Predizendo os Vencedores dos *Playoffs*: Um Estudo de Caso com Aprendizado de Máquina em Partidas de Futebol Americano

Danielle Regina Bernardes[1], **João Fernando V. Franciscon**[1], **Fernando Rafael Araújo**[1], **Marcos Paulo de Oliveira**[1], **Robson P. Bonidia**[1,*]

[1]Grupo BioFatecou - Faculdade de Tecnologia de Ourinhos (Fatec)
Ourinhos - SP - Brasil

danielle.bernardes@fatec.sp.gov.br, joao.franciscon01@fatec.sp.gov.br

fernando.araujo21@fatec.sp.gov.br, marcospaulo.dev@gmail.com, bonidia@usp.br

Resumo. *Com o avanço no desenvolvimento de tecnologias, a análise de informações sobre esportes tornou-se uma questão cada vez mais desafiadora. Além disso, as bases de dados disponíveis para estudos de previsão de resultados são limitadas. Considerando isso, neste artigo serão estudados conceitos relacionados ao futebol americano e modelagem de algoritmos de Aprendizado de Máquina (AM), aplicados para a predição de times ganhadores ou perdedores com base nos dados das últimas temporadas. Como resultados, é possível observar que técnicas de AM, quando combinadas com uma quantidade expressiva de dados e com os devidos tratamentos, podem fornecer bons resultados na previsão de resultados de partidas esportivas.*

Abstract. *With the advancement in technology development, the analysis of sports information has become an increasingly challenging issue. Furthermore, the available databases for prediction studies of outcomes are limited. Considering this, this article will explore concepts related to American footboall and the modeling of Machine Learning (ML) algorithms, applied to the prediction of winning or losing teams based on data from past seasons. As a result, it can be observed that ML techniques, when combined with a substantial amount of data and proper preprocessing, can yield promising results in forecasting sports match outcomes.*

1. Introdução

Com os avanços das tecnologias e aumento da capacidade de armazenamento e processamento de dados, uma imensa quantidade de dados relacionados a eventos esportivos é produzida, levando a um aumento de interesse do público [Horvat and Job 2020]. A variedade de dados, originada das várias modalidades, torna os sistemas de *big data* esportivos mais desafiadores [Patel et al. 2020]. De acordo com [Liu et al. 2020, Haiyun and Yizhe 2020], as análises de dados relacionadas à indústria esportiva incluem desde desempenho de atletas e saúde até as

[*]**Autor Correspondente:** rpbonidia@gmail.com, bonidia@usp.br
GitHub: https://github.com/Bonidia/DataBase-NFL

estatísticas de treinamento e de jogos, e podem auxiliar a equipe técnica e atletas no treino diário e no desenvolvimento de estratégias de jogo, tornando-se indispensáveis para vencer competições [Liu et al. 2020, Haiyun and Yizhe 2020].

Tais análises podem trazer inúmeros benefícios para esportes de massa e competições oficiais [Liu 2020]. Por exemplo, gerenciando e analisando a aptidão habitual e o desempenho físico de um atleta, pode-se detectar novos talentos. Como resultado, o conhecimento obtido desses dados esportivos pode ser utilizado para melhor atender treinadores e tomadores de decisão [Mataruna-Dos-Santos et al. 2020, Patel et al. 2020]. Além disso, a análise de dados esportivos concentra-se na descoberta do valor dos dados e fornece recursos e informações valiosos para empresas e empresários. Diversos esportes beneficiam-se dessas análises na literatura, principalmente utilizando inteligência artificial, especificamente algoritmos de Aprendizado de Máquina (AM), como basquetebol [Nguyen et al. 2022], voleibol [Muazu Musa et al. 2021], atletismo [Hoog Antink et al. 2021], futebol americano [Horvat and Job 2020] entre outras modalidades.

No entanto, a demanda por abordagens mais automatizadas em diversas modalidades esportivas está aumentando [Herold et al. 2019], incluindo esportes em equipe. Um exemplo é a *National Football League* (NFL), o principal campeonato de futebol americano, que movimenta bilhões de dólares anualmente [Durand et al. 2021]. Como resultado, diversos pesquisadores e apostadores têm se empenhado em enfrentar um grande desafio: a previsão dos resultados das partidas [Beal et al. 2020]. Conforme [Beal et al. 2020], tal desafio é fundamental para muitas partes interessadas no esporte, como as equipes para selecionar suas táticas, bem como as casas de apostas que definem probabilidades. No entanto, prever resultados de jogos esportivos é um desafio complexo que requer a consideração de diversos fatores, tais como estatísticas da equipe, o ânimo/moral dos jogadores e a localização do jogo, entre outros.

Sendo assim, este estudo propõe analisar algoritmos de AM para a previsão de resultados dos *playoffs* da NFL, utilizando como base de dados os últimos 51 anos de jogos. Até onde sabemos, essa é a maior base gerada até o momento para *playoffs*, com mais de 200 atributos por competidor da liga. Finalmente, este estudo busca responder à seguinte questão de pesquisa: **É possível prever os vencedores das partidas de *playoffs* da NFL?**

2. Trabalhos Relacionados

Durante o estudo, foram encontrados trabalhos relacionados que abordam temas similares, conforme apresentado na Tabela 1. Contudo, tais estudos apresentam poucos dados, quando comparados aos mais de 50 anos contidos no conjunto de dados proposto no presente artigo, além de não focarem nos *playoffs* do campeonato. Este estudo escolheu os *playoffs* devido ao expressivo volume financeiro movimentado por apostas nessa fase da competição. Em estimativas da *American Gaming Association* [Association 2022], cerca de 8 bilhões de dólares estavam envolvidos no mercado norte-americano, apenas na grande final.

Tabela 1. Comparativo dos Trabalhos Relacionados

Artigo	Objetivo	Temporadas
[Bosch 2018]	Comparar técnicas de *deep learning* com técnicas clássicas de AM na predição de resultados da NFL.	2009 a 2016
[Herold et al. 2019]	Discutir a aplicação do AM em jogadas ofensivas na NFL.	-
[Beal et al. 2020]	Comparar diferentes modelos de AM na classificação do resultado de partidas da NFL.	2015 a 2019
[Hsu 2020]	Predizer resultados de partidas da NFL utilizando dados do mercado de apostas.	1985 a 2017

3. Materiais e Métodos

3.1. Base de Dados

Conforme apresentado na seção anterior, este trabalho tem entre seus objetivos construir uma base de dados *benchmark* para trabalhar com a predição de resultados da NFL. Sendo assim, propomos a criação de um algoritmo para a raspagem de dados (do inglês *Web Scraping*) com o objetivo de construir uma base de dados dos últimos 51 anos de *playoffs* da NFL. Para o desenvolvimento desse algoritmo de raspagem, utilizamos a linguagem Python, em específico a biblioteca *Beautiful Soup*. Os dados foram extraídos do site *Pro-Football-Reference*[1], uma página da web que concentra diversas estatísticas sobre o futebol americano. Essa fonte tem sido reconhecida como confiável e é frequentemente utilizada por outras mídias renomadas, como a revista de negócios e finanças *Forbes*.

Com o algoritmo desenvolvido e a execução da raspagem de dados, surgiram alguns desafios relacionados à formatação dos dados extraídos. Por exemplo, ao longo dos anos, diversos times de futebol americano tiveram seus nomes alterados. Além disso, algumas estatísticas relacionadas ao regulamento do jogo foram criadas, o que ocasionava o tratamento de valores faltantes. Esses pontos serão explorados mais detalhadamente na próxima seção, onde será apresentado o fluxo de trabalho necessário até que fosse possível aplicar um algoritmo de AM. Como resultado inicial, foi gerada uma base de dados contendo o histórico de todos os jogos no período de 1970 até 2021.

Esses dados foram divididos em informações estatísticas do time da casa e do time desafiante. Ao todo, foram recuperados 262 atributos estatísticos, distribuídos em três grandes subgrupos de informações: (1) Dados estatísticos referentes às partidas (identificados pela sigla *"gs"*), (2) Dados estatísticos do time da casa (identificados pela sigla *"hts"*) e (3) Dados estatísticos do time desafiante (identificados pela sigla *"ats"*). O conjunto de dados completo pode ser obtido através do código disponibilizado no GitHub[2].

[1]https://www.pro-football-reference.com
[2]https://github.com/Bonidia/DataBase-NFL

3.2. Avaliação Experimental

A idealização deste projeto teve como objetivo demonstrar como seria uma previsão dos resultados de vitória das equipes de futebol americano, com base na coleção de dados extraídos por meio da técnica de raspagem de dados, conforme apresentado na Figura 1. O fluxo de trabalho proposto descreve as etapas realizadas para a criação do algoritmo de previsão dos resultados.

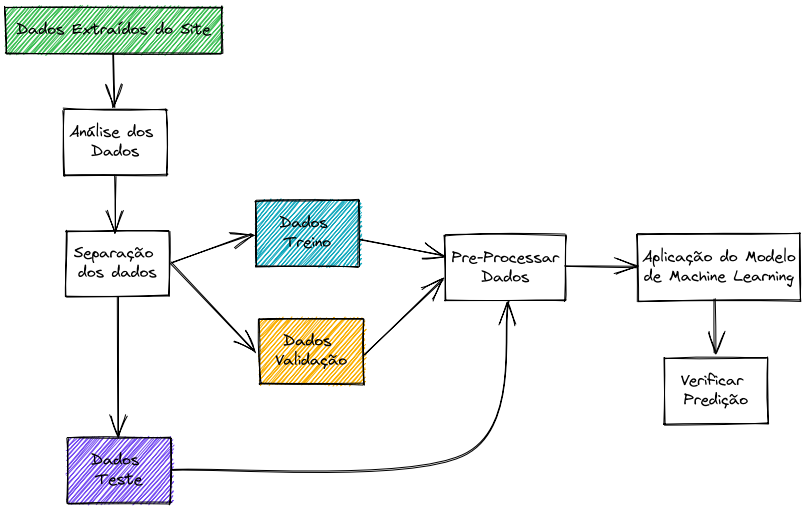

Figura 1. Fluxo de Trabalho para Modelagem do Algoritmo de ML.

A etapa de pré-processamento dos dados é representada pelos processos de preparação, organização e estruturação dos dados do conjunto de entrada, que contribui para as análises necessárias. As técnicas envolvidas nessa etapa são limpeza dos dados, para verificar os dados faltantes e também os dados que possuem algum valor que não pode ser interpretado pelo algoritmo; por exemplo, algum dado que veio com falha da etapa de raspagem dos dados. Para o trabalho apresentado, foram utilizadas técnicas simples de imputação pela média para valores numéricos e imputação pela moda para valores categóricos. Para normalização, foram utilizadas técnicas estatísticas que usam os quartis e a mediana dos dados, por estas serem mais robustas contra *outliers* do que as técnicas tradicionais baseadas na média e na variância.

Na sequência, a transformação dos dados com valores mais adequados para a modelagem do algoritmo de AM, e então por último, é verificada a possibilidade da redução dos dados, ou redução de dimensionalidade, pois o grande volume de atributos torna o processamento dos dados complexo. Para que o processamento do algoritmo de AM não seja tão complexo, devido à existência de um alto número de atributos, foi utilizada a redução de dimensionalidade, cujo objetivo é diminuir a dimensão dos dados, eliminando atributos irrelevantes ou redundantes. Considerando isso, podemos utilizar diferentes abordagens, entre elas a extração e seleção de atributos. Neste trabalho, foram testadas ambas as técnicas, com as de seleção de características apresentando melhores resultados.

A próxima etapa é de fato a execução do modelo de AM para iniciar a verificação das predições. Nesse caso, se o time irá vencer ou perder. Aqui, os algoritmos de AM selecionados para os experimentos foram as Florestas Aleatórias (RF, do inglês *Random Forest*), Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*), *Naive Bayes* (NB) e Redes Neurais Artificiais (ANN, do inglês *Artificial Neural Network*). Os experimentos foram realizados com a base de dados coletada, modelando cada algoritmo de AM. Os dados coletados e limpos foram divididos em conjuntos de treino e teste, usando o método *hold-out*, na proporção de 80% treino e 20% teste. Para cada modelo, foi realizado um *tuning* em seus parâmetros internos, para atingir o potencial máximo de cada experimento. Para avaliação, foram usadas as métricas de acurácia, acurácia balanceada, precisão, *recall* e *F1-score*. Em uma segunda fase de experimentos, foram realizadas análises temporais, considerando a evolução dos times durante a temporada.

4. Resultados e Discussão

Com base nos experimentos realizados, diversos resultados foram gerados, conforme explorado na Tabela 2. Como pode ser visto, o modelo com ANN possui o melhor desempenho, com uma acurácia de 63,5%, seguido pelo RF e NB com uma acurácia de 56,9%, respectivamente. Entre os modelos com as piores desempenhos, encontra-se o SVM, com 53,4% de acurácia. Na segunda fase de experimentos, realizada com a separação dos dados em temporadas, os seguintes resultados foram obtidos, conforme a Tabela 3, usando apenas o melhor modelo gerado, ANN.

Tabela 2. Resultados para Predição dos Vencedores dos *Playoffs*.

Classificador	Acurácia	Acurácia bal.	Precisão	Recall	F1-score
RF	0.569	0.505	0.500	0.040	0.070
SVM	0.534	0.513	0.450	0.360	0.400
NB	0.569	0.525	0.500	0.200	0.300
ANN	**0.635**	**0.604**	**1.000**	**0.208**	**0.345**

Tabela 3. Resultados com a Separação dos Dados em Temporadas.

Anos (1970-)	Acurácia	Acurácia bal.	Precisão	Recall	F1-Score
2017	0.590	0.540	0.484	0.280	0.355
2018	0.562	0.549	0.604	0.250	0.354
2019	0.517	0.517	0.526	0.378	0.440
2020	0.593	0.586	0.615	0.409	0.491

Ao considerar a evolução de cada time durante uma mesma temporada e entre temporadas diferentes, observamos ganhos marginais em todas as métricas utilizadas. Isso indica um potencial significativo nesse tipo de análise. Para uma melhor compreensão da tomada de decisão dos modelos, realizamos uma análise de interpretabilidade. O objetivo foi compreender as características mais importantes do conjunto de dados que influenciam na predição dos vencedores dos *playoffs*. Para

realizar essa análise de interpretabilidade, utilizamos a biblioteca SHAP (do inglês, *SHapley Additive exPlanations*) [Lundberg and Lee 2017].

Essa biblioteca fornece informações que possibilitam uma explicação do modelo de AM. Por meio dela, identificamos as 8 características mais importantes que contribuem para as previsões, incluindo: **(1)** Tentativa mais longa de ação tomada pelo ataque, que significa avançar a bola correndo com ela, em vez de passar ou chutar; **(2)** *First Downs* marcados após penalidade; **(3)** Defesa ganha posse da bola; **(4)** Percentual de tempo que uma equipe chega à zona vermelha e marca um *touchdown*; **(5)** Jogadas seguintes ao time possuir a posse da bola; **(6)** Pontos de *touchdowns* depois de o time entrar na zona vermelha; **(7)** Média de jardas de tentativas de passe de um *Quarterback* enquanto contabiliza *touchdowns* e interceptações; **(8)** Penalidades cometidas pelo time.

Vale ressaltar que essas características estão diretamente relacionadas à qualidade das jogadas e podem ser interpretadas como métricas de desempenho das equipes. Elas resumem em variáveis numéricas uma partida (ou temporada) inteira. A combinação das características mencionadas acima com os resultados de previsão do modelo de AM, que demonstrou superar uma abordagem de aposta aleatória, sugere que esses dados podem ser considerados relevantes para complementar a experiência das pessoas que realizam apostas. Essa abordagem pode contribuir para aumentar a probabilidade de apostas bem-sucedidas, o que é o objetivo deste trabalho.

5. Conclusão

A partir do estudo desenvolvido, é possível observar que eventos esportivos são extremamente complexos de serem modelados matematicamente e mesmo o melhor dos modelos continua longe de representar a realidade. Fatores emocionais dos participantes, por exemplo, têm um grande peso em seu desempenho e no resultado das partidas, deixando o problema mais desafiador. Quanto aos modelos utilizados, ANNs tendem a performar melhor do que algoritmos clássicos. No entanto, é importante notar que a qualidade dos dados se fez mais importante do que a escolha do modelo. Finalmente, concluímos que algoritmos de AM, quando combinadas com uma quantidade expressiva de dados e com os devidos tratamentos, têm um grande potencial na predição de resultados esportivos, auxiliando na tomada de decisão de apostadores, amadores ou profissionais.

Referências

Association, A. G. (2022). Super bowl lvi wagering estimates. Disponível em https://www.americangaming.org/resources/super-bowl-lvi-wagering-estimates/. Acesso em: 17 agosto 2022.

Beal, R. J., Norman, T., and Ramchurn, S. (2020). A critical comparison of machine learning classifiers to predict match outcomes in the nfl. *International Journal of Computer Science in Sport*, 19(2).

Bosch, P. (2018). Predicting the winner of nfl-games using machine and deep learning. *Vrije universiteit, Amsterdam*.

Durand, R. B., Patterson, F. M., and Shank, C. A. (2021). Behavioral biases in the nfl gambling market: Overreaction to news and the recency bias. *Journal of Behavioral and Experimental Finance*, 31:100522.

Haiyun, Z. and Yizhe, X. (2020). Sports performance prediction model based on integrated learning algorithm and cloud computing hadoop platform. *microprocessors and microsystems*, 79:103322.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., and Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6):798–817.

Hoog Antink, C., Braczynski, A. K., and Ganse, B. (2021). Learning from machine learning: prediction of age-related athletic performance decline trajectories. *GeroScience*, 43(5):2547–2559.

Horvat, T. and Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1380.

Hsu, Y.-C. (2020). Using machine learning and candlestick patterns to predict the outcomes of american football games. *Applied Sciences*, 10(13):4484.

Liu, G., Luo, Y., Schulte, O., and Kharrat, T. (2020). Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34(5):1531–1559.

Liu, Y. (2020). Teaching effect and improvement model of college basketball sports based on big data analysis. In *Journal of Physics: Conference Series*, volume 1533, page 042056. IOP Publishing.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Mataruna-Dos-Santos, L. J., Faccia, A., Helú, H. M., and Khan, M. S. (2020). Big data analyses and new technology applications in sport management, an overview. In *Proceedings of the 2020 International Conference on Big Data in Management*, pages 17–22.

Muazu Musa, R., Abdul Majeed, A. P., Suhaimi, M. Z., Mohd Razman, M. A., Abdullah, M. R., and Abu Osman, N. A. (2021). Performance indicators predicting medallists and non-medallists in elite men volleyball competition. In *Machine Learning in Elite Volleyball*, pages 43–49. Springer.

Nguyen, N. H., Nguyen, D. T. A., Ma, B., and Hu, J. (2022). The application of machine learning and deep learning in sport: predicting nba players' performance and popularity. *Journal of Information and Telecommunication*, 6(2):217–235.

Patel, D., Shah, D., and Shah, M. (2020). The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. *Annals of Data Science*, 7(1):1–16.