Universidade de Brasília
Departamento de Estatística

Modelos de Previsão de Resultados para o Campeonato Brasileiro de Futebol - Série A

Luana de Freitas Feijão

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Luana de Freitas Feijão

Modelos de Previsão de Resultados para o Campeonato Brasileiro de Futebol - Série A

Orientador: Prof. Eduardo Monteiro de Castro Gomes

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Para meus pais, vocês sempre acreditaram em mim e quero lhes deixar orgulhosos.

Agradecimentos

Aos meus pais, Viviana e Carlos, expresso minha gratidão pelos ensinamentos, incentivos e apoio constantes que me proporcionaram. Sempre depositaram sua confiança em mim e sou eternamente grata por serem pais maravilhosos.

Ao meu irmão, Lucas, que é minha fonte de inspiração e um exemplo a ser seguido, apesar de ser impossível igualá-lo. Agradeço por tornar minha vida mais leve e por estar sempre presente para mim.

À minha família como um todo, com destaque para meus avós, João, Vânia e Aleonira, que têm um papel fundamental em minha vida. Sou grata a Deus por tê-los ao meu lado e por poder compartilhar minhas conquistas com eles.

Ao meu namorado, Gabriel, que entrou em minha vida para somar em todos os aspectos. Seu amor e companheirismo me incentivam a ser a melhor versão de mim mesma.

Aos amigos que fiz durante o curso, Stephany, Marcelo, Julia e Sabrina, agradeço pelas risadas e pelos momentos compartilhados, que tornaram a jornada acadêmica mais leve e minha experiência na UnB mais feliz.

Aos amigos e colegas da ESTAT e dos estágios que fiz ao longo desses anos, expresso minha gratidão pelo aprendizado que contribuiu para meu crescimento como profissional e pelos inúmeros momentos vividos juntos.

Por fim, agradeço a todos os professores da UnB e ao Departamento de Estatística, em especial ao professor Eduardo, por me auxiliarem a chegar até aqui.

Muito obrigada a todos.

Resumo

O trabalho teve como objetivo ajustar e avaliar modelos de previsão de resultados para o Campeonato Brasileiro de Futebol - Série A, abrangendo o campeão, os quatro times que avançam para a Libertadores e os quatro times rebaixados para a Série B. Para isso, utilizou-se um banco de dados estatístico criado a partir dos resultados dos jogos obtidos do site da CBF. Foram aplicadas as técnicas de Regressão Logística, Árvore de Classificação e *Random Forest* para realizar as previsões dos campeonatos de 2020 a 2022, a fim de comparar os resultados com os resultados reais. Para isso, foram utilizados os jogos dos campeonatos de 2012 a 2019 como dados de treinamento. Toda a implementação computacional foi realizada utilizando a linguagem R.

Palavras-chaves: Modelagem, Previsões, Random Forest, Regressão Logística, Árvore de Classificação, CBF

Abstract

The objective of this study was to adjust and evaluate prediction models for the Brazilian Football Championship - Serie A, encompassing the champion, the four teams that advance to the Libertadores, and the four teams relegated to Serie B. To achieve this, a statistical database was created using game results obtained from the CBF website. The prediction models of Logistic Regression, Classification Tree, and Random Forest were applied to forecast the championships from 2020 to 2022, aiming to compare the results with the actual outcomes. The games from the championships between 2012 and 2019 were used as training data. All computational implementation was performed using the R programming language.

Keywords: Modeling, Predictions, Random Forest, Logistic Regression, Classification Tree, CBF

Lista de Tabelas

Lista de Quadros

Lista de Figuras

Sumário

1 Introdução

O futebol é o esporte mais popular do mundo, contando com mais de 3,5 bilhões de espectadores (VEROUTSOS, 2022). Além de ser um esporte que traz entusiasmo para aqueles que acompanham, é também um grande segmento para movimentação de dinheiro, não apenas com vendas de ingressos e produtos que estampam os brasões dos times, mas também com televisionamento e casas de apostas.

Com respeito a isto, segundo um estudo da Fact.MR (2022), prevê-se que o mercado de apostas esportivas seja avaliado em aproximadamente US$84,66 bilhões em 2022. Além disso, prevê-se que o mercado tenha um aumento de 10,3% ao ano entre 2022 e 2032, totalizando seu valor em cerca de US$225,65 bilhões em 2032. Logo, o aumento desse mercado acarreta em um aumento na demanda por estudos e modelos de previsão em jogos, uma vez que os apostadores buscam informações e análises mais assertivas para tomar decisões em suas apostas.

O presente trabalho visa estudar modelos que possam prever as chances de um time do Campeonato Brasileiro de Futebol - série A ou ser campeão, ou avançar para a Taça Libertadores da América ou ser rebaixado para a série B.

Desde 2003, os Campeonatos Brasileiros passaram a seguir uma estrutura de pontos corridos, com turno e returno. Cada campeonato da série A conta com 20 times, em que cada um joga duas vezes com um mesmo time adversário, sendo um como mandante e outro como visitante, totalizando 38 rodadas por time e 380 jogos ao todo. Para a pontuação, atribui-se 3 pontos para o time que vence, 0 para o que perde e 1 ponto para ambos os times que empataram. Ao final do campeonato, a classificação se dará pela soma dos pontos obtidos durante as 38 rodadas. O time que fica em primeiro lugar é declarado campeão. Os quatro primeiros times classificados avançam para à Taça Libertadores da América. Os últimos quatro colocados são rebaixados para a série B.

Neste contexto, o objetivo deste trabalho é propor modelos estatísticos e avaliar o desempenho dos mesmos para previsão de resultados de campeão, dos quatro primeiros e dos quatro últimos colocados no Campeonato Brasileiro de Futebol - série A. Para isso, serão utilizados dados extraídos do site da Confederação Brasileira de Futebol (CBF) dos campeonatos de 2012 a 2022 e testados modelos com técnicas de Regressão Logística, Árvore de Classificação e *Random Forest*, uma vez que esses são modelos bastante utilizados e já implementados e de fácil utilização. Os jogos dos campeonatos de 2012 a 2019 servirão como dados de treinamento para os modelo e a previsão será realizada nos cam-

peonatos de 2020 a 2022. Todos os cálculos e simulações serão realizados com o auxílio do software livre R Team (2023).

2 Referencial Teórico

Os métodos utilizados na modelagem para previsão no Campeonato Brasileiro de Futebol, que serão descritos, apresentam suposições fundamentais para sua aplicação. A modelagem será realizada para prever o campeão do campeonato, bem como os quatro times que se classificarão para a Libertadores e os quatro times que serão rebaixados.

2.1 Regressão Logística

A análise de regressão logística consiste em construir um modelo estatístico para prever a relação de uma variável resposta binária, isto é, que assume valores 0 e 1, a um conjunto de p variáveis independentes $X_1, \cdots, X_p$. A probabilidade de "sucesso" da variável resposta pode ser descrita como $P(Y_i = 1|X) = \pi(X)$ e de "fracasso" $P(Y_i = 0|X) = 1 - \pi(X)$ (JR; LEMESHOW; STURDIVANT, 2013).

A função matemática que representa a probabilidade de sucesso ou fracasso da variável resposta é dado por

$$P(Y_i|X_1, ..., X_p) = \pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}, \forall i = 1, ..., n, \tag{2.1.1}$$

em que $\beta_0, \cdots, \beta_p$ são os parâmetros do modelo.

Realizando algumas manipulações e aplicando o log em 2.1.1, pode-se estabelecer uma relação linear, chamada de logito (HUANG, 2014), dada por

$$logit(\pi(X)) = log\left(\frac{\pi(X)}{1 - \pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \forall i = 1, ..., n. \tag{2.1.2}$$

2.1.1 Seleção de Variáveis

Para modelos de regressão logística é possível aplicar um método específico para selecionar as variáveis. Esse método consiste em analisar diferentes modelos com base em uma medida específica e realizar a retirada ou adição de variáveis de forma iterativa, buscando obter o modelo com o melhor valor dessa medida analisada (CHAMBERS; HASTIE; PREGIBON, 1992).

No contexto deste trabalho, a função *stepAIC* do pacote *MASS* do *software* R

é utilizada para realizar essa seleção de variáveis. Essa função utiliza o critério AIC (Critério de Informação de Akaike) para determinar o melhor modelo. Vale ressaltar que existem três formas diferentes de aplicar a função *step*, porém, neste estudo, optou-se por utilizar apenas uma delas, a chamada "*forward*".

Ao empregar a abordagem *forward*, o método de seleção de variáveis consiste em adicionar variáveis uma por uma ao modelo, avaliando a melhoria do valor do critério AIC a cada adição. Dessa forma, é possível identificar quais variáveis têm maior impacto na qualidade do modelo em relação à medida analisada.

2.2 Árvores de Classificação

Os modelos de árvore são um método de aprendizado de máquina não-paramétrico utilizado para a classificação de dados em diferentes categorias ou classes (ver Figura 1). Essa técnica baseia-se na criação de uma estrutura em forma de árvore, na qual cada nó representa uma característica ou atributo dos dados e cada ramo representa uma possível resposta ou valor dessa característica. Vale lembrar que esses atributos podem ser tanto quantitativos quanto categóricos, e que a classificação ocorre de maneira binária.

Uma vantagem desse método é a sua alta interpretabilidade, além da capacidade de lidar com dados ausentes e múltiplas variáveis respostas. No entanto, os modelos de árvore podem ser instáveis e nem sempre produzem um desempenho preditivo ideal (KUHN et al., 2013).

Durante a construção da árvore, os dados de treinamento são divididos em subconjuntos com base nos valores dos atributos, buscando-se maximizar a pureza das classes em cada subconjunto resultante, isto é, fazer com que os nós contenham uma proporção maior de uma classe. Isso permite que a árvore seja capaz de fazer previsões precisas e classificar novos dados com base nas características relevantes identificadas durante o treinamento.

A pureza foi avaliada utilizando o índice de Gini (BREIMAN, 2017), uma medida que varia de 0 a 1 e que é utilizada para avaliar a desigualdade na distribuição de dados. Um valor de 0 indica uma distribuição perfeitamente igualitária, enquanto um valor de 1 indica uma distribuição completamente desigual.

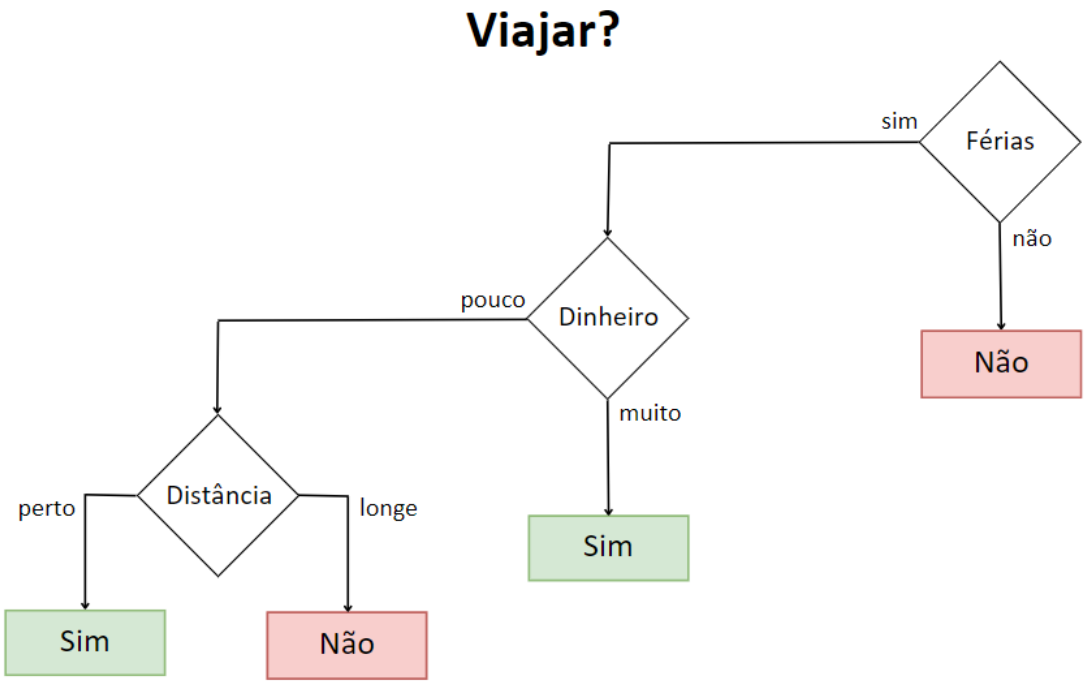

Figura 1: Exemplo de Árvore de Decisão para viajar ou não

2.3 *Random Forest*

Random Forest é uma técnica que combina várias árvores de classificação independentes. Cada árvore é construída usando diferentes variáveis e observações. Os resultados de previsão de cada árvore são combinados para formar um modelo final, selecionando as variáveis mais frequentes. Essa abordagem é útil para evitar o *overfitting*, característico em modelos que se ajustam excessivamente, mas não generalizam bem para novos dados. No entanto, modelos de floresta requerem maior capacidade computacional (BREIMAN, 2017).

Ao usar o modelo *Random Forest*, é necessário ajustar dois parâmetros: m_{try} e *ntree*. O parâmetro m_{try} determina o número de variáveis explicativas usadas em cada árvore, enquanto o parâmetro *ntree* define o número de amostras *bootstrap* criadas para treinar as árvores. A combinação de várias árvores forma o modelo *Random Forest*. Um valor maior para *ntree* geralmente resulta em um melhor desempenho do modelo, mas requer mais recursos computacionais (KUHN et al., 2013).

2.4 Medidas

Nesta seção, serão listadas as medidas utilizadas para a seleção do melhor modelo para a previsão dos dados.

2.4.1 Matriz de Confusão

A matriz de confusão é uma representação visual que mostra o número de previsões corretas e incorretas feitas pelo modelo (BRUCE; BRUCE, 2019), comparando-as com os valores reais dos dados.

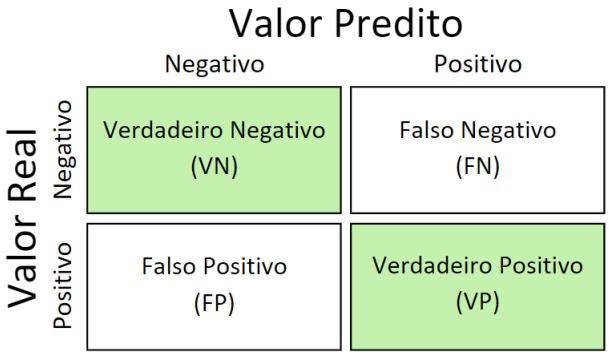

Figura 2: Matriz de confusão

- **Verdadeiro Positivo (VP):** valores positivos classificados como positivos;
- **Verdadeiro Negativo (VN):** valores negativos classificados como negativos;
- **Falso Positivo (FP):** valores negativos classificados como positivos;
- **Falso Negativo (FN):** valores positivos classificados como negativos.

2.4.2 Acurácia

A acurácia avalia o percentual de acertos, isto é, é a razão entre a quantidade de verdadeiros pelo o total de classificações:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.4.1}$$

2.4.3 Kappa

O coeficiente Kappa é uma medida que avalia a concordância entre as classificações observadas e esperadas. Quanto mais próximo de 1 for o valor do coeficiente, maior é a evidência de uma concordância nas previsões. Por outro lado, valores mais próximos de zero indicam que a concordância é puramente aleatória.

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.4.2}$$

Em que P_o é a é a proporção observada de concordância entre as classificações observadas e esperadas e P_e é a proporção esperada de concordância entre as classificações.

2.4.4 Precisão

A precisão avalia das classificações positivas, quantas foram verdadeiras.

$$Precisão = \frac{VP}{VP + FP} \tag{2.4.3}$$

2.4.5 *Recall* ou Sensibilidade

A *recall*, ou taxa de verdadeiros positivos, indica, das amostras positivas, quantas foram classificadas corretamente.

$$Recall = \frac{VP}{VP + FN} \tag{2.4.4}$$

2.4.6 F1 ou F-score

O F1-score, também conhecido como F-score, é uma medida estatística que combina a precisão a *recall* de um modelo. É apropriado para quando deseja-se avaliar o equilíbrio entre a precisão e a *recall*. O valor do F1-score varia de 0 a 1, onde 1 indica um desempenho perfeito e 0 indica um desempenho ruim. Um valor alto de F1-score indica um bom equilíbrio entre a precisão e a *recall* (SHUNG, 2018).

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{2.4.5}$$

3 Material e Métodos

Todo o trabalho foi desenvolvido utilizando o *software* R (versão 4.2.3), desde da extração dos dados e a confecção da nova base de dados que de fato será utilizada para a modelagem dos dados, até as análises estatísticas.

3.1 Campeonato Brasileiro de Futebol

O Campeonato Brasileiro de Futebol é a principal competição nacional de clubes no Brasil. Desde 2003, adotou-se um formato de pontos corridos com turno e returno, proporcionando uma disputa equilibrada ao longo da temporada. Cada campeonato da série A conta com a participação de 20 times. Cada equipe enfrenta todas as outras duas vezes: uma vez em casa, como mandante, e outra fora, como visitante. Isso resulta em um total de 38 rodadas por time e 380 jogos ao longo do campeonato.

O sistema de pontuação adotado é: a equipe vencedora de um jogo recebe 3 pontos, enquanto o time derrotado não pontua. Em caso de empate, ambas as equipes somam 1 ponto. Ao final das 38 rodadas, a classificação é determinada pela soma dos pontos obtidos. O time que alcança o maior número de pontos ao final do campeonato é declarado o campeão brasileiro. Além do título, os quatro primeiros colocados garantem vaga na Taça Libertadores da América, enquanto os quatro últimos colocados são rebaixados para a Série B.

Figura 3: Brasão dos times que competiram em 2022

Fonte: (CBF, 2022)

3.2 Banco de Dados

Para a modelagem dos dados, primeiramente realizou-se a extração dos dados do site da CBF (*link*: https://www.cbf.com.br/futebol-brasileiro/competicoes/campeonato-brasileiro-serie-a). Em seguida, foi criada uma nova base de dados, a qual foi efetivamente utilizada para os modelos de previsão.

3.2.1 Extração do Banco de Dados

Para a obtenção dos dados, foi utilizado a técnica de *web scraping*, a qual consiste numa técnica de extração de informações de uma página da Internet. Os dados foram retirados do site da CBF (CBF, 2022). Para isso, foi utilizado o pacote *rvest* do *software* R.

O banco de dados traz como informações os jogos dos campeonatos de 2012 a 2022 da série A, em que, a cada ano, ocorrem 380 jogos, sendo 38 partidas para cada time. Além disso, traz os times que jogaram, sendo 20 para cada ano, o resultado do placar para o time mandante e para o time visitante, o dia e horário em que ocorreu o jogo, a arena, cidade e estado em que sucedeu e o árbitro da partida e de que estado ele é. O banco de dados apresentou 4180 linhas, em que cada linha faz referência a um jogo, e 17 colunas.

O Quadro 1 é um exemplo do banco de dados e faz referência aos 5 primeiros jogos da base de dados, considerando apenas os times participantes, seus placares, o ano e número do jogo.

Quadro 1: Exemplo dos dados extraídos

Ano	**Jogo**	**Mandante**	**Placar do mandante**	**Visitante**	**Placar do visitante**
2012	1	Vasco da Gama	2	Grêmio	1
2012	2	Bahia	0	Santos	0
2012	3	Palmeiras	1	Portuguesa	1
2012	4	Figueirense	2	Náutico	1
2012	5	Corinthians	0	Fluminense	1

A partir desses dados, foi criada uma nova base de dados, a qual foi utilizada para a criação dos modelos.

3.2.2 Criação do Banco de Dados

A nova base de dados contém 220 linhas e 73 colunas, isto é, cada linha contém dados referentes de um time participante do campeonato de um ano.

Com os dados extraídos do site, foi possível criar variáveis derivadas referentes a pontos e gols, como a média de pontos/gols durante o campeonato, a quantidade de vezes que o time ganhou jogando em casa, a quantidade de vezes que ganhou jogando fora de casa, entre outras (ver Quadro 3). O Quadro 2 contém uma amostra de como ficou a nova base com as seis primeiras colunas:

Quadro 2: Exemplo da nova base de dados para a 10ª rodada

Ano	**Time**	**Vitória**	**Empate**	**Derrota**	**GP**	**GC**	**SG**
2012	Atlético Goianiense	1	2	7	7	18	-11
2012	Atlético Mineiro	8	1	1	19	7	12
2012	Bahia	1	4	5	7	16	-9
2012	Botafogo	5	2	3	21	15	6
2012	Corinthians	3	2	5	10	12	-2

Para cada modelo, foi utilizada a base contendo esses dados após algumas quantidades de rodadas para cada time. Optou-se por testar para 10, 12, 15, 17 e 20 rodadas. O objetivo é conseguir a melhor previsão com a menor quantidade possível de rodadas.

3.3 Modelagem dos Dados

Após a obtenção da tabela com os dados criados para cada time para um determinado número de rodadas, foram divididos os conjuntos de treinamento e de teste. Os dados de treinamento abrangiam os jogos dos campeonatos de 2012 a 2019, enquanto os de teste consistiam nos jogos de 2020 a 2022. Optou-se por mais de um ano para os dados de teste para ter uma maior robustez na conclusão da avaliação do desempenho dos modelos, uma vez que apenas um ano podia fazer com que o modelo acertasse ao acaso.

Foram criadas três versões distintas dos dados, baseadas no número de rodadas. A primeira contém todas as variáveis disponíveis (completo), a segunda exclui as relacionadas aos gols (sem.gols) e a terceira inclui somente as variáveis pertinentes ao time mandante (sem.vis). Essas variações foram implementadas como um palpite para otimizar o ajuste do modelo.

Para a construção dos modelos de previsão, foram utilizados os seguintes pacotes do *software* R: *stats*, *randomForest*, *rpart* e *c50*. O pacote *stats* foi empregado na modelagem de Regressão Logística, o pacote *randomForest* na modelagem de *Random Forest*, enquanto os pacotes *rpart* e *c50* foram utilizados na modelagem de árvores de classificação. Além desses, ainda utilizou-se o pacote *MASS* para realizar a técnica *step forward* no modelo de Regressão Logística.

Vale lembrar que o modelo C5.0 utilizado para árvore de classificação é uma versão mais avançada de outro modelo de árvore, o C4.5, que possui recursos adicionais, além de melhorias básicas. Essas melhorias podem resultar em árvores menores (KUHN et al., 2013).

A primeira etapa da modelagem consistiu na criação dos modelos utilizando os dados de treinamento. Nos casos dos modelos *Random Forest*, foram definidos previamente os parâmetros m_{try} e $ntree$. Em seguida, foram realizadas previsões para os dados de treinamento e construída uma matriz de confusão, utilizando o pacote *caret* para essa tarefa. A partir desses resultados, foram registradas as medidas de desempenho para formar uma tabela abrangendo todas as técnicas, variáveis utilizadas e número de rodadas. Foram geradas três tabelas, cada uma correspondendo a uma categoria de previsão: campeão, times que avançam para a Libertadores e times rebaixados para a série B. Com base nessas informações, foi possível selecionar o melhor modelo para cada caso.

A medida F1-*score* foi adotada como critério principal na escolha do melhor modelo, uma vez que combina as medidas relacionadas à classificação dos dados como positivos, que é o objetivo principal deste trabalho. As demais medidas apresentadas servem para fornecer uma análise complementar.

4 Resultados

Nesta seção, serão apresentadas análises descritivas e os resultados dos campeonatos e times ao longo dos anos de 2012 a 2022. Além disso, serão discutidos os modelos utilizados e suas performances na previsão dos resultados do campeonato, bem como as matrizes de confusão correspondentes aos melhores modelos de cada previsão.

Pela análise da Figura 4, observa-se uma constância na pontuação alcançada pelos times ao longo dos anos. Em média, os times acumulam cerca de 52 pontos ao final do campeonato, enquanto os times declarados campeões registram uma média de 79 pontos. Nota-se que o ano de 2019 apresentou uma maior dispersão na pontuação, sendo também o ano em que se obteve o maior valor pontual ao término do campeonato.

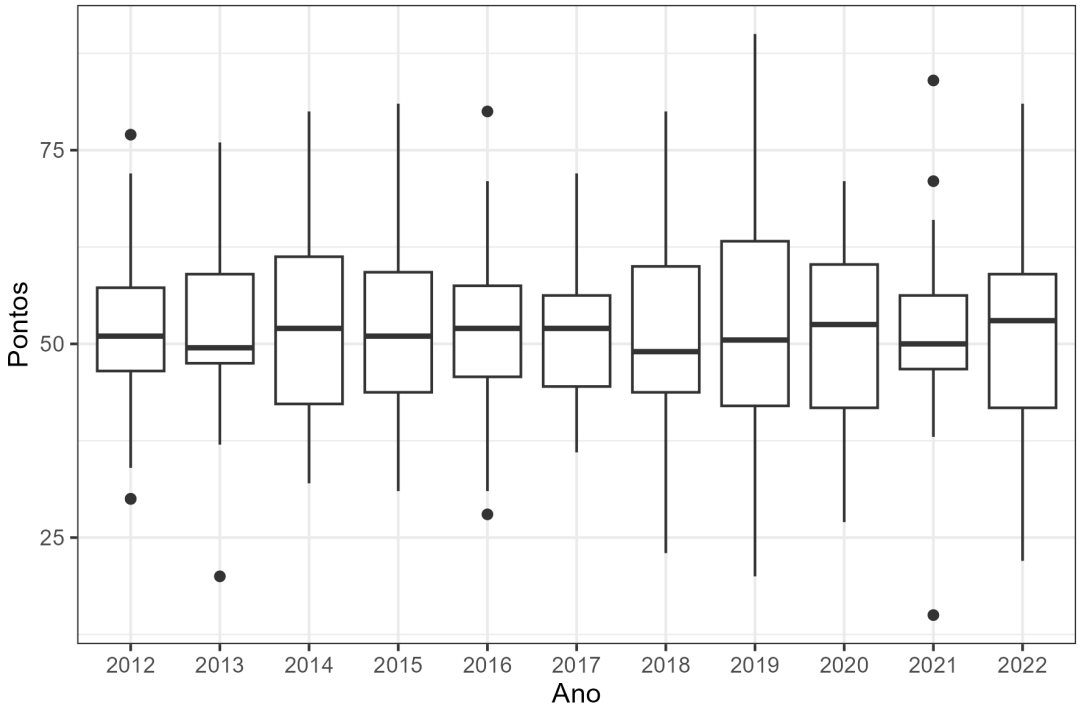

Figura 4: Boxplot dos pontos ao longo dos anos

Conforme apresentado na Tabela 1, verificou-se que os três times com maior pontuação média registraram aproximadamente 63 pontos cada. Esses times foram o Atlético Mineiro, Flamengo e Palmeiras. Ao comparar esses resultados com as informações do Quadro 4, nota-se que essas três equipes também se destacaram como algumas das mais vitoriosas ao longo do período de onze anos. Outros resultados podem ser consultados no Anexo A para fins de comparação.

Tabela 1: Médias de pontos, gols pró e gols contra por time

Time	**Campeonatos**	**Média de Pontos**	**Média de Gols Pró**	**Média de Gols Contra**
Athletico Paranaense	10	55.6	46.6	41.2
Atlético Goianiense	5	41.0	37.4	52.2
Atlético Mineiro	11	63.0	56.3	42.5
Avaí	4	35.0	29.8	57.5
Bahia	8	45.8	41.0	46.4
Botafogo	9	48.4	41.8	46.2
Ceará	5	44.4	39.0	41.8
Chapecoense	7	41.0	37.3	51.3
Corinthians	11	59.7	45.5	34.6
Coritiba	8	43.6	40.1	49.9
Criciuma	2	39.0	38.5	59.5
Cruzeiro	8	57.5	48.9	41.1
Csa	1	32.0	24.0	58.0
Cuiabá	2	44.0	32.5	39.5
Figueirense	4	39.2	35.5	54.8
Flamengo	11	63.0	56.0	41.3
Fluminense	11	55.2	47.8	43.8
Fortaleza	4	51.8	43.5	44.2
Goiás	6	46.5	42.0	52.2
Grêmio	10	61.3	49.1	36.1
Internacional	10	58.9	48.0	38.8
Joinville	1	31.0	26.0	48.0
Juventude	2	34.0	32.5	56.5
Náutico	2	34.5	33.0	65.0
Palmeiras	10	62.9	55.6	40.6
Paraná	1	23.0	18.0	57.0
Ponte Preta	5	45.6	40.0	48.6
Portuguesa	2	46.5	44.5	43.5
Red Bull Bragantino	3	51.0	51.3	48.3
Santa Cruz	1	31.0	45.0	69.0
Santos	11	57.3	49.1	39.1
Sport	8	45.8	39.1	49.6
São Paulo	11	58.5	48.4	39.5
Vasco da Gama	7	47.4	40.0	50.7
Vitória	5	44.4	46.6	56.2

4.1 Modelos

A seguir serão descritos os melhores modelos para cada resultado segundo a medida F1 e serão apresentadas as matrizes de confusão do melhor modelo elegido. Para os casos em que o melhor modelo for uma árvore de classificação, ainda será apresentado o gráfico da árvore.

Com o objetivo de selecionar o modelo mais adequado, foi elaborada uma tabela

contendo as medidas que auxiliam na avaliação e escolha dos melhores modelos.

Além disso, para cada método foram realizadas previsões utilizando diferentes conjuntos de variáveis: todas as variáveis (completo), apenas as variáveis referentes ao time mandante (sem vis), e todas as variáveis, exceto as relacionadas aos gols (sem gols). No total, foram gerados 75 resultados e, para análise, foram selecionados os 10 melhores modelos de acordo com a medida F1.

4.1.1 Modelo Previsão Campeão

A Tabela 2 apresenta os melhores 10 modelos para a previsão do campeão de acordo com o F1.

Tabela 2: TOP 10 modelos de previsão para o campeão ordenados por F1

Método	**Variáveis**	**Nº de Rodadas**	**Acurácia**	**Kappa**	**Sensibilidade**	**Precisão**	**F1**
Random Forest	sem vis	17	1.000	1.000	1.000	1.000	1.000
Regressão Logística com Step	completo	10	0.983	0.792	1.000	0.667	0.800
Random Forest	sem gols	12	0.983	0.792	1.000	0.667	0.800
Random Forest	completo	17	0.983	0.792	1.000	0.667	0.800
C5.0	completo	20	0.967	0.649	0.667	0.667	0.667
C5.0	sem vis	20	0.967	0.649	0.667	0.667	0.667
C5.0	sem gols	20	0.967	0.649	0.667	0.667	0.667
Random Forest	completo	20	0.950	0.545	0.500	0.667	0.571
Random Forest	sem vis	20	0.950	0.545	0.500	0.667	0.571
Random Forest	sem gols	20	0.950	0.545	0.500	0.667	0.571

A partir dos resultados apresentados na tabela acima, observa-se que a maioria dos modelos com melhor desempenho demonstra uma tendência previsível ao utilizar 20 rodadas. No entanto, é importante destacar que também foram identificados bons modelos que alcançaram resultados satisfatórios com apenas 10 e 12 rodadas, o que é especialmente relevante para o contexto de apostas. Além disso, é notável o destaque dos métodos *Random Forest* e C5.0, pois eles representam a maior parte dos modelos presentes no ranking dos 10 melhores.

Outro resultado constatado, o que será observado para os demais modelos nas próximas seções, é que o modelo de regressão logística não foi muito satisfatório quando comparado com os modelos de árvore. Apenas o modelo de regressão logística com *step* terá um desempenho relativamente melhor.

Para os modelos de *Random Forest* para prever o time campeão, foi considerado uma previsão de 0.25 como limiar para classificar um time como campeão ou não, uma vez que nenhum valor foi superior a 0.5, como é comumente observado.

O melhor modelo encontrado para a classificação do campeão foi o *Random Forest* após a 17ª rodada, considerando todas as variáveis referentes ao time mandante. Ele apresentou o F1 de 100%, isto é, todos as previsões para a classificação dos dados de treino como positivo foram corretas.

Tabela 3: Matriz de confusão do modelo *Random Forest* após a 17ª rodada, considerando apenas as variáveis relacionadas aos times mandantes

		Valor Predito	
		Negativo	**Positivo**
Valor Real	**Negativo**	57	0
	Positivo	0	3

A matriz de confusão na Tabela 3 exibe as previsões do modelo. Como foram considerados os dados de três campeonatos para teste, é notório que o modelo conseguiu prever corretamente o campeão em todos os casos, o que indica uma precisão e sensibilidade de 100%. Para este modelo, foi utilizado um m_{try} igual a 6 e um *ntree* igual a 500.

Para entender melhor esse modelo, é necessário avaliá-lo mais a fundo e quais variáveis foram importantes para a sua confecção.

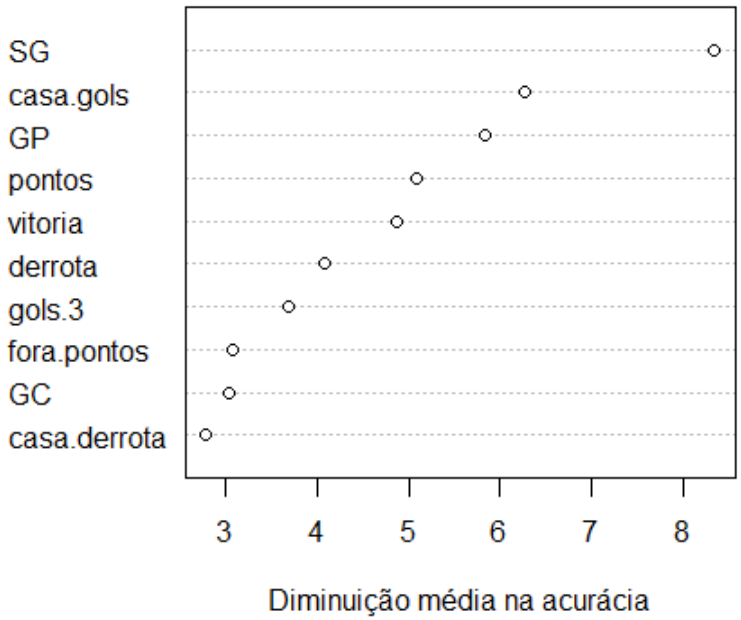

Figura 5: TOP 10 variáveis mais importantes do modelo para prever o campeão com base na diminuição média na acurácia

A Figura 5 apresenta as 10 variáveis mais relevantes do modelo, sendo apresenta-

das de cima para baixo. A partir dele, é possível observar a diminuição média da acurácia, a qual refere-se à redução média da precisão ou acurácia de um modelo de classificação após a remoção de uma variável específica. Essa medida avalia o impacto de cada variável na capacidade preditiva do modelo. Uma diminuição média da acurácia maior indica que a remoção da variável resulta em uma perda maior na precisão do modelo, enquanto uma diminuição média da acurácia menor indica que a variável tem menos influência na precisão do modelo.

É importante ressaltar que a variável *SG* se destacou como a mais relevante, apresentando uma diminuição média na acurácia de aproximadamente 8.5. Isso indica que a remoção dessa variável resultaria em uma perda significativa na precisão do modelo. Em seguida, tem-se a variável *casa.gols* como a segunda mais relevante, com uma diminuição média na acurácia de cerca de 6.5.

Esses resultados demonstram que variáveis mais "básicas", ou seja, aquelas que a próprio tabela do Campeonato Brasileiro apresenta ao entrar no *site* oficial, são relevantes para entender os principais fatores que contribuem para o sucesso de um time no campeonato

Esses resultados evidenciam que variáveis mais facilmente acessíveis, isto é, aquelas presentes na própria tabela do Campeonato Brasileiro disponível no *site* oficial, como o saldo de gols ou quantidade de vitórias, são relevantes para compreender os principais fatores que contribuem para o sucesso de um time no campeonato.

4.1.2 Modelo Previsão Libertadores

Como visto na seção 4.1, são apresentados na Tabela 4 os 10 melhores modelos para a previsão dos 4 times que se classificam para a Libertadores. O melhor modelo elegido será submetido a uma análise mais detalhada.

Tabela 4: TOP 10 modelos de previsão para os que avançam para a Libertadores ordenados por F1

Método	**Variáveis**	**Nº de Rodadas**	**Acurácia**	**Kappa**	**Sensibilidade**	**Precisão**	**F1**
Random Forest	completo	17	0.967	0.889	1.000	0.833	0.909
Random Forest	completo	15	0.950	0.828	1.000	0.750	0.857
Random Forest	sem gols	17	0.950	0.828	1.000	0.750	0.857
Rpart	completo	20	0.933	0.815	0.750	1.000	0.857
Rpart	sem vis	20	0.933	0.815	0.750	1.000	0.857
Rpart	sem gols	20	0.933	0.815	0.750	1.000	0.857
Random Forest	sem vis	20	0.933	0.778	0.900	0.750	0.818
C5.0	completo	20	0.917	0.762	0.733	0.917	0.815
C5.0	sem vis	20	0.917	0.762	0.733	0.917	0.815
Regressão Logística com Step	sem gols	15	0.933	0.762	1.000	0.667	0.800

Com base nos resultados apresentados na Tabela 4, nota-se que a técnica *Random Forest* e *Rpart* se destacaram entre as TOP 10. Quanto ao número de rodadas, verificou-se que os melhores resultados foram alcançados após a 20ª rodada, conforme era esperado.

O modelo mais eficaz para classificar os times que avançam para a Libertadores foi o *Random Forest* após a 17ª rodada, considerando todas as variáveis. Ele apresentou um F-score de 90.9%, evidenciando um ótimo equilíbrio entre precisão e sensibilidade.

Destaca-se também o modelo *Random Forest* após a 15ª rodada, considerando todas as variáveis, que apresentou um F1 de, aproximadamente, 86%. Embora não tenha sido o melhor modelo, ficando em quinto lugar, é importante ressaltar que, no contexto das apostas, acertar o resultado o mais cedo possível durante o campeonato é vantajoso.

A Tabela 5 mostra a matriz de confusão do melhor modelo escolhido para a classificação dos times que avançam para a Libertadores.

Tabela 5: Matriz de confusão do modelo *Random Forest* após a 17ª rodada, considerando todas as variáveis disponíveis

		Valor Predito	
		Negativo	**Positivo**
Valor Real	**Negativo**	48	0
	Positivo	2	10

Diante do exposto, é possível observar que o modelo obteve uma precisão de 83.3%, acertando corretamente 10 dos 12 resultados previstos. Além disso, todas as classificações feitas para os times que avançam para a Libertadores foram corretas, resultando em uma sensibilidade de 100%.

Para a construção desse modelo, foram utilizados os parâmetros $m_{try} = 12$ e $ntree = 100$, que determinam a seleção das variáveis explicativas e o número de amostras *bootstrap*, respectivamente.

Nesse modelo, as variáveis mais importantes são apresentadas na Figura 6.

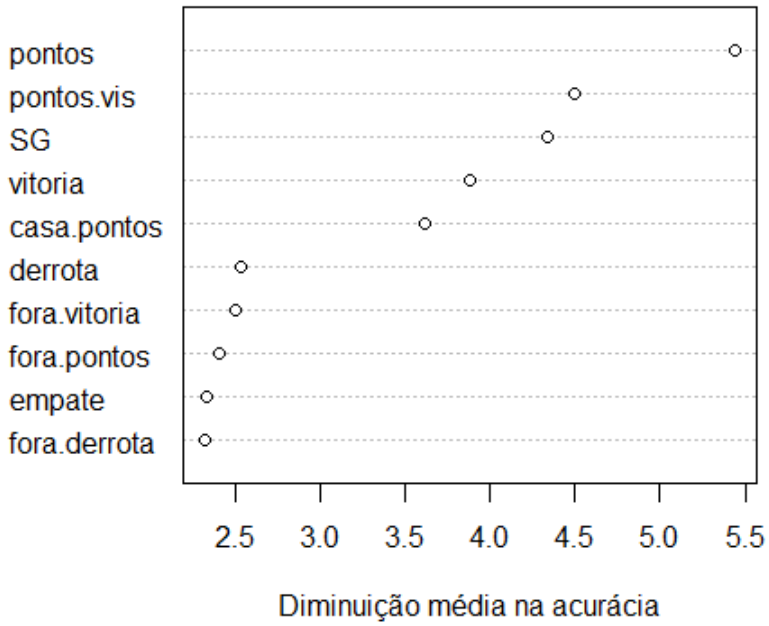

Figura 6: TOP 10 variáveis mais importantes do modelo para prever os times que avançam para a Libertadores com base na diminuição média na acurácia

Analogamente à análise realizada para a seleção de variáveis no modelo de previsão do campeão, é possível identificar as variáveis mais relevantes para o modelo final de previsão dos times que avançam para a Libertadores. Dentre as dez variáveis mais importantes nesse contexto, destacam-se *pontos*, *pontos.vis*, *SG*, *vitoria* e *casa.pontos*, as quais demonstraram forte influência na capacidade preditiva do modelo.

Além das variáveis previamente mencionadas, é interessante notar que outras variáveis também desempenham um papel relevante tanto na previsão do campeão quanto na previsão dos times que avançam para a Libertadores. Entre essas variáveis, destacam-se *SG*, *pontos*, *vitoria*, *derrota* e *fora.pontos*, as quais apareceram consistentemente entre as dez mais importantes em ambos os modelos.

Vale ressaltar que a variável *pontos*, que se destacou como a mais importante no modelo, apresentou uma diminuição média na acurácia de aproximadamente 5.5. Esse resultado indica que a remoção dessa variável do modelo resultaria em uma perda significativa na precisão das previsões. Isso reforça a relevância desse fator como um indicativo para determinar o sucesso de um time na busca pela classificação para a Libertadores, uma vez que os quatro times com maior pontuação irão avançar.

4.1.3 Modelo Previsão Rebaixados

Analogamente as seções 4.1.1 e 4.1.2, são apresentados na Tabela 6 os 10 melhores modelos para a previsão dos 4 times rebaixados para a série B. O melhor modelo elegido será submetido a uma análise mais detalhada.

Tabela 6: TOP 10 modelos de previsão para os rebaixados ordenados por F1

Método	**Variáveis**	**Nº de Rodadas**	**Acurácia**	**Kappa**	**Sensibilidade**	**Precisão**	**F1**
Rpart	sem gols	17	0.867	0.608	0.643	0.750	0.692
C5.0	-	17	0.833	0.561	0.556	0.833	0.667
Regressão Logística com Step	sem vis	20	0.867	0.583	0.667	0.667	0.667
Regressão Logística	sem gols	20	0.850	0.545	0.615	0.667	0.640
Random Forest	-	20	0.883	0.568	0.857	0.500	0.632
Random Forest	sem vis	20	0.883	0.568	0.857	0.500	0.632
Regressão Logística com Step	-	17	0.850	0.516	0.636	0.583	0.609
Regressão Logística com Step	sem gols	20	0.850	0.516	0.636	0.583	0.609
Random Forest	sem vis	17	0.867	0.524	0.750	0.500	0.600
Rpart	sem gols	15	0.833	0.479	0.583	0.583	0.583

Após analisar os resultados das medidas para cada modelo proposto, foi constatado que a maioria dos modelos com melhor desempenho utilizou a técnica de *Random Forest*. Ao comparar os valores de F1 nas Tabelas 2, 4 e 6, observa-se que os modelos para prever os times rebaixados apresentaram um desempenho inferior em comparação com os modelos para prever o campeão e os times que avançam para a Libertadores.

No entanto, o modelo *Rpart*, ao utilizar todas as variáveis, exceto as relacionadas a gols, após a 17ª rodada, obteve o melhor desempenho com um F-score de 66.7%. A matriz de confusão desse é apresentado na Tabela 7.

Tabela 7: Matriz de confusão do modelo *Rpart* após a 17ª rodada, considerando todas as variáveis, exceto aquelas relacionadas a gols

		Valor Predito	
		Negativo	**Positivo**
Valor Real	**Negativo**	43	5
	Positivo	3	9

Pela análise da matriz de confusão apresentada na Tabela 7, verifica-se que o modelo acertou 9 dos 12 resultados previstos, obtendo uma precisão de 75%. No entanto, o modelo classificou erroneamente que 14 times seriam rebaixados, quando na realidade apenas 9 foram, resultando em uma sensibilidade de aproximadamente 64%.

Como mencionado antes, todos os modelos desenvolvidos para prever os times

rebaixados apresentaram um desempenho inferior em comparação aos modelos utilizados para prever os times campeões e os que avançam para a Libertadores.

As variáveis utilizadas para montar esse modelo foram saldo de gols, pontos visitante, gols prós, gols contra e pontos. Sua árvore de decisão é ilustrado na Figura 7.

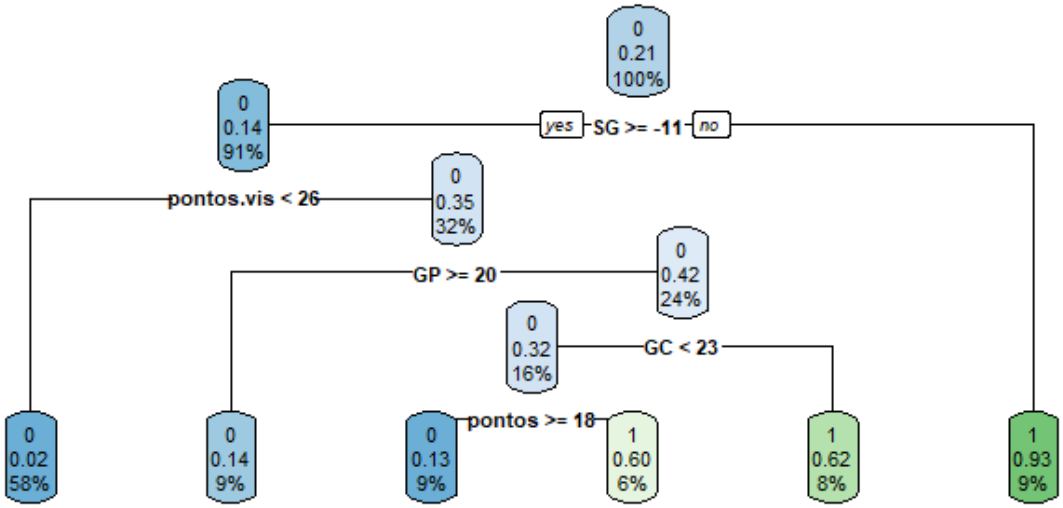

Figura 7: Árvore de decisões do modelo *Rpart* para classificar os times rebaixados

Na raiz da árvore, são apresentadas todas as observações. O valor de 0.21 indica a probabilidade inicial de um time ser classificado como "positivo", ou seja, rebaixado. Os nós à esquerda representam os times que possuem um saldo de gols igual ou maior do que -11, enquanto os nós à direita representam aqueles com saldo de gols menor do que -11. Essa interpretação se aplica aos demais nós da árvore.

Ao final da árvore, verifica-se que 23%, ou seja, 37 times do conjunto de treinamento foram classificados como rebaixados. É importante ressaltar que, nesses dados, um total de 32 times foram efetivamente rebaixados.

Um time que, na 17ª rodada, apresenta um saldo de gols menor do que -11 possui uma probabilidade de 93% de ser classificado como um time rebaixado. Por outro lado, um time que possui saldo de gols superior a -11 e ponto.vis menor do que 26 tem probabilidade de apenas 2% de ser classificado como um time rebaixado.

5 Conclusão

O objetivo deste trabalho foi estudar e avaliar os modelos para prever o time campeão, os quatro times que avançam para a Libertadores e os quatro times rebaixados para a Série B. Os melhores resultados foram obtidos utilizando as técnicas de *Random Forest* e árvore de classificação (*Rpart*).

Os métodos de *Random Forest* e Árvore de Classificação demonstraram ser os melhores em termos do valor de F1-score para as previsões. Além disso, o método de regressão logística com *step* apresentou um desempenho eficiente na previsão dos times rebaixados. No entanto, o método de regressão logística sem a técnica de seleção de variáveis teve um desempenho ruim na previsão, aparecendo entre os top 10 apenas uma vez e somente na previsão do time rebaixado.

A desvantagem da base de dados utilizada nas modelagens deste trabalho é a falta de informações sobre os jogadores e o desempenho dos times em outros campeonatos que ocorrem simultaneamente ao Campeonato Brasileiro de Futebol, como a Copa do Brasil. Isso impede a análise da influência de situações que poderiam ser relevantes, tais como lesões, número de cartões amarelos e vermelhos, desempenho do time em outros campeonatos, etc.

Para trabalhos futuros, seria possível realizar previsões utilizando um número diferente de rodadas, além das 10, 12, 15, 17 e 20 utilizadas neste trabalho. Além disso, seria possível empregar outras técnicas de seleção de variáveis para a regressão logística, bem como outros métodos de aprendizado de máquina mais robustos.

Referências

BREIMAN, L. *Classification and regression trees*. [S.l.]: Routledge, 2017.

BRUCE, A.; BRUCE, P. *Estatística prática para cientistas de dados*. [S.l.]: Alta Books, 2019.

CBF. *Brasileirão Série A*. 2022. Disponível em: ⟨https://www.cbf.com.br/futebol-brasileiro⟩.

CHAMBERS, J.; HASTIE, T.; PREGIBON, D. Statistical models in s. In: SPRINGER. *Compstat: proceedings in computational statistics, 9th symposium held at Dubrovnik, Yugoslavia, 1990*. [S.l.], 1992. p. 317–321.

FACT.MR. *Sports Betting Market*. 2022. Disponível em: ⟨https://www.factmr.com/report/sports-betting-market⟩.

HUANG, J. Z. *An introduction to statistical learning: With applications in r by gareth james, trevor hastie, robert tibshirani, daniela witten*. [S.l.]: Springer, 2014.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398.

KUHN, M. et al. Over-fitting and model tuning. *Applied predictive modeling*, Springer, p. 61–92, 2013.

SHUNG, K. P. Accuracy, precision, recall or f1. *Towards data science*, v. 15, n. 03, 2018.

VEROUTSOS, E. *The Most Popular Sports In The World*. 2022. Disponível em: ⟨https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html⟩.

Apêndice

A Descrição das variáveis

Quadro 3: Descrição das variáveis

Variável	**Descrição**
ano	ano do campeonato
time	nome do time participante
vitoria	quantidade de jogos que o time ganhou
empate	quantidade de jogos que o time empatou
derrota	quantidade de jogos que o time perdeu
pontos	soma dos pontos (3 se ganhou, 1 se empatou, 0 se perdeu)
GP	gols pró
GC	gols contra
SG	saldo de gols
gols.min	menor placar marcado
gols.max	maior placar marcado
pontos.vis	soma dos pontos dos times adversários
gols.min.vis	menor placar marcado pelo time adversário
gols.max.vis	maior placar marcado pelo time adversário
casa.jogos	quantidade de jogos em casa
casa.vitoria	quantidade de jogos que ganhou jogando em casa
casa.empate	quantidade de jogos que empatou jogando em casa
casa.derrota	quantidade de jogos que perdeu jogando em casa
casa.pontos	soma de pontos jogando em casa
casa.gols	soma de gols marcados jogando em casa
casa.gols.min	menor placar marcado jogando em casa
casa.gols.max	maior placar marcado jogando em casa
fora.vitoria	quantidade de jogos que ganhou jogando fora de casa
fora.empate	quantidade de jogos que empatou jogando fora de casa
fora.derrota	quantidade de jogos que perdeu jogando fora de casa
fora.pontos	soma de pontos jogando fora de casa
fora.gols	soma de gols marcados jogando fora de casa
fora.gols.min	menor placar marcado jogando fora de casa
fora.gols.max	maior placar marcado jogando fora de casa
casa.gols.vis	soma de gols marcados pelos times adversários jogando em casa

Continua na próxima página

Variável	**Descrição**
casa.gols.min.vis	menor placar marcado pelos times adversários jogando em casa
gols.min.7	menor placar marcado nos últimos 7 jogos
casa.gols.max.vis	maior placar marcado marcados pelo time visitante jogando em casa
fora.pontos.vis	soma de pontos dos times adversários jogando fora de casa
fora.gols.vis	soma de gols marcados pelos times adversários jogando fora de casa
fora.gols.min.vis	menor placar marcado marcados pelos times adversários jogando fora de casa
fora.gols.max.vis	maior placar marcado pelos times adversários jogando fora de casa
pontos.3	soma de pontos nos últimos 3 jogos
pontos.min.3	menor pontuação adquirida nos últimos 3 jogos
pontos.max.3	maior pontuação adquirida nos últimos 3 jogos
gols.3	soma de gols marcados nos últimos 3 jogos
gols.min.3	menor placar marcado nos últimos 3 jogos
gols.max.3	maior placar marcado nos últimos 3 jogos
pontos.min.3.vis	menor pontuação adquirida pelo time adversário nos últimos 3 jogos
pontos.max.3.vis	maior pontuação adquirida pelo time adversário nos últimos 3 jogos
gols.3.vis	soma de gols dos times adversários nos últimos 3 jogos
gols.min.3.vis	maior placar marcado pelo time adversário nos últimos 3 jogos
gols.max.3.vis	maior placar marcado pelo time adversário nos últimos 3 jogos
pontos.5	soma de pontos nos últimos 5 jogos
pontos.min.5	menor pontuação adquirida nos últimos 5 jogos
pontos.max.5	maior pontuação adquirida nos últimos 5 jogos
gols.5	soma de gols marcados nos últimos 5 jogos
gols.min.5	menor placar marcado nos últimos 5 jogos
gols.max.5	maior placar marcado nos últimos 5 jogos
pontos.min.5.vis	menor pontuação adquirida pelo time adversário nos últimos 5 jogos
pontos.max.5.vis	maior pontuação adquirida pelo time adversário nos últimos 5 jogos
gols.5.vis	soma de gols dos times adversários nos últimos 5 jogos

Continua na próxima página

Variável	**Descrição**
gols.min.5.vis	maior placar marcado pelo time adversário nos últimos 5 jogos
gols.max.5.vis	maior placar marcado pelo time adversário nos últimos 5 jogos
pontos.7	soma de pontos nos últimos 7 jogos
pontos.min.7	menor pontuação adquirida nos últimos 7 jogos
pontos.max.7	maior pontuação adquirida nos últimos 7 jogos
gols.7	soma de gols marcados nos últimos 7 jogos
gols.min.7	menor placar marcado nos últimos 7 jogos
gols.max.7	maior placar marcado nos últimos 7 jogos
pontos.min.7.vis	menor pontuação adquirida pelo time adversário nos últimos 7 jogos
pontos.max.7.vis	maior pontuação adquirida pelo time adversário nos últimos 7 jogos
gols.7.vis	soma de gols dos times adversários nos últimos 7 jogos
gols.min.7.vis	maior placar marcado pelo time adversário nos últimos 7 jogos
gols.max.7.vis	maior placar marcado pelo time adversário nos últimos 7 jogos
forca.media	força média do time (média da soma dos pontos dos times com o qual jogou)
campeao1	campeão do Campeonato
campeao4	os 4 times que avançam para a Libertadores
derrotado4	os 4 times rebaixados

Anexo

A Resultados dos Campeonatos da Série A de 2012 a 2022

Quadro 4: Campeões da série A de 2012 a 2022

Campeão		
Ano	**Time**	**Pontos**
2012	Fluminense	77
2013	Cruzeiro	76
2014	Cruzeiro	80
2015	Corinthians	81
2016	Palmeiras	80
2017	Corinthians	72
2018	Palmeiras	80
2019	Flamengo	90
2020	Flamengo	71
2021	Atlético Mineiro	84
2022	Palmeiras	81

Quadro 5: Times classificados para a libertadores de 2012 a 2022

Libertadores								
Ano	**Time**	**Pontos**	**Time**	**Pontos**	**Time**	**Pontos**	**Time**	**Pontos**
2012	Fluminense	77	Atlético Mineiro	72	Grêmio	71	São Paulo	66
2013	Cruzeiro	76	Grêmio	65	Athletico Paranaense	64	Botafogo	61
2014	Cruzeiro	80	São Paulo	70	Corinthians	69	Internacional	69
2015	Corinthians	81	Atlético Mineiro	69	Grêmio	68	São Paulo	62
2016	Palmeiras	80	Flamengo	71	Santos	71	Atlético Mineiro	62
2017	Corinthians	72	Palmeiras	63	Santos	63	Grêmio	62
2018	Palmeiras	80	Flamengo	72	Internacional	69	Grêmio	66
2019	Flamengo	90	Palmeiras	74	Santos	74	Grêmio	65
2020	Flamengo	71	Internacional	70	Atlético Mineiro	68	São Paulo	66
2021	Atlético Mineiro	84	Flamengo	71	Palmeiras	66	Fortaleza	58
2022	Palmeiras	81	Internacional	73	Fluminense	70	Corinthians	65

Quadro 6: Times rebaixados para a série B de 2012 a 2022

Rebaixados								
Ano	**Time**	**Pontos**	**Time**	**Pontos**	**Time**	**Pontos**	**Time**	**Pontos**
2012	Sport	41	Palmeiras	34	Atlético Goianiense	30	Figueirense	30
2013	Portuguesa	44*	Vasco da Gama	44	Ponte Preta	37	Náutico	20
2014	Vitória	38	Bahia	37	Botafogo	34	Criciuma	32
2015	Avaí	42	Vasco da Gama	41	Goiás	38	Joinville	31
2016	Internacional	43	Figueirense	37	Santa Cruz	31	América	28
2017	Avaí	43	Coritiba	43	Ponte Preta	39	Atlético Goianiense	36
2018	América	40	Sport	39*	Vitória	37	Paraná	23
2019	Cruzeiro	36	Chapecoense	32	Csa	32	Avaí	20
2020	Vasco da Gama	41	Goiás	37	Coritiba	31	Botafogo	27
2021	Bahia	43	Grêmio	43	Sport	38	Chapecoense	15
2022	Ceará	37	Atlético Goianiense	36	Avaí	35	Juventude	22

Observação: Os valores destacados com um asterisco (*) indicam casos em que houve alteração devido a decisões judiciais.