a precisão e o desempenho do modelo (HASTIE et al., 2009). Para as três técnicas preditivas analisadas, foram exploradas diversas combinações de hiper parâmetros, buscando encontrar a configuração que

maximizasse o desempenho, no entanto, interessantemente, as variações nos resultados, decorrentes das diferentes combinações de hiper parâmetros, não apresentaram diferenças significativas em termos de métricas de avaliação. Isso sugere uma estabilidade notável dessas técnicas em relação ao conjunto de dados analisado. Para os modelos finais, os hiper parâmetros adotados foram: para a Regressão Linear, utilizou-se uma regularização L2 com coeficiente de 0,01; para o K-Vizinhos Mais Próximos, escolheu-se um valor de k=5 com uma métrica de distância euclidiana; e para a Floresta Aleatória, optou-se por um total de 100 árvores com uma profundidade máxima de 10 e um mínimo de 2 amostras para divisão de um nó.

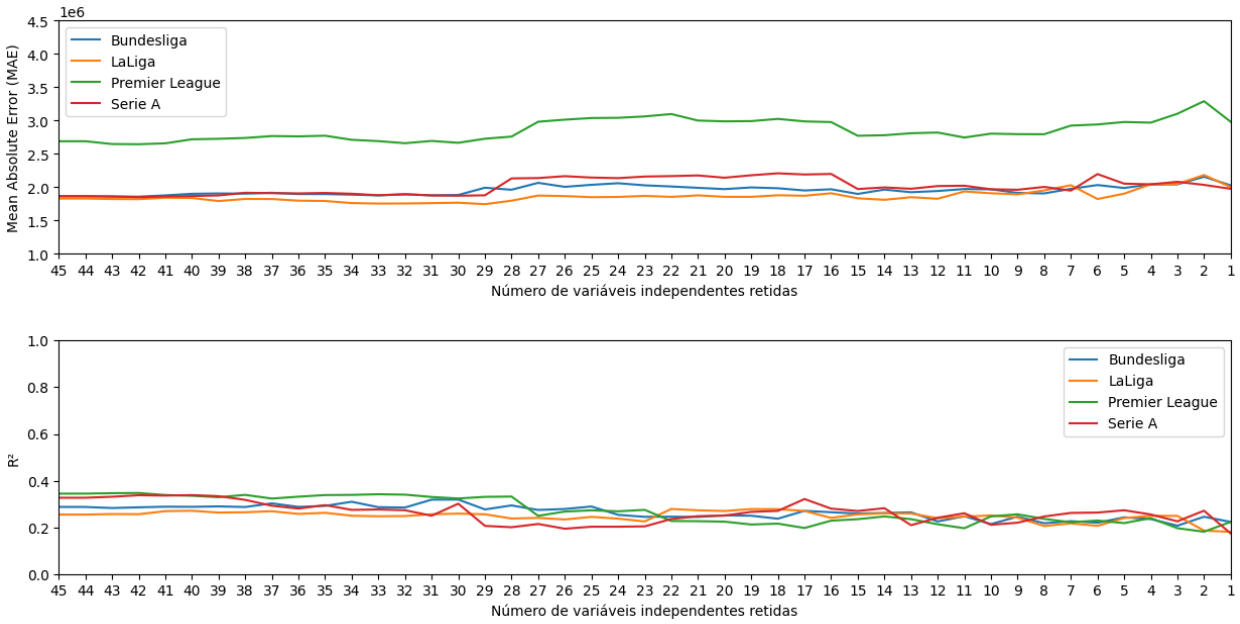


Figura 4 – Resultados do K-Vizinhos Mais Próximos

Nas análises que seguem, o escopo de predição de valorização foi restrito à posição "meio-campista" de atuação dos jogadores, com vistas a realizar uma avaliação mais focada e direcionada com base na técnica de predição selecionada (Floresta Aleatória). A escolha desta posição se justifica por contar com maior disponibilidade de dados em relação às outras posições, englobar diferentes perfis (como volantes, alas, meias-armadores e meias-atacantes), e por possuir a maior média de precificação entre todas as posições. Esses são fatores que aumentam o potencial para descobrir percepções significativas, logo a abordagem proposta tem maior potencial de contribuição.

4.3. Análise da posição meio-campista via Floresta Aleatória

Através da seleção e análise do subconjunto específico de meio-campistas, foi possível realizar uma investigação mais aprofundada sobre as características e desempenho dos jogadores nessa posição, além de permitir possíveis associações de diferentes perfis em cada campeonato.

Assim como detalhado nos procedimentos metodológicos, a Floresta Aleatória foi treinada iterativamente para os quatro campeonatos, a fim de se identificar o número recomendado de variáveis a serem utilizadas. Este processo de seleção de variáveis é essencial para simplificar o modelo, evitar o superajuste do modelo preditivo e melhorar a precisão da previsão, descartando variáveis que têm uma pequena contribuição para a explicação da variável alvo.

A Tabela 2 sintetiza os resultados das iterações ótimas para cada subconjunto de dados. As variações nos resultados entre os campeonatos podem ser atribuídas às suas características particulares, incluindo o estilo de jogo preferido, bem como peculiaridades nos próprios dados, como cardinalidade, variabilidade, presença de outliers, entre outros aspectos. Nesse cenário, o modelo treinado para o campeonato da *Bundesliga* se destaca, demonstrando uma maior capacidade de generalização da regressão em análise. Esse modelo apresentou os valores mais altos de R^2 tanto para os dados de treinamento quanto para os de teste, além do menor valor de MAE, indicando uma adequação superior do modelo aos dados desse campeonato em comparação com os outros subconjuntos analisados.

	Bundesliga	La Liga	Premier League	Serie A
# Variáveis retidas	35	42	39	31
# Variáveis removidas	10	3	6	14
R^2	0,89	0,87	0,89	0,87
MAE	1.703.015	1.897.731	2.569.441	1.724.347

Tabela 2 – Resultados da iteração ótima para cada campeonato

O primeiro passo na análise via Floresta Aleatória envolveu o cálculo da Informação Mútua (IM) para cada campeonato (Anexo C), servindo como base para os treinamentos iterativos subsequentes, permitindo refinar o modelo de forma iterativa e garantindo que as variáveis mais influentes sejam incorporadas nas fases subsequentes do processo de modelagem.

O melhor desempenho do modelo para o campeonato da *Bundesliga* foi alcançado pelo modelo retendo 35 variáveis, quando o MAE atingiu o valor mínimo de €1.703.015. Nesta iteração, as 10 variáveis com menor IM foram removidas do modelo, tendo a *continent_code_EU* como linha de

corte possuindo um valor de IM de 0,05160 (Figura 5). Isso significa que todas as variáveis com IM abaixo deste valor foram consideradas menos relevantes e, portanto, removidas do modelo. Essa linha de corte é uma métrica crucial para determinar a importância das variáveis no modelo e direcionar a seleção de características.

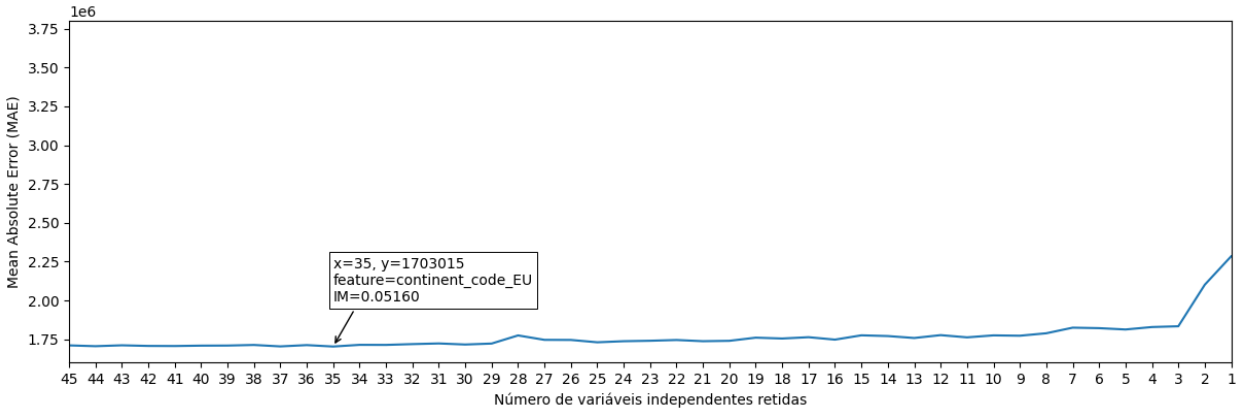


Figura 5 – Treinamento iterativo do campeonato Bundesliga

Seguindo o mesmo procedimento, a análise foi então aplicada ao campeonato *La Liga*. O modelo apresentou o melhor desempenho ao reter 42 variáveis, tendo sido removidas as 3 com menor IM; o MAE atingiu seu valor mínimo de €1.897.731. O ponto de corte nesse caso foi a *feature preferred_foot_both*, que obteve um valor de IM de 0,01923 (Figura 6).

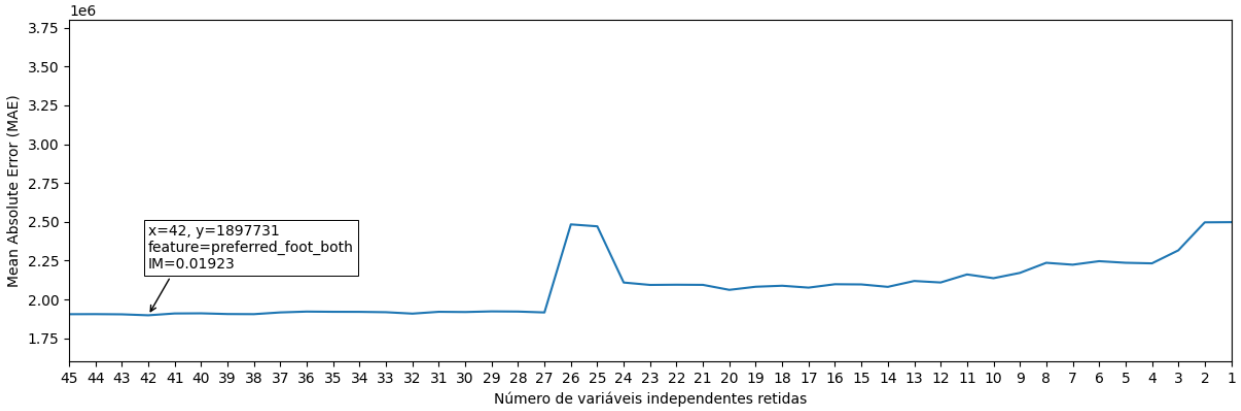


Figura 6 – Treinamento iterativo do campeonato La Liga

Para a *Premier League*, a iteração que resultou no menor valor de MAE, €2.569.441, ocorreu ao reter 39 variáveis, sendo as 6 variáveis de menor IM excluídas do modelo. Notavelmente, o valor de corte para esta iteração foi estabelecido pela *feature continent_code_SA*, com um valor de IM de 0,03625 (Figura 7).

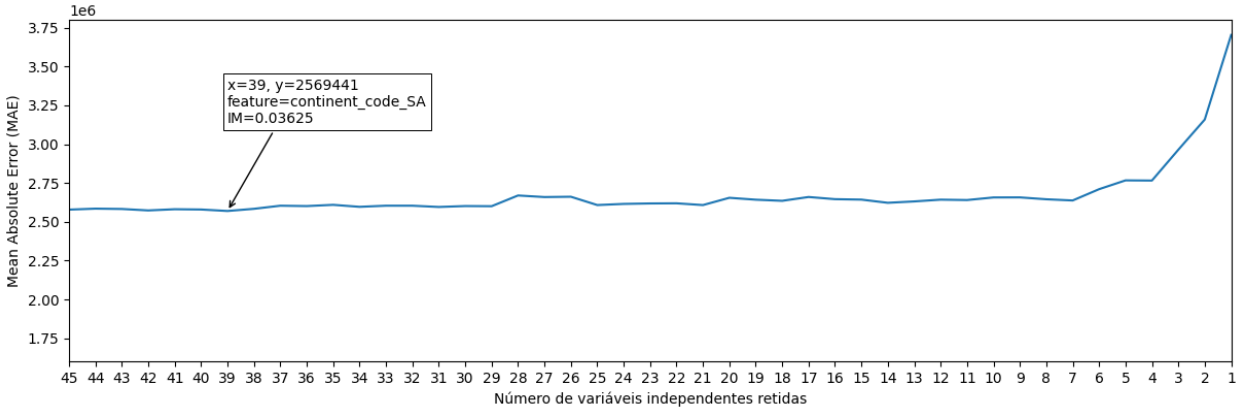


Figura 7 – Treinamento iterativo do campeonato Premier League

Finalmente, a análise da posição de meio-campista no campeonato *Serie A* resultou em um valor de MAE mínimo de €1.724.347 retendo 31 variáveis. Nesta modelagem, as 14 variáveis de menor IM foram removidas do modelo, sendo a feature *goals_by_penalty* como linha de corte com IM=0,19477 (Figura 8).

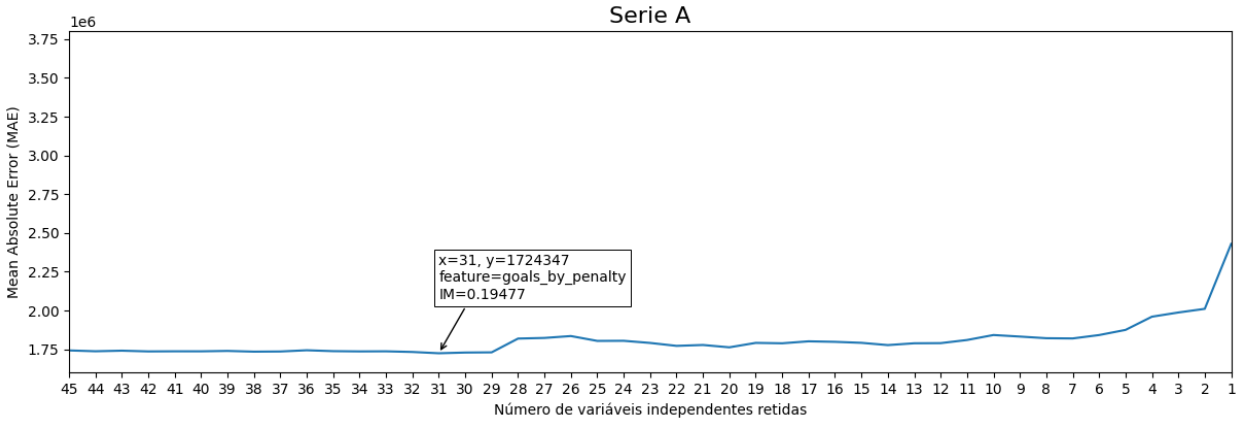


Figura 8 – Treinamento iterativo do campeonato Serie A

A análise inicial dos valores de IM calculados para as características dos jogadores em cada um dos campeonatos sugerem padrões de similaridade entre as variáveis. As variáveis que indicam o pé de preferência do jogador (*preferred_foot_both*, *preferred_foot_left* e *preferred_foot_right*) apresentaram valores de IM muito próximos de zero, sendo algumas delas as primeiras a serem removidas nos treinamentos iterativos dos modelos. Isso contrasta com as conclusões do estudo de Herm et al. (2014), que sugeria um impacto positivo da habilidade de jogar com ambos os pés no valor de mercado de um jogador. Similarmente, as variáveis que indicam o continente de origem do jogador (*continent_code_AF*, *continent_code_AS*, *continent_code_EU*, *continent_code_NA*,

continent_code_OC e *continent_code_SA*) também tiveram valores de IM baixos e foram consequentemente removidas. Embora essas variáveis possam parecer intuitivamente importantes, os dados sugerem que elas não adicionam informação relevante ao modelo e, ao invés disso, podem introduzir ruído, afetando negativamente as métricas de MAE e R^2.

4.4. Análise da importância das variáveis para meio-campistas

Para aprofundar o entendimento sobre a relevância de cada característica nos modelos elaborados, foi conduzida uma análise utilizando a métrica *Mean Decrease in Impurity* (MDI), reconhecida por avaliar a importância das variáveis em modelos de aprendizado de máquina baseados em árvores de decisão. Como explicado por Louppe et al. (2013), a MDI mede a contribuição de uma variável para a acurácia do modelo, calculando a média da diminuição na impureza nos nós da árvore onde a variável é empregada (ou seja, uma redução maior na impureza sinaliza uma variável de elevada importância). Nesta seção, comparam-se os valores de MDI entre os modelos aplicados aos campeonatos da *Bundesliga*, *La Liga*, *Premier League* e *Serie A*, permitindo identificar e discutir as diferenças e semelhanças nas variáveis que desempenham papéis significativos em cada campeonato.

Baseando-se nos valores da MDI, avaliou-se e categorizou-se a importância das variáveis em cada modelo de campeonato. Os gráficos das Figuras 9 a 12 mostram a MDI de cada característica, além de representarem o desvio padrão dessa métrica. Nota-se que, apesar de algumas discrepâncias entre os campeonatos, certas variáveis exibem elevada importância em todos os conjuntos de dados avaliados.

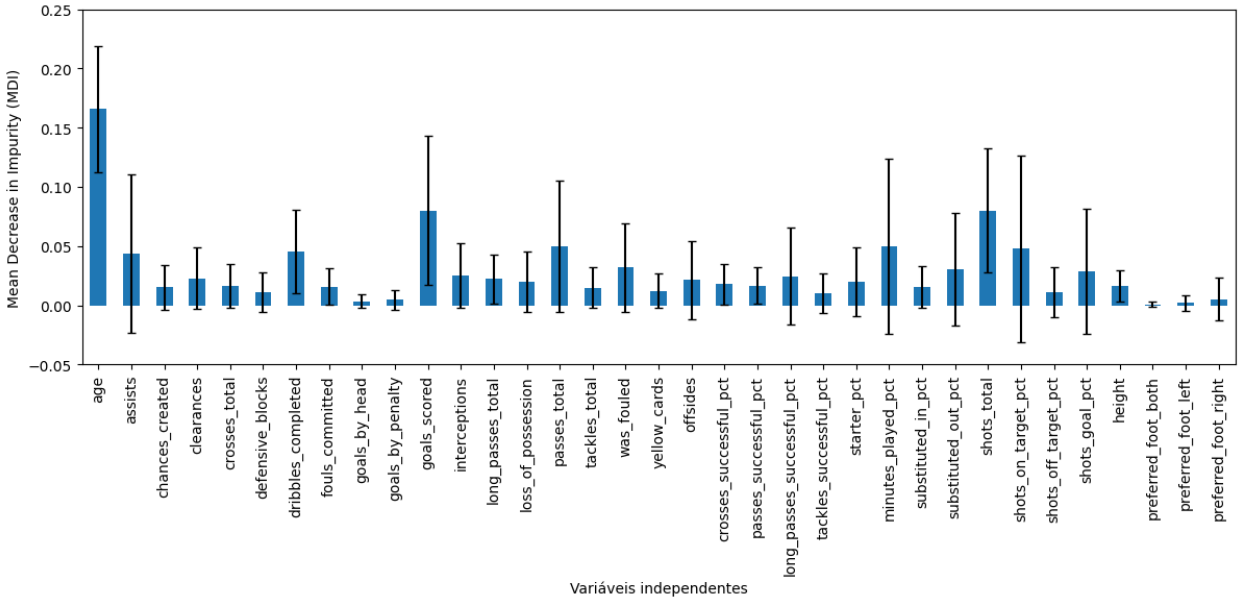


Figura 9 – Importância das variáveis (Bundesliga)

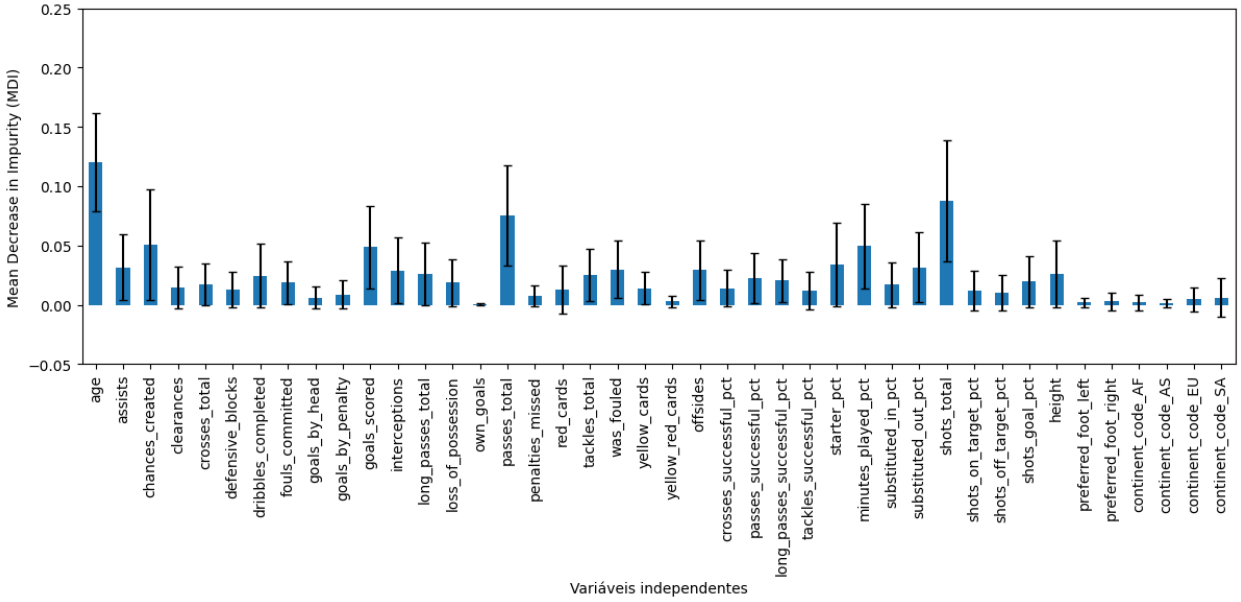


Figura 10 – Importância das variáveis (La Liga)

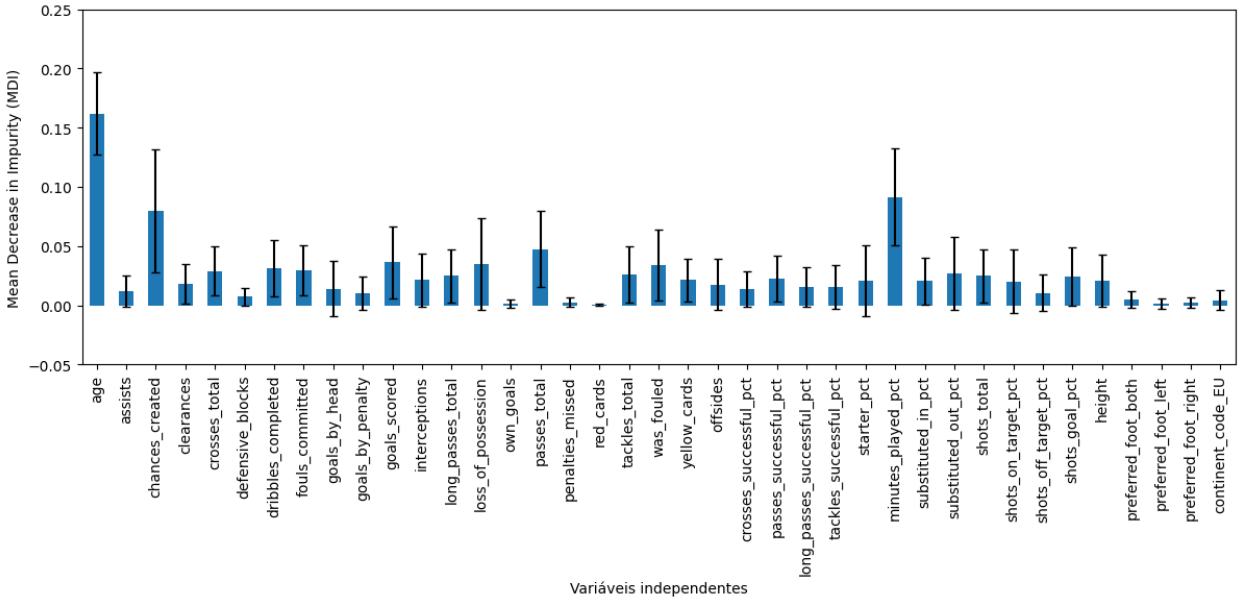


Figura 11 – Importância das variáveis (Premier League)

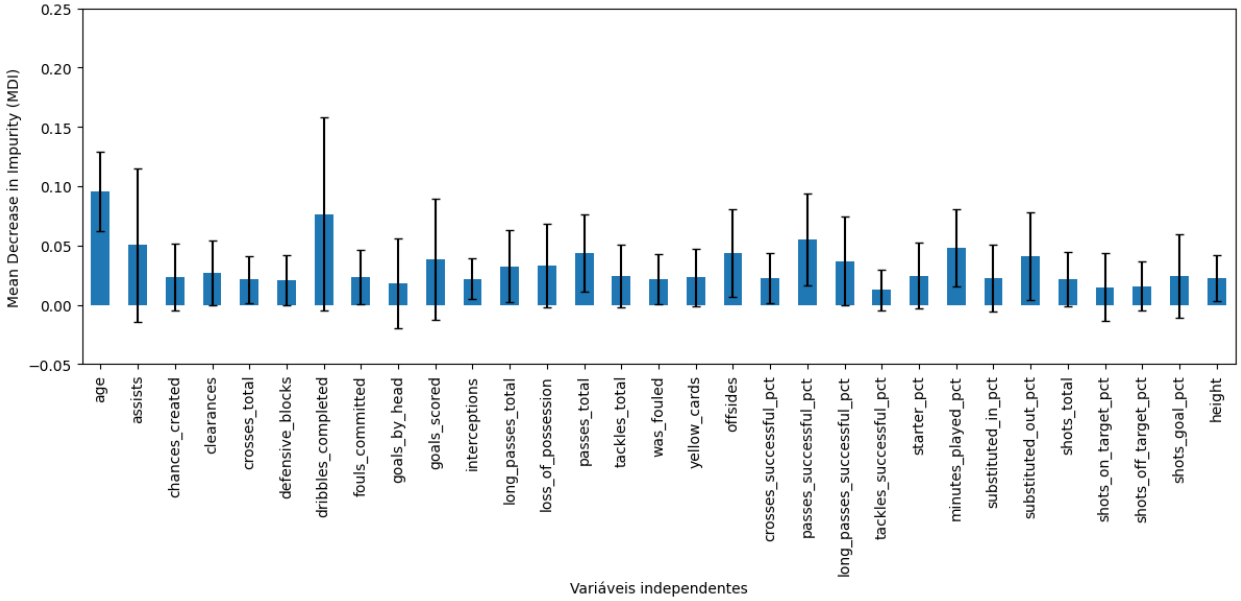


Figura 12 – Importância das variáveis (Serie A)

Para facilitar a comparação entre os diferentes campeonatos, os valores de MDI obtidos foram compilados na Tabela 3, viabilizando uma visualização abrangente e direta das diferenças e semelhanças na relevância das variáveis e assim possibilitando uma análise comparativa das

peculiaridades de cada campeonato. O valor percentual da tabela representa a importância de cada fator (linhas) em relação a variável de maior relevância em cada campeonato (colunas).

	Bundesliga	La Liga	Premier League	Serie A
age	100%	100%	100%	100%
assists	27%	27%	7%	52%
chances_created	9%	43%	50%	24%
clearances	14%	13%	11%	28%
continent_code_AF	0%	2%	1%	0%
continent_code_AS	0%	1%	1%	0%
continent_code_EU	0%	4%	1%	0%
continent_code_NA	0%	0%	0%	0%
continent_code_OC	0%	0%	0%	0%
continent_code_SA	0%	5%	4%	0%
crosses_successful_pct	11%	12%	9%	23%
crosses_total	10%	14%	19%	22%
defensive_blocks	7%	11%	4%	22%
dribbles_completed	28%	21%	20%	79%
fouls_committed	10%	16%	17%	24%
goals_by_head	2%	5%	9%	19%
goals_by_penalty	3%	8%	6%	0%
goals_scored	42%	41%	20%	41%
height	10%	22%	12%	24%
interceptions	15%	24%	14%	23%
long_passes_successful_pct	15%	17%	9%	39%
long_passes_total	13%	22%	15%	33%
loss_of_possession	12%	16%	21%	34%
minutes_played_pct	30%	42%	57%	50%
offsides	13%	24%	11%	45%
own_goals	0%	1%	1%	0%
passes_successful_pct	10%	19%	14%	57%
passes_total	30%	63%	29%	46%
penalties_missed	0%	6%	2%	0%
preferred_foot_both	1%	0%	2%	0%
preferred_foot_left	1%	2%	1%	0%
preferred_foot_right	3%	3%	1%	0%
red_cards	0%	11%	0%	0%
shots_goal_pct	17%	17%	16%	25%
shots_off_target_pct	7%	8%	6%	17%
shots_on_target_pct	29%	10%	11%	16%
shots_total	48%	73%	17%	23%
starter_pct	12%	28%	12%	25%
substituted_in_pct	9%	14%	12%	23%
substituted_out_pct	18%	26%	17%	43%
tackles_successful_pct	6%	10%	9%	13%
tackles_total	9%	21%	15%	25%
was_fouled	19%	25%	23%	23%
yellow_cards	7%	12%	13%	24%
yellow_red_cards	0%	3%	0%	0%

Tabela 3 – Representação tabular dos valores de MDI

Observa-se que a variável idade (*age*) do jogador é altamente relevante em todos os modelos, independente do campeonato considerado, impactando negativamente a variação do valor de mercado do jogador. Esse comportamento corrobora vários estudos similares, como os de Carmichael e Thomas (1993), Bryson et al. (2012), Herm et al. (2014), Müller et al. (2017) e Gyimesi e Kehl (2021), reforçando a importância de um planejamento estratégico a longo prazo tanto para a gestão de clubes de futebol no recrutamento de jogadores, quanto para os próprios jogadores em relação à sua carreira.

Ao examinar a *Bundesliga*, as características que mais se sobressaem no modelo treinado são *shots_total* e *goals_scored*. Isso pode sugerir que meio-campistas com uma orientação ofensiva, contribuindo ativamente com tentativas de finalização e com a concretização de gols, apresentam uma maior probabilidade de obter destaque econômico neste campeonato.

Na *La Liga*, *shots_total* e *passes_total* são as variáveis que mais influenciam a variação do valor de um meio-campista. Pode-se entender que jogadores que estão frequentemente engajados em ambas as ações de finalização de jogadas ofensivas e ações de manutenção da posse de bola e distribuição do jogo estão mais propensos a terem uma maior valorização de valor de mercado.

No contexto da *Premier League*, *minutes_played_pct* e *chances_created* são as características que mais impactam a variação do valor de um meio-campista. A porcentagem do tempo total de jogo emerge como o elemento mais relevante, assim como na Bundesliga, além do número de oportunidades criadas pelo jogador, indicando que meio-campistas que consistentemente criam oportunidades de gol para seus colegas de equipe têm maior propensão para experimentar mudanças significativas em seu valor.

Finalmente, no caso da *Serie A*, *dribbles_completed*, *passes_successful_pct* e *assists* são as características que mais impactam a variação do valor de um meio-campista. Isto pode sugerir que os jogadores com habilidades individuais notáveis e capacidade de manter a posse de bola sob pressão, além da eficácia tanto na criação de jogadas (através da concretização de assistências) quanto na distribuição da bola (através de passes completados com sucesso) têm mais chances de serem valorizados no campeonato italiano.

5. CONCLUSÕES

Este estudo propôs uma sistemática para análise das variáveis que influenciam a variação do valor de mercado de meio-campistas em quatro dos principais campeonatos de futebol do mundo (Premier League, Serie A, La Liga e Bundesliga) utilizando a técnica de aprendizado de máquina Floresta Aleatória para construir modelos de previsão de valorização e gerar inferências com base nas variáveis mais relevantes.

Em termos do desempenho das técnicas de predição testadas, a Floresta Aleatória mostrou-se superior em comparação a outros modelos, além de reforçar a eficácia da Informação Mútua como método de seleção de variáveis, pois a remoção de variáveis com menor IM resultou em um Erro Médio Absoluto igual ou menor.

A análise direcionada à posição de meio-campista revelou que, independentemente do campeonato, a idade dos jogadores desempenha um papel fundamental na determinação do valor de mercado, corroborando o que é comumente observado na prática do futebol. Porém, é evidente que cada campeonato possui características próprias que se refletem nas variáveis que mais influenciam a variação do valor dos jogadores. Na Bundesliga e La Liga, características relacionadas à ofensividade dos jogadores se destacam como determinantes, enquanto na Premier League e Serie A as variáveis que indicam a participação na construção de jogadas e criação de oportunidades de gol são preponderantes.

Estes resultados têm implicações significativas para os clubes de futebol e os próprios jogadores. Para os clubes, essas evidências podem ser úteis na tomada de decisões estratégicas, como a contratação de novos jogadores, levando em conta as características que mais contribuem para o valor de mercado em seu campeonato específico. Para os jogadores e seus agentes, essas informações podem orientar o desenvolvimento de habilidades específicas para maximizar o valor de mercado, ou até mesmo serem utilizadas como embasamento estatístico para justificar o potencial de valorização do jogador ao ser transferido para um clube de um determinado campeonato.

No entanto, vale lembrar que a aplicação desses insights deve ser cuidadosa e contextualizada. Os resultados apresentados são baseados em modelos estatísticos e, portanto, podem não refletir perfeitamente a realidade. Além disso, o futebol é um esporte complexo e dinâmico, e as

preferências dos clubes e os estilos de jogo podem mudar com o tempo. Portanto, é recomendado que esses resultados sejam usados em conjunto com outras fontes de informação e análise para informar decisões de recrutamento e desenvolvimento.

Quanto as limitações deste estudo, destaca-se que a análise é restrita apenas à posição de meio-campista e aos quatro principais campeonatos de futebol da Europa, sendo assim, os resultados podem não se aplicar a outras posições ou ligas. Além disso, os modelos de Floresta Aleatória podem capturar interações complexas entre as variáveis, mas podem não ser capazes de explicar a totalidade da variação no valor de mercado dos jogadores, dado que fatores intangíveis e subjetivos (como carisma do jogador, popularidade ou percepção da mídia) não podem ser facilmente quantificados e incluídos no modelo. Por fim, a análise é restrita aos dados disponíveis e, portanto, há uma dependência da qualidade e abrangência desses dados, sendo possível que algumas variáveis que possam influenciar o valor de mercado dos jogadores não tenham sido incluídas neste estudo, seja porque não foram registradas ou porque não estavam disponíveis nos conjuntos de dados utilizados.

Com base em tais limitações, sugere-se que, em estudos futuros, sejam incluídas estatísticas mais detalhadas de desempenho dos jogadores nas partidas, incluindo posição do jogador em campo no momento em que as ações foram executadas, número absoluto e taxa de sucesso de disputas aéreas, erros cometidos que resultaram em finalização do adversário e a posição inicial e final de passes, cruzamentos, dribles e condução de bola. Além de dados acerca das partidas, informações extracampo podem também ter um certo impacto na valorização e, portanto, a inserção de dados como número de lesões, suspensões de longo prazo (por julgamento após expulsão, *doping* ou mau-comportamento), partidas em que o jogador não foi relacionado e popularidade do jogador (por meio da análise de sentimentos de redes sociais, por exemplo) podem apresentar uma melhora nos resultados obtidos.

Em relação ao treinamento dos modelos de machine learning, nota-se um potencial significativo para ampliar a análise além da posição de meio-campista, explorando as demais posições macro no futebol, que incluem os goleiros, defensores e atacantes. Cada uma dessas posições tem um conjunto único de responsabilidades e características no campo que podem influenciar o valor de mercado de um jogador de maneiras diferentes, o que torna importante estudá-las individualmente,

sendo assim, esta expansão permitiria desenvolver uma compreensão mais completa dos fatores que impulsionam o valor de mercado dos jogadores de futebol.

Além disso, é recomendável considerar a avaliação de outras técnicas de machine learning não exploradas neste estudo, como redes neurais e *gradient boosting*. As redes neurais, que são inspiradas pelo funcionamento do cérebro humano, têm demonstrado um desempenho excepcional em uma variedade de tarefas de aprendizado de máquina, especialmente quando há muitos dados disponíveis (GOODFELLOW et al., 2016). Por outro lado, o *gradient boosting* é um método poderoso e flexível que constrói um modelo preditivo de forma incremental e é especialmente bom para lidar com dados heterogêneos (FRIEDMAN, 2001), o que pode ser viável em aplicações de futebol. Assim, o uso dessas técnicas pode melhorar ainda mais a precisão das previsões e proporcionar novos insights sobre a influência de várias variáveis no valor de mercado dos jogadores, possibilitando também a comparação dos resultados obtidos com diferentes técnicas de aprendizado de máquina para proporcionar uma visão mais completa das possíveis variações e nuances na determinação do valor de mercado dos jogadores.

REFERÊNCIAS

Ashley, K. Applied Machine Learning for Health and Fitness: A Practical Guide to Machine Learning with Deep Vision, Sensors and IoT. Editora Apress, 2020.

Brandes, L. e Franck, E. Social preferences or personal career concerns? Field evidence on positive and negative reciprocity in the workplace. Journal of Economic Psychology, v. 33, n. 5, p. 925-939, 2012.

Breiman, L. Random Forests. Machine Learning, v 45, p. 5-32, 2001.

Breiman, L., Friedman, J., Stone, C.J. e Olshen, R.A. Classification and Regression Trees. Routledge, Nova Iorque, 1984.

Bruce, P., Bruce, A. e Gedeck, P. Practical Statistics for Data Scientists, 2nd Edition. Editora Packt Publishing, 2020.

Bryson, A., Frick, B. e Simmons, R. The Returns to Scarce Talent: Footedness and Player Remuneration in European Soccer. Journal of Sports Economics, v 14, n 6, p. 606-628, 2013.

Carmichael, F. e Thomas, D. Bargaining in the transfer market: theory and evidence. Applied Economics, v 25, n 12, p. 1467-1476, 1993.

Cerri, R. e Carvalho, A. C. P. de L. F. de. Aprendizado de máquina: breve introdução e aplicações. Cadernos de Ciência & Tecnologia, v 34, n 3, p. 297-313, 2017.

Charnet, R., Freire, C. A. L., Charnet, E. M. R. e Bonvino, H. Análise de modelos de regressão linear com aplicações. UNICAMP. Campinas. 1999.

Constantinou, A. C., Fenton, N. E. e Neil, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. Knowledge-Based Systems, v. 50, p. 60-86, 2013.

Cotta, L., de Melo, P., Benevenuto, F., e Loureiro, A. A. Using FIFA soccer video game data for soccer analytics, 2016.

Cover, T. M. e Thomas, J. A. Elements of Information Theory. 2nd edn. Wiley-Interscience, New Jersey, 2006.

Dangeti, P. Statistics for Machine Learning. Editora Packt Publishing, 2017.

Faraway, J. J. Does Data Splitting Improve Prediction? Stat Comput, v 26, p. 49-60, 2016.

Faceli, K., Lorena, A. C., Gama, J. e Carvalho, A. C. P. L. F. De. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

Fonseca, J. J. S. Metodologia da pesquisa científica. Fortaleza: UEC, 2002.

Franck, E. e Nüesch, S. Talent and/or popularity: what does it take to be a superstar? Economic Inquiry, v 50, n 1, p. 202-216, 2012.

Frick, B. The football players' labor market: Empirical evidence from the major European leagues. Scottish J. Political Economy, v 54, n 3, p. 422–446, 2007.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. Ann. Statist., v 29, n 5, p. 1189-1232, 2001.

Fry, T. R. L., Galanos, G. e Posso, A. Let's get Messi - Top-scorer productivity in the European champions league. Scottish Journal of Political Economy, v 61, n 3, p. 261-279, 2014.

Garcia-del-Barrio, P. e Pujol, F. Hidden Monopsony Rents in Winner-take-all Markets - Sport and Economic Contribution of Spanish Soccer Players. Managerial and Decision Economics, v 28, n 1, p. 1-82, 2007.

Gazola, S. Construção de um modelo de regressão para avaliação de imóveis. Dissertação de Mestrado – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, UFSC, Florianópolis, SC, 2002.

George, N. Practical Data Science with Python. Editora Packt Publishing, 2021.

Gerhardt, T. E. e Silveira, D. T. Métodos de Pesquisa. 1. ed. Porto Alegre: Editora da UFRGS, 2009.

Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. Editora O'Reilly Media, 2019.

Gil, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2007.

Goodfellow, I., Bengio, Y., e Courville, A. Deep Learning. Editora MIT Press, 2016.

Guyon, I. e Elisseeff, A. An introduction to variable and feature selection. Journal of machine learning research, v 3, p. 1157-1182, 2003

Gyimesi, A. e Kehl, D. Relative age effect on the market value of elite European football players: a balanced sample approach. European Sport Management Quarterly, 2021.

Hastie, T., Tibshirani, R. e Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Editora Springer, 2009.

He, M., Cachucho, R. e Knobbe, A. Football player's performance and market value. 2nd workshop of sports analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), p. 1-9, 2015.

Herm, S., Callsen-Brack, H.-M. e Kreis, H. When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. Sport Manage. Rev., v 17, n 4, p. 484-492, 2014.

Huyen, C. Designing Machine Learning Systems. Editora O'Reilly Media, 2022.

Kaplan, T. When It Comes to Stats, Soccer Seldom Counts. Disponível em: https://www.nytimes.com/2010/07/09/sports/soccer/09soccerstats.html. Acesso em: 04/06/2022.

Kennard, R. W. e Stone, L. A. Computer aided design of experiments. Technometrics, v 11, n 1, p. 137-148, 1969.

Kiefer, S. The impact of the Euro 2012 on popularity and market value of football players. International Journal of Sport Finance, v 9, n 2, p. 95-110, 2014.

Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. Department of Computer Science and Technology, University of Peloponnese, Grécia, 2007.

Louppe, G., Wehenkel, L., Sutera, A., e Geurts, P. Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems, v 26, p. 431-439, 2013.

Lu, S. Measuring dependence via mutual information. Tese (Mestrado) - Departamento de Matemática e Estatística, Queen's University, Kingston, 2011.

Lucifora, C. e Simmons, R. Superstar Effects in Sport: Evidence From Italian Soccer. Journal of Sports Economics, v 4, n 1, p. 35-55, 2003.

Markovitch, S. e Rosenstein, D. Feature Generation Using General Constructor Functions. Machine Learning, v 49, n 1, p. 59–98, 2002.

Medcalfe, S. English league transfer prices: is there a racial dimension? A re-examination with new data. Applied Economics Letters, v 15, n 11, p. 865-867, 2008.

Montgomery, D. C., Peck, E. A. e Vining, G. G. Introduction to linear regression analysis. Editora John Wiley & Sons, 2012.

Müller, O., Simons, A. e Weinmann, M. Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, v. 263, n. 2, p. 611-624, 2017.

Nassis, G. P., Verhagen, E., Brito, J., Figueiredo, P. e Krustrup, P. A review of machine learning applications in soccer with an emphasis on injury risk. Biology of Sport, v. 40, n. 1, p. 233-239, 2023.

Pappalardo, L. e Cintia, P. Quantifying the relation between performance and success in soccer. Advances in Complex Systems, v 21, n 4, 2018.

Pappalardo, L., Cintia P., Ferragina P., Massucco E., Pedreschi D. e Giannotti F. PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Trans. Intell. Syst. Technol. (TIST), v. 10, n. 5, p. 1-27, 2019.

Rajer-Kanduč, K., Zupan, J. e Majcen, N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. Chemometrics and Intelligent Laboratory Systems, v 65, n 2, p. 221-229, 2003.

Reilly, B. e Witt, R. Disciplinary sanctions in English Premiership Football: Is there a racial dimension? Applied Economic Letters, v. 18, n. 3, p. 360-370, 1995.

Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F. M. e Baca, A. Machine learning application in soccer: A systematic review. Biology of Sport, v. 40, n. 1, p. 249-263, 2023.

Schneider, C. F. Machine learning aplicado na previsão de resultados de partidas de futebol: um estudo de caso para comparação de diferentes classificadores. Tese (Graduação) - Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

Seger, C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. Tese (Graduação) - School of Electrical Engineering and Computer Science (EECS), KTH Royal Institute of Technology, Estocolmo, 2018.

Shannon, C. E. A mathematical theory of communication. Bell System Technical Journal, v 27, n 1, p. 379–423, 623–656, 1948.

Silva, L. M. O. Da. Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais. Tese (Doutorado) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

Singh, P. e Lamba, P. S. Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. Journal of Discrete Mathematical Sciences and Cryptography, v 22, n 2, p. 113–126, 2019.

Winemiller, S., Love, A. e Stamm, J. Recruiting Reporters' Perceptions of Ethical Issues. Communication & Sport, v. 10, n. 3, p. 456-476, 2020.

Yu, L. e Liu, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, v. 5, p. 1205–1224, 2004.

Zhu, F., Lakhani, K. R., Schmidt, S. L. e Herman, K. TSG Hoffenheim: football in the age of analytics. Harvard Business School Case 616–010, 2015.

ANEXOS

ANEXO A – VARIÁVEIS DO CONJUNTO DE DADOS ESTATÍSTICOS DE EVENTOS DE PARTIDAS DE FUTEBOL

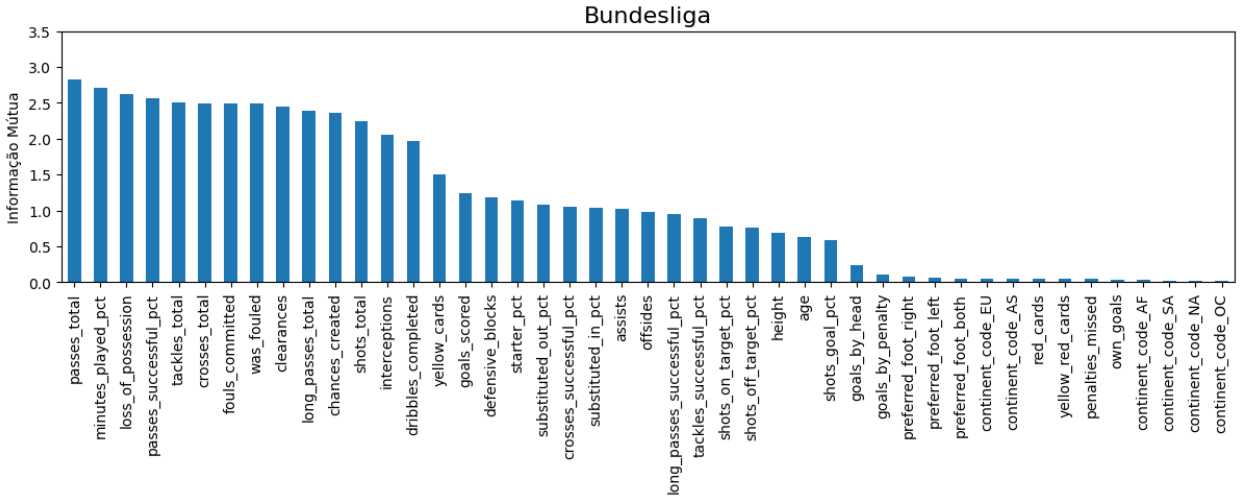
starter	Indicador se o jogador iniciou a partida como titular
assists	Número de contribuições de um jogador que ajuda diretamente um companheiro de equipe a marcar um gol
chances_created	Número de passes que levaram diretamente a um chute
clearances	Número de ações defensivas onde um jogador limpa a bola para longe de sua própria área de gol
crosses_successful	Número de cruzamentos bem-sucedidos, recebidos por algum companheiro de equipe
crosses_total	Número total de cruzamentos realizados
defensive_blocks	Número de bloqueios defensivos de chutes adversários
dribbles_completed	Número de ações para vencer um adversário onde o jogador com a bola avança para o território adversário
fouls_committed	Número de faltas cometidas pelo jogador ao utilizar meios não permitidos
goals_by_head	Número de gols marcados utilizando a cabeça
goals_by_penalty	Número de gols marcados de pênalti
goals_scored	Número de gols marcados pelo jogador
interceptions	Número de interceptações de passes feitas pelo jogador, que causa uma mudança de posse de bola
long_passes_successful	Número de passes longos bem-sucedidos feitos pelo jogador
long_passes_total	Número de passes longos totais feitos pelo jogador
long_passes_unsuccessful	Número de passes longos malsucedidos feitos pelo jogador
loss_of_possession	Número de posses de bola perdidas pelo jogador
minutes_played	Minutos jogados pelo jogador na partida
own_goals	Número de gols contra marcados pelo jogador

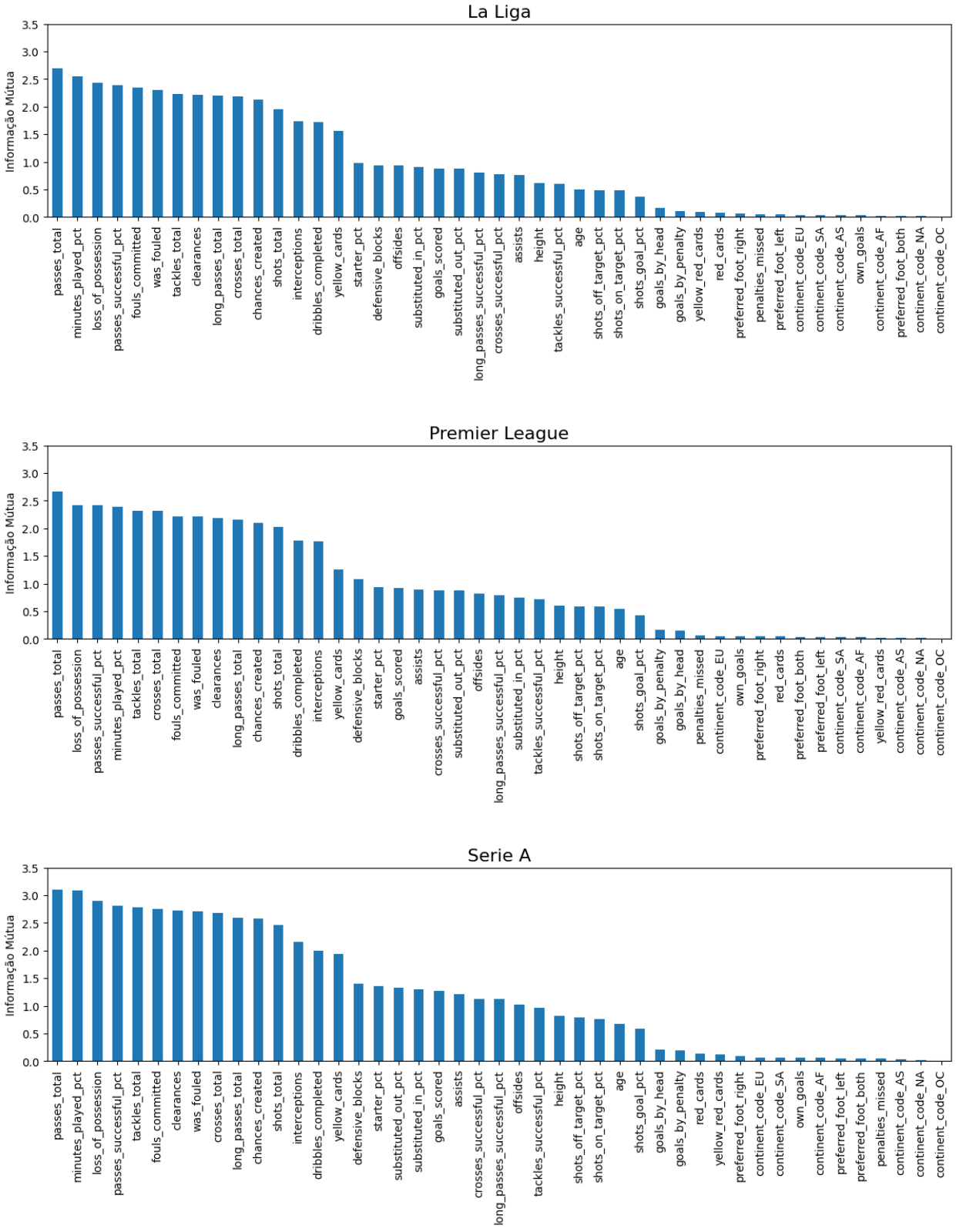
passes_successful	Número de passes bem-sucedidos feitos pelo jogador
passes_total	Número de passes totais feitos pelo jogador
passes_unsuccessful	Número de passes malsucedidos feitos pelo jogador
penalties_missed	Número de pênaltis não convertidos pelo jogador
red_cards	Número de cartões vermelhos recebidos pelo jogador
substituted_in	Indicador se o jogador entrou no decorrer da partida
substituted_out	Indicador se o jogador foi substituído no decorrer da partida
tackles_successful	Número de desarmes defensivos bem-sucedidos realizados pelo jogador
tackles_total	Número de desarmes defensivos totais realizados pelo jogador
was_fouled	Número de faltas recebidas pelo jogador
yellow_cards	Indicador se o jogador recebeu um primeiro cartão amarelo
yellow_red_cards	Indicador se o jogador recebeu um segundo cartão amarelo e, consequentemente, um cartão vermelho
offsides	Número de jogadas impedidas cometidas pelo jogador
shots_on_target	Número de finalizações em direção ao gol feitas pelo jogador
shots_off_target	Número de finalizações para fora do gol feitas pelo jogador
corner_kicks	Número de escanteios cobrados pelo jogador
shots_blocked	Número de finalizações bloqueadas feitas pelo jogador

ANEXO B – VARIÁVEIS DO CONJUNTO DE DADOS DE CARACTERÍSTICAS DOS JOGADORES DE FUTEBOL

position	Posição em campo que o jogador atua (defensor, meio-campista ou atacante)
country_code	Código do país de origem do jogador
height	Altura do jogador (em centímetros)
preferred_foot	Pé de preferência do jogador (esquerdo, direito ou ambos)

ANEXO C – INFORMAÇÃO MÚTUA ENTRE AS VARIÁVEIS INDEPENDENTES E A VARIÁVEL DEPENDENTE EM CADA CAMPEONATO PARA A POSIÇÃO MEIO-CAMPISTA




La Liga
Informação Mútua
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
passes_total
minutes_played_pct
loss_of_possession
passes_successful_pct
fouls_committed
was_fouled
tackles_total
clearances
long_passes_total
crosses_total
chances_created
shots_total
interceptions
dribbles_completed
yellow_cards
starter_pct
defensive_blocks
offsides
substituted_in_pct
goals_scored
substituted_out_pct
long_passes_successful_pct
crosses_successful_pct
assists
height
tackles_successful_pct
age
shots_off_target_pct
shots_on_target_pct
shots_goal_pct
goals_by_head
goals_by_penalty
yellow_red_cards
red_cards
preferred_foot_right
penalties_missed
preferred_foot_left
continent_code_EU
continent_code_SA
continent_code_AS
own_goals
continent_code_AF
preferred_foot_both
continent_code_NA
continent_code_OC
Premier League
Informação Mútua
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
passes_total
loss_of_possession
passes_successful_pct
minutes_played_pct
tackles_total
crosses_total
fouls_committed
was_fouled
clearances
long_passes_total
chances_created
shots_total
dribbles_completed
interceptions
yellow_cards
defensive_blocks
starter_pct
goals_scored
assists
crosses_successful_pct
substituted_out_pct
offsides
long_passes_successful_pct
substituted_in_pct
tackles_successful_pct
height
shots_off_target_pct
shots_on_target_pct
age
shots_goal_pct
goals_by_penalty
goals_by_head
penalties_missed
continent_code_EU
own_goals
preferred_foot_right
red_cards
preferred_foot_both
preferred_foot_left
continent_code_SA
continent_code_AF
yellow_red_cards
continent_code_AS
continent_code_NA
continent_code_OC
Serie A
Informação Mútua
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
passes_total
minutes_played_pct
loss_of_possession
passes_successful_pct
tackles_total
fouls_committed
clearances
was_fouled
crosses_total
long_passes_total
chances_created
shots_total
interceptions
dribbles_completed
yellow_cards
defensive_blocks
starter_pct
substituted_out_pct
substituted_in_pct
goals_scored
assists
crosses_successful_pct
long_passes_successful_pct
offsides
tackles_successful_pct
height
shots_off_target_pct
shots_on_target_pct
age
shots_goal_pct
goals_by_head
goals_by_penalty
red_cards
yellow_red_cards
preferred_foot_right
continent_code_EU
continent_code_SA
own_goals
continent_code_AF
preferred_foot_left
preferred_foot_both
penalties_missed
continent_code_AS
continent_code_NA
continent_code_OC