UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Adenilson Borba Lopes da Silva

Modelagem via árvore de decisão para previsão de jogos de futebol

Campina Grande - PB

09 de Fevereiro de 2020

Adenilson Borba Lopes da Silva

Modelagem via árvore de decisão para previsão de jogos de futebol

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Kleber Napoleão Nunes de Oliveira Barros

Campina Grande - PB

09 de Fevereiro de 2020

S586m Silva, Adenilson Borba Lopes da.

Modelagem via árvore de decisão para previsão de jogos de futebol [manuscrito] / Adenilson Borba Lopes da Silva. - 2020.

30 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2020.

"Orientação : Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros , Coordenação do Curso de Estatística - CCT."

1. Futebol . 2. Casa de apostas. 3. Estatística . 4. Árvore de decisão . I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

Adenilson Borba Lopes da Silva

Modelagem via árvore de decisão para previsão de jogos de futebol

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 28 de feveriro de 2020.

BANCA EXAMINADORA

Kleber Napoleão Nunes de Oliveira Barros
Universidade Estadual da Paraíba

Mácio Augusto de Albuquerque
Universidade Estadual da Paraíba

Alexandro Bezerra Cavalcanti
Universidade Federal de Campina Grande

"Dedico essa monografia a Deus, que me deu força, coragem e foco durante toda esta longa caminhada e a todos que acreditaram em minha pessoa."

Agradecimentos

A minha Mãe Denise, meus irmãos Ardson e Anderson e meus dois preciosos sobrinhos Davi e Margareth meus melhores presentes.

Agradeço também a minha noiva Simone, que de forma especial e carinhosa me deu força e coragem para lutar nos momentos de dificuldades.

A todos os professores do curso de Estatística da UEPB, que foram tão importantes na minha vida acadêmica e no desenvolvimento dessa monografia.

Ao professor Kleberr N. N. O. Barros pela paciência e incentivo que tornaram possível a conclusão desta monografia.

Ao meu professor e amigo Walber Pereira Nery por seus conselhos e insistência em minha pessoa.

A todos meus amigos e colegas pelo o apoio constante, em especial a Jhonathan, José Francisco, Ronaldo, Victor, Ramon, Eduardo, Jamilton, Apollo e Rafael que juntos formamos a turma do Ximbó, também a Ravylla, Maria Taynan e Marcela que sempre me motivaram e que nunca me abandonaram quando precisei.

E a todos meus Colegas de Curso, em especial Gustavo, João Paulo, Sandro e Fábio que iniciaram essa jornada juntos comigo.

"Quando alguém te machuca, você sente ódio e quando você machuca alguém, você fica amargo, mas se sente culpado também. Conhecer a dor nos ajuda a crescer, a madurecer e crescer significa ser capaz de pensar e tomar decisões próprias."

(Naruto Uzumaki)

Resumo

Após avanço tecnológico a analise 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 vêm 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 a partir de Odds geradas por 6 diferentes bancas de apostas, onde a variável resposta pode ser 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. No estudo foi utilizado três softwares, um para coleta e armazenamento dos dados, um para análises dos dados e um para construção do artigo. Primeiramente, foi coletado em um site de bancas de apostas no período de aproximadamente 3 meses, as Odds de 6 diferentes bancas de apostas, referentes a 600 jogos de 9 campeonatos de futebol. Após coleta dos dados foi feita uma analise estatística dos dados, onde foi encontrado um modelo de teste via árvore de decisão que acertou 69 dos 120 jogos (57,5%). Por fim foi feita a comparação de qual seria a melhor casa de aposta a se apostar e o quão lucrativo o modelo seria em diferentes ocasiões simuladas.

Palavras-chaves: Futebol. Casa de apostas. Estatística. Árvore de decisão.

Abstract

After a technological advance, the analysis which is focused on sportive ends became fundamentally important to tactics evolution and achievement of good results. At soccer, the use of these analyses have been growing up and bring innumerable benefits not only to tactic development but also to the physical part of athletes. Further on the tactic collaboration and technique to soccer, the statistic is also frequently used on predictions, that comprehend since a penalty kick until the final result of the game. This paper has the purpose to find a result predicting model in which the variant answer can be Home (home team wins), Draw or Visitor (visitor team wins) from the position on the classification table and Odds generated by bookmakers using the Decision Tree Method that after data modeling and analysis of the accuracy model which was analyzed which bookmaker would be more profitable. In the study was used three softwares, the first to the collection and data storage, the second one to date analyses, and the last one to the construction of this paper. Firstly, it was collected in a bookmaker site in a period of approximately three months, the Odds from six different bookmakers, referring to 600 games of 9 soccer championships. After the date collects it was made a statistic analysis where was found a test model by way of a decision tree which hit 69 of 120 games (57,7%). Lastly, compared which of the bookmakers would be the best to bet in and how much profit would be the model in different occasions simulated.

Key-words: Soccer. Bookmakers. Statistic. Decision Tree.

Lista de ilustrações

Lista de tabelas

Sumário

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 vêm crescendo e trazendo inúmeros benefícios tanto para o desenvolvimento tático, quanto para o físico 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.

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 bastante. 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 final da partida. Esses fatores podem ser preditos usando um determinado tipo de modelo estatístico, sendo que o tipo e modelo varia de acordo com algumas características que vão depender do fator que está sendo observado. São nesses modelos, onde de fato, elas conseguem o retorno financeiro através de taxas que são empregadas nas Odds.

As Odds são quantidades derivadas da probabilidade de um determinado evento. Elas são calculadas a partir de análises estatísticas feita 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 $P = (1/Odd_a)$, onde P é a probabilidade 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 nessa Odd, então se o resultado coincidir com o da Odd que foi apostada ele receberá R$17,70 (Odd × Aposta). 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 (também chamada de juice) da casa de apostas. Ou seja, é a comissão que a casa de apostas receberá. Essas margens variam de acordo com cada casa de apostas, porém, estima-se que na grande parte das casas de apostas esteja entre 3% e 6% (ODDSSHARK, 2018).

O presente trabalho tem como objetivo principal 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 no qual o modelo final é testado em diferentes contextos e ocasiões para saber o quão rentável seria o modelo.

2 Fundamentação Teórica

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 da sua equipe. 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 de sua 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 consequentemente quem sai com a vitória. O futebol está pronto para a mudança e na estatística é onde se encontra a base para essas mudanças. Através da estatística que irá mudar a forma, as ideias preconcebidas e ditar novas normas, renovar as práticas e derrubar as antigas táticas de jogo. São os números que vão nos permitir ver o jogo como nunca o vimos antes (ANDERSON; SALLY, 2013).

Neste trabalho será realizado o método árvore de decisão para modelagem do resultado (vitória mandante, empate e vitória 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. Os 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 proposito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema (BREIMAN et al., 1984). Wilkinson (2004) afirma que 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

Segundo 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 $X_1, X_2, ..., X_p$, em J regiões distintas e não sobrepostas, $R_1, R_2, ..., R_p$.

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 $R_1, ..., R_J$ 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.

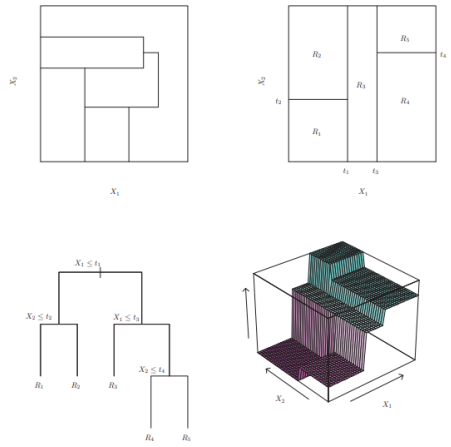

Figura 1 – Exemplo de regiões em forma de retângulos (JAMES et al., 2013).

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 segundo 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 $R_1, ..., R_J$ que minimizam a SQR,

dado por

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

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 seleciona-se o preditor X_j e o ponto de corte que dividem o espaço do preditor nas regiões $\{X \mid X_j < s\}$ e $\{X \mid X_j \geq s\}$ que levam a máxima redução possível da SQR. A notação $\{X \mid X_j < s\}$ significa a região do espaço preditor em que X_j assume um valor menor que s. Ou seja, considera-se todos preditores $X_1, ..., X_p$ 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

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

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

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

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 recursos p não é muito 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 $R_1, ..., R_J$

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 o treinamento definido, mas 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 $R_1, ..., R_J$) pode levar a menores 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 subárvore. Para determinar a melhor poda, intuitivamente, seleciona-se uma subárvore que leva à menor taxa de erro de teste. Dada uma subárvore, 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 subárvore possível seria muito complicado, uma vez que existe um número extremamente grande de subárvores possíveis. Weiss e Indurkhya (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 subárvores possíveis, considera uma sequência de árvores indexadas por um parâmetro de ajuste não negativo α, para cada valor de α corresponde uma subárvore $T \subset T_0$ tal que

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

é o menor possível. Aqui $|T|$ indica 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 $y_i - \hat{y}_{R_j}$ 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 subárvore e sua adequação aos dados de treinamento. Quando $\alpha = 0$, a subárvore 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 subárvore 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 subárvores 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 subárvore correspondente a α. Breiman (2017) resume esse processo assim:

1. 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.

2. Aplique a poda da complexidade de custos à árvore grande para obter uma sequência das melhores subárvores, em função de α.

3. Use a validação cruzada com dobra K para escolher α. Ou seja, divida o treinamento em K dobras. Para cada k = 1,..., 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.

   Faça a média dos resultados para cada valor de α e escolha o α que minimiza o erro médio.

4. Retorne a subárvore 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. 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, utiliza-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 planeja-se 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 :

$$E = 1 - \underset{k}{max}\left(\hat{p}_{mk}\right). \tag{2.5}$$

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

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

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

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

Como $0 \leq \hat{p}_{mk} \leq 1$, segue-se que $0 \leq -\hat{p}_{mk} log\left(\hat{p}_{mk}\right)$. 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 dessas três abordagens podem ser usadas 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).

2.5 Métodos de Reamostragens

Os métodos de reamostragem são ferramentas indispensáveis nas estatísticas modernas. Eles envolvem repetidamente tirar amostras de um conjunto de treinamento e reajustar um modelo. As abordagens de reamostragem eram computacionalmente bastante demoradas, porque elas envolve o ajuste do mesmo método estatístico várias 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 que chegue a ser proibido. 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.

2.5.1 Método K-Fold de Validação Cruzada (K-Fold Cross-Validation)

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 preferencia 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, $EQM_1, EQM_2, ..., EQM_K$. A estimativa da CV da dobra k é a por

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

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).

3 Metodologia

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 refentes as probabilidades de vitória do time mandante, vitoria 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, esses valores forma 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.

3. Para construção do artigo o TeXstudio 2.12.22, um editor LaTeX de código aberto para várias plataformas. Seus recursos incluem um verificador ortográfico interativo, dobragem de código e destaque de sintaxe.

.

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, nota-se que a maioria dos campeonatos estão próximo dos 60 jogos, com exceção do campeonato Brasileiro, que tem 110 jogos. 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.

Tabela 1 – Número de jogos dos campeonato 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%

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.

Nela nota-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 minima entre as observações foi de 11% e a máxima foi de 89%. Para a variável Total de Chutes Mandante e Visitantes 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 minimo atingidos 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

Tabela 2 – Análise descritiva das estatísticas dos jogos

Variável	Média	Minimo	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)

maior 7 e os valores máximo e mínimo atingido por esta variável foi 15 e 0. A médias para cartões amarelos foi de 2,05 e 2,40 com variância de 1,86 e 2,24 paras 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 referente as 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 1° Quartil e 3° Quartil das 6 casas de apostas.

Na Tabela 3, 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 maior variância, ao observamos os 1° e 3° Quartis percebe se que as casas que tiveram melhor intervalo de confiança (considerando o maior 3° 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 um dos outros, dado que a variância ficou entre 2,26 (na casa C) e 1,79 (na casa A). 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, No Empate todas as médias foram representativas, já nos visitantes que foi a categoria que teve a maior variância no geral temos que nenhuma das casas teve

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

Casa de aposta	Resultado	Média	Variância	1° Quartil	3° 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	1° Quartil	3° 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	1° Quartil	3° 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

média representativa.

A Tabela 4 mostra algumas estatísticas referentes ao comportamento da média e da variância também separadas por classes, Mandante, Visitante e Empate e Geral. Nela observa-se o quanto varia 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, 1° e 3° 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	1° Quartil	3° 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

Comparando os valores da Tabela 3 com a Tabela 4, percebe-se que apenas a casa

E é quem 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, no 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, nota-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 priores de 0.53, 0.22 e 0.23 para Mandantes, Visitante e Empate respectivamente. O modelo final foi feito com as 480 observações iniciais (80%) e logo em seguida testado nas 120 observações finais (20%).

Na Figura 2 observa-se os testes e as suas respectivas conclusões, no qual se a variável de um determinado jogo for de acordo com o teste ela seguira o ramo da esquerda, se não o da direita, até chegar ao um nó folha que seria a estimativa do resultado da partida. Junto aos teste tem-se inicialmente uma priori de cada Resultado (E, M e V) no nó raiz e depois de cada bifurcação realizada as posteriores atualizadas junto ao resultado estimado. Após escolha do melhor modelo foi realizado as estimativas com o banco de dados teste para saber a taxa de acerto do modelo. A Tabela 5 mostra os resultados dos valores Ajustado vs Observados.

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

A diagonal principal são os acertos, e as demais combinações linha coluna são os erros, ao fazer a contagem temos 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

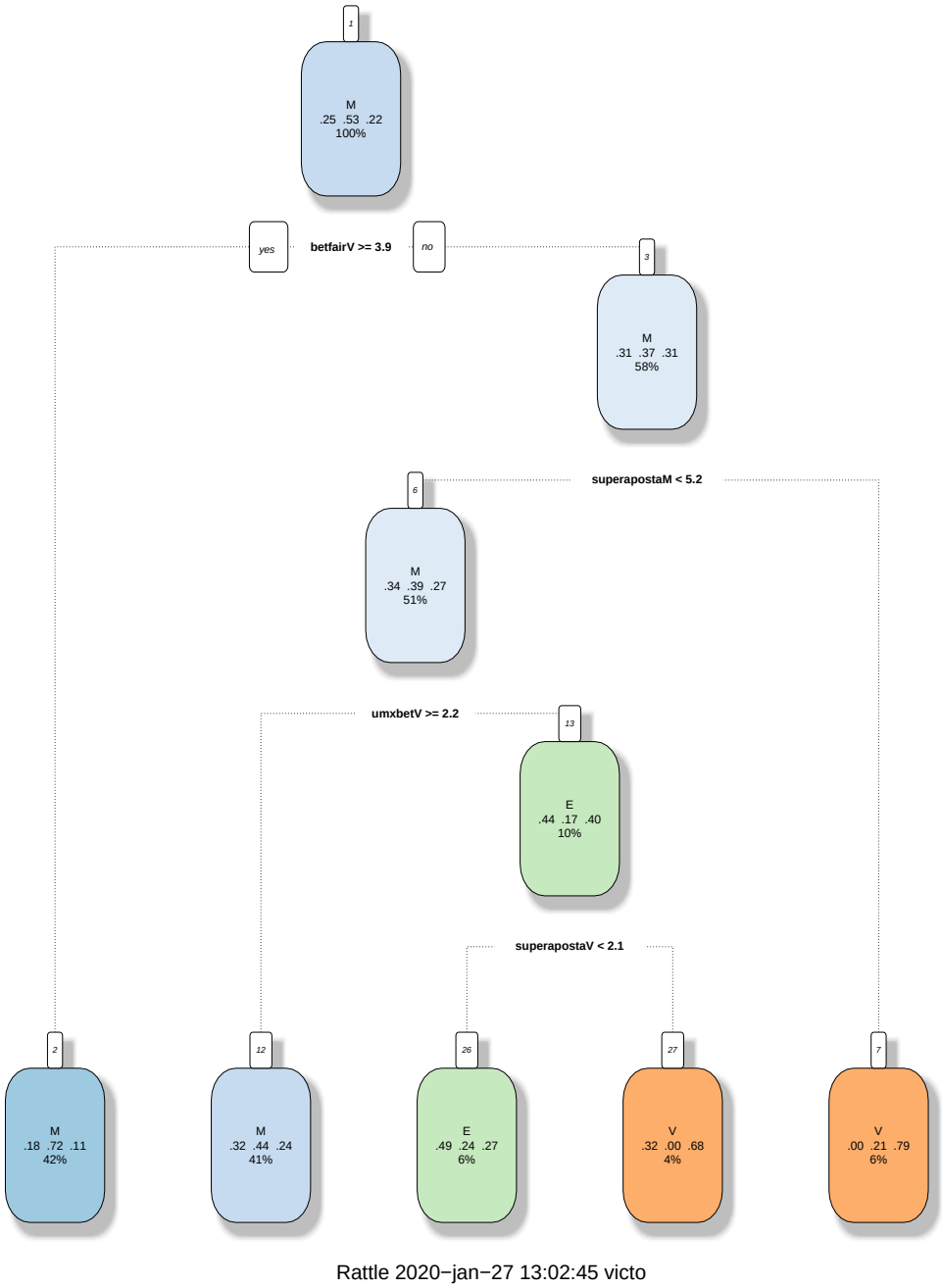

Figura 2 – Árvore de Decisão e Probabilidades.

(30,77%). Com relação a importância das variáveis dentro do modelo a Tabela 6 mostra as variáveis seguidas de suas respectivas importâncias.

Na Tabela 6, nota-se que das 5 variáveis mais importantes 4 delas são de Visitantes, dos Visitantes apenas a Casa A não foi significante, 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).

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%

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 contextos, onde é 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/120jogos) para os 120 jogos, 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%

Ao analisar a Tabela 7, observa-se que em todas as casas obteve-se lucro, 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 contextos testados foram:

1. O Apostador inicia com R$100,00 e sempre apostava um mesmo valor em todos os jogos. Os valores testados 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 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 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 Contextos × Ocasiões

Contexto \ Ocasião	1	2	3	4
1	663	1226	1789	2352
2	816,77	468,18	81,34	0,07
3	723	1246	1769	2292

A Tabela 8 mostra os lucros das possíveis combinações dos Contextos e Ocasiões. No contexto 1 para as 4 ocasiões, 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 ocasião 1 o saldo chegou a ser R$820,50 e teve um valor mínimo de R$106,75, para a ocasião 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 ocasião 4 o saldo variou entre R$2982,00 e R$127,00. No contexto 2 percebe-se que nas ocasiões 3 e 4 se saiu no prejuízo, o máximo e minimo que o saldo atingiu nesses dois contextos foi de R$1600,81 e R$34,08 na ocasião 3 e de R$578,12 e R$0,05 para a ocasião 4, já nas ocasiões 1 e 2 teve-se um bom lucro de 308% e 134,09% respectivamente. Já no contexto 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 ocasiões, o lucro cresce a medida que o valor apostado por jogo aumenta, para essas ocasiões especificas os valores do saldo final foi de R$723,00, R$1246,00, R$1769,00 e R$2292,00.

5 Conclusão

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 enfase, dado que a quantia de dinheiro 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 de Odds geradas pelas próprias bancas de apostas para assim prever jogos, comparar qual das casas de apostas seria mais vantajoso apostar e é claro 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 priores de 0,53, 0,22 e 0,23 para os resultados Mandante, empate e Visitante respectivamente. Após feito o modelo teste, foi testado sua precisão, no geral o modelo acertou 57,5% dos jogos(69 dos 120), onde em proporção ao número de jogos que foi ajustado ele acertou 57 dos 95 jogos que o modelo ajustou como Mandante (60%), 4 dos 11 jogos que ajustou como Empate (36,36%) e 8 dos 14 jogos que o modelo ajustou como Visitante (57,14%). Logo em seguida para saber qual das casas de aposta é 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 ficaram 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 contextos combinados com diferentes ocasiões para ter a noção de quanto lucrativo o modelo pode ser. O contexto 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. Com isso vendo 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 com bancos de dados bem maiores.

Referências

ANDERSON, C.; SALLY, D. Os números do jogo: por que tudo o que você sabe sobre futebol está errado. *São Paulo: Paralela*, 2013. Citado na página 13.

BREIMAN, L. *Classification and regression trees*. [S.l.]: Routledge, 2017. Citado 3 vezes nas páginas 15, 16 e 17.

BREIMAN, L. et al. Classification and regression trees. wadsworth int. *Group*, v. 37, n. 15, p. 237–251, 1984. Citado 3 vezes nas páginas 13, 16 e 18.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, Oxford University Press, v. 76, n. 3, p. 503–514, 1989. Citado na página 19.

CARLING, C. et al. The role of motion analysis in elite soccer. *Sports medicine*, Springer, v. 38, n. 10, p. 839–862, 2008. Citado na página 13.

GROCHTMANN, M.; GRIMM, K. Classification trees for partition testing. *Software Testing, Verification and Reliability*, Wiley Online Library, v. 3, n. 2, p. 63–82, 1993. Citado na página 17.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 13.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado 4 vezes nas páginas 14, 15, 18 e 19.

LOH, W.-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 1, n. 1, p. 14–23, 2011. Citado na página 17.

MOISEN, G. Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, p. 582–588, 2008. Citado na página 18.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. *Sistemas Inteligentes. Rezende, SO Editora Manole Ltda*, p. 115–140, 2003. Citado na página 16.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis Group, v. 58, n. 302, p. 415–434, 1963. Citado na página 14.

ODDSSHARK. *O que são odds*. 2018. Disponível em: <https://www.oddsshark.com/br/como-apostar/o-que-sao-odds>. Acesso em: 21 jan. 2020. Citado na página 12.

ROKACH, L.; MAIMON, O. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Ieee, v. 35, n. 4, p. 476–487, 2005. Citado na página 18.

SEHNEM, R.; FROZZA, R. Análise de variáveis em partidas de futebol para previsão de resultados. *Anais do Salão de Ensino e de Extensão*, p. 217, 2019. Citado na página 11.

WEISS, S. M.; INDURKHYA, N. Small sample decision tree pruning. In: *Machine Learning Proceedings 1994*. [S.l.]: Elsevier, 1994. p. 335–342. Citado na página 16.

WILKINSON, L. Classification and regression trees. *Systat*, v. 11, p. 35–56, 2004. Citado na página 13.