ANÁLISE DE DADOS DOS JOGOS DO BRASILEIRÃO NOS ANOS DE 2018 A 2024

Gustavo Camargo[1], Carla Christina de Oliveira Pinto[2], Darlon Vasata[3], Thiago Berticelli Ló[4]
[1] *Instituto Federal do Paraná - Campus Cascavel (gustavocamargo203@gmail.com)*
[2,3,4] *Instituto Federal do Paraná - Campus Cascavel.*

Resumo: O estudo consistiu na análise quantitativa de jogos do Brasileirão, campeonato brasileiro de futebol, entre os anos 2018–2024, foram utilizadas bibliotecas Python e a metodologia aplicada foi a OSEMN. Após limpeza e exploração da base, foram identificados padrões de público, gols e desempenho dos clubes, evidenciando o potencial da Ciência de Dados no esporte, mesmo apesar das limitações da qualidade dos dados.

Palavras-chave: Futebol; Análise de dados; Python.

INTRODUÇÃO

O futebol é um esporte de equipe e grande espetáculo cultural no Brasil, cativando o público a cada temporada. Cerca de 68% dos brasileiros com acesso a internet são fãs de futebol (Kantar, 2025).

A principal competição futebolística do país é o Campeonato Brasileiro de futebol (Brasileirão), campeonato que abrange times de todos os estados do país, tendo uma alta competitividade entre as torcidas. Com isso, o interesse da população pelo esporte tem grande impacto na mídia: a última rodada do Brasileirão de 2024 atingiu cerca de 143,1 milhões de espectadores nos canais televisivos Globo e SporTV.

Os fãs gostam de discutir e saber todos detalhes dos seus times, como: Qual é a média de gols do meu time? Qual time é mais atacante? Qual time é mais defensivo?. Portanto, neste contexto, este trabalho visa explorar os dados do banco de dados do Brasileirão e apresentar os resultados da análise.

O esporte gera um volume de dados massivo a cada temporada, e possibilitou compreender a maneira que poderia ser utilizada para responder algumas perguntas por meio da análise dos dados disponibilizados.

MATERIAL E MÉTODOS

A metodologia utilizada neste trabalho foi o fluxo OSEMN (do inglês *Obtain, Scrub, Explore, Model, Interpret*), que corresponde em português às etapas: Obter, Limpar, Explorar, Modelar e Interpretar, conforme vista na Figura 1:

Figura 1. Demonstração da metodologia.
Fonte: Data Science-PM(2025).

1. *Obtain*: Fase para obter os dados brutos, nosso banco de dados.

2. *Scrub*: Limpar e preparar o banco de dados, verificar valores faltantes, duplos, nulos e etc.

3. *Explore*: Realizar a exploração dos dados com estatística.

4. *Model*: Treinar modelos com o banco de dados já tratados.

5. *Interpret*: Interpretar a análise feita.

Para a limpeza, preparação, exploração e interpretação dos dados do trabalho foi utilizada a linguagem de programação Python. A escolha da linguagem foi feita pelo fato de possuir uma sintaxe clara e de simples compreensão, além de utilizar uma forma de indentação, com uma hierarquia de elementos, deixando o código organizado e limpo (Mckinney, 2014). Além disso, a linguagem possui bibliotecas muito bem estruturadas especialmente para a análise de dados, como a Pandas e Numpy, estas projetadas para manipulação de dados, que permitem a manipulação e a transformação de dados de maneira eficiente, possibilitando a leitura de arquivos em diversos formatos e a estruturação de dados.

Para a criação dos gráficos foram utilizadas as bibliotecas Seaborn e Matplotlib. Seaborn foi usada para gráficos estáticos, de maior facilidade, e Matplotlib para a personalização dos gráficos. O trabalho foi construído no ambiente de desenvolvimento Google Colaboratory, serviço hospedado do notebook Jupyter, especialmente adequado para Aprendizado de Máquina e Ciência de Dados (Google, 2024).

A base de dados utilizada foi retirada da plataforma *Kaggle*, site que compartilha dados, análises, explorações e entre outras funções. Com um volume de 1.83 MB, sendo 35 colunas e 8453 linhas, onde cada linha é um jogo de uma rodada. A base possui informações relacionadas a ano, rodada, nomes dos times mandante e visitante, além de diversas informações numéricas como números de faltas, impedimentos, valor das equipes, idade média dos times, capacidade dos estádios, entre outras. Além destas colunas, no decorrer do trabalho foram adicionadas *features*, que são colunas extraídas a partir das que já existem na base, com a finalidade de completar informações necessárias para a análise, como por exemplo o resultado da partida.

A etapa 4 da metodologia (*Model*) não foi abordada, visto que se trata de um trabalho com objetivo exploratório.

RESULTADOS E DISCUSSÃO

Inicialmente após a coleta dos dados, observou-se a necessidade de realizar o processo de limpeza da base, devido a inconsistências nos dados, com muitos dados nulos, duplicados e a falta de colunas fundamentais para a análise, sendo esta considerada a parte mais complexa e demorada do trabalho. Nesta etapa, como resultados parciais obtivemos uma base de dados adequada para análise, a partir da qual foi possível realizar a análise objetivo deste trabalho.

Na etapa de exploração, foram realizadas divisões de temas, o primeiro tema foi nomeado de (Análises Iniciais), consiste na exploração dos dados sobre público, gols e jogos.

No público buscamos por padrões durante os anos, considerando o fato da pandemia entre 2020 e 2021. Obtivemos, assim, gráficos que demonstram a queda do público nos respectivos anos, e a quantidade de jogos com um certo número de público. Nos gols e nos jogos, foi feita a análise básica também, com informações sobre as médias, medianas, modas, desvios padrões, e coeficientes de variação, além de pequenas tendências.

No segundo tema (Análises dos Times), foca na análise de informações sobres os times, objetivo deste trabalho. Inicialmente foi feita uma análise ampla de todos os times que participaram ao menos uma vez da série A do campeonato brasileiro. No período de

2018 a 2024, obtivemos uma série de tendências, respostas e percentuais. Conforme os gráficos das Figuras 2, 3 e 4, observa-se quais times foram mais ofensivos, quais foram mais eficientes em seus ataques, e o percentual de pontos que cada time fez por rodada.

No gráfico da Figura 2, (times mais ofensivos), o eixo X é o score ofensivo dos times e o eixo Y são os times, sendo os mais ofensivos do topo e os menos abaixo.

No gráfico da Figura 3, (times mais eficientes), o eixo X significa é o percentual de eficiência (gols/chutes) dos times e o eixo Y são os times, seguindo a ideia dos mais eficientes acima.

No gráfico da Figura 4 (percentual distribuído de resultados), a primeira coluna representa o percentual de pontos de derrota, a segunda, o percentual de pontos de empate, a terceira de vitória e cada linha representa um times.

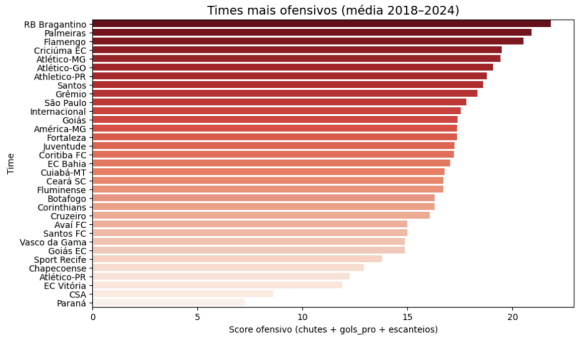

Figura 2. Times mais ofensivos.
Fonte: Autores.

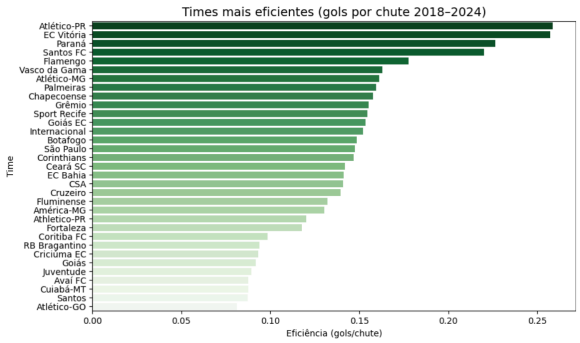

Figura 3. Times mais eficientes.
Fonte: Autores.

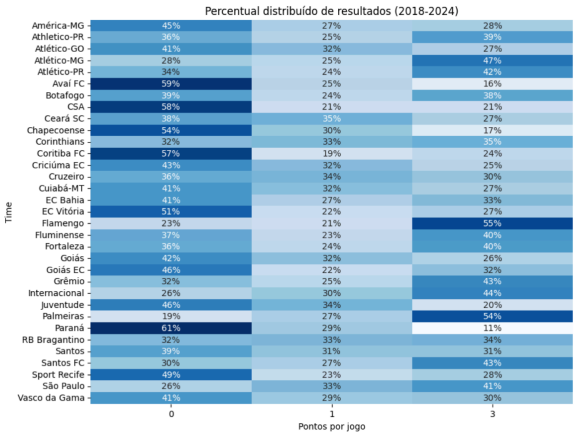

Figura 4. Percentual de pontos por time.
Fonte: Autores.

Dentro do tema Análises dos Times, foi feito uma série histórica de rivalidades, divididas em grupos como times de São paulo e do Sul, nesta parte obtivemos os resultados (vitórias, empates e média de gols de cada time) dos confrontos de 2 times específicos por vez.

No gráfico da esquerda da Figura 5 é apresentado o número de vezes que cada time obteve a vitória sobre o outro, e o número de empates. No gráfico da direita da Figura 5, é apresentado a média de gols marcados por cada time em seus confrontos.

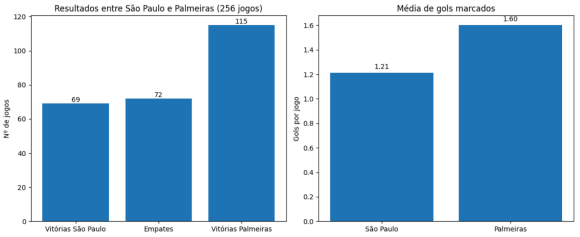

Figura 5. Estatísticas de confrontos entre São Paulo e Palmeiras. Fonte: Autores.

Seguindo assim, foi feita a exploração mais aprofundada de um time específico, no caso escolhemos o clube São Paulo. Conseguimos responder uma série de perguntas, com estatísticas temporais como: qual a média de gols, pontos, faltas, chutes, impedimentos e idade do clube neste período; o quanto o desempenho do time variou jogando como mandante e como visitante; se a idade pesa no número de

vitórias; correlações de ataques, e defesas; quais foram as evoluções dos pontos por ano; análise de estatísticas fora dos padrões.

No gráfico da Figura 6 (Estatísticas de Gols do São Paulo), o eixo Y representa a média de gols marcados, e o eixo X são os respectivos anos. Além disso, tem 3 linhas, onde a amarela representa a mediana dos gols, a linha verde é o desvio padrão, e a linha vermelha é o coeficiente de variação.

No gráfico da Figura 7 (Pontos obtidos sendo Mandante x Visitante), foi feita uma análise para comparação, que busca observar a quantidade de pontos que o time escolhido fez durante os anos, jogando como time mandante e visitante, onde o eixo Y é o número de pontos, e o eixo X são os anos.

O gráfico da Figura 8 (Desempenho do São Paulo por Ano), busca explorar o quanto o time teve de vitórias, empates e derrotas em cada ano, sendo o eixo Y o número de jogos, e o eixo X os anos.

O gráfico da Figura 9 (Linhas de evolução dos pontos), possui uma linha para cada temporada de jogos, nele o eixo Y representa a quantidade de pontos, e o X as rodadas, podendo assim observar como foi a evolução de pontos do time em cada rodada, e em cada temporada.

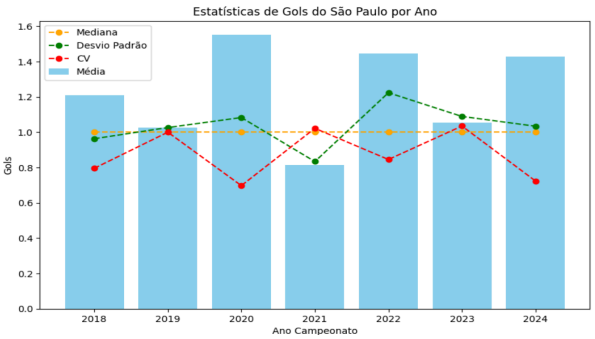

Figura 6. Estatísticas de Gols do São Paulo.
Fonte: Autores.

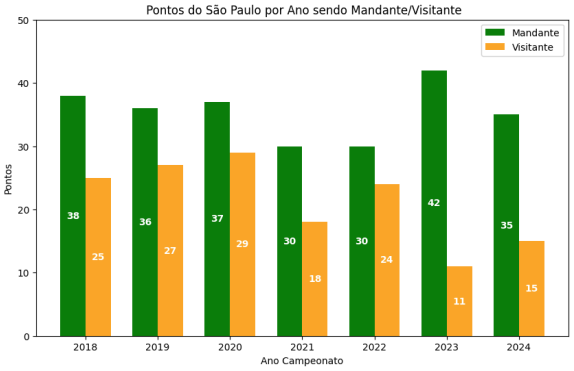

Figura 7. Pontos obtidos sendo Mandante x Visitante.
Fonte: Autores.

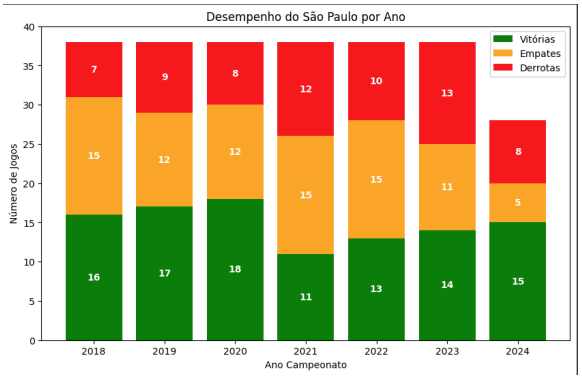

Figura 8. Desempenho de jogos por temporada.
Fonte: Autores.

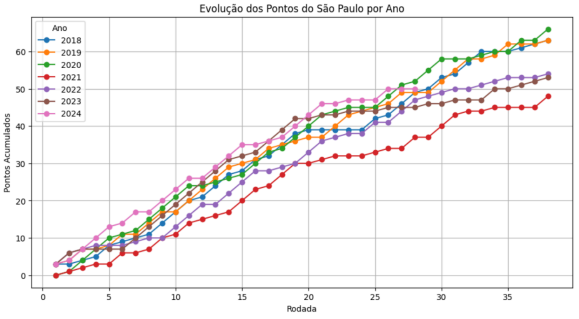

Figura 9. Linhas de evolução dos pontos.
Fonte: Autores.

CONCLUSÃO

A análise do banco de dados dos jogos do brasileirão foi feita, padrões foram identificados, como os dados dos públicos que tiveram uma nos anos da pandemia. Na Tabela 1 foram respondidas as perguntas, e apresentados os *insights* que foram extraídos.

Tabela 1: Insights extraídos.

Perguntas	Respostas
Quais são os dois times que mais marcam pontos?	Palmeiras e o Flamengo
Quais são os times mais ofensivos?	Palmeiras e o Flamengo
Quais são os times mais eficientes nos chutes?	Atlético Paranaense e o Santos
Quais são os times mais defensivos?	Santos e Goiás

Além disso, sobre o São Paulo, time que foi escolhido para uma análise mais aprofundada, percebeu-se que jogar como mandante ou visitante tem uma importância significativa, obtendo muito mais pontos jogando em casa. Contudo, o time apresenta uma variância de pontos, vitórias e derrotas durante os anos, não apresentando uma tendência sequencial de melhora ou piora.

Os resultados permitem compreender, visualizar e analisar *insights* importantes que servem de apoio tanto para torcedores, quanto para analistas e clubes.

O trabalho teve como principal restrição o fato de que o banco de dados possui algumas inconsistências cuja qualidade foram um desafio e exigiu um esforço no tratamento do banco de dados.

Recomenda-se, para trabalhos futuros, a utilização de modelos de aprendizagem de máquina, para a realização de previsões, como o resultado de uma partida para um time específico.

REFERÊNCIAS

DATA SCIENCE-PM. **OSEMN Framework – The Data Science Process.** https://www.datascience-pm.com/osemn/.

GOOGLE. **Google Colaboratory**. https://colab.research.google.com/.

The Matplotlib Development Team. **Pyplot tutorial.** https://matplotlib.org/stable/tutorials/pyplot.html.

The NumPy Developers. **About NumPy.** 2024. https://numpy.org/about/.

The Pandas Development. **About Pandas.** 2024. https://pandas.pydata.org/about/index.html.

MCKINNEY, Wes. **Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython.** Novatec Editora, 2018.

KAGGLE. *Kaggle*. 2025. https://www.kaggle.com/.

Kantar IBOPE Media. **"68% dos brasileiros com acesso à internet são fãs de futebol"**. Kantar IBOPE Media, 26 mai. 2022. https://kantaribopemedia.com/conteudo/68-dos-brasileiros-com-acesso-a-internet-sao-fas-de-futebol/.