

Modelagem via árvore de decisão para previsão de jogos de futebol

Decision tree modeling for football game prediction

Modelo de árbol de decisión para la predicción de juegos de fútbol

Recebido: 19/07/2020 | Revisado: 06/08/2020 | Aceito: 10/08/2020 | Publicado: 16/08/2020

Adenilson Borba Lopes Silva

ORCID: <https://orcid.org/0000-0001-8019-8966>

Universidade Estadual da Paraíba, Brasil

E-mail: 123ninoia@gmail.com

Kleber Napoleão Nunes de Oliveira Barros

ORCID: <https://orcid.org/0000-0003-2515-3292>

Universidade Estadual da Paraíba, Brasil

E-mail: kleberbarros@cct.uepb.edu.br

Mácio Augusto de Albuquerque

ORCID: <https://orcid.org/0000-0002-0113-9130>

Universidade Estadual da Paraíba, Brasil

E-mail: marcioaa@uepb.edu.br

Resumo

Após avanço tecnológico a análise de dados voltada para fins esportivos se tornou de fundamental importância para evolução tática e obtenção de bons resultados. No futebol, a utilização dessas análises vem crescendo e trazendo inúmeros benefícios, tanto para o desenvolvimento tático, quanto na parte física dos atletas. Além da colaboração tática e técnica para o futebol, a estatística também é bastante utilizada em previsões, que abrange desde uma cobrança de pênalti até o resultado final do jogo. O Objetivo deste trabalho é encontrar um modelo para previsão de resultados de partidas de futebol. Mandante (Time Mandante sair vencedor) Empate ou Visitante (Time Visitante sair vencedor) usando o método de Árvore de decisão, onde, após modelagem dos dados e análise da precisão do modelo foi analisada qual casa seria mais rentável.

Palavras-chave: Futebol; Casa de apostas; Estatística; Árvore de decisão.

Abstract

After technological advances, data analysis for sports purposes has become of fundamental importance for tactical evolution and obtaining good results. In football, the use of these analyzes has been growing and bringing numerous benefits, both for the tactical development, as well as in the physical part of the athletes. In addition to tactical and technical collaboration for football, statistics are also widely used in predictions, ranging from a penalty kick to the final result of the game. The objective of this work is to find a model for predicting the results of soccer matches. Mandante (Mandante Team wins) Draw or Visitor (Visiting Team wins) using the Decision Tree method, where, after modeling the data and analyzing the accuracy of the model, which house would be more profitable was analyzed.

Keywords: Soccer; Bookmakers; Statistic; Decision tree.

Resumen

Después de los avances tecnológicos, el análisis de datos para fines deportivos se ha convertido en una importancia fundamental para la evolución táctica y la obtención de buenos resultados. En el fútbol, el uso de estos análisis ha ido creciendo y trayendo numerosos beneficios, tanto para el desarrollo táctico como para la parte física de los atletas. Además de la colaboración táctica y técnica para el fútbol, las estadísticas también se usan ampliamente en las predicciones, que van desde un tiro penal hasta el resultado final del juego. El objetivo de este trabajo es encontrar un modelo para predecir los resultados de los partidos de fútbol. Mandante (gana el equipo Mandante) Empate o visitante (gana el equipo visitante) utilizando el método del árbol de decisión, donde, después de modelar los datos y analizar la precisión del modelo, se analizó qué casa sería más rentable.

Palabras clave: Fútbol; Corredor de apuestas; Estadística; Árbol de decisión.

1. Introdução

O uso da análise de dados direcionada para o esporte vem crescendo a cada ano. Com ela se tem acesso a informações confiáveis que podem ser utilizadas para realizar uma série de verificações que podem trazer diversos benefícios para as mais diversas áreas esportivas. Além disso, fornece aos gestores, olheiros e empresários de atletas, importantes informações de análise temporal de desempenho e, com isso, a possibilidade para o atleta de ser contratado por um clube melhor ou ter uma proposta melhor de contrato.

No futebol, após o avanço da tecnologia, a utilização dessas análises vem crescendo e trazendo inúmeros benefícios tanto para o desenvolvimento tático, quanto para o físico (Afonso et al., 2020) dos atletas. Além da colaboração tática e técnica para o futebol, a estatística também é bastante utilizada em previsões, que abrange desde uma cobrança de pênalti até o resultado do jogo.

Uma área estatística que vem ganhando destaque nos esportes junto a análise de desempenho é a previsão de resultados. O futebol, como um dos esportes mais populares do mundo, movimenta grande parte dessas apostas. Apesar de ser um esporte coletivo e depender de diversos fatores como, por exemplo, expulsões, lesões e erros de arbitragem, é possível calcular a probabilidade do resultado, principalmente, analisando dados históricos de partidas anteriores (Sehnem & Frozza, 2019).

Visando o retorno financeiro, o número de bancas de aposta está aumentando. Existem apostas para os mais diversos tipos de fatores que envolvem uma partida de futebol. Por exemplo, número de escanteios na partida, total de chutes a gol, posse de bola, cartões amarelos e vermelhos, assim como, obviamente, o resultado da partida. Esses fatores podem ser preditos usando um modelo estatístico, de acordo com algumas características que vão depender do fator que está sendo observado. Através destes modelos, as casas de apostas conseguem o retorno financeiro através de taxas que são empregadas nas odds.

As odds são quantidades derivadas do cálculo das probabilidades de um evento. Elas são calculadas a partir de análises estatísticas feitas pelas casas de apostas e apontam a perspectiva do evento em questão. No ramo das apostas a Odd é a proporção referente ao quanto terá de retorno caso o evento que foi apostado venha a acontecer. A probabilidade aproximada de um evento usando a Odd é dada por $P = (1/Odd_a)$, em que P é a probabilidade, baseada na frequência dos resultados passados e Odd_a é a Odd referente ao evento A.

Por exemplo, se uma Odd para um jogo qualquer for de 1,77 para o mandante e uma determinada pessoa apostar R\$10,00 nela, se o resultado coincidir com o da Odd que foi apostada, a pessoa receberá R\$17,70 ($Odd \times Aposta$). No entanto, caso erre ela perderá o valor apostado. Ao fazer a soma de todas as probabilidades usando como base as odds que são geradas pelas casas de apostas, o valor sempre excederá 1. Isso acontece graças a margem, ou *juice*, da casa de apostas. Ou seja, é a comissão que a casa de apostas receberá. Essas margens variam de acordo com cada casa, porém, estima-se que na grande parte das casas de apostas esteja entre 3% e 6% (Oddsshark, 2018).

O presente trabalho tem como objetivo a predição de resultados de partidas de futebol a partir de odds geradas por 6 diferentes Banca de apostas e de algumas estatísticas

utilizando o método de Árvore de decisão (*decision tree*). Além disso, pretende-se analisar qual das casas de aposta é mais rentável para, logo após, testar diferentes situações, das quais o modelo final é testado em diferentes contextos para determinar o quão rentável seria o modelo.

2. Metodologia

No futebol, a evolução seja ela técnica ou tática, é fundamental para a obtenção de títulos, e para tanto é necessário identificar quais os pontos fortes e fracos das equipes. Uma das maneiras de identificar esses pontos é pelas estatísticas da partida. Carling et al. (2008) afirmam que o principal objetivo da análise de partidas é identificar os pontos fortes e fracos da equipe, que podem ser desenvolvidos e melhorados respectivamente. Da mesma forma, um treinador pode analisar o desempenho de uma equipe rival e usar os dados para identificar maneiras de combater os pontos fortes dessa equipe e explorar suas fraquezas.

Nos últimos anos, diversos trabalhos vêm sendo desenvolvidos com o intuito de analisar os diversos fatores que compõem uma partida de futebol com o objetivo de prever o principal fator do esporte, o número de gols, e conseqüentemente quem sai com a vitória. O futebol está cada vez mais aberto para mudanças baseadas estatísticas para grandes bases (Rein e Memmert, 2016).

Através da estatística que se muda a forma, as ideias preconcebidas, ditando novas normas, renovando práticas e revisando antigas táticas de jogo. São os números que permitem ver o jogo como nunca o vimos antes (Anderson & Sally, 2013).

Neste trabalho, ajustou-se o modelo de árvore de decisão para a predição do resultado (vitória do mandante, empate ou vitória do visitante) utilizando como principal fator as odds coletadas em um site voltado a apostas.

2.1 Árvore de Decisão

A árvore de decisão é um modelo representado graficamente por nós e ramos, parecidos com uma árvore, mas no sentido invertido (Han; Pei & Kamber, 2011). O nó raiz é o primeiro da árvore no topo da estrutura. Nós internos, incluindo o nó raiz, são nós de decisão. Cada um contém um teste sobre uma variável independente e os resultados desse teste formam os ramos das árvores.

Os nós folhas representam valores de predição para as variáveis dependentes ou

distribuições de probabilidade desses valores. O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição acurado ou descobrir a estrutura preditiva do problema (Breiman et al., 1984). Conforme Wilkinson (2004) afirma: existem dois tipos de árvores de decisão, as árvores de regressão que tem sua variável dependente de valores numéricos e as árvores de classificação no qual as variáveis dependente são categóricas.

2.2 Árvore de Regressão

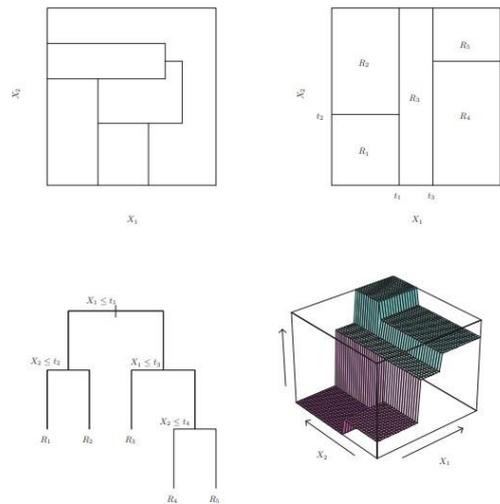
Conforme James et al.(2013) o processo para construção de uma árvore de regressão de maneira geral é feita em dois passos:

1. Particionar o espaço do preditor, ou seja, o conjunto de valores possíveis para, X_1, X_2, \dots, X_p , em J regiões distintas e não sobrepostas, R_1, R_2, \dots, R_p .
2. Para cada observação que cai na região R_j , obtêm-se a mesma previsão, que é simplesmente a média dos valores de resposta para o observações de treinamento em R_j .

Em teoria, as regiões podem ter qualquer forma. No entanto, ao dividir o espaço do preditor em retângulos ou caixas, simplifica e facilita a interpretação do modelo preditivo resultante.

O objetivo é encontrar as caixas R_1, \dots, R_J que minimizam a soma de quadrado do resíduo (SQR), fornecido por X_j (Morgan & Sonquist, 1963). A Figura 1 mostra alguns exemplos de divisão das regiões em retângulos.

Figura 1 – Exemplo de regiões em forma de retângulos.



Fonte: James et al. (2013).

Por exemplo, suponha que na Etapa 1 sejam obtidas duas regiões, R_1 e R_2 , e que a resposta média das observações de treinamento na primeira região seja 10, enquanto a média de resposta das observações de treinamento na segunda região é 20. Então, para uma dada observação $X = x$, se $x \in R_1$, será previsto um valor de 10 e, se $x \in R_2$, será previsto um valor de 20.

Após elaboração da Etapa 1, a construção das regiões da Etapa 2 podem assumir qualquer forma. No entanto, ao optar por dividir o espaço do preditor em retângulos de alta dimensão ou caixas, a simplificação facilita a interpretação dos resultados preditivos do modelo resultante. O objetivo é encontrar as caixas R_1, \dots, R_J que minimizam a SQR, dado por

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \quad (2.1)$$

em que \hat{y}_{R_j} é a resposta média para as observações de treinamento dentro do j -ésima caixa. Infelizmente, é inviável computacionalmente considerar todas as possíveis partições do espaço do recurso em J caixas (James et al., 2013).

Por esse motivo, realiza-se uma abordagem alternativa de cima para baixo, também conhecida como divisão binária recursiva. A abordagem é descendente, porque começa no topo da árvore (nesse ponto todas as observações pertencem a uma única região) e,

sucessivamente, divide o espaço preditor; cada divisão é indicada através de duas novas ramificações mais abaixo na árvore.

A cada passo do processo de construção de árvores, a melhor divisão é feita nessa etapa específica, em vez de olhar para o futuro e escolher uma divisão que levará a uma árvore melhor em alguma etapa futura.

Breiman (2017) realiza a divisão binária recursiva, primeiramente selecionando o preditor X_j e o ponto de corte que dividem o espaço do preditor nas regiões $\{X / X_j < s\}$ e $\{X / X_j \geq s\}$ que levam a máxima redução possível da SQR (2.1). A notação $\{X / X_j < s\}$ significa que a região do espaço preditor em que X_j assume um valor menor que s . Ou seja, considera-se que todos os preditores X_1, \dots, X_p e todos os valores possíveis dos pontos de corte s para cada um dos preditores e escolhe-se o preditor e os pontos de cortes de modo que a árvore resultante tenha a menor SQR. Em maiores detalhes, para qualquer j e s , defini-se o par de semiplanos,

$$\begin{aligned} R_1(j, s) &= \{X \mid X_j < s\} \\ R_2(j, s) &= \{X \mid X_j \geq s\} \end{aligned} \quad (2.2)$$

e busca-se o valor de j e s que minimizem a expressão

$$\sum_{i: x_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 - \sum_{i: x_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2, \quad (2.3)$$

em que \hat{y}_{R_1} é a resposta média para as observações de treinamento em $R_1(j, s)$, e \hat{y}_{R_2} é a resposta média para as observações de treinamento em $R_2(j, s)$. A localização dos valores de j e s que minimizam (2.3) pode ser feita rapidamente, especialmente quando o número de variáveis independentes p não é grande.

Em seguida, repete-se o processo, procurando o melhor preditor e o melhor ponto de corte para dividir ainda mais os dados, a fim de minimizar a SQR dentro de cada uma das regiões resultantes. Entretanto, desta vez, em vez de dividir todo o espaço preditivo, dividi-se uma das duas regiões identificadas anteriormente. Agora, têm-se três regiões.

Após isto, procura-se dividir uma dessas três regiões, para minimizar a SQR. O processo continua até um critério de parada ser atingido; por exemplo, pode-se continuar até que nenhuma região contenha mais de cinco observações.

Uma vez que as regiões R_1, \dots, R_J do tipo (2.2) foram criadas, estima-se a resposta para uma determinada observação de teste usando a média das observações de treinamento a região à qual essa observação de teste pertence.

2.3 Podas das Árvores

O processo de Árvore de Regressão pode produzir boas previsões sobre os dados de treinamento, porém provavelmente superestima os dados, levando a um desempenho ruim do conjunto de testes. Isso ocorre porque a árvore resultante pode ser muito complexa. Uma árvore menor com menos divisões (ou seja, menos regiões R_1, \dots, R_J) pode levar a menor variação e melhor interpretação ao custo de um pequeno viés (Monard & Baranauskas, 2003).

Uma possível alternativa é construir a árvore apenas por determinado tempo, pois a diminuição na SQR devido a cada divisão excede algum limite. Essa estratégia resultará em árvores menores, mas é arriscado demais desde que uma divisão aparentemente inútil no início da árvore pode ser seguida por uma boa divisão, ou seja, uma divisão que leva a uma grande redução na SQR posteriormente (Breiman, 2017).

Portanto, uma estratégia melhor é cultivar uma árvore T_0 muito grande e depois podá-la de volta para obter uma sub-árvore. Para determinar a melhor poda, intuitivamente, seleciona-se uma sub-árvore que leva à menor taxa de erro de teste. Dada uma sub-árvore, pode-se estimar seu erro de teste usando a validação cruzada ou a abordagem do conjunto de validação. Contudo, estimar o erro de validação cruzada para cada sub-árvore possível seria muito complicado, uma vez que existe um número extremamente grande de sub-árvores possíveis. Weiss & Indurkha (1994) mostra que a poda de ligação mais fraca (*weakest link pruning*) fornece uma maneira de fazer exatamente isso.

Na poda de ligação mais fraca em vez de considerar todas as sub-árvores possíveis, considera uma sequência de árvores indexadas por um parâmetro de ajuste não negativo α , para cada valor de α corresponde uma sub-árvore $T \subset T_0$ tal que

$$\sum_{j=1}^{|T|} \sum_{i: x_i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T|, \quad (2.4)$$

é o menor possível. $|T|$ representa o número de nós do terminal da árvore T , R_j é o retângulo (isto é, o subconjunto do espaço do preditor) correspondente ao m -ésimo nó terminal e

$y_i - \hat{y}_{RJ}$ é a resposta prevista associada a R_j , ou seja, a média das observações de treinamento em R_j . O parâmetro de ajuste α controla uma troca entre a complexidade da sub-árvore e sua adequação aos dados de treinamento. Quando $\alpha = 0$, a sub-árvore T simplesmente será igual a T_0 , porque então (2.4) apenas mede o erro de treinamento (Breiman et al., 1984).

No entanto, à medida que α aumenta, há um preço a pagar por ter uma árvore com muitos nós terminais e, portanto, a quantidade (2.4) tenderá a ser minimizada para uma sub-árvore menor. Acontece que, à medida que aumenta α de zero em (2.4), os nós terminais da árvore são podados de maneira aninhada e previsível, obtendo assim toda a sequência de sub-árvores em função de α . É propício selecionar um valor de α usando um conjunto de validação cruzada. Voltamos então ao conjunto de dados completo e obtendo-se a sub-árvore correspondente a α , Breiman (2017) resume esse processo assim:

3. Use a divisão binária recursiva para cultivar uma grande árvore no treinamento dos dados, parando apenas quando cada nó do terminal possui menos do que algum número mínimo de observações.
4. Aplique a poda da complexidade de custos à árvore grande para obter uma sequência das melhores sub-árvores, em função de α .
5. Use a validação cruzada com dobra K para escolher α . Ou seja, divida o treinamento em K dobras. Para cada $k = 1, \dots, K$:
 - a) Repita as etapas 1 e 2 em todas, exceto na k -ésima posição dos dados de treinamento.
 - b) Avalie o erro médio de previsão ao quadrado nos dados na k -ésima dobra à esquerda, em função de α .
 - c) Encontre a média dos resultados para cada valor de α e escolha o α que minimiza o erro médio.
6. Retorne a sub-árvore da Etapa 2 que corresponde ao valor escolhido de α .

2.4 Árvore de Classificação

Uma árvore de classificação é muito semelhante a uma árvore de regressão, exceto que é usada para prever uma resposta qualitativa em vez de quantitativa. O modelo de árvore de classificação pertence à categoria dos métodos de aprendizado de máquina por classificação supervisionada, como o modelo de regressão logístico (Albuquerque et al., 2020) ou redes

neurais (Vogado, 2019), em que se deseja mapear uma variável dependente categórica y (e.g., cães ou gatos; homens ou mulheres; baixa renda, classe média ou alta) com um conjunto de características ou variáveis independentes X . Para exemplo de modelo de classificação não supervisionada, consultar (Barros, et al. 2020).

Lembre-se de que para uma árvore de regressão, a resposta prevista para uma observação é dada pela resposta média das observações de treinamento que pertencem ao mesmo nó terminal. Por outro lado, para uma árvore de classificação, estima-se que cada observação pertença à classe de treinamento mais comum entre as observações na região a qual à pertence (Grochtmann & Grimm, 1993).

Na interpretação dos resultados de uma árvore de classificação, geralmente se está interessado não apenas na previsão da classe correspondente a uma região de nó terminal específica, mas também nas proporções entre as observações de treinamento que se enquadram nessa região (Loh, 2011).

A tarefa de cultivar uma árvore de classificação é bastante semelhante à tarefa de cultivar uma árvore de regressão. Assim como na configuração de regressão, utilizam-se recursos da divisão binária para aumentar uma árvore de classificação. No entanto, na configuração da árvore de classificação, a SQR não pode ser usada como critério para fazer as divisões binárias.

Uma alternativa natural a SQR é a taxa de erro de classificação. Como se planeja atribuir uma observação em uma determinada região à classe de observações de treinamento mais comum nessa região, a taxa de erro de classificação é simplesmente a fração das observações de treinamento naquela região que não pertencem à classe mais comum:

$$E = 1 - \max_k(\hat{p}_{mk}), \quad (2.5)$$

em que \hat{p}_{mk} representa a proporção de observações de treinamento na m -ésima região que pertencem à k -ésima classe. No entanto, verifica-se que o erro de classificação não é suficientemente sensível para o cultivo de árvores e, na prática, outras duas medidas são preferíveis (James et al., 2013).

Breiman et al. (1984) afirmam que o índice Gini é uma medida da variação total entre as classes K e é definido por

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}). \quad (2.6)$$

Não é difícil ver que o índice Gini assume um pequeno valor se todos os \hat{p}_{mk} estiverem próximos de zero ou um. Por esse motivo, o índice de Gini é referido como uma medida de pureza do nó, um pequeno valor indica que um nó contém predominantemente observações de uma única classe. Ou seja, um nó puro é um nó que tem predominância total de uma única classe. Moisen (2008) mostra que uma alternativa ao índice de Gini é a entropia cruzada, que é dada por

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk}). \quad (2.7)$$

Como $0 \leq \hat{p}_{mk} \leq 1$, segue-se que $0 \leq -\hat{p}_{mk} \log(\hat{p}_{mk})$. Pode-se mostrar que a entropia cruzada assumirá um valor próximo de zero se os \hat{p}_{mk} estiverem todos próximos de zero ou próximo a um. Portanto, como o índice de Gini, a entropia cruzada levará em um valor pequeno se o m-ésimo nó for puro. De fato, acontece que o índice Gini e a entropia cruzada são bastante semelhantes numericamente.

Ao construir uma árvore de classificação, o Índice de Gini ou a entropia normalmente são para avaliar a qualidade de uma determinada divisão, uma vez que essas duas abordagens são mais sensíveis à pureza do nó do que a taxa de erro de classificação (Rokach & Maimon, 2005). Qualquer uma das três abordagens pode ser usada quando se poda uma árvore, mas se a previsão da árvore final após a poda for o principal objetivo, a taxa de classificação é a preferível (James et al., 2013).

Neste trabalho foi adotada a metodologia da poda da árvore de decisão por taxa de classificação, conforme James et al. (2013).

2.5 Métodos de reamostragens

Os métodos de reamostragem são ferramentas indispensáveis na estatística moderna. Eles envolvem repetidamente tirar amostras de um conjunto de treinamento e reajustar um modelo. As abordagens de reamostragem eram computacionalmente custosas, porque elas envolvem o ajuste do mesmo método estatístico milhares de vezes usando diferentes subconjuntos dos dados de treinamento. No entanto, devido aos recentes avanços na computação, os requisitos computacionais dos métodos de reamostragem geralmente não são

mais um problema proibitivo. Os métodos mais comuns de reamostragem são Validação Cruzada (CV) e Bootstrap (James et al.,2013).

Ambos os métodos são ferramentas importantes na aplicação prática de muitos procedimentos de aprendizagens estatísticas. Por exemplo, a validação cruzada pode ser usada para estimar o erro associado a um determinado método estatístico de aprendizagem para avaliar seu desempenho ou para selecionar o nível apropriado de flexibilidade.

De acordo com (Burman,1989) o método K-Fold de Validação Cruzada (CV) consistem em dividir o conjunto de observações em k grupos, ou dobras, de preferência com tamanhos iguais. Dos k grupos, um único grupo é retido como dados de validação para testar o modelo, e os $k - 1$ grupos restantes são usados como dados de treinamento. Esse processo é repetido k vezes, com cada um dos k grupos usados exatamente uma vez como dados de validação, em cada um desses processos é calculado o erro quadrado médio (EQM). Esse processo resulta em k estimativas do erro de teste, $EQM_1, EQM_2, \dots, EQM_k$. A estimativa da CV da dobra k é a por

$$CV_k = \frac{1}{k} \sum_{i=1}^k EQM_i, \quad (2.8)$$

no qual, quando a abordagem vai para o cenário de classificação a validação cruzada funciona exatamente como descrito anteriormente, exceto que em vez de usar o EQM para quantificar o erro de teste, usamos o número de observações classificadas incorretamente (James et al., 2013).

A validação cruzada não é capaz de aumentar a acurácia das estimativas, porém permite realizar uma média amostral (2.8) dos erros quadráticos médios (EQM), aumentando a certeza sobre o EQM.

3. Metodologia

Conforme Pereira et al. (2018), este estudo se apresenta como trabalho descritivo de ênfase no método quantitativo.

O banco de dados é composto por 36 variáveis coletadas de 9 diferentes Campeonatos de futebol: Brasileirão Serie A, Superliga Argentina de Futebol (SAF), Primeira Liga de Portugal, Série A Italiana, LaLiga da Espanha, Eredivise da Holanda, Premier League da Inglaterra, Bundesliga da Alemanha e Ligue 1 da França. A coleta foi entre o período de 13 de

Agosto à 28 de Outubro de 2019 no site www.resultados.com. Das 36 variáveis que compõem o banco, 18 são odds de 5 diferentes casas de apostas e uma bolsa de apostas, as odds coletadas são referentes as probabilidades de vitória do time mandante, vitória do time visitante e empate. As 5 casas de apostas que tiveram suas odds coletadas foram Bet365, 1XBet, BetSon, Superapostas e UniBet, a bolsa de apostas que teve suas odds coletadas foi a BETFair e que a partir daqui serão chamadas de casa A, B, C, D, E e F, não necessariamente nessa ordem.

As demais variáveis são: time mandante, time visitante, posição do time mandante, posição do time visitante, posse de bola do mandante, posse de bola do visitante, total de chutes do mandante, total de chutes do visitante, total de chutes certos do mandante, total de chutes certos do visitante, número de cartões amarelos e vermelhos do mandante, número de cartões amarelos e vermelhos do visitante, gols do mandante na partida, gols do visitante na partida, resultado da partida. Para alguns jogos não se tinha as odds de todas as casas de apostas, estes valores foram imputados pelas médias das odds das outras casas.

Os Softwares utilizados para a construção desenvolvimento do trabalho foram:

1. Para armazenagem de dados: Microsoft Office Excel versão 2019 (16.0), que é um editor de planilhas produzido pela Microsoft para computadores que utilizam o sistema operacional Microsoft Windows, utilizado em computadores e dispositivos móveis para armazenamento e algumas análises.
2. Para as análises estatísticas o R versão 3.6.1 (05 de julho de 2018), ambiente computacional e também uma linguagem de programação que vem progressivamente se especializando em manipulação, análise e visualização gráfica de dados com ambiente disponível para diferentes sistemas operacionais: Unix/Linux, Mac e Windows. Junto ao RStudio, um software livre de ambiente de desenvolvimento integrado para R.

4. Resultados e Discussão

Após coleta dos dados foram feitas as análises descritivas e objetivas das 36 variáveis. Inicialmente, será feita uma análise descritiva de algumas variáveis, após análise descritiva, será feito a árvores de decisão para assim chegar as conclusões e discussão dos resultados.

4.1 Análise descritiva

Inicialmente, foi realizada uma contagem da quantidade de jogos que cada campeonato teve durante o período de coleta de dados, que foi descrito na Tabela 1. Ao observamos a Tabela 1, verifica-se que a maioria dos campeonatos está próximo dos 60 jogos, com exceção do campeonato Brasileiro, que tem 110 jogos.

Tabela 1 – Número de jogos dos campeonatos de 2019.

Campeonato	Números de jogos	Proporção
Alemão	53	8,83%
Argentino	83	13,33%
Brasileiro	110	18,33%
Espanhol	49	8,16%
Francês	67	11,16%
Holandês	58	9,66%
Inglês	68	11,33%
Italiano	58	9,66%
Português	54	9,00%
Total	600	100%

Fonte: Autores.

Isso aconteceu devido ao tempo que foi coletado os dados, o campeonato brasileiro estava entrando em sua reta final (tinha dois jogos por semana), já os outros campeonatos (que seguem o calendário europeu) estavam em início de temporada. As análises referentes as estatísticas dos jogos (Posse de bola, total de chutes, finalizações certas, cartão amarelo, cartão vermelho, número de gols marcados) estão representadas na Tabela 2.

Tabela 2 – Análise descritiva das estatísticas dos jogos.

Variável	Média	Mínimo	Máximo	Variância (Desvio Padrão)
Posse bola M	51,12	11	81	111,54 (10,56)
Posse bola V	48,75	19	89	111,85 (10,57)
Total Chutes M	10,96	0	28	17,27 (4,20)
Total de Chutes V	8,65	1	21	2113,35 (3,65)
Finalizações Certas M	4,90	0	14	6,54 (2,55)
Finalizações Certas V	3,84	0	15	4,84 (2,20)
Cartão Amarelo M	2,05	0	6	1,86 (1,36)
Cartão Amarelo V	2,40	0	8	2,24 (1,49)
Cartão Vermelho M	0,09	0	2	0,109 (0,33)
Cartão Vermelho V	0,152	0	2	0,14 (0,38)
Gols Marcados M	1,42	0	9	1,44 (1,21)
Gols Marcados V	1,09	0	5	1,19 (1,09)

Fonte: Autores.

Nela observa-se que a posse de bola para o time Mandante (M) teve média de 51,12% e variância 111,54 e média de 48,74% e variância de 111,85 para os times Visitantes (V), a posse de bola mínima entre as observações foi de 11% e a máxima foi de 89%. Para a variável total de chutes do mandante e visitante as respectivas médias e variâncias foram de 10,96 e 8,65 para a média e 17,27 e 2113,35 as variâncias, onde o máximo e mínimo alcançados por essa variável foram 21 e 0. A variável finalizações certas teve média de 4,90 e 3,84 para mandante e visitante respectivamente, nenhuma das variáveis teve variância maior 7 e os valores máximo e mínimo atingido por esta variável foi 15 e 0. A média para cartões amarelos foi de 2,05 e 2,40 com variância de 1,86 e 2,24 para mandantes e visitantes respectivamente, os cartões vermelhos tiveram as médias de 0,09 para mandante e 0,152 para visitante, as variâncias foram de 0,109 para mandante e 0,14 para visitante. Já para o número de gols marcados os mandantes tiveram a maior média com 1,42 contra 1,09 dos visitantes, a variância foi de 1,44 para os mandantes e 1,19 para os visitantes, o maior número de gols em uma partida foi de 9 gol e o menor foi 0.

Com as variáveis referentes às odds, foi feita uma análise mais detalhada, essas análises estão representadas nas Tabelas 3 e 4. A Tabela 3 é dividida por resultado (mandante, empate e visitante), no qual foi analisado a média a variância e o 1º Quartil e 3º Quartil das 6 casas de apostas.

Na Tabela 3 são encontradas as principais estatísticas para as odds das 6 casas agregadas por vitória do mandante, empate ou vitória do visitante.

Tabela 3 – Média, 1º Quartil, 3º Quartil, variância das odds referentes as 6 casas de apostas.

Casa de aposta	Resultado	Média	Variância	1º Quartil	3º	Quartil
A	M	2,73	5,84	1,65		2,90
B	M	2,20	4,96	1,69		2,98
C	M	2,70	5,52	1,67		2,88
D	M	2,73	6,73	1,68		2,85
E	M	2,85	7,83	1,70		2,84
F	M	2,69	5,11	1,65		2,85
Casa de aposta	Resultado	Média	Variância	1º Quartil	3º	Quartil
A	E	3,91	1,79	3,20		4,00
B	E	4,03	2,04	3,26		4,14
C	E	3,89	2,26	3,10		4,00
D	E	3,98	2,19	3,15		4,15
E	E	4,08	2,13	3,26		4,25
F	E	3,90	1,90	3,15		4,00
Casa de aposta	Resultado	Média	Variância	1º Quartil	3º	Quartil
A	V	4,60	13,58	2,50		5,25
B	V	4,66	15,38	2,54		5,40
C	V	4,71	19,02	2,50		5,50
D	V	4,54	15,32	2,50		5,20
E	V	4,85	19,54	2,60		5,51
F	V	4,49	13,74	2,50		5,10

Fonte: Autores.

Para os mandantes a casa de aposta que teve melhor média foi a E com 2,85 e a que teve a média mais baixa foi B com 2,20, ou seja, em média a casa mais rentável para se apostar no time mandante é a casa E, porém, ela é casa que teve a maior variância, ao observamos os 1º e 3º Quartis percebe se que as casas que tiveram melhor intervalo de confiança (considerando o maior 3º Quartil) foi a B com intervalo de [1,69 ; 2,98] seguido da casa de apostas A, C, D, F e E. Para os Empates, a classificação da maior para a menor médias foi: E, B, D, A, F e C com médias 4,08; 4,03; 3,98; 3,91; 3,90 e 3,89 respectivamente, os intervalos de confianças foram bem próximos uns dos outros, dado que a variância ficou entre 1,79 (na casa A) e 2,26 (na casa C). Já nos visitantes a variância foi relativamente alta comparando com os dois resultados anteriores, a maior média ficou para casa E (4,85) enquanto a pior (4,49) ficou com a casa F, as variâncias entre as casas foi de 13,74 (casa F) até 19,02 (casa C). Com relação a média ser representativa ($Média > 2 \times Desvio\ Padrão$) nos mandantes nenhuma das médias foram representativas. Para empate, todas as médias foram representativas, já para os visitantes que foi a categoria que teve a maior variância no geral nenhuma das casas teve média representativa.

A Tabela 4 mostra algumas estatísticas referentes ao comportamento da média e variância também separadas por classes, mandante, visitante, empate e geral. Nela, observa-se o quanto variam as médias das odds para mandante, empate e visitante, e também tem as informações da menor e maior média das odds das 6 casas de apostas para cada um dos 3 casos, 1º e 3º Quartil das médias, variâncias e também a média e variância geral.

Tabela 4 – Análise estatística da média e variância das odds para os mandantes, visitantes, empate e geral.

Variável	Média	Variância	Mínimo	Máximo	Quartil	Quartil
Média M	2,74	5,48	1,063	32,33	1,66	2,90
Média E	3,96	2,061	2,832	13,95	3,18	4,09
Média V	4,64	15,38	1,053	31,00	2,53	5,34
Variância M	0,6441	76,73	0,000017	150,06	0,001	0,008
Variância E	0,0373	0,013	0,001	1,24	0,006	0,018
Variância V	0,8976	136,83	0,000067	278,61	0,003	0,070
Média G	3,78	8,25	1,053	32,33	2,37	4,08
Variância G	0,52	71,24	0,000017	278,81	0,0026	0,024

Fonte: Autores.

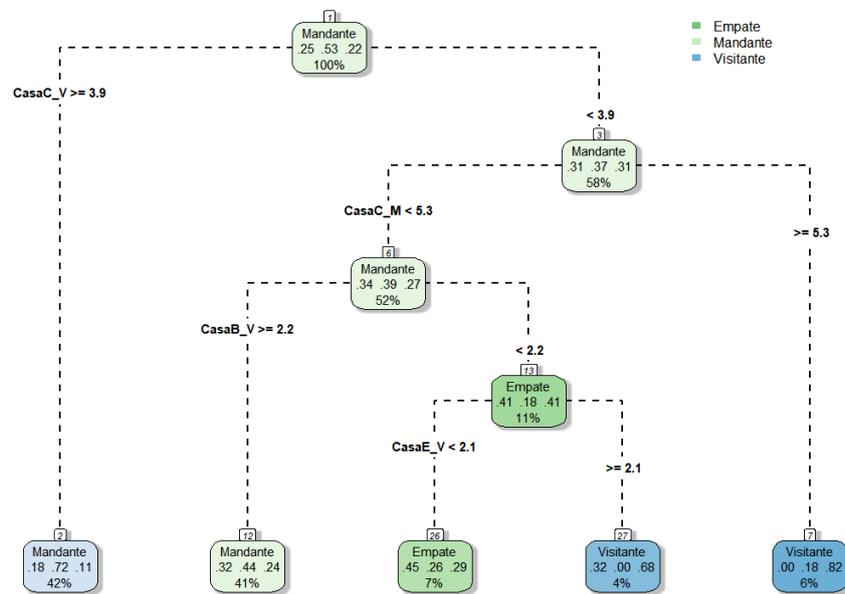
Comparando os valores da Tabela 3 com a Tabela 4, percebe-se que apenas a casa E está acima da média geral para os Mandantes, para Empates 3 casas (B, D e E) estão acima e as outras 3 (A, C e F) abaixo, para visitantes as casas B, C e E ficaram acima da média e as casas A, D e F ficaram abaixo. Analisando a variância entre a casas de apostas por jogo, verifica-se que o empate tem a menor média de variação (0,0373) entre as odds, seguido do mandante (0,6441) e por fim os visitantes (0,8976) e analisando de uma forma geral a média foi de 3,78 e variância 8,25 onde a menor média foi 1,053 encontrada nos visitantes e a maior foi 32,33 encontradas no mandante, já a variância teve uma média geral de 0,52, no qual a menor variância entre as casas de apostas em um jogo foi de 0,000017 enquanto a maior foi de 278,81.

4.2 Árvore de decisão

Após as análises descritivas dos dados, foram criados diferentes modelos. Utilizando o método de validação cruzada, o modelo final e de melhor precisão foi o modelo representado pela árvore de classificação da Figura 2 com probabilidades a posteriori de 0,25; 0,53 e 0,22 para empate, mandante e visitante, respectivamente. As probabilidades dentro do gráfico devem ser lidas na ordem alfabética (empate, mandante e visitante). O modelo final foi

obtido com as 480 observações de treino (80%) e verificado nas 120 observações de teste (20%).

Figura 2 – Árvore de Decisão e Probabilidades.



Fonte: Autores.

Observa-se na figura a árvore de decisão e as suas respectivas conclusões. As decisões devem ser tomadas em base nas folhas das árvores (nós finais):

- Se a odd do visitante da casa de apostas C for maior ou igual que 3,9, deve-se apostar na vitória do mandante com 72% de acerto;
- Se a odd do visitante da casa de apostas C for menor que 3,9 e a odd do mandante da casa de apostas C for menor que 5,3 e a odd do visitante da casa de apostas B for maior ou igual a 2,2, deve-se apostar no mandante com 44% de acerto;
- Se a odd do visitante da casa de apostas C for menor que 3,9 e a odd do mandante da casa de apostas C for menor que 5,3 e a odd do visitante da casa de apostas B for menor que 2,2 e a odd do visitante da casa de apostas E for menor que 2,1, deve-se apostar no empate com 45% de acerto.
- Se a odd do visitante da casa de apostas C for menor que 3,9 e a odd do mandante da casa de apostas C for menor que 5,3 e a odd do visitante da casa

de apostas B for menor que 2,2 e a odd do visitante da casa de apostas E for maior ou igual a 2,1, deve-se apostar no visitante com 68% de acerto.

- Se a odd do visitante da casa de apostas C for menor que 3,9 e a odd do mandante da casa de apostas C for maior ou igual a 5,3, deve-se apostar no visitante com 82% de acerto.

A Tabela 5 mostra os resultados dos valores ajustados versus valores observados no conjunto de teste, a partir da árvore de decisão da Figura 2.

Tabela 5 – Tabela de acertos da Árvores Teste.

Ajustados\Observado	Mandante	Empate	Visitante	total
Mandante	57	22	16	95
Empate	5	4	2	11
Visitante	2	4	8	14
Total	64	30	26	120

Fonte: Autores.

A diagonal principal da tabela apresenta os acertos, e as demais combinações linha x coluna são os erros. Ao realizar a contagem, obtém-se que a árvore de teste acertou 57 dos 95 jogos o resultado estimado foi mandante (60%), 4 dos 11 jogos que foram estimados como empate (36,36%), 8 dos 14 jogos que o visitante venceu (57,14%), assim acertando num total de 69 de 120 jogos (57,5%). Observando com relação aos resultados observados observa-se que dos jogos que realmente o mandante foi o vencedor o modelo acertou 57 dos 64 jogos (89,06%), 4 dos 30 jogos que foi empates (13,33%) e 8 dos 26 jogos que o visitante saiu com o resultado (30,77%). Com relação a importância das variáveis dentro do modelo a Tabela 6 mostra as variáveis seguidas de suas respectivas importâncias.

Tabela 6 – Importância das variáveis em %.

Variável	Resultado	Importância
Casa E	V	100.000%
Casa C	V	95.656%
Casa D	V	95.363%
Casa B	M	74.140%
Casa F	V	73.738%
Casa F	M	49.590%
Casa D	M	47.677%
Casa E	M	39.709%
Casa C	M	38.064%
Casa A	M	35.977%
Casa B	V	16.576%
Casa D	E	10.189%
Casa A	E	9.828%
Pos mandante	-	0.000%
Casa F	E	0.000%
Pos visitante	-	0.000%
Casa C	E	0.000%
Casa E	E	0.000%
Casa B	E	0.000%
Casa A	V	0.000%

Fonte: Autores.

Na Tabela 6, nota-se que das 5 variáveis mais importantes 4 são de visitantes, dos visitantes apenas a Casa A não foi significativa, outro ponto relevante é que para esse modelo a variável posição na tabela tanto mandante quanto visitante não tiveram importância junto a quatro variáveis referentes a empate (apenas casa A e casa D tem significância).

4.3 Aplicação: Simulação de retorno de apostas

Após escolha do modelo e analisar sua precisão, foram criadas situações fictícias para testar o quão seria vantajoso financeiramente a precisão desse modelo e qual seria a casa de aposta a dar mais lucros, também apostando sempre na melhor odd entre as casas (representado como melhor retorno). Foram testados alguns cenários, onde foi alterado o valor que o apostador fictício teria para investir, e o quanto ele apostaria por jogo. A Tabela 7 mostra os resultados para apostas fixas (investimento/120 jogos) para os 120 jogos simulados, onde os investimentos iniciais foram de R\$120,00, R\$240,00, R\$480,00 e R\$600,00.

Tabela 7 – Tabela retorno financeiro da Árvores Teste.

Casa\Investimento	R\$ 120	R\$ 240	R\$ 480	R\$ 600	Lucro (%)
A	136,85	273,70	547,40	684,25	14,04%
B	146,36	280,70	561,44	701,80	16,96%
C	136,58	273,20	546,30	682,90	13,8%
D	138,07	276,14	552,30	690,30	15,1%
E	141,09	282,18	564,30	705,40	17,6%
F	138,45	276,90	553,80	692,30	15,37%
Melhor Retorno	142,52	285,04	570,10	712,60	18,76%

Fonte: Autores.

Ao analisar a Tabela 7, observa-se que em todas as casas obteve-se lucro. O lucro que variou de 13,8% até 17,6%. A casa de aposta que deu maior lucro foi a casa E e a que menos deu lucro foi a casa C. O modelo de apostas utilizando as melhores odds obteve um lucro de 18,76%. Os outros cenários testados foram:

1. O apostador inicia com R\$100,00 e sempre apostava um mesmo valor em todos os jogos. Os valores testados (formas de aposta) foram R\$25,00, R\$50,00, R\$75,00 e R\$100,00.
2. O apostador inicia com R\$200,00 e sempre apostava uma proporção fixa de seu Saldo em cada jogo. As proporções testadas (formas de aposta) foram de 1/5, 1/4, 1/3 e 1/2.
3. O apostador inicia com R\$200,00 e sempre apostava um mesmo valor, porém, apenas nos jogos que o modelo estimava que o time Mandante seria o vencedor. Os valores testados (formas de aposta) foram R\$25,00, R\$50,00, R\$75,00 e R\$100,00.

Utilizando as melhores odds entre as 6 casas, a Tabela 8 mostra como ficou o saldo final nas combinações de cada contexto com as 4 ocasiões que foram propostas.

Tabela 8 – Tabela Cenários × Formas de aposta.

Cenário \ Formas de aposta	1	2	3	4
1	663,00	1226,00	1789,00	2352,00
2	816,77	468,18	81,34	0,07
3	723,00	1246,00	1769,00	2292,00

Fonte: Autores.

A Tabela 8 mostra os lucros das possíveis combinações dos cenários e formas de apostas. No cenário 1 para as 4 formas de aposta, o saldo final foi de R\$663,00, R\$1226,00, R\$1789,00 e R\$2352,00, respectivamente. Porém ao analisar separadamente o

comportamento do saldo, na forma de aposta 1 o saldo chegou a ser R\$820,50 e teve um valor mínimo de R\$106,75, para a forma de aposta 2 a variação foi entre R\$1541,00 e R\$113,50, para a 3 foi R\$2261,50 e R\$120,25 e por fim, na forma 4 o saldo variou entre R\$2982,00 e R\$127,00. No cenário 2, percebe-se que nas formas de aposta 3 e 4 o apostador saiu no prejuízo, o máximo e mínimo que o saldo atingiu nesses dois contextos foi de R\$1600,81 e R\$34,08 na forma de aposta 3 e de R\$578,12 e R\$0,05 para a forma de aposta 4, já nas formas 1 e 2, teve-se um bom lucro de 308% e 134,09%, respectivamente. Já no cenário 3 que foi apostar apenas nos jogos que o modelo estimou como mandante (categoria que o modelo teve melhor precisão), obteve-se lucros em todas as formas, o lucro cresce a medida que o valor apostado por jogo aumenta, para essas formas de aposta específicas os valores do saldo final foi de R\$723,00, R\$1246,00, R\$1769,00 e R\$2292,00.

5. Considerações Finais

Junto ao avanço computacional a previsão de resultados voltada para o ramo de aposta de partidas de futebol vem ganhando destaque no ramo da estatística e computação. Com essas previsões em alta e com a alta popularidade do futebol essa área vem ganhando muita ênfase, dado que a quantia que geralmente está envolvida nesses jogos é gigante. Visando essa área de previsão de resultados, o presente estudo teve como objetivo a criação de um modelo a partir das odds geradas pelas próprias bancas de apostas para assim prever jogos, comparar qual das casas de apostas seria mais vantajoso apostar e analisar o quão lucrativo pode ser este modelo.

Após coleta, armazenamento e análise dos dados, foi criado um modelo teste (a partir das 480 observações iniciais) de árvore de decisão utilizando o método de validação cruzada com probabilidades a priores de 0,53, 0,22 e 0,23 para os resultados favoráveis ao time mandante, empate e favorável ao visitante, respectivamente. Após ajustado o modelo com os dados de treino, foi verificada sua precisão com um conjunto de teste. No geral, o modelo acertou 57,5% dos jogos teste (69 dos 120), em que se acertou 57 dos 95 jogos que o modelo indicou como resultado favorável ao mandante (60%), 4 dos 11 jogos que indicou como empate (36,36%) e 8 dos 14 jogos que o modelo indicou como resultado favorável ao visitante (57,14%). Em seguida para saber qual das casas de aposta seria a mais vantajosa apostar foi comparado o retorno financeiro de cada uma das 6 casas de apostas ao apostar um determinado investimento (R\$120,00, R\$240,00, R\$480,00 e R\$600,00) dividido entre os 480 jogos, os lucros simulados ficaram compreendidos entre 13,8% e 17,6%, sendo que o lucro

máximo que poderia ser alcançado usando as odds mais altas entre as 6 casas de apostas foi de 18,76%. Por fim foram testados alguns cenários combinados com diferentes formas de aposta para ter a noção de quanto lucrativo o modelo pode ser. O cenário mais lucrativo foi o que o apostador iniciava com R\$200,00 apostava apenas nos jogos que o modelo ajustava como mandante o vencedor, ao apostar R\$25,00, R\$50,00, R\$75,00 e R\$100,00 por jogo. Os lucros finais seriam de R\$723,00, R\$1246,00, R\$1769,00 e R\$2292,00 respectivamente.

Observando que a modelagem via árvore de decisão é um bom método para previsão de jogos de futebol, em pesquisas futuras pretendemos aumentar a precisão de acerto do modelo, utilizando técnicas mais avançadas de modelagem via árvore de decisão e regressão multinomial com bases de dados maiores.

Referências

Afonso, M. S., Barros, S. S., Koth, A. P., Rodrigues, V. L., Neves, F. B., & Lourenção, L. G. (2020). Sports physiotherapy in program of prevention of injury in professional football. *Research, Society and Development*, 9(3), 72932434.

Albuquerque, M. A., Lucena, S. L. L., & Barros, K. N. N. O. (2020). Comparação de modelo clássico e Bayesiano para dados de óbitos perinatais no ISEA, Campina Grande-PB. *Research, Society and Development*, 9(8), e464985477-e464985477.

Anderson, C., & Sally, D. (2013). Os números do jogo: porque tudo o que você sabe sobre futebol está errado. *São Paulo: Paralela*.

Barros, K. N. N. O., Albuquerque, M. A., Gomes, A. S., & Dantas, D. R. G. (2020). Análise de agrupamentos exploratória dos usuários do Programa Multidisciplinar de Tratamento do Tabagismo do HUAC, Campina Grande-PB. *Research, Society and Development*, 9(8), e825986532-e825986532.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (Wadsworth, Belmont, CA). ISBN-13, 978-0412048418.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3), 503-514.

Carling, C., Bloomfield, J., Nelsen, L., & Reilly, T. (2008). The role of motion analysis in elite soccer. *Sports medicine*, 38(10), 839-862.

Grochtmann, M., & Grimm, K. (1993). Classification trees for partition testing. *Software Testing, Verification and Reliability*, 3(2), 63-82.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. 112, 18. New York: springer.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14-23.

Moisen, G. G. (2008). Classification and regression trees. In: *Jørgensen, Sven Erik; Fath, Brian D. (Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. 582-588.*, 582-588.

Monard, M. C., & Baranauskas, J. A. (2003). Indução de regras e árvores de decisão. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1, 115-139.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302), 415-434.

Oddsshark. O que são odds. 2018. Recuperado de:<<https://www.oddsshark.com/br/como-apostar/o-que-sao-odds>>.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka, R. (2018). Metodologia da pesquisa científica. [e-book]. Santa Maria. Ed. UAB/NTE/UFSM. Recuperado de

https://repositorio.ufsm.br/bitstream/handle/1/15824/Lic_Computacao_Metodologia-Pesquisa-Cientifica.pdf?sequence=1

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1), 1-13.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 476-487.

Sehnem, R., & Frozza, R. (2019). Análise de variáveis em partidas de futebol para previsão de resultados. *Anais do Salão de Ensino e de Extensão*, 217.

Vogado, L. H., Veras, R. M., Araujo, F. H., Silva, R. R., & Aires, K. R. (2019, June). Rede Neural Convolutacional para o Diagnóstico de Leucemia. In *Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, 46-57.

Weiss, S. M., & Indurkha, N. (1994). Small sample decision tree pruning. In *Machine Learning Proceedings 1994*. 335-342. Morgan Kaufmann.

Wilkinson, L. (2004). Classification and regression trees. *Systat*, 11, 35-56.

Porcentagem de contribuição de cada autor no manuscrito

Adenilson Borba Lopes Silva – 35%

Kleber Napoleão Nunes de Oliveira Barros– 35%

Mácio Augusto de Albuquerque– 30%