**Previsão de gols marcados em partidas do Campeonato Brasileiro com uso redes neurais artificiais.**

Leandro Augusto Diniz Andrade1\*;Prof. Dr. Rodrigo Rosalis da Silva2

**Previsão de gols marcados em partidas do Campeonato Brasileiro com uso redes neurais artificiais.**

**Palavras-chave:** Aprendizagem de Máquina; Futebol; Previsão de Resultados; Análise de Componentes Principais; Análise de Correspondência;

1. **Introdução**

O futebol é considerado o esporte mais popular do mundo, visto que ao redor do planeta existem em torno de 270 milhões de pessoas entre atletas e árbitros de futebol (CONMEBOL, 2013). No Brasil o futebol foi classificado como o esporte favorito e principal modalidade praticada por 15,3 milhões de pessoas, o que representa 39,9% da população brasileira que regularmente pratica atividades físicas, segundo dados levantados no PNAD de 2015 (IBGE, 2016).

Dentro do cenário de futebol nacional, a Série A do Campeonato Brasileiro é considerado o principal torneio, onde todas as equipes se enfrentam em duelos de turno e returno no formato de “pontos corridos”, em que o time que acumular o maior número de pontos ao final do returno se sagrará campeão.

A predição de resultados do futebol, assim como a maioria dos demais esportes, possui natureza estocástica, visto que o resultado de uma partida de futebol é influenciado por diversos fatores, tais como mas não limitado à mando de campo, sequência dos últimos jogos, média de gols marcados e sofridos, entre muitas outras variáveis.

Dado o grande interesse popular pelo futebol e a complexidade de predição dos resultados das partidas, existe uma crescente demanda pelo mercado de casas de apostas esportivas no Brasil e no mundo. Segundo um levantamento feiro pela H2 Gambling Capital, o mercado de apostas esportivas faturou em 2020 cerca de 12,5 bilhões de reais no Brasil e 59,6 bilhões de dólares em todo o mundo (El País, 2021).

De acordo com uma pesquisa publicada pelo Grand View Research, o mercado de apostas esportivas poderá atingir a marca de 140 bilhões de dólares anuais no mundo em 2028 (Folha de São Paulo, 2022).

A exploração de bancas de apostas no Brasil é considerada uma contravenção penal desde 1946, porém a lei 13.756/2018 autorizou este tipo de atividade de forma temporária até que seja formalizado a sua regulamentação (Outrasmídia, 2022).

Este trabalho consistirá em um modelamento de predição de resultados (número de gols do mandante e visitante) da Série A do Campeonato Brasileiro baseado no uso de redes neurais artificiais, amparado pelo histórico de informações de edições do campeonato nacional entre 2008 à 2020.

1. **Material e Métodos**
2. **Coleta de Dados**

O conjuntos de dados utilizados nesta pesquisa foi formado por um banco de dados histórico do Campeonato Brasileiro, disponível no portal Base dos Dados (BD, 2021), ao qual foram acrescentadas novas informações como por exemplo “média acumulada de gols marcados” e “aproveitamento das últimas 5 partidas”. A esta base de dados processada foram adicionadas informações de outras competições nacionais, estas informações adicionais foram coletadas do site “Campeões do Futebol” (Campeões do Futebol, 2021).

A base de dados consolidada contém duas variáveis de saída (“Gols do clube mandante” e “Gols do clube visitante”) e trinta variáveis de entrada, das quais somente duas são compartilhadas entre os clubes mandantes e visitantes, enquanto que catorze variáveis são dedicadas aos clubes mandantes e outras catorze variáveis são dedicadas aos clubes visitantes da rodada, conforme pode ser observado na (Tabela 1).

Tabela 1. Variáveis de entrada

|  |  |  |
| --- | --- | --- |
| Descrição | Tipo | Compartilhamento |
| Ano do Campeonato | Qualitativa | Compartilhada |
| Número da rodada do campeonato | Qualitativa | Compartilhada |
| Nome do clube | Qualitativa | Dedicada |
| Estado brasileiro de origem do clube | Qualitativa | Dedicada |
| Posição atual do time no campeonato | Qualitativa | Dedicada |
| Participação do clube no campeonato Libertadores do ano corrente | Qualitativa | Dedicada |
| Participação do clube na final da Copa do Brasil do ano anterior | Qualitativa | Dedicada |
| Participação do clube na final do campeonato estadual do ano corrente | Qualitativa | Dedicada |
| Posição do clube como campeão ou vice na última edição do Campeonato Brasileiro | Qualitativa | Dedicada |
| Média acumulada de gols marcados pelo clube no ano corrente | Quantitativa | Dedicada |
| Média acumulada de gols sofridos pelo clube no ano corrente | Quantitativa | Dedicada |
| Valor da equipe titular do clube | Quantitativa | Dedicada |
| Média de idade da equipe | Quantitativa | Dedicada |
| Percentual de aproveitamento do clube na última rodada | Quantitativa | Dedicada |
| Percentual de aproveitamento do clube nas 3 últimas rodadas | Quantitativa | Dedicada |
| Percentual de aproveitamento do clube nas 5 últimas rodadas | Quantitativa | Dedicada |

Fonte: Dados originais da pesquisa

Na seleção das variáveis de entrada deste estudo foram considerados trabalhos anteriores, como Freitas et al. (2019) que verificou a influência do valor de mercado de um clube em seu desempenho ao decorrer do Campeonato Brasileiro.

Outro aspecto influente neste estudo é o fator mando de campo (mandante / visitante) visto a presença de estudos na literatura que indicam um desempenho favorável a equipe mandante no Campeonato Brasileiro ( Kumagawa, 2021; Oliveira et al. 2020).

1. **Análise de Correspondência Simples (Anacor)**

Segundo Fávero e Belfiore (2017), a análise de correspondência simples é uma técnica bivariada utilizada para investigar a associação entre duas variáveis categóricas, cujo resultado permite a avaliação da intensidade dessa associação através de mapas perceptuais.

1. **Principal Component Analysis (PCA)**

Conforme apresentado Fawcett e Provost (2016), conjuntos muito grandes de dados em termos de quantidades de variáveis de entrada podem representar um problema computacional, exigindo equipamentos de alto desempenho, problema este que pode ser resolvido com a seleção de atributos e/ou redução da dimensionalidade antes da modelagem do algoritmo preditivo.

Segundo Araújo et al. (2018), o uso de variáveis de entrada que não impactam o resultado do jogo ou variáveis redundantes (diferentes variáveis que expressam a mesma ideia) só provocaria mais “ruído” ao modelo, desta forma deteriorando o resultado obtido pelo modelo.

A técnica de PCA gera um novo conjunto de variáveis, entre elas não correlacionadas, chamadas de componentes principais, onde cada componente principal é uma combinação linear das variáveis originais. Este método permite explicar a quantidade máxima de variância com o menor número de componentes principais (Ciaburro e Venkateswaran, 2017).

Este trabalho utilizará PCA como forma de se reduzir a dimensionalidade das variáveis de entrada através da criação de novas variáveis não correlacionadas formadas pela combinação linear das variáveis originais.

1. **Otimização de Hiperparâmetros**

Neste trabalho será realizada uma pesquisa de grade aleatória (random search) com validação cruzada 5-folds para avaliar valores distintos de taxa de aprendizado [0,1 / 0,05 / 0,01], neurônios [5 / 10 / 20], camadas [3 / 5 /10] e diferentes funções de ativação [Sigmóide / ReLU / Softmax].

1. **Redes Neurais Artificiais (RNA)**

Uma rede neural artificial pode ser definida como um grupo de neurônios artificiais interconectados e organizados em ao menos três camadas distintas (camada de entrada, camada escondida e camada de saída) que transformam um conjunto de entrada em uma saída desejada, podendo ser classificada como rede neural rasa ou rede neural profunda a depender da quantidade de camadas escondidas conforme indicado na (Figura 1).

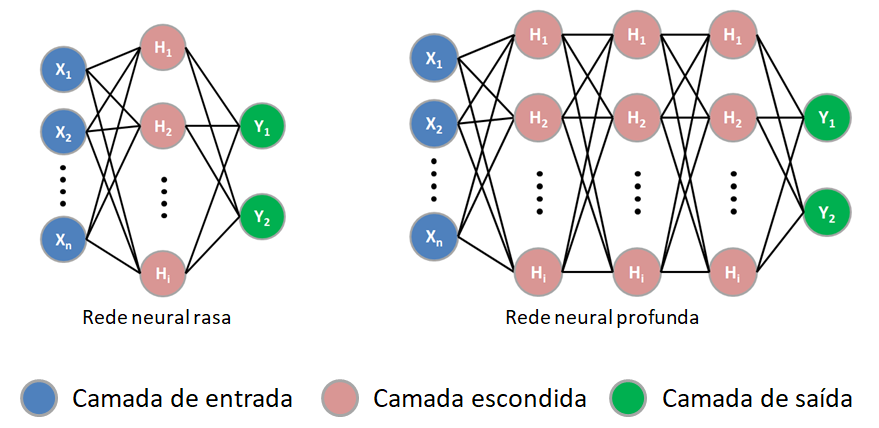


Figura 1 Representação de Rede Neural Artificial

Segundo explicado por Ciaburro e Venkateswaran (2017), os neurônios das redes neurais artificiais funcionam como uma unidade central de processamento onde em seu interior são realizadas operações matemáticas com as variáveis de entrada resultando nas variáveis de saída.

Neste trabalho serão utilizadas três arquiteturas de redes neurais artificiais distintas, as quais serão comparadas com outras três arquiteturas de redes neurais recorrentes em termos de erro quadrático médio.

1. **Avaliação e Desempenho de Modelos**
2. **Validação Cruzada**

O método de validação cruzada, também conhecido como cross-validation, estabelece que o conjunto de dados seja dividido em k subconjuntos de tamanhos aproximadamente iguais, onde em cada iteração um subconjunto será utilizado para teste enquanto que os demais subconjuntos serão utilizados para treino. Ao término das k iterações, o desempenho do modelo será definido pela média dos desempenhos observados em cada iteração (Costa, 2021).

Neste trabalho o banco de dados será dividido em cinco subconjuntos conforme ilustrado na (Figura 2).

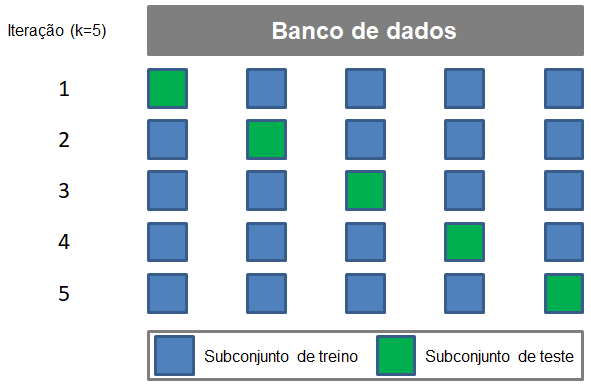


Figura 2 Representação da divisão de subconjuntos

1. **Erro Quadrático Médio (EQM)**

O desempenho dos modelos propostos neste trabalho serão verificados pelo método de erro quadrático médio conforme eq.(1)

(1)

onde, n: representa o tamanho da amostra; yp: representa o valor estimado pelo modelo; e yr: representa o valor real da observação.

1. **Resultados Preliminares**
2. **Análise de Correspondência Simples (Anacor)**

O emprego da técnica de Análise de Correspondência Simples neste trabalho teve como objetivo principal verificar se as variáveis de entrada do banco de dados apresentam associação estatisticamente significativa com as variáveis de saída (“Gols do time mandante” e “Gols do time visitante”) que serão utilizada no modelo de rede neural artificial.

Como requisito para uso da técnica de Anacor, as variáveis quantitativas foram convertidas para variáveis qualitativas de quatro categorias com base em seus respectivos quartis.

Para este trabalho foi considerado um nível de significância 95%, onde o valor crítico do *P-value* é 0,05 e o teste de hipótese pode ser representado da seguinte forma:

H0: Ausência de associação estatisticamente significativa (*P-value* ≥ 0,05)

H1: Presença de associação estatisticamente significativa (*P-value* < 0,05)

A (Tabela 2) indica o resultado da aplicação da técnica de análise de correspondência simples, onde as variáveis de entrada classificadas como pouco significativas em decorrência de seu elevado valor de *P-Value* foram removidas do banco de dados.

Tabela 2. Análise de significância com base na Anacor

|  |  |  |  |
| --- | --- | --- | --- |
| Variável de entrada | P-Value (Gols - man) | P-Value (Gols - vis) | Resultado da análise |
| Número da rodada do campeonato | 0,2009 | 0,6524 | Pouco significativa |
| Média de idade (man) | 0,4619 | 0,2165 | Pouco significativa |
| Média de idade (vis) | 0,2597 | 0,061 | Pouco significativa |
| Finalista da Copa do Brasil (vis) | 0,5687 | 0,0503 | Pouco significativa |
| Finalista do Campeonato Estadual (man) | 0,4414 | 0,3119 | Pouco significativa |
| Finalista do Campeonato Estadual (vis) | 0,1182 | 0,8715 | Pouco significativa |
| Ano do Campeonato | 0,0007 | 0,0031 |  |
| Time (man) | 0,0005 | 0,0007 |  |
| Time (vis) | 0 | 0,6394 |  |
| Posição atual do time no campeonato (man) | 0 | 0,0093 |  |
| Posição atual do time no campeonato (vis) | 0 | 0,0031 |  |
| Valor da equipe titular (man) | 0 | 0 |  |
| Valor da equipe titular (vis) | 0 | 0 |  |
| % de aproveitamento da última rodada (man) | 0,0205 | 0,0125 |  |
| % de aproveitamento das 3 última rodada (man) | 0,0106 | 0,0611 |  |
| % de aproveitamento das 5 última rodada (man) | 0,0192 | 0,1618 |  |
| % de aproveitamento da última rodada (vis) | 0,0009 | 0,0926 |  |
| % de aproveitamento das 3 última rodada (vis) | 0,0209 | 0,0368 |  |
| % de aproveitamento das 5 última rodada (vis) | 0,0004 | 0,0425 |  |
| Participação na Libertadores (man) | 0,0131 | 0,1782 |  |
| Participação na Libertadores (vis) | 0 | 0,0747 |  |
| Finalista da Copa do Brasil (man) | 0,3636 | 0,0061 |  |
| Finalista do Brasileiro (man) | 0 | 0,5431 |  |
| Finalista do Brasileiro (vis) | 0,0468 | 0,9861 |  |
| Média acumulada de gols marcados (man) | 0 | 0,2202 |  |
| Média acumulada de gols marcados (vis) | 0,2363 | 0 |  |
| Média de gols sofridos (man) | 0,0001 | 0 |  |
| Média de gols sofridos (vis) | 0 | 0,552 |  |

Resultados originais da pesquisa

Os resultados obtidos de *P-Value* permitiram a comparação entre variáveis que carregam certo grau redundância de informação, servindo como parâmetro para a decisão de se remover as variáveis classificadas como redundantes nas duas tabelas a seguir:

Tabela 3. Resultado da Anacor – Variáveis reduntantes (% de aproveitamento visitante)

|  |  |  |  |
| --- | --- | --- | --- |
| Variável de entrada | P-Value (Gols - man) | P-Value (Gols - vis) | Resultado da análise |
| % de aproveitamento da última rodada (vis) | 0,0009 | 0,0926 | Variável redundante |
| % de aproveitamento das 3 última rodada (vis) | 0,0209 | 0,0368 | Variável redundante |
| % de aproveitamento das 5 última rodada (vis) | 0,0004 | 0,0425 |  |

Resultados originais da pesquisa

Tabela 4. Resultado da Anacor – Variáveis reduntantes (% de aproveitamento mandante)

|  |  |  |  |
| --- | --- | --- | --- |
| Variável de entrada | P-Value (Gols - man) | P-Value (Gols - vis) | Resultado da análise |
| % de aproveitamento da última rodada (man) | 0,0205 | 0,0125 |  |
| % de aproveitamento das 3 última rodada (man) | 0,0106 | 0,0611 | Variável redundante |
| % de aproveitamento das 5 última rodada (man) | 0,0192 | 0,1618 | Variável redundante |

Resultados originais da pesquisa

1. **Principal Component Analysis (PCA)**

Como requisito para uso da técnica de redes neurais artificiais, as variáveis qualitativas foram transformadas em variáveis quantitativas através da criação de variáveis binárias *dummy* o que, devido ao elevado número de categorias das variáveis “Nome time mandante” e “Nome time visitante”, resultou em um aumento expressivo da quantidade de variáveis de entrada, subindo de duas variáveis para setenta e quatro variáveis.

A técnica de PCA foi empregada neste trabalho com o intuito de se reduzir a dimensionalidade provocada pela conversão das variáveis qualitativas “Nome time mandante” e “Nome time visitante” em variáveis quantitativas.

Segundo Fávero e Belfiore (2017), o critério da raiz latente, o qual consiste na seleção dos fatores derivados de autovalores maiores do que 1, é considerado como um parâmetro recomendável para a definição da quantidade de componentes.

Para que o custo computacional não seja elevado de forma considerável devido ao aumento das variáveis de entrada da rede neural artificial, neste trabalho foi utilizado como referência a quantidade de dez fatores obtidos através da técnica de PCA.

O resultado obtido no PCA pode ser observado na (Tabela 5) e (Figura 3):

Tabela 5. Resultado do PCA

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Quantidade de componentes | Autovalores | Variação acumulada | | 10 | 1,1023 | 15% | | 38 | 1,0016 | 55% | | 39 | 0,9998 | 56% | |  |

Resultados originais da pesquisa

Figura 3 Variação acumulada x Quantidade de Componentes

Fonte: Resultados originais da pesquisa

Conforme pode ser observado na tabela e gráficos acima, o critério da raiz latente indica o uso de trinta e oito componentes representando 55% de variação acumulada das variáveis “Nome time mandante” e “Nome time visitante”.

A referência de dez componentes sugerida neste trabalho representa uma variação acumulada de apenas 15% destas mesmas variáveis.

Dado o custo computacional elevado que seria utilizar os trinta e oito componentes recomendados pelo critério da raiz latente e a baixa variação acumulada no caso em que fossem utilizados somente dez componentes conforme sugerido neste trabalho, o estudo prosseguiu sem a utilização das variáveis de entrada “Nome time mandante” e “Nome time visitante” a partir desta etapa.

1. **Redes Neurais Artificiais**
2. **Referências Bibliográficas**

Araújo, R. T. A.; Ventura, T. B.; Santos, M.; Gomes, C. F. S.; Longaray, A. A. 2018. Proposta de ferramenta de predição utilizando machine learning. In: VIII Seminário e Workshop em Engenharia Oceânica, 2018, Rio Grande, RS, Brasil.

Base dos Dados [BD]. 2021. Competições de Futebol (Organização Transfermarket). Disponível em: <https://basedosdados.org/dataset/e97395c3-e845-4790-86d9-7e91467531c1>. Acesso em: 10 out. 2022

Campeões do Futebol. 2021. Disponível em: < https://www.campeoesdofutebol.com.br>. Acesso em: 10 out. 2022

Ciaburro, G.; Venkateswaran, B. 2017. Neural Networks with R. 1ed. Packt Publishing Ltd., Birmingham, B3 2PB, United Kingdom.

CONMEBOL. 265 milhões de pessoas jogam futebol no mundo inteiro. Disponível em: <https://www.conmebol.com/pt-br/notas-pt-br/265-milhoes-de-pessoas-jogam-futebol-no-mundo-inteiro/>. Acesso em: 10 out. 2022.

Costa, I. B. 2021. Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina. Tese de Doutorado em Ciência da Computação. Campus I, Universidade Federal de Campina Grande, Campina Grande, PB, Brasil. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/20618>. Acesso em: 10 out. 2022.

El País. 2021. Casas de aposta esportiva tomam o Brasil, mas movimentam seus bilhões de reais fora do país. Disponível em: <https://brasil.elpais.com/esportes/2021-09-25/casas-de-aposta-esportiva-tomam-o-brasil-mas-movimentam-seus-bilhoes-de-reais-fora-do-pais.html>. Acesso em: 10 out. 2022.

Fávero, L. P.; Belfiore, P. 2017. Manual de Análise de dados: Estatística e Modelagem Multivariada com Excel, SPSS e Stata. 1ed. Elsevier, Rio de Janeiro, RJ, Brasil.

Fawcett, T.; Provost, F. 2016. Data Science para Negócios – O que você precisa saber sobre mineração de dados e pensamento analítico de dados. 1ed. Alta Books, Rio de Janeiro, RJ, Brasil.

Freitas, W.A.; Santos, Y. D. L.; Ross, S. D.; Alcoforado, L. F. 2019. Investir para conquistar: Uma análise estatística sobre o Campeonato Brasileiro de futebol. In: IV Seminário Internacional de Estatística com R & Python e as tendências de colaboração, 2019, Niterói, RJ, Brasil.

FOLHA DE SÃO PAULO. 2022. Com alto investimento no futebol, sites de apostas esperam regulamentação. Disponível em: <https://www1.folha.uol.com.br/esporte/2022/05/com-alto-investimento-no-futebol-sites-de-apostas-esperam-regulamentacao.shtml>. Acesso em: 10 out. 2022.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2016. Pesquisa Nacional por Amostra de Domicílios. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9127-pesquisa-nacional-por-amostra-de-domicilios.html?=&t=destaques>. Acesso em: 10 out. 2022.

Kumagawa, R. M. S. 2021. A vantagem do mando de campo no futebol: diferenças entre os clubes e o efeito da distância viajada para o Campeonato Brasileiro. REGEN – Revista de Gestão, Economia e Negócios. Volume II, número I: 20-31.

Oliveira, P. V. S. R; Assis, J. V.; Mota Júnior, R. J.; Silva, O. W.; Lavaroto, V. N.; Silva, D. C. 2020. Vantagem de jogar em casa na Série A do Campeonato Brasileiro e na Copa do Brasil. RBFF - Revista Brasileiro de Futsal e Futebol. Volume 12, número 48: 180-186.

OUTRASMIDIAS. 2022. Apostas: das bancas clandestinas à elite do futebol. Disponível em: <https://outraspalavras.net/outrasmidias/apostas-esportivas-da-clandestinidade-a-elite-do-futebol/>. Acesso em: 10 out. 2022.

Silva, A. B. L.; Barros, K. N. N. O.; Albuquerque, M. A. A. 2020. Modelagem via árvore de decisão para previsão de jogos de futebol. Reaserch, Society and Development. Volume 9, número 9.