UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
GRADUAÇÃO DE ENGENHARIA DE PRODUÇÃO MECÂNICA

André Pegoraro Neto

Avaliação de Desempenho de Jogadores da Série A do Campeonato Brasileiro de Futebol: estudo de caso aplicando o modelo de Gols Esperados (xG)

Florianópolis
2024

André Pegoraro Neto

Avaliação de Desempenho de Jogadores da Série A do Campeonato Brasileiro de Futebol: estudo de caso aplicando o modelo de Gols Esperados (xG)

Trabalho de Conclusão de Curso submetido ao curso de Engenharia de Produção Mecânica do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel em Engenharia Mecânica com habilitação em Produção

Orientador(a): Prof. Ricardo Villarroel Dávalos, Dr.

Florianópolis
2024

Neto, André Pegoraro
Avaliação de Desempenho de Jogadores da Série A do Campeonato Brasileiro de Futebol : estudo de caso aplicando o modelo de gols esperados (xG) / André Pegoraro Neto ; orientador, Ricardo Villarroel Dávalos, 2024.
87 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2024.

Inclui referências.

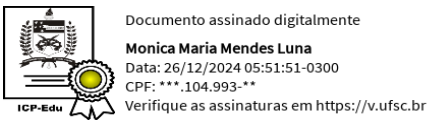
1. Engenharia de Produção Mecânica. 2. aprendizado de máquina. 3. gols esperados. 4. probabilidade. 5. futebol. I. Dávalos, Ricardo Villarroel. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

André Pegoraro Neto

Avaliação de Desempenho de Jogadores da Série A do Campeonato Brasileiro de Futebol: estudo de caso aplicando o modelo de Gols Esperados (xG)

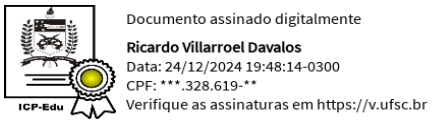
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Engenharia Mecânica com habilitação em Produção e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Produção Mecânica.

Florianópolis, 16 de dezembro de 2024.



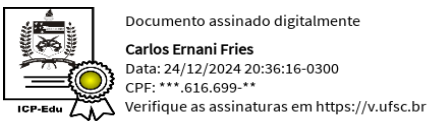
Coordenação do Curso

Banca examinadora



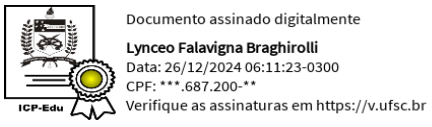
Prof. Ricardo Villarroel Dávalos, Dr.

Orientador(a)



Prof. Carlos Ernani Fries, Dr.

Universidade Federal de Santa Catarina



Prof. Lynceo Falavigna Braghirolli, Dr.

Universidade Federal de Santa Catarina

Florianópolis, 2024

Para minha mãe, Neide.

AGRADECIMENTOS

Agradeço, primeiramente, ao meu pai, Aroldo, que me apoiou e investiu nas minhas decisões. Agradeço à minha mãe de coração, Luciana, que compartilhou muitas histórias comigo desde minha infância.

Aos meus professores, meus mestres, que me ensinaram coisas o suficiente que não sou capaz de mensurar, cada qual com sua forma de ensinar e agir, buscaram extrair o melhor de mim.

Agradeço aos meus familiares, em especial às minhas tias. Cristina e Adriana, irmãs da minha mãe, que me acolheram durante meus três anos de ensino médio. Cristina sempre dizia que eu precisava estudar mais. Elenir, Maristela e Adriane, irmãs do meu pai, que – desde criança – me deram confiança na forma como eu enxergo o mundo.

Agradeço à minha companheira, Bruna, que compartilha dos meus sonhos, das minhas metas de vida e que esteve ao meu lado nos momentos mais difíceis desse semestre.

Aos meus amigos que estiveram comigo durante todo o curso, nas vitórias e nas derrotas. Lucca, meu braço direito. Bruno, apesar de distante, presente nos momentos mais importantes.

Por fim, agradeço ao futebol. Esporte esse que já me presenteou com tantas alegrias, mas também com tantas frustrações. Mais uma vez o futebol me surpreende. Das coisas menos importantes, o futebol é a mais importante.

“Procurem o Senhor de Moivre;
ele sabe essas coisas melhor do que eu.” (Newton *apud* Bernstein, 2018, p. 131)

RESUMO

Este trabalho apresenta um estudo de caso de um modelo preditivo de gols esperados (xG) voltado para a avaliação de desempenho de jogadores da Série A do Campeonato Brasileiro de Futebol. O objetivo principal é avaliar uma ferramenta preditiva capaz de mensurar a probabilidade de uma finalização resultar em gol, com base em variáveis e características contextuais do chute. A pesquisa adota uma abordagem de estudo de caso, utilizando dados de finalizações coletados por *web scraping*. Dois modelos foram estudados e comparativos entre eles foram feitos. Os resultados mostram que um dos modelos (V2) obteve melhor desempenho em comparação ao modelo básico mas inferior a um modelo comercial, evidenciando-se como uma ferramenta eficaz para avaliar a criação e conversão de oportunidades de gol. As análises destacaram os atacantes, meio-campistas e defensores que melhor performaram segundo as métricas estabelecidas para análise.

Palavras-chave: futebol; aprendizado de máquina; regressão logística.

ABSTRACT

This project presents a case study of an expected goals (xG) predictive model aimed at evaluating the performance of players in the Brazilian Serie A Football League. The main objective is to assess a predictive tool capable of measuring the probability of a shot resulting in a goal, based on variables and contextual characteristics of the shot. The research adopts a case study approach, using shot data collected through web scraping. Two models were studied, and comparisons were made between them. The results show that one of the models (V2) achieved better performance compared to the basic model but inferior to a commercial model, proving to be an effective tool for evaluating the creation and conversion of goal opportunities. The analyses highlighted the forwards, midfielders, and defenders who performed best according to the metrics established for analysis.

Keywords: football; machine learning; logistic regression.

LISTA DE FIGURAS

Figura 1 – Diagrama de funcionamento do *CRISP-DM*……..……26
Figura 2 – Requisição *GET* às *API*s do SofaScore em Python…….28
Figura 3 – Tipos de Algoritmos de Aprendizagem de Máquina…….35
Figura 4 – Mapeamento do Processo de modelagem do modelo de xG……..…..42
Figura 5 – Dimensões de um campo de futebol no padrão FIFA……..47
Figura 6 – Dimensões de campos de futebol do Campeonato Brasileiro……48
Figura 7 – Exemplo prático de ângulo formado por finalização no campo……..50
Figura 8 – Diagrama do processo para extração de dados……53
Figura 9 – Partidas adiadas no Campeonato Brasileiro de 2022…….....54
Figura 10 – Coordenadas cartesianas da posição de um jogador no campo…….56
Figura 11 – Coordenadas cartesianas da posição-destino da bola no gol após…….57
Figura 12 – Amostra dos dados obtidos do dataset…….57
Figura 13: Finalizações do S.C. Internacional contra o S.E. Palmeiras em partida válida pelo Brasileirão 2024…….58
Figura 14 – Gráfico do percentual de finalizações por parte do corpo……..…..60
Figura 15 – Gráfico do percentual de gols marcados por situação de jogo…….61
Figura 16 – Distribuição das finalizações que resultaram em gol no Campeonato Brasileiro de 2022……..62
Figura 17 – Distribuição dos Gols pela Largura do Campo (metros).....63
Figura 18 – Distribuição dos Chutes pela Largura do Campo (metros).....64
Figura 19 – Distribuição dos chutes por ângulo de finalização….....65
Figura 20 – Distribuição dos gols por ângulo de finalização….....65
Figura 21 – Distribuição dos Chutes por Coordenada Y da boca do gol…….66
Figura 22 – Gráfico de dispersão da relação entre xG e ângulo……..……..…67
Figura 23 – Gráfico de dispersão da relação entre xG a distância……..…….68
Figura 24 – Comparativo de Gols por Total de Finalizações…….75
Figura 25 – Comparativo de Gols por Total de Finalizações…….76
Figura 26 – Comparativo de Gols por Finalizações no Gol……77
Figura 27 – Comparativo de gols por xG acumulado……..80

LISTA DE QUADROS

Quadro 1 – Matriz de Confusão de Aprendizado de Máquina……..38
Quadro 2 – Métricas para análise de desempenho de jogadores ofensivamente….39

LISTA DE TABELAS

Tabela 1 – Amostra dos dados com xG do Modelo Básico calculado……..44
Tabela 2 – Total de chutes por posição do jogador taticamente……59
Tabela 3 – Total de finalizações por tipo…….60
Tabela 4 – Média das variáveis quantitativas por fator “gol”.....62
Tabela 5 – Coordenadas médias da boca do gol para as finalizações do Campeonato Brasileiro Série A 2022…….68
Tabela 6 – Matriz de correlação entre as variáveis quantitativas e xG do modelo básico…….69
Tabela 7 – Matriz de correlação entre as variáveis quantitativas e xG do modelo V2……..…70
Tabela 8 – Resultado da avaliação do modelo básico pelas métricas de desempenho……..71
Tabela 9 – Comparativo de desempenho entre os Modelos Básico e V2 pelas métricas de desempenho..……...……71
Tabela 10 – Comparativo de desempenho entre os três modelos pelas métricas de desempenho……...……72
Tabela 11 – Os 10 jogadores que mais finalizaram no campeonato……...…….74
Tabela 12 – As 10 melhores taxas de gols por chutes no alvo do campeonato…..…76
Tabela 13 – Os 10 maiores gols esperados acumulados do campeonato……...…78
Tabela 14 – Os 10 jogadores com maior taxa de Gols por xG no campeonato……..79

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
xG	*Expected Goals*
xGOT	*Expected Goals on Target*
DF	*DataFrame*
EDA	*Exploratory Data Analysis* (Análise Exploratória de Dados)
CRISP-DM	*Cross Industry Standard Process for Data Mining*
NaN	*Not a Number*
CI/CD	*Continuous Integration/Continuous Delivery* (Integração e Entrega Contínua)
JSON	JavaScript Object Notation
HTML	Hypertext Markup Language
CSV	Comma-separated value
ML	Machine Learning
DS	Data Science

SUMÁRIO

1 INTRODUÇÃO..... 17
1.1 PROBLEMA..... 19
1.2 OBJETIVOS..... 21
1.3 JUSTIFICATIVA DA PESQUISA..... 21
2 FUNDAMENTAÇÃO TEÓRICA..... 24
2.1 ESTUDOS SOBRE MODELOS DE GOLS ESPERADOS..... 24
2.2 O PROCESSO PADRÃO PARA MINERAÇÃO DE DADOS EM DIVERSAS INDÚSTRIAS (CRISP-DM)..... 25
2.3 EXTRAÇÃO DE DADOS PARA MODELOS DE GOLS ESPERADOS..... 26
2.3.1 Técnica de Extração: Web Scraping..... 26
2.3.2 Interface de Extração dos Dados..... 27
2.3 ANÁLISE EXPLORATÓRIA DE DADOS..... 29
2.3.1 Visualização de Dados..... 29
2.3.2 Estatística Descritiva..... 29
2.3.3 Identificação de Outliers..... 30
2.4 PRÉ-PROCESSAMENTO DE DADOS..... 30
2.4.1 Limpeza de Dados..... 30
2.4.2 Transformação de Dados..... 31
2.4.3 Engenharia de Características..... 32
2.4.4 Divisão dos Dados..... 33
2.5 SELEÇÃO E TREINAMENTO DO MODELO..... 33
2.5.1 Modelos Básicos de Aprendizado de Máquina..... 34
2.5.2 Escolha do Modelo..... 35
2.6. AVALIAÇÃO DO MODELO..... 36
2.7 AVALIAÇÃO DE DESEMPENHO DE JOGADORES..... 39
2.8 CONSIDERAÇÕES FINAIS..... 40
3 PROCEDIMENTOS METODOLÓGICOS..... 41
3.1 TIPO DE PESQUISA..... 41
3.2 ETAPAS METODOLÓGICAS..... 41
3.3 DELIMITAÇÕES DO TRABALHO..... 42
4 ESTUDO DE CASO..... 43
4.1 MODELAGEM..... 43
4.1.1 Modelo Básico de Gols Esperados..... 43
4.1.2 Modelo Avançado de Gols Esperados..... 44
4.2 PRÉ-PROCESSAMENTO DE DADOS..... 45
4.2.1 Limpeza e Tratamento de Dados Inconsistentes..... 46
4.3 TRANSFORMAÇÃO DE DADOS PARA A MODELAGEM..... 46
4.3.1 Desnormalização de Coordenadas..... 46

4.3.2 Cálculo da Distância de Finalização....48
4.3.3 Ângulo da Finalização.... 49
4.4. DADOS DE PARTIDAS DE FUTEBOL.... 52
4.4.1 Dados Espaço-temporais.... 52
4.4.2 Extração dos Dados.... 53
4.4.3 Formato Geral dos Dados.... 54
4.5 ANÁLISE EXPLORATÓRIA DOS DADOS....58
4.5.1 Análises de Finalizações.... 59
4.5.1.1 Total de Finalizações por Posição do Jogador....59
4.5.1.2 Total de Finalizações por Tipo....59
4.5.1.3 Finalizações por Parte do Corpo Utilizada....60
4.5.2 Análises de Gols e Distribuição....61
5 RESULTADOS.... 67
5.1 ANÁLISE DE CORRELAÇÃO ENTRE VARIÁVEIS....67
5.2 AVALIAÇÃO DOS MODELOS.... 70
5.3 VALIDAÇÃO COM MODELO COMERCIAL.... 72
5.4 AVALIAÇÃO DE JOGADORES....73
5.4.1 Taxa de Conversão Geral....73
5.4.2 Taxa de Gols por Finalização no Gol....76
5.4.3 Taxa de Gols por xG.... 78
5.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO....80
6 CONCLUSÃO.... 81

1 INTRODUÇÃO

O conservadorismo no cenário esportivo tem historicamente limitado o surgimento de mudanças significativas, especialmente no futebol. Entretanto, atualmente percebe-se um movimento revolucionário nos esportes, de busca por novas soluções além do tradicional. Nesse contexto, os dados emergem como elementos-chave na formulação de decisões estratégicas, como aquelas relacionadas à compra ou empréstimo de jogadores (Eggels, 2016).

Um dos precursores da análise de dados no esporte é Bill James, estatístico e personagem central do livro *Moneyball: The Art of Winning an Unfair Game.* Hakes e Sauer (2006) fornecem uma avaliação econômica da história de Moneyball, abrangendo o desempenho dos jogadores e sua contribuição para a organização que os emprega dentro da *Major League Baseball*. O livro é um dos fundamentos da análise de dados esportiva e conta a história de como Bill Beane, um *manager* (treinador) que trabalhava no Oakland Athletics, explorou uma lacuna no mercado de transferências do *baseball* americano. Beane, juntamente a Paul dePodesta, seu assistente de escritório, utilizaram análise de dados para direcionar decisões na contratação de atletas, levando à equipe ao vice-campeonato da temporada de 2002 da *Major League Baseball* (Primeira Liga de Beisebol dos EUA), contra adversários que possuíam investimentos mais robustos (Eggels, 2016).

A revolução das estatísticas no beisebol também foi notada em outros esportes, como no futebol. A origem das análises de dados no futebol remonta à década de 1930, quando Charles Reep começou a realizar estudos deste tipo durante partidas de futebol na Inglaterra (Eggels, 2016). Reep anotou dados referentes a mais de 2000 partidas, entre as décadas de 1930 e 1960, o que eventualmente o levou a seu artigo "*Skill and Chance in Association Football*" publicado no Journal of the Royal Statistical Society em 1968 (Eggels, 2016).

Com o aumento do interesse pelo setor de ciência de dados em outros esportes, o futebol está gradualmente adotando esse tipo de abordagem para tomadas de decisões. Segundo Eggels (2016), atualmente, a coleta de dados não é mais realizada por indivíduos como Reep, mas por empresas especializadas, como a Wyscout, Opta e SofaScore. Este processo de coleta de dados do jogo para um *dataset*, conhecido como *scouting*, tem sido visto de forma cada vez mais frequente, envolvendo a observação de eventos nas partidas para obtenção de informações a

níveis individual e coletivo. O *scouting,* no futebol, é o processo de identificação, avaliação e monitoramento de jogadores, utilizando observações diretas, com o objetivo de ajudar clubes e treinadores a tomar decisões informadas sobre contratações, táticas e desenvolvimento de jogadores (Hamil, Walters & Watson, 2010).

Neste contexto futebolístico, destaca-se o Brasil. Dono de cinco títulos mundiais, o Brasil (ainda) é conhecido mundialmente como o “país do futebol”. Mesmo que não tenha levantado o troféu nenhuma vez nos últimos 20 anos, o país do futebol destaca-se por ter uma liga nacional forte. Segundo ranqueamento feito pela IFFHS (2024, 2023, 2022, 2021), o Brasileirão Série A tem figurado entre as quatro ligas mais fortes do mundo, sendo a vencedora em duas ocasiões, em 2021 e 2022. O ranqueamento leva em consideração a performance dos clubes de cada liga em competições internacionais ao longo dos anos. O futebol brasileiro tem dominado a Copa Libertadores da América, competição mais importante do continente americano. Das últimas cinco finais disputadas, somente um clube não-brasileiro figurou entre os finalistas. Ainda, o Brasileirão é um campeonato extremamente competitivo. Apesar de ser possível dizer quem serão os figurantes ao título no início de cada temporada, é difícil prever com exatidão quem será o campeão. O Brasileirão é o segundo campeonato mais imprevisível do planeta, somente atrás do Campeonato Argentino (CNN, 2024). Das últimas dez temporadas, cinco clubes diferentes ergueram a taça de campeão nacional.

A Série A do Brasileirão é a primeira divisão de futebol masculino do Brasil. Organizado pela CBF (Confederação Brasileira de Futebol), é um torneio de pontos corridos composto por 20 clubes em um formato em que todos se enfrentam duas vezes por temporada, em casa e fora de casa. Dessa maneira, o campeonato é composto por dois turnos, cada um com 19 rodadas, totalizando 38 rodadas e 380 jogos. O torneio possui esse formato desde 2006. Em 2023, o bolo total de premiação foi de R$475 milhões (Lance, 2024).

Considerando esses dois cenários, da revolução dos dados no futebol e da alta competitividade na liga nacional brasileira, os modelos baseados em dados surgem como uma alternativa para os clubes brasileiros buscarem tomar decisões mais assertivas. A utilização de modelos baseados em dados por parte dos clubes de futebol apoia tanto a análise de partidas quanto a avaliação de jogadores individuais. Essa análise, que abrange jogadores das próprias equipes e de outras,

permite aos clubes identificar atletas a serem negociados, bem como encontrar os mais adequados para preencher posições vulneráveis em suas próprias equipes.

Nesse sentido, propõe-se realizar um estudo de caso com modelo baseado em dados para analisar o desempenho de jogadores da Série A do Brasileirão, com o objetivo de mostrar o poder de uma ferramenta desta categoria no quesito auxílio à tomadas de decisões no alto escalão dos clubes brasileiros.

1.1 PROBLEMA

Considerando os investimentos realizados pelos clubes brasileiros, faz-se necessário manter boas colocações nas competições nacionais e internacionais ao final de cada temporada. Somente na janela de transferências de 2024, os clubes brasileiros da Série A somaram mais de 1 bilhão de reais em investimentos (Transfermarkt, 2024). No futebol, os resultados das partidas, e portanto as conquistas, são influenciados por uma variedade de fatores, incluindo a sorte. Uma equipe pode dominar completamente seu adversário e ainda assim empatar ou até perder uma partida (Eggels, 2016).

Um clube é uma entidade que emprega milhares de pessoas, transformando a vida de profissionais dentro e fora de campo, fomentando o esporte. Segundo pesquisa Ernst & Young, o futebol brasileiro em 2018 teve impacto de 0,72% no PIB nacional, ao movimentar R$52,9 bilhões (Correio do Estado, 2020). Nesse contexto, ferramentas analíticas podem ser usadas para auxiliar os gestores a tomarem decisões mais assertivas, visando o melhor cenário futuro para os clubes.

A aplicação da ciência de dados no futebol tem ganhado destaque tanto na Europa quanto no Brasil, apesar de possuírem níveis de maturidade distintos. Na Europa, Liverpool e Ajax são exemplos de clubes que possuem departamentos dedicados à análise de dados, liderados por profissionais com formação técnica na área, que contribuem para melhorias no desempenho das equipes. Segundo coluna de Álvarez (2021), físicos teóricos, matemáticos e programadores lideram os departamentos de análise de dados de alguns dos principais clubes da Europa, que atribuem a eles a melhora nos jogos e nas contratações.

No Brasil, o cenário é diferente. De acordo com estudo realizado por Gonçalves e Menezes (2023, p. 4), com dirigentes de clubes de futebol paulistas, “clubes utilizam sistemas de terceiros para analisar contratações ou mesmo os

próprios atletas, enquanto outros focam em ter seus próprios analistas". Ou seja, os clubes brasileiros utilizam dados para alimentar seu processo de tomadas de decisões, mas de forma limitada à contratos com terceiros. O Palmeiras, servindo como exemplo nessa ocasião, comprava dados de outras empresas e os recebia em forma de relatórios prontos e padronizados, como acontece com outros clubes no Brasil (Globo Esporte, 2024). Entretanto, esse cenário está mudando.

O Palmeiras, clube com mais títulos da Série A do Campeonato Brasileiro, fundou, em 2024, o Centro de Ciência de Dados em sua Academia de Futebol. O setor usa tecnologia e matemática para registrar dados personalizados com três objetivos: estudar adversários, avaliar a própria forma de jogar e, ainda, mapear reforços a nível mundial no mercado (Globo Esporte, 2024). O clube compra os dados brutos e, a partir disso, o novo setor fica responsável por desenvolver análises e relatórios personalizados para os tomadores de decisões.

Com isso, entende-se que há uma lacuna que separa a maturidade analítica dos clubes europeus dos clubes brasileiros. Ao passo que existe uma cultura bem consolidada em volta desses aspectos nos clubes sediados no velho mundo, aqui ainda há caminho a ser percorrido para que a ciência de dados esteja plenamente estabelecida. Isso não deve ser encarado como um problema, muito pelo contrário, mas uma oportunidade para desenvolver tecnologias ainda melhores nesta área, justificando o status do Brasil de "país do futebol" citado anteriormente, mas de forma ainda mais contemporânea.

Segundo matéria do Globo Esporte (2024), a equipe do centro de ciência de dados do Palmeiras tem desenvolvido modelos matemáticos para a busca por jogadores. Esses modelos são fórmulas matemáticas aplicadas aos números, que trazem alguma conclusão. "Um modelo conhecido é o Expected Goals" (Oki *apud* Globo Esporte, 2024). O trabalho dos *scouts*, embora invisível, é de extrema importância, contribuindo ativamente para a sustentabilidade e sucesso dos clubes, seja ao nível desportivo ou financeiro, sendo a sua figura já vista como um "gerador de milhões" (Ciência da Bola, 2024).

Ainda, além de auxiliar treinadores e comissões técnicas de clubes, a análise de dados no futebol auxilia jogadores na percepção do seu desempenho nas partidas. Um exemplo disso é Kevin De Bruyne, meio-campista belga que atua pelo Manchester City. Em 2021, o atleta acertou a renovação de seu contrato por £83,2 milhões. De forma incomum, segundo entrevista feita pelo jornal inglês Mirror (2021),

De Bruyne contratou analistas de dados para avaliar sua influência no City e ter poder de barganha no negócio. “O atleta não usou um agente para negociar o novo contrato, conduzindo ele mesmo as negociações com o City” (McDonnell *apud* Mirror, 2021, tradução própria).

Por fim, há necessidade de adaptação do modelo aos dados do futebol brasileiro. Apesar de ser o mesmo esporte, a qualidade, nível técnico e escolas de desenvolvimento dos atletas difere do Brasil para o velho mundo. Dessa forma, faz-se necessário desenvolver um modelo direcionado para o Brasileirão buscando obter melhores resultados.

1.2 OBJETIVOS

A seguir, são descritos os objetivos geral e específico deste TCC, que delineiam as metas a serem alcançadas no desenvolvimento deste estudo de caso.

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é estudar o desenvolvimento um um modelo de gols esperados para avaliar o desempenho de jogadores da Série A do Campeonato Brasileiro de Futebol.

1.2.2 Objetivos Específicos

a) Avaliar, a partir do modelo proposto, como os jogadores se destacam em termos de eficácia na criação de oportunidades de gol e na conversão dessas oportunidades em gols reais;
b) Identificar, também a partir do modelo proposto, os jogadores com melhor desempenho no campeonato no quesito conversão de chances em gol, a partir das probabilidades calculadas;
c) Validar os resultados do modelo proposto a partir de um modelo comercial;
d) Aprimorar previsões de gols a partir de métricas específicas que considerem finalizações realizadas no Campeonato Brasileiro de Futebol de 2022.

1.3 JUSTIFICATIVA DA PESQUISA

Segundo a CBF Academy (2024), hoje, a maior parte das decisões tomadas pelas organizações esportivas - sejam elas dentro ou fora de campo - passam pela área de inteligência das mesmas. Não é mais segredo que a ascensão de clubes e ligas de futebol está relacionada com a eficácia do seu departamento de análise de dados (CBF Academy, 2024).

Entretanto, como abordado na descrição do problema, há uma lacuna ainda não preenchida no que diz respeito à consolidação da cultura *data driven* dentro dos clubes brasileiros de futebol. As decisões passam pela área de inteligência das instituições, mas o apoio às tomadas de decisões está limitado, majoritariamente, a serviços prestados por terceiros. Nesse cenário, surge a influência do uso de modelos preditivos de *expected goals* (xG) dentro dos clubes.

O xG auxilia a capturar o processo subjacente de criação de chances, indo além de simplesmente contar gols. Isso permite avaliar o desempenho com maior precisão e eliminar a aleatoriedade inerente ao futebol, onde o resultado de uma partida pode depender de fatores aleatórios (como um gol isolado), sem refletir necessariamente a qualidade das oportunidades criadas (Mead *et al.*, 2023). Ao usar o xG, clubes podem tomar decisões estratégicas mais informadas, tanto em nível tático quanto de transferências, avaliando, por exemplo, se um jogador está abaixo ou acima da expectativa em termos de conversão de chances em gols (Anzer *et al.*, 2021).

O xG é utilizado para fins financeiros, em negociações de contratos e transferências de jogadores, auxiliando os gestores na determinação do valor de mercado de um atleta com base em seu desempenho previsto, em vez de depender exclusivamente de números de gols. As contratações de Mohamed Salah e Roberto Firmino pelo clube inglês Liverpool FC foram feitas por recomendações do departamento de análise de dados do clube, contratações essas que culminaram em um trio de ataque (juntamente com Sadio Mané) vencedor da UEFA Champions League (Época Negócios, 2019). Isso traz um suporte mais confiável à tomada de decisões baseada em dados dentro dos clubes (Mead *et al*., 2023).

Ou seja, a análise de desempenho de jogadores a partir de um estudo de caso aplicando modelo de gols esperados no campeonato brasileiro busca salientar o poder dessa ferramenta, assim como a urgência de seu uso por departamentos

especializados em análise e ciência de dados dentro dos clubes de futebol brasileiros. Não se trata apenas de ganhos marginais, mas sim de possíveis melhoras significativas. Elevar o rendimento em 2% a 3% pode representar premiações adicionais de €60 milhões (aproximadamente R$380 milhões na cotação atual) para um clube que dispute a Champions League (Graham *apud* Álvarez, 2021).

1.4 ESTRUTURA DO TRABALHO

Este trabalho está organizado em capítulos que abordam de forma sequencial os aspectos teóricos, metodológicos e práticos deste estudo. A estrutura foi assim definida para orientar o leitor, facilitar a compreensão dos passos e das análises desenvolvidas ao longo do projeto.

Na introdução, foram apresentados o contexto do trabalho, os objetivos, a justificativa e a relevância da análise do xG para o desempenho de jogadores no Campeonato Brasileiro de Futebol.

No segundo capítulo são abordados os conceitos teóricos necessários para a compreensão do modelo de gols esperados. Esta seção fundamenta a base teórica utilizada ao longo do trabalho, de forma a permitir ao leitor estar alinhado com as metodologias aplicadas.

Em seguida, na metodologia, são descritos os procedimentos metodológicos adotados no estudo, incluindo o tipo de pesquisa, a coleta e o pré-processamento dos dados, além das técnicas de modelagem utilizadas.

No quarto capítulo é apresentada a aplicação prática do modelo de gols esperados no contexto do Campeonato Brasileiro. Nesta seção, são descritas as etapas de desenvolvimento e avaliação dos modelos de xG básico e avançado, permitindo que o leitor compreenda como esses modelos foram implementados e adaptados para o contexto específico da pesquisa.

No quinto capítulo, são discutidos os resultados obtidos com a aplicação dos modelos, incluindo a comparação entre os modelos básico e avançado, e a interpretação dos dados à luz dos objetivos propostos – a análise de desempenho dos jogadores.

Por fim, o último capítulo traz as conclusões do trabalho e sugestões para estudos futuros. Este último capítulo visa resumir as contribuições do estudo,

destacando pontos de melhoria e possíveis aprofundamentos para pesquisas futuras no campo da ciência de dados no futebol.

A estrutura foi elaborada dessa forma para proporcionar ao leitor uma visão clara do desenvolvimento e das contribuições deste trabalho, facilitando o entendimento do processo.

2 FUNDAMENTAÇÃO TEÓRICA

Para fundamentar teoricamente este trabalho, são apresentados estudos acadêmicos que desenvolveram modelos de gols esperados de maneiras diferentes, juntamente à metodologia escolhida para projetos de mineração, análise e ciência de dados – o Processo Padrão para Mineração de Dados em Diversas Indústrias (CRISP-DM, do inglês *Cross-Industry Standard Process for Data Mining*).

2.1 ESTUDOS SOBRE MODELOS DE GOLS ESPERADOS

O modelo padrão de gols esperados (xG) frequentemente utiliza apenas a distância e o ângulo do chute em relação ao gol como variáveis preditoras principais. Esta abordagem é simples, mas bastante eficaz para uma estimativa inicial da probabilidade de um chute resultar em gol. De acordo com Eggels (2016), os modelos básicos de xG se concentram nas variáveis de distância e ângulo do chute, devido à sua influência direta na probabilidade de gol.

Eggels (2016) utilizou um modelo de regressão logística para prever os gols esperados em seu modelo. O estudo focou nas variáveis distância e ângulo do chute em relação ao gol, situação do jogo (por exemplo, se o jogador estava sob pressão) e tipo de assistência (passe, cruzamento, lançamento). Nesta abordagem, o modelo mostrou que as variáveis distância e ângulo eram os preditores mais significativos para a probabilidade de gol. Mas a inclusão de variáveis contextuais também melhorou a precisão do modelo (Eggels, 2016).

Como objetivo da pesquisa, Eggels (2016) buscou prever resultados de partidas em um estudo de caso considerando a LaLiga, primeira divisão do campeonato espanhol de futebol (temporada 2015/2016). Na ocasião, o modelo mostrou limitações na previsão de resultados de partidas equilibradas (particularmente empates). Outro resultado da pesquisa focou na questão

estratégica. De acordo com Eggels (2016), percepções do modelo permitem que equipes de futebol tomem decisões mais informadas ao longo do tempo, avaliando o desempenho em uma temporada, partidas individuais ou contribuições de jogadores.

Em trabalho desenvolvido na Universidade Federal do Paraná, Withoeft (2020) utilizou os modelos de Regressão Logística, *Random Forest* e XGBoost para um modelo preditivo de gols esperados. Após avaliação do desempenho, os modelos de Regressão Logística e XGBoost se mostraram os melhores. Pela maior simplicidade e interpretabilidade, o modelo de Regressão Logística foi o escolhido para realizar análises (Withoeft, 2020).

Em outro trabalho, Fairchild *et al.* (2021) utilizaram uma abordagem de machine learning para ajustar o modelo de xG, incorporando variáveis como a velocidade da bola, posição do goleiro, posição dos jogadores no campo, além de distância e ângulo do chute. O estudo utilizou técnicas avançadas de aprendizado de máquina, como XGBoost e Random Forest. A análise demonstrou que a inclusão de variáveis adicionais (posição do goleiro e a velocidade da bola) aumentou a precisão do modelo, que resultou em uma visão mais detalhada do processo de tomada de decisão dos jogadores.

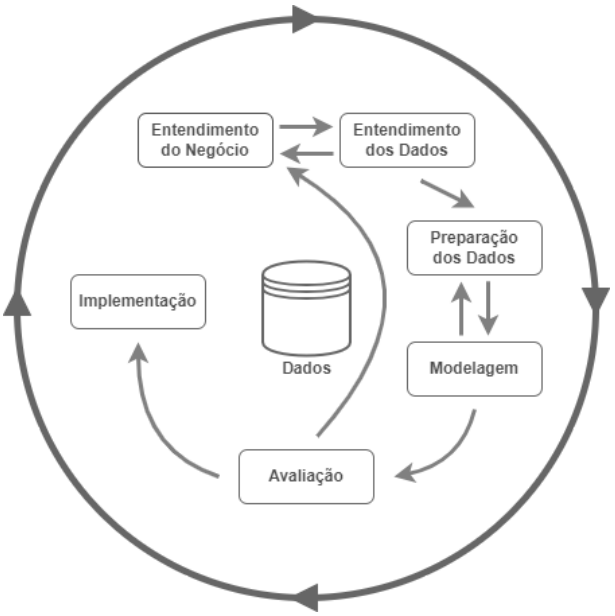
2.2 O PROCESSO PADRÃO PARA MINERAÇÃO DE DADOS EM DIVERSAS INDÚSTRIAS (CRISP-DM)

Os dados necessários para a alimentação do modelo de gols esperados precisam ser extraídos de fontes confiáveis que reúnem essas informações em um único repositório. A extração de informação útil de grandes bases de dados é o processo conhecido como *data mining*, ou mineração de dados. São muitas as metodologias e técnicas usadas na mineração de dados, pois envolvem modelos de estatística, inteligência artificial e machine learning, que não são únicos (Insper, 2022).

Publicado em 1999 para padronizar os processos de mineração de dados em diversas indústrias, o Processo Padrão para Mineração de Dados em Diversas Indústrias (CRISP-DM, do inglês *Cross-Industry Standard Process for Data Mining*) tornou-se desde então a metodologia mais comum para projetos de mineração de dados, análise e ciência de dados (DSPA, 2024). Essa metodologia inclui as

seguintes etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação.

Figura 1 – Diagrama de funcionamento do *CRISP-DM*.



Fonte: adaptado pelo autor (Eggels, 2016).

O entendimento do negócio concentra-se em compreender os objetivos do projeto do ponto de vista empresarial, convertendo esse conhecimento em uma definição de problema de mineração de dados e, em seguida, desenvolvendo um plano preliminar projetado para alcançar os objetivos;

A fase de entendimento dos dados tem como objetivo o recolhimento de dados e início de atividades para familiarização com os dados, identificando problemas ou conjuntos interessantes. A preparação dos dados abrange todas as atividades para construir o conjunto de dados final. Na modelagem é feita uma seleção de técnicas de modelagem e calibração dos parâmetros dessas técnicas para otimização. Na avaliação revisa-se a construção do modelo para garantir que ele alcance adequadamente os objetivos empresariais. Por fim, na etapa de implementação, é feita a organização do conhecimento adquirido de forma apresentável para que o cliente possa usá-lo.

2.3 EXTRAÇÃO DE DADOS PARA MODELOS DE GOLS ESPERADOS

2.3.1 Técnica de Extração: *Web Scraping*

A técnica usada para realizar a extração dos dados será o *web scraping*. *Web scraping* é uma forma de extrair dados de uma página web, convertendo-os em um formato estruturado que pode ser armazenado e analisado. Este processo envolve o envio de requisições *HTTP* para obter o conteúdo das páginas e, em seguida, o uso de ferramentas e bibliotecas específicas para parsear o *HTML* e extrair as informações desejadas (Mitchell, 2018).

Segundo Khder (2021), *web scraping* e *web crawling* são processos automatizados de extração de dados de sites. Eles são vistos em *Business Intelligence*, permitindo a coleta de dados que não estão em formatos facilmente acessíveis, como *JSON* ou *XML*.

A técnica de web scraping pode ser usada para extrair dados de APIs através de requisições feitas a partir de sites. Muitas *API*s retornam os dados em formato *JSON*, que podem ser acessados ao enviar uma requisição *HTTP* usando bibliotecas de Python, como *requests*. O processo envolve enviar um pedido *HTTP* para o endpoint da *API* e, em resposta, a *API* fornece os dados em *JSON*, que são então processados e manipulados pelo *script* de *web scraping* para extrair as informações desejadas.

JSON (*JavaScript Object Notation*) é um dos formatos padrão para o envio de dados por meio de requisições *HTTP* entre navegadores e outras aplicações. O *JSON* é um formato de dados muito mais flexível do que formas tabulares como *CSV*, e seus tipos básicos são objetos (dicionários), *arrays* (listas), *strings*, números, *booleanos* e nulos (McKinney, 2017).

Mitchell (2018) também cita que, sendo amplamente utilizado em *API*s, mais popular que o *XML* devido à sua eficiência e menor tamanho, o arquivo de formato *JSON* transforma objetos e listas em dicionários e listas no Python, de maneira a facilitar o acesso e manipulação dos valores ali armazenados.

2.3.2 Interface de Extração dos Dados

Uma *API* (*Application Programming Interface*) é uma interface que permite que diferentes sistemas se comuniquem, oferecendo dados de forma estruturada, em formatos como JSON ou XML. No contexto do *web scraping, API*s são usadas para acessar dados diretamente de um site de forma automatizada. Entretanto, nem todos os sites disponibilizam *API*s completas ou gratuitas, o que pode justificar o uso de web scraping para acessar essas informações. Conforme abordado por Khder (2021), *API*s facilitam a extração de dados, mas nem sempre fornecem acesso completo ou ilimitado.

De forma geral, o *web scraping* transforma dados não estruturados da *web* em informações estruturadas, permitindo a análise em larga escala. É importante ressaltar que este processo facilita a coleta de dados a partir de grandes volumes de páginas *web*, algo que seria inviável manualmente (Sirisuriya, 2015).

Existem diferentes bibliotecas em Python para *web scraping*: *Requests*, *BeautifulSoup, Scrapy e Selenium*. Para este projeto, será utilizada a biblioteca *Requests.* Segundo Mitchell (2018), a biblioteca *Requests* no Python é usada para lidar com requisições *HTTP* de forma mais simples. Ainda, Mitchell (2018) cita que suas vantagens incluem melhor suporte a autenticação e maior facilidade em enviar dados através da web.

Figura 2 – Requisição *GET* às *API*s do SofaScore em Python.

```python
for url in urls:
    response = requests.get(url, headers=headers)

    if response.status_code == 200:
        data = response.json()

        for shot in data['shotmap']:
            shot_info = {
                'event_id': url.split('/')[-2]
                , 'player_name': shot['player']['name']
                , 'player_position': shot['player'].get('position', '')
                , 'player_jersey_number': shot['player'].get('jerseyNumber', '')
                , 'player_id': shot['player']['id']
                , 'is_home': shot['isHome']
                , 'shot_type': shot['shotType']
                , 'situation': shot['situation']
                , 'player_x': shot['playerCoordinates']['x']
                , 'player_y': shot['playerCoordinates']['y']
                , 'player_z': shot['playerCoordinates'].get('z', 0)
                , 'body_part': shot['bodyPart']
                , 'goal_mouth_location': shot['goalMouthLocation']
                , 'goal_mouth_x': shot['goalMouthCoordinates']['x']
                , 'goal_mouth_y': shot['goalMouthCoordinates']['y']
                , 'goal_mouth_z': shot['goalMouthCoordinates']['z']
                , 'xg': shot.get('xg', None)
                , 'shot_id': shot['id']
                , 'time': shot['time']
                , 'added_time': shot.get('addedTime', 0)
                , 'time_seconds': shot['timeSeconds']
                , 'incident_type': shot['incidentType']
            }
            all_shots.append(shot_info)
    else:
        print(f"Erro na requisição para {url}: {response.status_code}")

dataframe = pd.DataFrame(all_shots)
```

Fonte: do autor, 2024.

2.3 ANÁLISE EXPLORATÓRIA DE DADOS

A Análise Exploratória de Dados, ou EDA (*Exploratory Data Analysis*), é a etapa em que se entende a estrutura dos dados, identifica-se padrões, detecta-se anomalias e verifica-se os pressupostos para o modelo preditivo. Nesse contexto, dos dados extraídos da *API* do SofaScore, em que os dados são fornecidos no formato *JSON*, a *EDA* é a ferramenta que, em uma analogia, serve para “desbravar uma tabela como se fosse uma terra desconhecida”. Ela ocorre em paralelo com o pré-processamento de dados, uma vez que os resultados ali obtidos guiam a limpeza e a transformação dos dados. McKinney (2017) ressalta que a *EDA* é uma ferramenta que auxilia na compreensão inicial dos dados e permite que os analistas identifiquem possíveis *outliers* e relacionamentos a serem explorados nos modelos.

2.3.1 Visualização de Dados

As visualizações são um dos principais componentes da *EDA*. Elas permitem a representação gráfica dos dados de maneira a visualizar padrões e tendências. Neste projeto, gráficos de barras, histogramas e *scatter plots* (gráficos de dispersão) são utilizadoss para entender a distribuição das variáveis-chave ao longo do conjunto de dados. McKinney (2017) explica que histogramas são úteis para verificar a distribuição de variáveis contínuas, enquanto que gráficos de dispersão revelam a relação entre duas variáveis numéricas. Além disso, mapas de calor (*heatmaps*), conforme descrito por VanderPlas (2016), são outra ferramenta para identificar correlações entre variáveis.

2.3.2 Estatística Descritiva

Além das visualizações, pode-se utilizar o cálculo de estatísticas descritivas de forma a obter uma visão geral dos dados. Média, mediana e o desvio padrão resumem variáveis numéricas e explicitam sua distribuição. Neste trabalho, tais cálculos são aplicados para compreender a distribuição das variáveis básicas constituintes de um modelo de gols esperados. No caso de dados JSON, que contém informações complexas e aninhadas, é utilizada a biblioteca *Pandas* do Python para o cálculo dessas estatísticas. Como mencionado por VanderPlas

(2016), a agregação de dados é uma ferramenta utilizada para extrair informações úteis de grandes datasets.

2.3.3 Identificação de Outliers

Um terceiro aspecto da EDA é a identificação de *outliers*, que são pontos de dados que se desviam significativamente do padrão geral. Esses pontos, de certa forma, podem influenciar negativamente o modelo de xG se não forem tratados adequadamente. VanderPlas (2016) sugere que a visualização por meio de *box plots* ou gráficos de dispersão detectam esses outliers, que (posteriormente) podem ser removidos ou transformados antes do treinamento do modelo.

2.4 PRÉ-PROCESSAMENTO DE DADOS

O pré-processamento de dados, no desenvolvimento de modelos de aprendizado de máquina, assegura que os dados estejam organizados e em um formato apropriado para a análise e modelagem. Segundo abordado por McKinney (2017), o pré-processamento inclui técnicas que transformam dados brutos em um conjunto adequado para modelagem, abordando dados ausentes, duplicados e inconsistentes. No contexto deste projeto, em que os dados são extraídos da *API* do SofaScore no formato *JSON*, o pré-processamento é necessário para lidar com a complexidade das estruturas de dados aninhadas e variáveis categóricas, que exigem etapas específicas de limpeza e transformação antes que possam ser analisadas.

2.4.1 Limpeza de Dados

A limpeza de dados é a primeira etapa do pré-processamento, pois garante que os dados estejam em condições de serem usados na modelagem. De acordo com McKinney (2017), a limpeza de dados trata da remoção de valores ausentes, duplicados ou inconsistentes para evitar que informações incorretas prejudiquem a análise e a construção de modelos.

A identificação e o tratamento de dados faltantes (*missing values*) podem ser feitos removendo-se linhas ou colunas com muitos valores ausentes ou

substituindo-os por valores adequados, a depender da situação e da estratégia utilizada. Esse processo é descrito por McKinney (2017), que sugere o uso da biblioteca *Pandas* no Python para realizar imputações consistentes.

Ademais, é necessário remover duplicatas nos dados para evitar redundância e distorções nos resultados, portanto é preciso garantir a unicidade do grão da tabela – neste projeto, de análise de desempenho esportivo através do xG, a granularidade da tabela é um chute feito por um jogador. VanderPlas (2016) sugere que os registros com colunas contendo valores “NaN” (Not a Number) sejam tratados ou retirados da base de dados. Deve-se estar ciente de que o NaN é um pouco como um "vírus de dados" - ele infecta qualquer outro objeto com o qual entra em contato, e o resultado de uma operação aritmética com *NaN* será outro *NaN* (VanderPlas, 2016, p. 122).

2.4.2 Transformação de Dados

A transformação dos dados é a próxima etapa do pré-processamento e envolve aplicar técnicas de normalização, padronização e codificação para que os dados estejam no formato adequado para o treinamento do modelo. McKinney (2017) destaca que a normalização é essencial quando os atributos possuem escalas diferentes, como a distância do chute (em metros) e o ângulo do chute (em graus). A normalização ajusta os valores para um intervalo comum, geralmente entre 0 e 1, o que facilita o aprendizado do modelo, evitando que variáveis com maiores magnitudes dominem o processo de ajuste.

Por outro lado, a padronização transforma os dados para que tenham média zero e desvio padrão um, o que é particularmente importante para modelos que assumem uma distribuição normal dos dados. Segundo VanderPlas (2016), a padronização ajuda a estabilizar o processo de otimização, resultando em modelos mais consistentes. Esse caso foi observado nos dados usados para este trabalho e posteriormente é abordado com mais profundidade na etapa de transformação de dados.

Por fim, chega a etapa de codificação de variáveis categóricas. Segundo orientado por VanderPlas (2016, p.183), devem ser criadas novas variáveis a partir de uma mesma coluna. Por exemplo, pode-se ter um conjunto de dados que contém informações na forma de códigos, como A = "nascido na América", B = "nascido no

Reino Unido", C = "gosta de queijo", D = "gosta de spam" (VanderPlas, 2016, p. 183). são usadas para converter variáveis categóricas em representações numéricas, permitindo que o modelo interprete essas variáveis corretamente. Em relação aos dados do Sofascore, essas técnicas serão particularmente úteis para representar diferentes tipos de finalização e situações precedentes à finalização.

2.4.3 Engenharia de Características

A engenharia de características (do termo em inglês *Feature Engineering*) é um dos processos mais importantes para melhorar a performance do modelo, principalmente em um modelo de gols esperados. Como explicado por VanderPlas (2016), a criação de novas características (ou features) é necessária para capturar relações mais complexas entre as variáveis. No contexto do xG, características como a distância do chute ao gol, ângulo do chute e posição do jogador no campo são utilizadas para determinar a probabilidade de um gol.

A engenharia de características envolve também a remoção de variáveis irrelevantes ou redundantes, o que é crucial para evitar que o modelo seja sobrecarregado com informações desnecessárias. Ao focar apenas nas variáveis mais relevantes, como as mencionadas anteriormente, é possível melhorar significativamente a eficiência e a precisão do modelo de xG. McKinney (2017) ressalta que a redução de dimensionalidade e a seleção de variáveis são etapas críticas no processo de engenharia de características.

Em resumo, o pré-processamento de dados no contexto de dados JSON do SofaScore e a aplicação de um modelo de gols esperados envolve uma série de etapas essenciais para garantir que os dados estejam prontos para análise. A limpeza, transformação e engenharia de características desempenham papeis importantes na preparação dos dados, buscando melhorar a precisão e a capacidade de generalização do modelo.

2.4.4 Divisão dos Dados

A divisão dos dados é uma etapa do desenvolvimento de modelos de machine learning que permite a avaliação de seu desempenho em dados não vistos (ou vistos) durante o treinamento. Em geral, os dados são divididos em dois

conjuntos: treinamento e teste. Segundo VanderPlas (2016), essa proporção pode variar dependendo do tamanho do dataset. Quando o conjunto de dados é pequeno, pode ser necessário ajustar a divisão para garantir que o modelo tenha dados suficientes tanto para treinamento quanto para teste.

Além da divisão tradicional em conjuntos de treinamento e teste, técnicas de validação cruzada, como a validação cruzada k-fold, são frequentemente utilizadas para avaliar a performance do modelo de forma mais robusta. A validação cruzada envolve dividir o dataset em k partes (folds), treinando o modelo em k-1 partes e testando-o na parte restante. Esse processo é repetido k vezes, com cada parte sendo usada como conjunto de teste uma vez. Isso ajuda a garantir que o modelo não esteja "super ajustado" ao conjunto de treinamento e que ele generalize bem para novos dados (Mak & Joseph, 2018).

2.5 SELEÇÃO E TREINAMENTO DO MODELO

Antes de partir para a seleção do modelo, faz-se necessário entender o contexto por trás dessa etapa e por quê essa é a ferramenta utilizada para atingir o objetivo deste trabalho. Machine learning é frequentemente classificado como uma subárea da inteligência artificial (IA), mas essa categorização pode ser enganosa em um primeiro momento (Vanderplas, 2016, p. 332).

Segundo Vanderplas (2016), o estudo de machine learning surgiu a partir de pesquisas nesse contexto (da IA), mas, na aplicação em ciência de dados, é mais útil pensar em machine learning como uma maneira de construir modelos de dados.

> Fundamentalmente, machine learning envolve a construção de modelos matemáticos para ajudar a entender dados. O termo “aprendizado” surge quando damos a esses modelos parâmetros ajustáveis que podem ser adaptados aos dados observados; dessa forma, o programa pode ser considerado como “aprendendo” a partir dos dados. Uma vez ajustados aos dados previamente observados, esses modelos podem ser usados para prever e entender aspectos de novos dados observados (VanderPlas; Jake, 2016, p. 332).

2.5.1 Modelos Básicos de Aprendizado de Máquina

Os modelos de ML podem ser classificados em duas categorias principais: aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado, o modelo é treinado com dados rotulados, o que significa que os

dados incluem tanto as variáveis de entrada quanto o resultado esperado, também chamado de variável alvo (*target variable*) (VanderPlas, 2016). Esse tipo de aprendizado é subdividido posteriormente em tarefas de classificação e regressão (VanderPlas, 2016).

Na classificação, o objetivo é prever uma das categorias da variável alvo. Esses modelos são apropriados para problemas onde a saída é discreta. Segundo Eggels (2016), a classificação é uma forma de aprendizado supervisionado, onde o objetivo é aprender uma função que mapeia atributos para classes predefinidas, como "gol" ou "não gol" no contexto de oportunidades de finalização.

Em contrapartida, em problemas de regressão, o modelo matemático visa prever valores contínuos, sendo útil em contextos em que a saída esperada é numérica. De acordo com Montgomery e Runger (2018), na regressão linear, o objetivo é modelar a relação entre uma variável dependente contínua (Y) e uma ou mais variáveis independentes (X), assumindo uma relação linear entre elas. Esse modelo é pela equação 2.1.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.1}$$

O termo ϵ representa o erro aleatório. O método de ajuste mais comum é o método dos mínimos quadrados, que minimiza a soma dos quadrados dos resíduos para encontrar a linha de melhor ajuste aos dados observados (Montgomery e Runger, 2018).

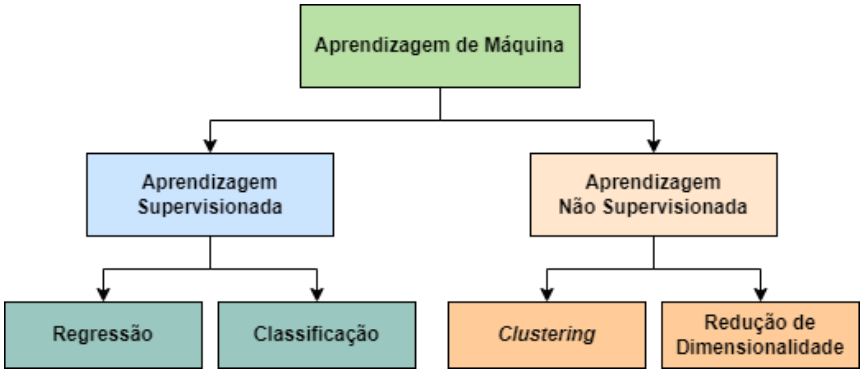
Na regressão logística, segundo Montgomery e Runger (2018), procura-se modelar a probabilidade de uma variável dependente binária, ou seja, uma variável com dois resultados possíveis, como "sucesso/falha" ou "sim/não". A função logística, que é uma função sigmoide, transforma a combinação linear das variáveis independentes em um valor entre 0 e 1, fornece a estimativa da probabilidade de ocorrência de um evento em função do número de variáveis (Yen, 2009). Esse modelo é expresso na equação 2.2.

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{2.2}$$

Isso permite que o modelo seja aplicado em problemas de classificação binária, onde se deseja prever a probabilidade de um evento específico ocorrer (Montgomery & Runger, 2018).

O aprendizado não supervisionado lida com dados sem rótulos, de maneira que o modelo precisa identificar padrões ou estruturas dentro dos dados sem informações predefinidas sobre as categorias ou resultados (VanderPlas, 2016). Técnicas descritivas, como o agrupamento (clustering), são abordagens de aprendizado não supervisionado (Eggels, 2016). O algoritmo k-means é um método utilizado para detectar agrupamentos naturais nos dados, enquanto a Análise de Componentes Principais (PCA) é uma técnica empregada para reduzir a dimensionalidade do dataset, mantendo as características mais importantes (VanderPlas, 2016).

Figura 3 – Tipos de Algoritmos de Aprendizagem de Máquina



Fonte: adaptado de Dutta et. al (2020)

2.5.2 Escolha do Modelo

A escolha do modelo de machine learning mais adequado envolve considerar fatores como o tipo de dados disponível, o objetivo do projeto e os requisitos de performance e interpretabilidade. Segundo VanderPlas (2016), a seleção de um modelo deve considerar o tipo de problema e a natureza dos dados. Complexidade e interpretabilidade também são fatores que devem ser considerados. VanderPlas (2016) menciona que modelos mais complexos, como redes neurais, apresentam alta capacidade de ajuste mas tendem a ser mais difíceis de interpretar, sendo um obstáculo em áreas onde a compreensão do processo de tomada de decisão do modelo é essencial. Modelos lineares, como a regressão logística, são

frequentemente preferidos em cenários que demandam interpretabilidade, por conta de suas simplicidade e transparência (VanderPlas, 2016).

A regressão logística, apesar de ser uma técnica de regressão, é utilizada em modelos de classificação. Segundo a equação (2.2), modelos de regressão logística podem ser aplicados em problemas onde se deseja prever a probabilidade de um evento específico ocorrer. Para esses modelos, o objetivo é prever o resultado de uma variável dependente binária. E segundo o que foi desenvolvido por Eggels (2016), isso enquadra a metodologia em um problema de classificação.

Assim, de forma análoga à tese desenvolvida por Eggels (2016), o tipo de modelo escolhido para este trabalho foi o de Regressão Logística. O problema deste estudo de caso pede um resultado numérico no qual a qualidade de uma oportunidade de marcar gol é fornecida. Portanto, pode-se argumentar que técnicas de regressão logística são as mais adequadas para esse problema (Eggels, 2016).

2.6. AVALIAÇÃO DO MODELO

Para avaliar modelos preditivos, são utilizadas métricas estatísticas pré-definidas e já amplamente abordadas em projetos desta natureza. Inicialmente, por tratar-se de um modelo de regressão, pensou-se em utilizar o valor de R^2 para avaliação do modelo de gols esperados por se tratar de uma regressão. Porém, Hosmer e Lemeshow (2000) explicam que o uso do R^2 tradicional não é adequado para modelos de regressão logística devido à natureza binária de seu desfecho. Em modelos de regressão linear, o R^2 mede a proporção da variabilidade da variável resposta que é explicada pelo modelo, aplicando-se melhor à variáveis dependentes contínuas. No entanto, como a regressão logística trabalha com probabilidades e resultados binários, o R^2 tradicional não capta adequadamente o ajuste do modelo.

Para contornar essa limitação, *pseudo* R^2 foi desenvolvido como uma métrica alternativa, sendo útil para estimar o "poder explicativo" de modelos logísticos sem se basear em variabilidade linear. Segundo Hosmer e Lemeshow (2000), o *pseudo* R^2 inclui métricas como o R^2 *de Cox e Snell* e o R^2 *de Nagelkerke*. O R^2 *de Cox e Snell* é definido de modo a comparar a verossimilhança do modelo ajustado com a de um modelo nulo (um modelo sem preditores) e oferece uma medida da "melhoria"

do ajuste em relação ao modelo nulo. O R^2 *de Nagelkerke* é uma versão ajustada que escala o valor de modo que seu máximo seja exatamente 1, o que facilita a interpretação em termos de poder explicativo (Hosmer & Lemeshow, 2000).

A performance será avaliada conforme a capacidade de prever gols esperados. Para cada finalização, o modelo calcula uma probabilidade efetiva de gol, a qual é o valor de xG. Como exemplo, uma finalização com uma probabilidade de 0.2 xG indica que há uma probabilidade de 20% de que aquela finalização resulte em gol. O xG será utilizado para avaliar o desempenho de jogadores durante toda a temporada do Campeonato Brasileiro de 2022. O objetivo é comparar os gols reais com os gols esperados, de maneira a avaliar a eficácia dos atletas nesse quesito.

Em sua tese, Eggels (2016) utiliza precisão, *recall*, *F-score* e a área sob a curva ROC como métricas avaliativas para seu modelo. Eggels (2016) argumenta que o objetivo do modelo não é prever, de forma correta, todas as tentativas, mas sim classificar as oportunidades de gol de forma que as chances de sucesso recebam pontuações mais altas. Dessa forma, a *AUC* (Área sob a Curva ROC) é a métrica preferida, pois oferece uma visão sobre a eficácia do modelo em classificar corretamente as oportunidades de gol (Eggels, 2016). A precisão (*precision*), a recuperação (*recall*) e o *F1-score* resultam em interpretações sobre o desempenho relacionado aos verdadeiros positivos, falsos positivos e falsos negativos, enquanto a *AUC* fornece uma avaliação baseada em probabilidade da qualidade de classificação do modelo (VanderPlas, 2016).

Por fim, VanderPlas (2016) ainda sugere o uso de validação cruzada (*cross-validation*) para avaliar a performance do modelo em dados não observados, o que permite testar o modelo em diferentes divisões do dataset. Essa técnica não só auxilia na redução do risco de *overfitting*, mas também fornece uma estimativa mais precisa do desempenho do modelo, facilitando a escolha de um modelo alinhado às necessidades do projeto. “Esses fatores contribuem para a seleção de modelos que são eficazes e adaptados ao contexto específico do problema, promovendo previsões consistentes e precisas ao longo do tempo” (VanderPlas, 2016).

Para este trabalho, o método utilizado para avaliar o desempenho dos modelos de gols esperados será, como abordado por Eggels (2016), pautado nas

métricas calculadas a partir da Matriz de Confusão dos resultados, explicada no quadro 1.

Quadro 1 – Matriz de Confusão de Aprendizado de Máquina.

		Real	
		Positivo (gol)	Negativo (não-gol)
Predito	Positivo (gol)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo (não-gol)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: adaptado de Eggels (2016)

A matriz de confusão é uma ferramenta útil para identificar onde ocorrem os erros de classificação e entender melhor o desempenho de um modelo de aprendizado de máquina, principalmente em problemas de classificação (VanderPlas, 2016). Com isso, precisão, revocação, F1-score e área sob a curva ROC são definidas, respectivamente, pelas equações 2.3, 2.4, 2.5 e 2.6:

$$Precisão = \frac{VP}{VP + FP} \tag{2.3}$$

$$Recuperação = \frac{VP}{VP + FN} \tag{2.4}$$

$$F - score = \frac{2 \bullet Precisão \bullet Revocação}{Precisão + Revocação} \tag{2.5}$$

$$AUC = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right) \tag{2.6}$$

O valor da área sob a curva ROC varia entre 0 e 1, sendo que, quanto mais próximo de 1, melhor a performance do modelo. A *AUC*, portanto, oferece percepções sobre a capacidade do classificador de classificar oportunidades de gol melhores com pontuações de fato mais altas (Eggels, 2016).

2.7 AVALIAÇÃO DE DESEMPENHO DE JOGADORES

Uma análise efetiva de desempenho dos jogadores de futebol pode incluir diferentes métricas, como a taxa de conversão de chutes, bem como as finalizações no alvo e os gols esperados (Eggels, 2016).

a) No caso gols por total de finalizações (taxa de conversão geral), é avaliado o quão eficiente o jogador é em transformar suas tentativas em gols, uma métrica básica que auxilia na comparação entre jogadores com diferentes estilos de jogo e volumes de finalizações (Eggels, 2016);
b) Na análise de “gols por finalização no gol”, Eggels (2016) destaca que essa métrica refina a (primeira) análise ao focar apenas nas tentativas que foram realmente direcionadas à meta, proporcionando um entendimento mais preciso da capacidade do jogador em finalizar com precisão. Aqui pode-se discernir a habilidade do jogador em acertar a meta;
c) Por último, de forma ainda mais refinada, a métrica "gols por xG" compara os gols marcados com o valor de xG, de forma a avaliar a qualidade das oportunidades de gol criadas. O uso de xG auxilia a avaliar a eficácia do jogador de acordo com a expectativa média de gol para cada tentativa, refletindo se ele está superando ou ficando aquém da expectativa, dada a dificuldade das finalizações (Eggels, 2016).

Quadro 2 – Métricas para análise de desempenho de jogadores ofensivamente.

Métrica	Qualidade Avaliada
Taxa de conversão geral	Habilidade do jogador em capitalizar tentativas
Taxa de gols por finalização no gol	Habilidade do jogador em acertar a meta
Total de gols por xG acumulado	Habilidade do jogador de converter chances em gols

Fonte: do autor, 2024

Em outras palavras, se um jogador tem uma taxa de gols por xG superior a 1, isso indica que ele está excedendo a expectativa de conversão das chances, o que indica uma habilidade acima da média em converter chutes difíceis em gols. Essa é uma questão chave a ser compreendida: o fato de um atleta possuir taxa superior a 1 nesta métrica indica, de forma quantitativa, que ele tem a habilidade de transformar situações improváveis de gol em resultado. Por outro lado, se a taxa é

inferior a 1, o jogador está marcando menos gols do que o esperado, o que pode supor dificuldades do atleta em aproveitar melhor as chances, especialmente aquelas de alta qualidade.

Portanto, a combinação das três métricas avalia a eficiência e a tomada de decisões dos jogadores em campo, pois considera não apenas a quantidade de chutes, mas a qualidade e a probabilidade de cada tentativa resultar em gol. Ressalta-se que esta última é a proposta de análise que gera o resultado principal esperado, ou seja, o que fundamenta o título deste trabalho.

2.8 CONSIDERAÇÕES FINAIS

Neste capítulo, foi apresentada a fundamentação teórica para embasar o que foi proposto no estudo de caso deste trabalho. Não somente isso, como também os pontos-chave para compreender a teoria por trás de um modelo preditivo de gols esperados. Três obras foram comentadas, cada uma com suas particularidades, porém mais profundidade na obra de Eggels (2016), que será diretriz no desenvolvimento deste TCC.

Ressalta-se a importância do diagrama de processos definido pelo CRISP-DM, método estruturado para conduzir projetos de ciência de dados aplicados ao futebol. Adicionalmente, já dentro das etapas propostas pelo organograma de mineração de dados em diversas indústrias, a descrição do web scraping como ferramenta de extração de dados, que será aplicado para a API do SofaScore. Ainda, a EDA e o pré-processamento dos dados, sua importância antes da modelagem. Por fim, as métricas escolhidas para avaliar o modelo preditivo e o desempenho dos atletas no torneio.

Em resumo, a fundamentação teórica aqui discutida formula a base para a análise aprofundada e modelagem que serão conduzidas nos capítulos subsequentes deste trabalho, visando à aplicação e validação do modelo de xG no contexto da Série A do Campeonato Brasileiro de Futebol.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 TIPO DE PESQUISA

O presente estudo configura-se como um estudo de caso, focado na aplicação e adaptação de modelos de *Expected Goals* (xG) para a análise de desempenho de jogadores de futebol da Série A do Campeonato Brasileiro. A escolha por um estudo de caso justifica-se pela natureza específica dos dados e dos métodos empregados, que exigiram adaptações em modelos existentes para atender às particularidades da base de dados e das características da competição em análise.

Neste contexto, o estudo de caso permitiu uma abordagem detalhada e contextualizada, explorando a viabilidade e eficácia dos modelos de regressão logística aplicados à previsão de gols esperados. Por meio de análises empíricas e ajustes metodológicos, buscou-se evidenciar variáveis específicas do futebol: posição do jogador, contexto da finalização e outras características derivadas dessas, impactam as previsões de xG.

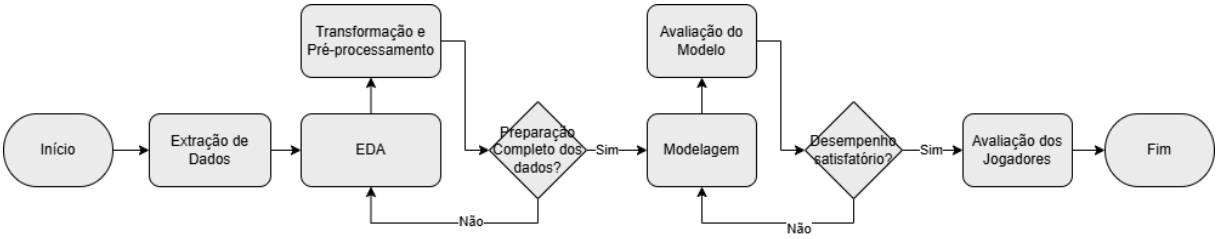
Ademais, o processo de adaptação dos pacotes e métodos utilizados no presente trabalho reflete a necessidade de ajustes técnicos que não se enquadram em um processo de desenvolvimento de novas tecnologias, mas sim na adaptação prática e análise aprofundada de uma aplicação existente.

3.2 ETAPAS METODOLÓGICAS

Os procedimentos metodológicos para o desenvolvimento do modelo de xG seguiram a sequência de etapas do *CRISP-DM* até a obtenção dos resultados. Entretanto, para este trabalho, algumas adaptações foram feitas. Dessa forma, os procedimentos metodológicos foram divididos em 5 etapas, que podem ser visualizadas em um mapeamento de processo na Figura 4.

a) Extração de dados;
b) Análise Exploratória de Dados;
c) Transformação e Pré-processamento de dados;
d) Modelagem;
e) Avaliação do modelo;

Figura 4 – Mapeamento do Processo de modelagem do modelo de xG.



Fonte: do autor (2024)

Embora o CRISP-DM tenha sido originalmente desenvolvido para data mining, suas fases e princípios se aplicam a projetos de ML (*Machine Learning*). O CRISP-DM oferece uma estrutura clara e bem definida para todo o ciclo de vida de um projeto de mineração de dados. Sua flexibilidade permite que seja adaptado para diferentes tipos de projetos e setores industriais (Shearer et al., 2000).

3.3 DELIMITAÇÕES DO TRABALHO

Este trabalho está delimitado a um estudo de caso que visa o desenvolvimento e aplicação de um modelo de previsão de gols esperados, que inclui: a descrição do método de desenvolvimento; avaliação de desempenho do modelo; a análise de desempenho dos jogadores. Este projeto possui caráter acadêmico, o que significa que o presente se concentra na avaliação e na validação de um modelo preditivo para fins de estudo e pesquisa. Portanto, não abrange aspectos técnicos como ajustes de hiperparâmetros, *deploy*, monitoramento ou manutenção do modelo como em um *pipeline* de dados.

Integração Contínua e Entrega Contínua (CI/CD) é uma prática central em *pipelines* de dados e desenvolvimento de *software* que visa automatizar e melhorar o processo de desenvolvimento, teste, entrega e implantação de código e dados (Kleppmann, 2015). Em um *pipeline* de dados, essa prática assegura que todas as alterações feitas no código (ou nos dados) sejam integradas e testadas com uma frequência pré-programada para detectar problemas antes da implantação final. Segundo Kleppmann (2015), o conceito de *CI/CD* é frequentemente utilizado em operações de desenvolvimento de software (*DevOps*) e engenharia de dados para reduzir erros manuais e aumentar a confiabilidade dos fluxos de dados e das análises.

Dessa forma, dado que este não é um aplicativo destinado ao uso contínuo, entende-se que não há necessidade de implementação de um pipeline de *CI/CD*, monitoramento constante ou manutenção. Em resumo, este trabalho não possui objetivos de aplicação prática periódica e foi planejado para atender a propósitos puramente acadêmicos.

4 ESTUDO DE CASO

Este capítulo apresenta o estudo de caso desenvolvido para avaliar o desempenho dos jogadores da Série A do Campeonato Brasileiro de 2022 utilizando um modelo de Gols Esperados (xG). O estudo foi conduzido em duas etapas principais de modelagem, com o objetivo de aprimorar a precisão das previsões e capturar características específicas das finalizações realizadas ao longo do campeonato. Aqui serão abordadas todas as etapas do estudo de caso: uma descrição conceitual dos modelos e as etapas seguidas sequencialmente como definidas no CRISP-DM.

4.1 MODELAGEM

A modelagem foi feita em duas etapas. Dois modelos foram feitos. A primeira etapa de modelagem foi para um modelo básico de xG, e após essa modelagem, foi feito um modelo mais robusto, com mais variáveis sendo utilizadas para a aprendizagem da máquina.

4.1.1 Modelo Básico de Gols Esperados

Segundo Scholtes & Karakus (2024) um modelo básico de xG incorpora apenas distância com relação ao gol, ângulo do chute e sua interação. Para o desenvolvimento do modelo, como abordado na seção 2.5 deste trabalho, foi escolhido um modelo de Regressão Logística por ser adequado ao tipo de problema que está sendo atacado, de classificação, em que se pode prever um valor probabilístico na variável de saída.

As métricas foram calculadas na etapa de pré-processamento dos dados e serão descritas de forma mais detalhada no decorrer deste capítulo. Juntamente a isso, foram definidas as variáveis de entrada do modelo, enquanto que a variável de

saída foi a coluna “goal”, criada a partir dos dados da coluna shot_type (tipo do chute) – os chutes podem ser do tipo “gol”, “defesa do goleiro”, “na trave”, “para fora”, entre outros. A coluna goal possui valores binários, com 1 indicando que a finalização resultou em gol, e 0 caso contrário. O modelo de regressão logística foi treinado com os dados de entrada e de saída usando o método de *fit* da biblioteca *scikit-learn*, uma biblioteca python específica para desenvolvimento de modelos preditivos. Após o treinamento, o modelo faz uma predição das probabilidades de gol (xG) para cada observação e assim é obtido o valor de xG. A Tabela 1 é uma amostra do resultado final do modelo, com as coordenadas do jogador no campo, distância, ângulo da finalização e o resultado final da probabilidade de gol (xG). Posteriormente, neste capítulo, os procedimentos de pré-processamento e transformação de dados são mais detalhados.

Tabela 1 – Amostra dos dados com xG do Modelo Básico calculado

	Coordenada X	Coordenada Y	Distância	Ângulo	Gol	xG Logístico
0	4.830	45.900	12.842854	13.132408	0	0.077197
1	9.240	31.552	9.558782	40.939015	0	0.172889
2	20.790	45.152	23.592185	15.650856	0	0.033271
3	25.305	53.652	32.039727	10.360881	0	0.014327
4	22.260	18.564	27.088331	12.745262	0	0.023155

Fonte: do autor (2024)

4.1.2 Modelo Avançado de Gols Esperados

O desenvolvimento do modelo avançado de expected goals (xG), ou modelo V2, foi realizado com o objetivo de aumentar a precisão das previsões ao incorporar variáveis adicionais. Em comparação com o modelo básico de xG, que utiliza apenas a distância até o gol e o ângulo de finalização como variáveis de entrada, o modelo v2 inclui outras variáveis qualitativas que refletem diferentes contextos e características do chute. Essa abordagem permite ao modelo capturar aspectos adicionais que influenciam a probabilidade de um chute resultar em gol.

Primeiramente, foi criada uma cópia do conjunto de dados original, contendo as variáveis player_x_meters, player_y_meters, is_home (indicador se o time estava

jogando em casa), shot_type (tipo de chute), situation (situação do chute, como assistência, escanteio, etc.), body_part (parte do corpo utilizada), além da distância até o gol e do ângulo de finalização. Em seguida, as variáveis categóricas body_part e situation foram transformadas em variáveis dummies (ou variáveis indicadoras) usando a função pd.get_dummies, permitindo que o modelo trabalhasse com informações qualitativas de forma quantitativa.

As variáveis de entrada (X) deste modelo incluíram:

a) As coordenadas cartesianas e características geométricas do chute: player_x_meters, player_y_meters, distance_to_goal, angle_degrees;
b) Indicador de contexto: is_home;
c) A partir da variável body_part dos dados brutos e abordada na seção 4.4.3 deste trabalho, foram extraídas variáveis *dummies* que representam a parte do corpo usada para o chute: body_part_head, body_part_left-foot, body_part_right-foot, e body_part_other. Essas novas variáveis possuem formato binário, de maneira análoga à outras variáveis desse formato apresentadas anteriormente;
d) Da variável sitação de jogo (situation): situation_assisted, situation_corner, situation_fast-break, situation_free-kick, situation_penalty, situation_regular, situation_set-piece e situation_throw-in-set-piece.

O modelo foi treinado utilizando regressão logística, com limite de iterações para garantir convergência. No treinamento, os registros com valores ausentes foram preenchidos com zero para evitar problemas no ajuste do modelo. Em seguida, foi ajustado para prever a probabilidade de gol, com base nas variáveis de entrada descritas.

Comparativamente, Eggels (2016) possuia uma base de dados mais robusta, com ainda mais variáveis adicionais. Com isso, além de ângulo e distância, ainda utilizou número de atacantes em linha, número de defensores em linha (total de defensores entre o gol e a finalização), posição do goleiro, qualidade do jogador, qualidade do goleiro e a situação de jogo.

Portanto, a semelhança entre este modelo e o desenvolvido por Eggels (2016) reside nas variáveis em comum: ângulo, distância e situação de jogo.

4.2 PRÉ-PROCESSAMENTO DE DADOS

Antes da modelagem de dados, é necessário que estes passem por um pré-processamento e transformação, caso necessário, para garantir que as informações estejam estruturadas e em um formato apropriado para a próxima etapa. Conforme discutido por McKinney (2017) na Seção 2.4, esse processo envolve uma série de técnicas que convertem dados brutos em conjuntos prontos para utilização.

Após as análises exploratórias realizadas na EDA, o pré-processamento foi focado na remoção de dados inconsistentes e no cálculo de variáveis importantes para o treinamento do modelo. Esses passos garantiram que o conjunto de dados estivesse devidamente preparado para a modelagem subsequente, assegurando a organização dos dados relevantes para a modelagem.

4.2.1 Limpeza e Tratamento de Dados Inconsistentes

O tratamento dos dados inconsistentes foi realizado por meio da eliminação dos registros problemáticos, já que esses representavam menos de 1% do total. Esse percentual reduzido de dados inconsistentes permitiu que a remoção fosse uma prática adequada, evitando complicações no modelo sem impacto significativo na quantidade e na qualidade das informações utilizadas para a análise e modelagem.

4.3 TRANSFORMAÇÃO DE DADOS PARA A MODELAGEM

4.3.1 Desnormalização de Coordenadas

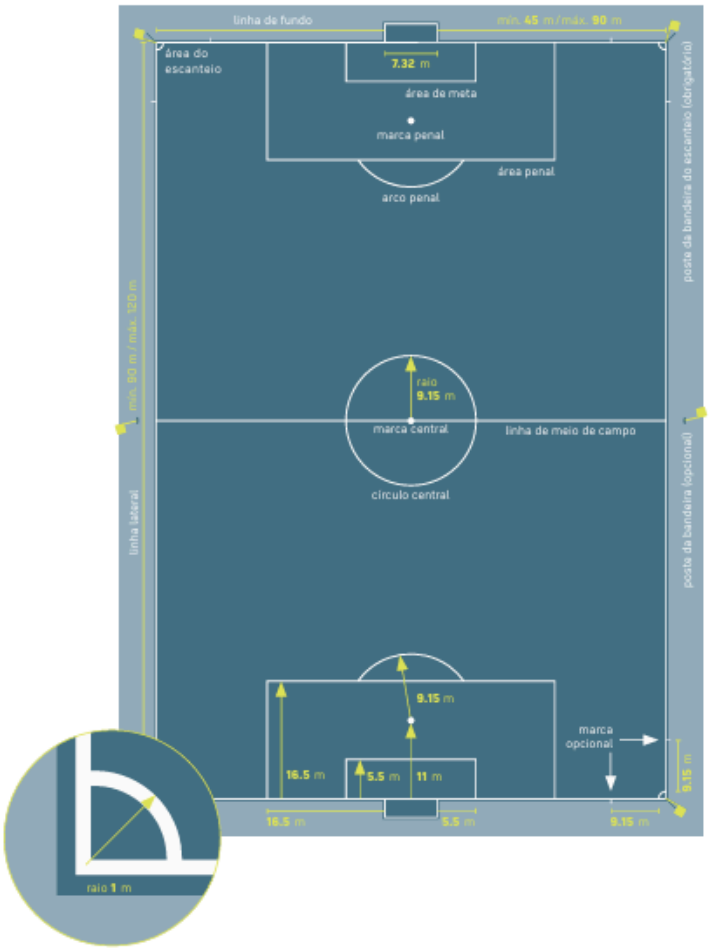
Os dados espaço-temporais brutos extraídos da API estavam normalizados, o que significa que os valores de coordenadas eram representados dentro de um intervalo fechado, como discutido na Seção 3.1.3. Do ponto de vista da modelagem de aprendizado de máquina, a normalização é um aspecto positivo, pois mantém os valores das variáveis dentro de uma escala consistente, evitando problemas de distorção que poderiam confundir o modelo durante o treinamento e a previsão (McKinney, 2017). A normalização, em termos técnicos, refere-se ao processo de

transformar os dados para que fiquem dentro de uma escala específica (como de 0 a 1), facilitando a comparação entre variáveis de diferentes magnitudes.

No caso dos dados fornecidos pela API do SofaScore, as coordenadas X e Y do campo de futebol vieram normalizadas, o que significa que os valores estavam restritos a um intervalo padrão, representando uma escala relativa do campo. No entanto, para análises exploratórias mais aprofundadas, principalmente na geração de gráficos e na interpretação estatística descritiva, a normalização das coordenadas introduziu uma certa distorção, dificultando a visualização e interpretação dos dados em relação às dimensões reais de um campo de futebol.

Segundo o *International Football Association Board* (2024, p. 39), nas dimensões padrão de um campo oficial de futebol, a linha lateral deve ter, no mínimo, 90 metros e, no máximo, 120 metros de comprimento. Já as linhas de fundo devem ter entre 45 e 90 metros de comprimento.

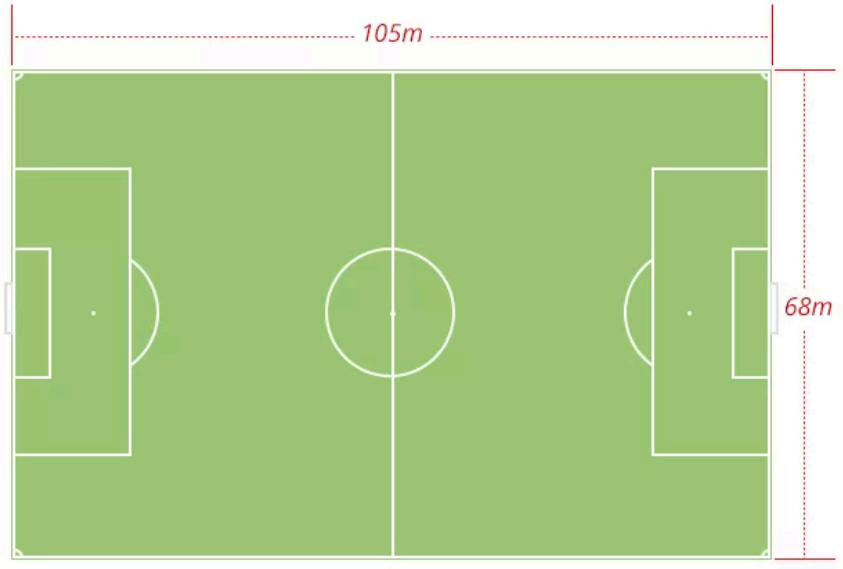
Figura 5 – Dimensões de um campo de futebol no padrão FIFA.



Fonte: IFAB (2024)

Entretanto, na Série A do Campeonato Brasileiro de Futebol, desde 2016, as dimensões padrão dos campos de futebol são de 105 metros de comprimento por 68 metros de largura (CBF *apud* Globo Esporte, 2016). A padronização foi efetivada após constantes reclamações de treinadores, e foi efetuada em três frentes: promoção de treinamento com os responsáveis pela manutenção dos gramados dos estádios brasileiros, diagnóstico de cada um dos vinte campos e subsídio (Globo Esporte, 2016).

Figura 6 – Dimensões de campos de futebol do Campeonato Brasileiro.



Fonte: Globo Esporte (2016)

Ou seja, para trazer os dados para as dimensões padrão oficiais do futebol brasileiro e possibilitar uma análise mais precisa, foi necessário realizar a desnormalização dos dados brutos. Esse processo consistiu em multiplicar as coordenadas X e Y por fatores de escala.

$$player\_x\_meters = player\_x \times \frac{105}{100} \quad (4.1)$$

$$player\_y\_meters = player\_y \times \frac{68}{100} \quad (4.2)$$

4.3.2 Cálculo da Distância de Finalização

Uma das variáveis necessárias para a concepção do modelo básico de xG é a distância do chute com relação à meta. Para o cálculo de tal variável, utilizou-se a fórmula euclidiana para o cálculo da distância entre dois pontos.

De acordo com Anton (2012, p. 132), considerando-se dois pontos, $P_1(x_1, y_1)$ e $P_2(x_2, y_2)$, no espaço vetorial R^2, nesse caso o campo de futebol, a fórmula para o cálculo da distância d entre dois pontos implica na equação 4.3.

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.3}$$

Nas coordenadas cartesianas do campo, de acordo com as dimensões padrão da CBF, sabe-se que o gol sempre estará localizado no ponto de origem da linha lateral e no ponto médio da linha de fundo, $x_1 = 0$ e $y_1 = 34$. No espaço vetorial R^2, o ponto pode ser definido por $P(0, 34)$.

$$P_1(x_1, y_1) = P(0, 34) \tag{4.4}$$

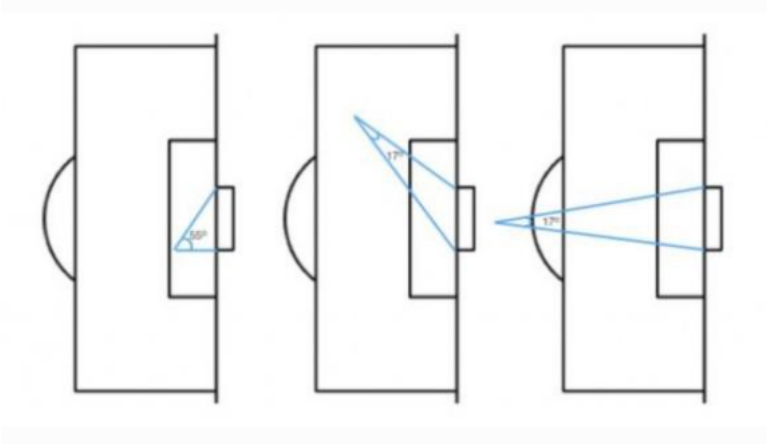
$$d(P_1, P_2) = \sqrt{(x_2 - 0)^2 + (y_2 - 34)^2} \tag{4.5}$$

Com isso, dentro do conjunto de dados foi criada uma nova variável para armazenar as distâncias (em metros) de todos os chutes realizados durante o campeonato, que foi utilizada tanto para a etapa de EDA, como para a modelagem.

4.3.3 Ângulo da Finalização

A segunda variável necessária para a construção de um modelo de xG básico é o ângulo da finalização. O termo “ângulo”, nessa circunstância, pode ser compreendido como o ângulo formado entre o ponto no campo de onde ocorreu a finalização e as duas traves que limitam o tamanho da goleira.

Figura 7 – Exemplo prático de ângulo formado por finalização no campo.



Fonte: Soccermatics (2022)

Como definido por *International Football Association Board* (2024, p. 38), a goleira possui um comprimento de 7,32 metros, com seu ponto médio coincidindo com o ponto médio da linha de fundo (à 34 metros da origem). Dessa forma, o ângulo pode ser definido através do ângulo formado entre dois vetores que partem da origem da finalização $P(0, y)$ e atingem dois pontos distintos no eixo y em $y = 30.34$ e $y = 37.66$, com x variando conforme a posição do jogador (variável player_x_meters). Em outras palavras, o cálculo utiliza as coordenadas do jogador no campo para obter o ângulo formado por esses vetores com relação à origem, usando o conceito de produto escalar entre vetores e as magnitudes dos mesmos.

Os vetores A e B são definidos pelas coordenadas $(player\_x\_meters, player\_y\_meters)$ do jogador, subtraindo os pontos $(0, 30.34)$ e $(0, 37.66)$, respectivamente. Assim, tem-se as equações 4.6 e 4.7.

$$A = (A_x, A_y) = (player\_x\_meters - 0,\ player\_y\_meters - 30.34) \quad (4.6)$$

$$B = (B_x, B_y) = (player\_x\_meters - 0,\ player\_y\_meters - 37.66) \quad (4.7)$$

Com isso, O produto escalar entre os vetores A e B é dado pela equação 4.8.

$$A \cdot B = A_x \cdot B_x + A_y \cdot B_y \quad (4.8)$$

Segundo Anton (2012), o produto escalar permite calcular o ângulo entre dois vetores usando a relação do cosseno. Assim, as magnitudes dos vetores A e B, conforme Anton (2012), são calculadas nas equações 4.9 e 4.10.

$\|A\| = \sqrt{A_x^{\,2} + A_y^{\,2}}$ e

(4.9)

$$\|B\| = \sqrt{B_x^{\,2} + B_y^{\,2}} \tag{4.10}$$

O ângulo θ entre os vetores A e B foi obtido pela fórmula do produto escalar.

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{4.11}$$

Dessa maneira, o ângulo em radianos foi calculado pela equação 4.12.

$$\theta = arccos\left(\frac{A \cdot B}{\|A\|\|B\|}\right) \tag{4.12}$$

Por fim, o ângulo em graus foi encontrado utilizando-se uma relação trigonometria simples.

$$angle\_degrees = \theta \times \left(\frac{180}{\pi}\right) \tag{4.13}$$

Esse processo permitiu adicionar uma coluna ao DF que guardasse o valor do ângulo em graus formado entre os vetores, que, de forma análoga à distância do chute, também foi utilizado para análises descritivas na EDA e para o desenvolvimento dos modelos na etapa de modelagem. Todos os cálculos seguiram as fórmulas para produto escalar e magnitude, conforme descrito por Anton (2012).

4.4. DADOS DE PARTIDAS DE FUTEBOL

Para aplicar a metodologia abordada neste capítulo, é necessário ter entendimento dos dados. Dessa forma, esta seção tem o objetivo de explorar e esclarecer os dados disponíveis. Primeiramente, abordam-se questões envolvidas com a fonte de dados, em que três aspectos dessa fonte de dados são comentados. Depois, os métodos de extração de dados são apresentados de forma prática, assim como o formato geral dos dados é discutido. Por fim, ainda são discutidos os possíveis problemas de qualidade nos mesmos. Com isso, explica-se como os dados são combinados em um mesmo repositório.

4.4.1 Dados Espaço-temporais

O *SofaScore* é uma aplicação para acompanhamento de resultados esportivos. A ferramenta é gerenciada e desenvolvida pela empresa *SofaIT*, que tem domicílio na Croácia. O aplicativo conta com avaliação de mais de 3 milhões de atletas, cobrindo mais de 20 mil ligas e torneios de 20 esportes diferentes (SofaScore, 2024). Ademais, o aplicativo disponibiliza tanto dados coletivos como dados individuais das partidas, de forma a possibilitar análises à nível de equipe ou a nível pessoal do desempenho de cada jogador

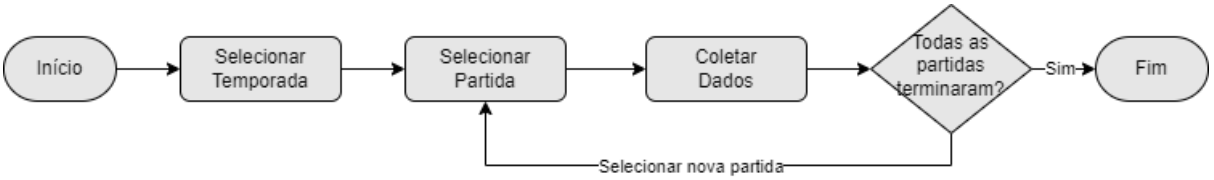
A ferramenta foi escolhida por ser mundialmente reconhecida e por fornecer dados de código aberto (*open source*), facilitando seu acesso para o propósito acadêmico deste trabalho, diferentemente de outras plataformas que requerem assinaturas pagas ou possuem acesso restrito aos dados.

Além de relatórios padrão disponibilizados dentro da plataforma, a aplicação do SofaScore pode disponibilizar seus dados para análises personalizadas dentro dos clubes. O SofaScore Editor é uma ferramenta que digitaliza súmulas de torneios de base, futebol feminino, escolinhas e ações para sócios. Dentro do Campeonato Brasileiro de Futebol, o SofaScore é parceiro do Fluminense Football Club (Fluminense, 2024). Deste ponto em diante, em analogia à Eggels (2016), os dados do SofaScore referentes às partidas serão referidos à dados espaço-temporais.

4.4.2 Extração dos Dados

Os dados das partidas foram coletados através da API do SofaScore. Primeiro, o usuário acessou o site oficial da plataforma e determinou a competição da qual coletar dados. Cada competição é separada por edições, que consistem na temporada referente ao torneio. Uma vez definida a temporada, no caso do Campeonato Brasileiro de Futebol, foram selecionadas as partidas por rodada.

Figura 8 – Diagrama do processo para extração de dados



Fonte: adaptado de Eggels (2016)

Seguindo a orientação de Mitchell (2018) abordada na seção 2.3, os dados foram extraídos de APIs do Sofascore através de uma requisição GET da biblioteca *Requests*. Entretanto, durante o processo de *web scraping*, houve empecilhos.

Inicialmente, no desenvolvimento do código, após a importação das bibliotecas necessárias para a realização do processo de extração de dados, foi definido um dicionário de cabeçalhos de rede para requisições HTTP, simulando o comportamento de um navegador. Isso incluiu os parâmetros user-agent e sec-fetch para evitar bloqueios do site ao tentar realizar a coleta de dados.

Após validado um teste inicial com a *URL* de uma partida aleatória, foi criada uma lista em formato de texto com todos os links relacionados às 380 partidas da Série A do Campeonato Brasileiro de 2022. Feito isso, no *script* do Google Colab, foi definida uma função em python para ler essa lista, extrair os identificadores dos eventos e formatar as *URLs* apropriadas para acessar os mapas de chutes via API.

Depois de coletados os links, o código foi programado para realizar um *loop* sobre estes (com um operador *for*), utilizando a biblioteca *Requests* para obter os dados de cada partida. Caso a resposta fosse bem-sucedida, os dados em formato *JSON* eram processados. Obteve-se duas respostas mal-sucedidas devido à duas partidas remarcadas no campeonato.

Figura 9 – Partidas adiadas no Campeonato Brasileiro de 2022.

Fonte: SofaScore (2022)

Com isso, os dados extraídos foram armazenados em uma lista, que posteriormente foi convertida para um *dataframe* utilizando-se a biblioteca *Pandas*.

4.4.3 Formato Geral dos Dados

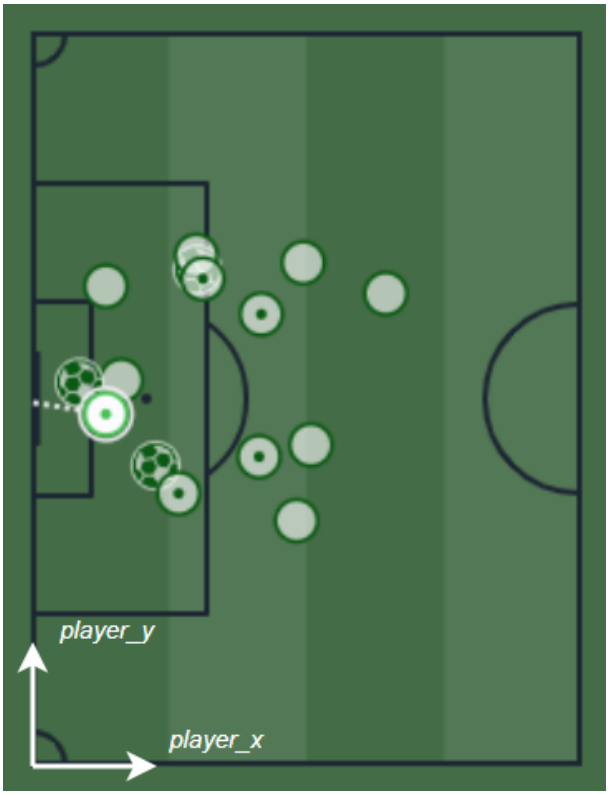
Os dados espaço-temporais consistem em incidentes que ocorreram em um determinado momento de uma partida de futebol. Cada um desses incidentes possui atributos de maneira a fornecer informações mais detalhadas sobre o incidente. Esses incidentes incluem finalizações, passes, faltas, cartões e outros (Eggels, 2016). A seguir foram descritas as variáveis e atributos extraídos da API do SofaScore.

a) event_id: é o identificador do evento na base de dados do SofaScore. Nesse caso, o identificador da partida de futebol;
b) player_name: é o nome do jogador envolvido no incidente; nesse caso, o jogador que finaliza;
c) player_position: a posição do jogador classificada de forma tática. Pode ser F (atacante, do termo em inglês *forward*), M (meio-campista, do inglês *midfielder*), D (defensor, do inglês *defender*) e G (goleiro, do inglês *goalkeeper*);
d) player_jersey_number: número da camisa do jogador;
e) player_id: identificador único do atleta na base de dados do SofaScore;
f) is_home: variável binária que identifica se o incidente está associado a um jogador que estava jogando em casa ou fora de casa: *True* ou *False*;

g) shot_type: o tipo da finalização. Essa é uma variável categórica, que pode ser classificada em diferentes classes:
    - Bloqueado (*block*);
    - Gol (*goal*);
    - Fora (*miss*);
    - Trave (*post*);
    - Defesa do Goleiro (*save*);
h) situation: finalização classificada por situação de jogo. Na base de dados espaço-temporais utilizada neste trabalho, existem oito classificações distintas para situações de jogo que, segundo a Stats Perform (2024), são definidas da seguinte forma:
    - Regular (*regular*): uma tentativa criada a partir de um ataque em jogada aberta (open play);
    - Bola parada (*set-piece*): uma tentativa criada onde a jogada começa a partir de uma situação de falta indireta com bola parada;
    - Lateral (*throw-in-set-piece*): uma tentativa criada a partir de um arremesso lateral;
    - Falta direta (*free-kick*): uma tentativa criada a partir de uma situação de falta direta;
    - Escanteio (*corner*): uma tentativa criada a partir de uma situação de escanteio;
    - Contra-ataque (*fast-break*): uma tentativa criada após a defesa rapidamente transformar defesa em ataque, recuperando a bola em seu próprio campo (contra-ataque);
    - Pênalti (*penalty*) – o próprio pênalti em si; qualquer chute de rebote originado pelo pênalti passa a ser classificado como bola parada;
    - Assistido (*assisted*): apesar de não ser explicitamente definida pela Opta como situação padrão de jogo, essa situação está presente na base de dados do SofaScore, e sugere que a finalização foi originada por uma assistência. Uma assistência é o toque final de um companheiro de equipe que leva o destinatário da bola a marcar um gol (Stats Perform, 2024).
i) player_x, player_y e player_z são as coordenadas cartesianas que indicam a posição do jogador dentro do campo. Deve-se imaginar o campo de futebol

como um plano cartesiano visto de cima, com os eixos das abscissas e ordenadas, e com a origem deste sendo no canto inferior esquerdo do campo. No caso dos dados espaço-temporais extraídos da *API* do SofaScore, $player\_x \in [0, 100]$ e $player\_y \in [0, 100]$. Ainda, é importante ressaltar que todos os dados das partidas são relacionados à uma mesma metade do campo. Pode-se compreender essa lógica a partir da Figura 10, em que o time defensor se encontra na metade esquerda da cancha, enquanto o time atacante vem da metade direita e, portanto, ataca a metade esquerda do campo.

Figura 10 – Coordenadas cartesianas da posição de um jogador no campo.



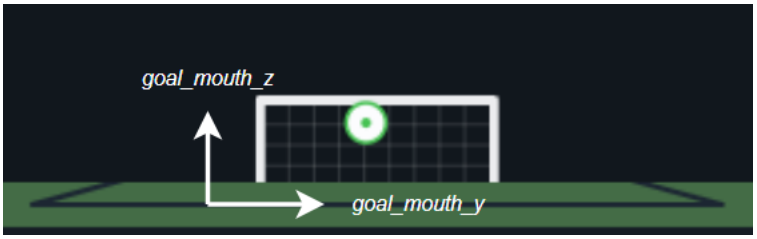
Fonte: adaptado de SofaScore (2022).

j) body_part: a parte do corpo que o jogador utilizou para executar a finalização. Essa coluna se divide em quatro variáveis:pé direito (*right-foot*), pé esquerdo (*left-foot*), cabeça (*head*) e outro (*other*). Nesse caso, o termo “pé” inclui qualquer tipo de conexão com a perna (Stats Perform, 2024)

k) goal_mouth_x, goal_mouth_y, goal_mouth_z são coordenadas da localização da finalização no gol. As variáveis goal_mouth_y e goal_mouth_z estão

definidas no intervalo [0, 100], enquanto que a variável goal_mouth_x está definida no intervalo [0, 0];

Figura 11 – Coordenadas cartesianas da posição-destino da bola no gol após uma finalização.



Fonte: adaptado de SofaScore (2022).

l) xg: resultado do modelo de gols esperados do próprio SofaScore;
m) shot_id: identificador único da finalização;
n) time: tempo da partida (em minutos) em que ocorreu a finalização e está definido no intervalo [0, 90].
o) added_time: tempo adicional da partida, em minutos. Caso o incidente tenha acontecido durante o tempo regulamentar, o added_time é igual a zero. Dessa forma, é diferente de zero caso o incidente tenha acontecido durante o tempo adicional;
p) time_seconds: tempo da partida em segundos;
q) incident_type: tipo do incidente, que nesse caso será somente um: finalização (do termo em inglês, *shot*).

Uma amostra dos dados espaço-temporais é fornecida na Figura 12.

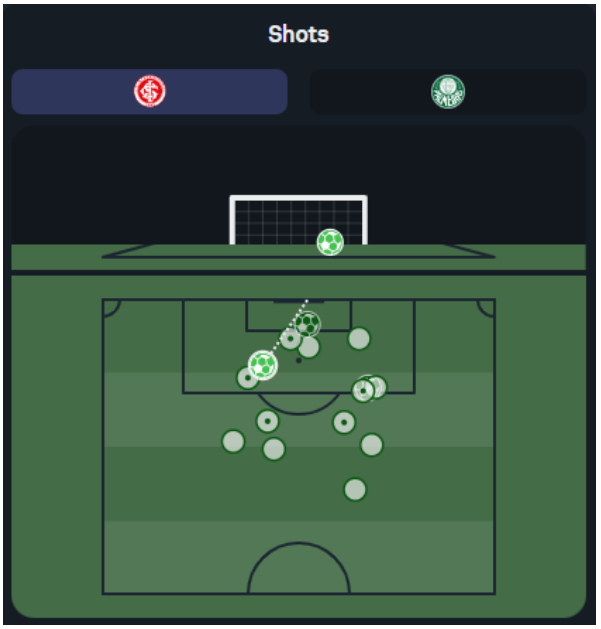
Figura 12 – Amostra dos dados obtidos do dataset.

	event_id	player_name	player_position	player_jersey_number	player_id
0	10113678	Fred	F		12764
1	10113678	Fred	F		12764
2	10113678	Fred	F		12764
3	10113678	Luiz Henrique	M	7	1035995
4	10113678	Cristiano	D	12	886239
5	10113678	Jhon Arias	M	21	844096
6	10113678	Germán Cano	F	14	33238
7	10113678	Nonato	M	16	922566
8	10113678	Nonato	M	16	922566
9	10113678	Nathan	M	14	358538

Fonte: do autor, 2024.

A API do SofaScore disponibiliza os dados de finalizações de forma visual em um mapa para utilização do usuário. O mapa mostra tanto os dados das coordenadas do campo, como também da boca do gol. Um exemplo pode ser visto na Figura 13.

Figura 13: Finalizações do S.C. Internacional contra o S.E. Palmeiras em partida válida pelo Brasileirão 2024.



Fonte: SofaScore, 2022.

4.5 ANÁLISE EXPLORATÓRIA DOS DADOS

“O objetivo da Análise Exploratória de Dados (EDA) é explorar os dados sem ideias claras sobre os aspectos que se estão buscando” (Eggels, 2016, p. 5). Essa etapa, como abordado na fundamentação teórica deste trabalho, envolveu análises gerais buscando entender características e padrões na base de dados. Após o processo de extração da API do SofaScore, o conjunto de dados foi armazenado em uma única tabela, um *dataframe* no script, que também será referido pelos termos “tabela” e “DF”.

Primeiramente, a tabela foi inspecionada para obter uma visão geral dos dados coletados. Com isso, foi possível calcular o número de partidas distintas presentes nos dados através da contagem distinta da variável event_id, confirmando o número total de jogos da temporada: 380.

4.5.1 Análises de Finalizações

Nesta subseção, são apresentadas as análises relacionadas às finalizações realizadas pelos jogadores da Série A do Campeonato Brasileiro de Futebol de 2022. Essas análises visam compreender de que forma as finalizações se distribuíram por diferentes características, tais quais a posição do jogador, a parte do corpo utilizada e o tipo de finalização. Para isso, foram utilizadas visualizações gráficas e também estatísticas descritivas, de maneira a identificar padrões que influenciassem o desempenho dos jogadores.

4.5.1.1 Total de Finalizações por Posição do Jogador

A primeira análise considerou o total de finalizações por posição dos jogadores. O agrupamento dos dados pela posição tática (Defensor, Meio-Campista, Atacante, Goleiro) revelou que os jogadores de meio-campo e ataque foram responsáveis pela maior parte das finalizações, sendo 5.007 e 3.393, respectivamente. Isso é esperado, visto que atacantes e meio-campistas têm papel mais ofensivo nas partidas, enquanto defensores, com 1.953 finalizações, tendem a participar de forma mais contida no ataque, preocupados com a defesa. Os goleiros apresentaram apenas duas finalizações, reflexo de ocasiões raras.

Tabela 2 – Total de chutes por posição do jogador taticamente.

Posição	Total de Chutes
Defensor	1953
Atacante	3393
Goleiro	2
Meio-Campista	5007

Fonte: do autor (2024)

4.5.1.2 Total de Finalizações por Tipo

A avaliação do total de finalizações por tipo envolveu categorizar as finalizações pela classificação da definição: bloqueado, gol, fora, trave e defesa do goleiro. O tipo mais frequente foi "Fora" (41%), seguido de "Defesa" (23%) e "Bloqueado" (26%). Apenas cerca de 9% das finalizações resultaram em gol,

destacando a dificuldade dos jogadores de converter as oportunidades em resultados efetivos.

Tabela 3 – Total de finalizações por tipo.

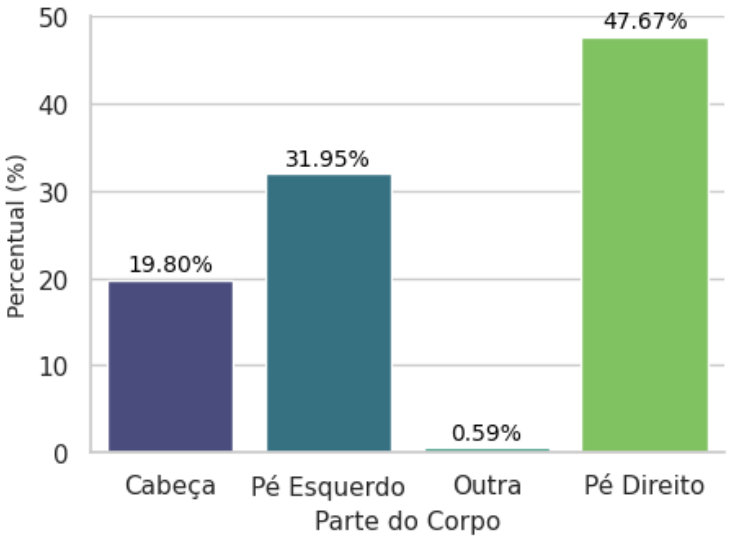
Tipo de Finalização	Total de Finalizações	Percentual do Total
Fora	4178	40.347658
Bloqueada	2683	25.910188
Defesa do Goleiro	2409	23.264124
Gol	905	8.739739
Trave	180	1.738291

Fonte: do autor (2024)

4.5.1.3 Finalizações por Parte do Corpo Utilizada

Para entender melhor como as finalizações variam conforme a parte do corpo utilizada, analisamos as categorias "Pé Direito", "Pé Esquerdo", "Cabeça" e "Outra". As finalizações com o pé direito representaram quase metade do total (47,7%), seguidas pelo pé esquerdo (31,9%) e pela cabeça (19,8%). Isso reforça a prevalência de jogadores destros no campeonato, enquanto o uso da cabeça mostra uma proporção significativa, evidenciando a importância de cruzamentos e bolas aéreas na criação de oportunidades de gol.

Figura 14 – Gráfico do percentual de finalizações por parte do corpo.



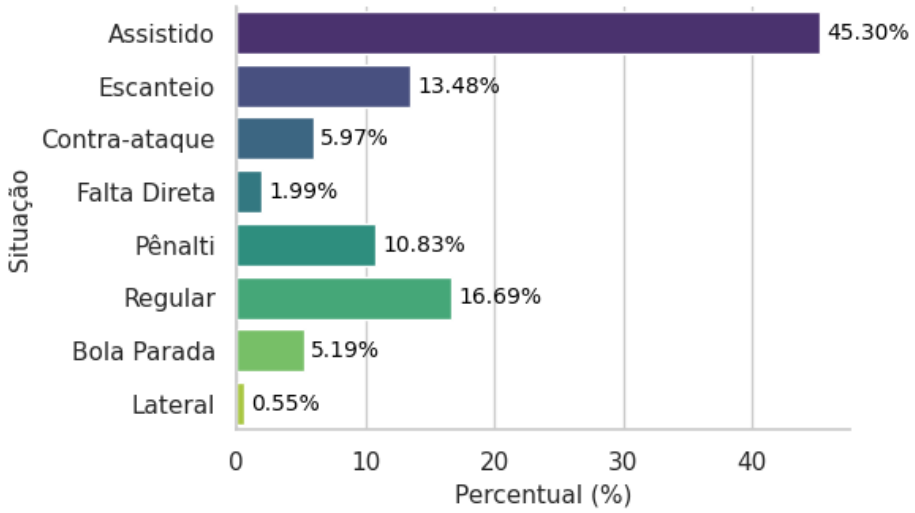
Fonte: do autor (2024)

O que deve ser compreendido aqui é a parte do todo que corresponde à cada finalização. Essa etapa da EDA está informando ao leitor que, basicamente, de cada 5 finalizações realizadas no campeonato, 1 delas é de cabeça.

4.5.2 Análises de Gols e Distribuição

Seguidamente das análises de finalizações, foram feitas análises relacionadas a todos os gols marcados durante o campeonato. Outra variável categórica relevante analisada foi a situação de jogo. Como mencionado na seção 4.4.3, as situações de jogo são definidas em 8 categorias diferentes. Percebe-se como no caso dos gols marcados, quase metade destes foram categorizados na situação “Assistido”, uma boa parte em “Regular” e outra parte considerável em “Escanteio”.

Figura 15 – Gráfico do percentual de gols marcados por situação de jogo.



Fonte: do autor (2024)

Por outro lado, percebe-se como somente 2% dos gols do torneio são marcados através de faltas diretas. Ou seja, a cada 50 gols marcados, somente 1 é de falta.

Além da análise de distribuição de variáveis categóricas dentro do DF, também foi feita a análise de distribuição das variáveis quantitativas. Como abordado anteriormente na etapa de estatística descritiva, a média é uma das

ferramentas aliadas ao analista durante a etapa de EDA. Para a presente base de dados, foram analisadas as médias dos valores das coordenadas X e Y do jogador dentro do campo, juntamente à distância e o ângulo calculados. Os dados foram agrupados por resultado da finalização: gol ou não.

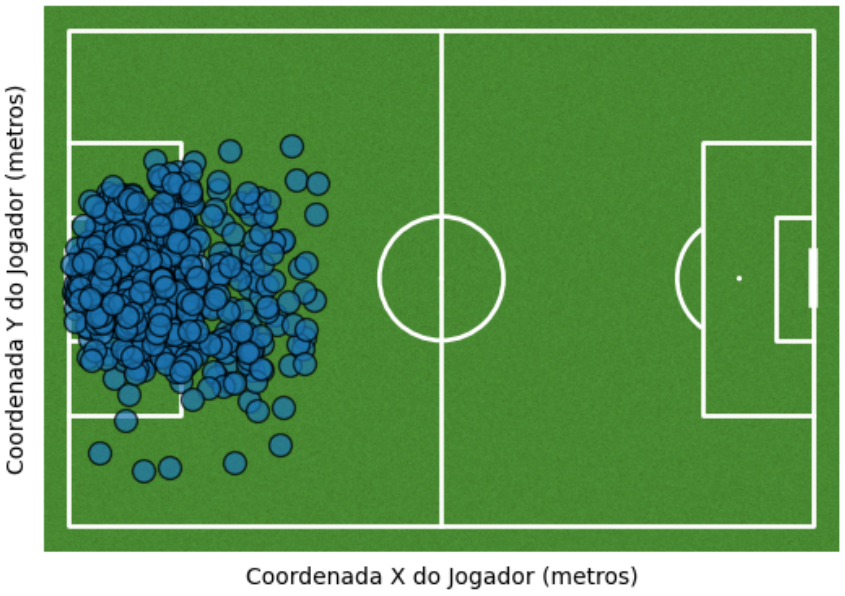
Tabela 4 – Média das variáveis quantitativas por fator “gol”.

Gol	Coordenada X	Coordenada Y	Distância	Ângulo
0	17.331933	33.631908	19.701616	22.741938
1	10.541304	33.605600	11.986590	40.305582

Fonte: do autor (2024)

A média de ambos os valores de gol e não-gol para a coordenada Y não é coincidência. Ambos os valores se aproximam de 34m, que é o ponto médio da largura de um campo de futebol (68m). À primeira vista, isso pode gerar o questionamento: como é a distribuição geográfica dos gols no campo? Para responder a essa pergunta, o gráfico da Figura 16 transcreve essa informação de maneira visual.

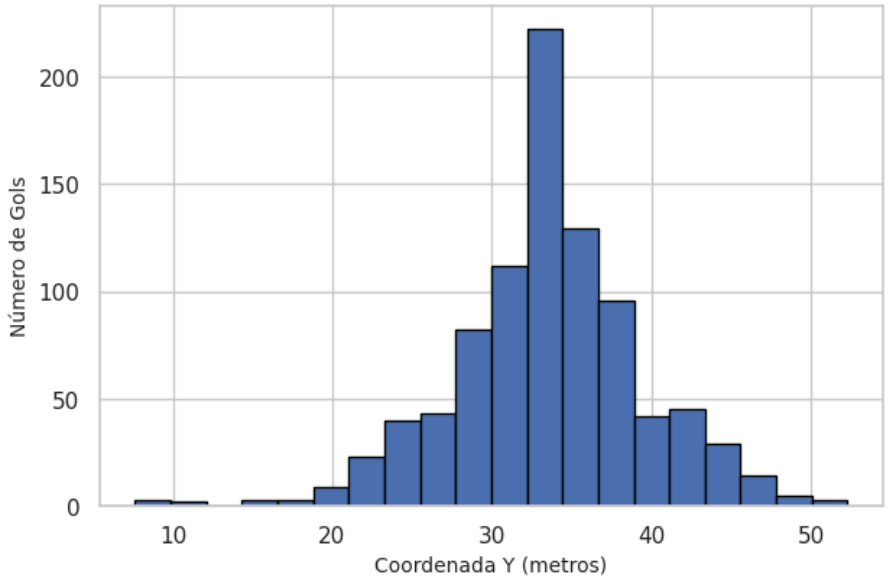
Figura 16 – Distribuição das finalizações que resultaram em gol no Campeonato Brasileiro de 2022.



Fonte: do autor (2024)

A partir da distribuição geográfica de todos os gols do campeonato, surge o questionamento de como seria a distribuição volumétrica dos gols pela largura do campo, coordenada Y do jogador, variável que possui média muito próxima ao ponto médio do campo. A Figura 17 responde a essa pergunta.

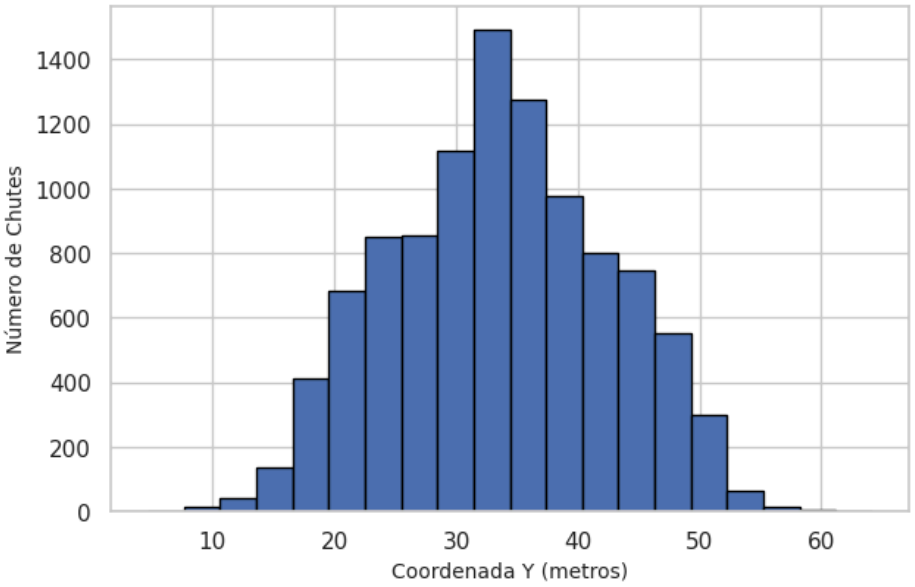
Figura 17 – Distribuição dos Gols pela Largura do Campo (metros).



Fonte: do autor (2024)

Percebe-se que a distribuição de gols através da coordenada Y se aproxima de uma distribuição normal de probabilidade. O resultado do gráfico de distribuição mostrado na Figura 17 é, de certa forma, esperado. Quando executado de um ponto mais próximo ao ponto médio da largura do campo, um chute possui maior ângulo de abertura com relação ao gol. É algo trivial, se observado de forma matemática; empírico e intuitivo, se observado pelo ponto de vista dos jogadores. Os atletas buscam o meio campo para realizar finalizações. Os pontas invertem suas posições e buscam um ponto mais claro da cancha para realizar o arremate. Tal fenômeno pode ser observado na Figura 18.

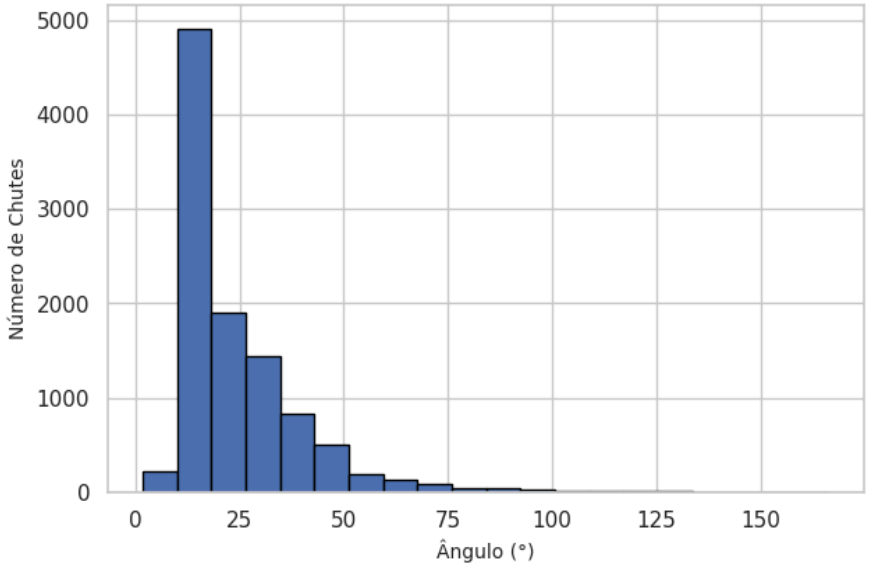
Figura 18 – Distribuição dos Chutes pela Largura do Campo (metros).



Fonte: do autor (2024)

Em contrapartida, ao analisar a distribuição de uma variável quantitativa dependente das coordenadas do campo, nesse caso, o ângulo da finalização, percebeu-se um resultado brevemente distinto. Como esperado, por também depender da variável coordenada X (com média 17.33 para não-gols), a distribuição dos chutes por ângulo apresentou um gráfico com outro formato, sendo mais frequentes as finalizações próximas de 20°.

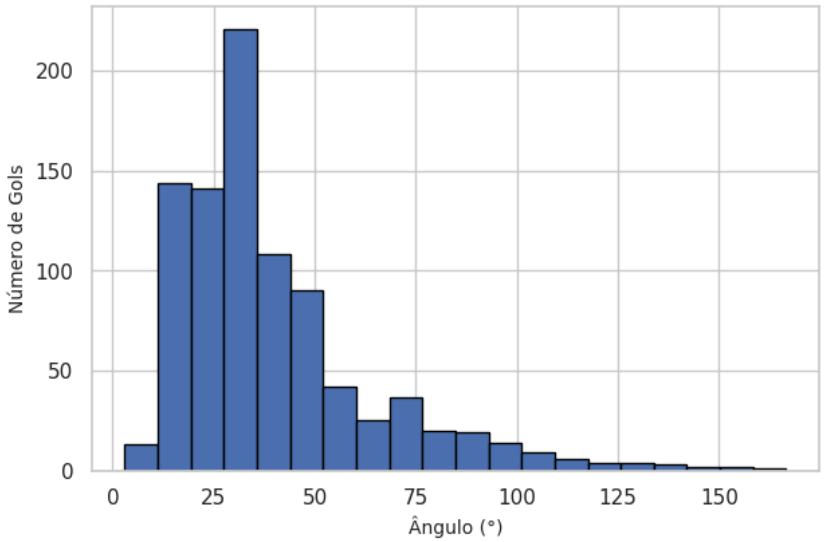
Figura 19 – Distribuição dos chutes por ângulo de finalização



Fonte: do autor (2024)

Para o caso dos gols, ainda dentro do contexto do ângulo da finalização, o resultado também é diferente. Espera-se que um jogador tenha maior probabilidade de marcar quanto maior o ângulo, uma vez que assim o atleta possui mais clareza para acertar a meta. Tal resultado pode ser validado na Figura 20.

Figura 20 – Distribuição dos gols por ângulo de finalização



Fonte: do autor (2024)

Outro ponto observado foi a média das coordenadas cartesianas dos chutes com relação às coordenadas de onde foram parar no gol. Essa análise é feita de forma superficial pois esses dados não entram no escopo do modelo de xG. As variáveis independentes de goal_mouth podem ser utilizadas para o desenvolvimento de um outro modelo, o xGOT, que retorna uma probabilidade pós finalização. Este modelo expande o modelo original de xG ao atribuir créditos para chutes no alvo com base em uma combinação da qualidade da chance subjacente (xG) e da qualidade de sua execução (StatsPerform, 2024).

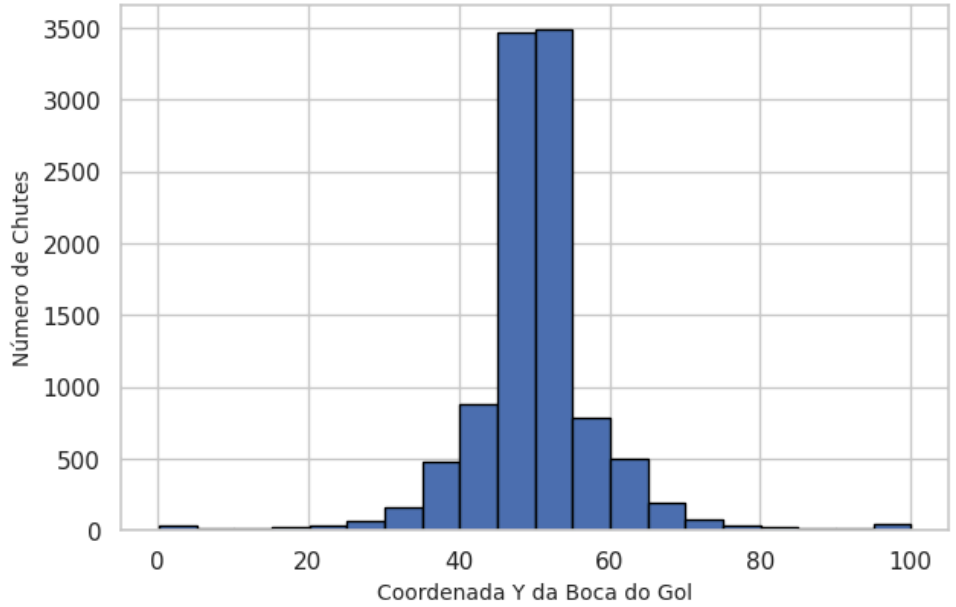
Tabela 5 – Coordenadas médias da boca do gol para as finalizações do Campeonato Brasileiro Série A 2022.

Gol	Boca do Gol X	Boca do Gol Y	Boca do Gol Z	Distância (m)	Ângulo (°)	xG SofaScore
0	0.0	50.201788	26.008804	19.701616	22.741938	0.073610
1	0.0	50.001436	12.432486	11.986590	40.305582	0.308772

Fonte: do autor (2024)

Percebe-se que, novamente, de maneira análoga ao Y das coordenadas cartesianas do campo, o valor médio de Y nas coordenadas da boca do gol se aproxima do ponto médio da mesma. A distribuição de chutes do campeonato pode ser vista na Figura 21.

Figura 21 – Distribuição dos Chutes por Coordenada Y da boca do gol.



Fonte: do autor (2024)

4.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

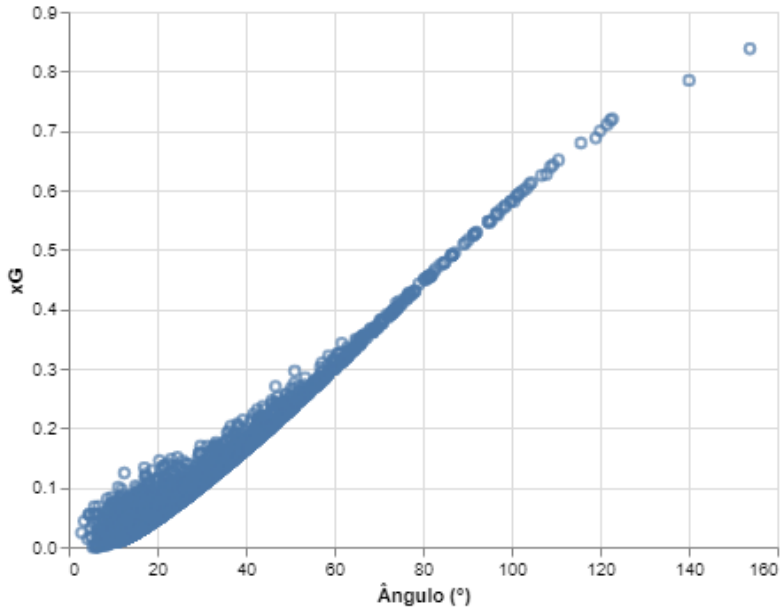
A partir da fundamentação e metodologia propostas, foram apresentados os procedimentos do estudo de caso, com todas as etapas de modelagem até a aplicação da ferramenta proposta, o modelo de xG. Desde a etapa de coleta e pré-processamento de dados, até o desenvolvimento dos modelos básico e avançado, evidenciou-se a necessidade de variáveis adicionais a partir das etapas de transformação de dados e a engenharia de características. A apresentação das dinâmicas das partidas reforça o potencial do xG como ferramenta estratégica para aprofundar a compreensão do jogo e auxiliar gestores nas tomadas de decisões pelos clubes.

5 RESULTADOS

5.1 ANÁLISE DE CORRELAÇÃO ENTRE VARIÁVEIS

Com o modelo básico considerando apenas a distância e o ângulo para a previsão, antes de calcular as métricas de avaliação propriamente ditas, buscou-se compreender a correlação entre ambas as variáveis com os resultados obtidos pelo modelo. A interpretação de gráficos de relação entre as variáveis, a seguir, buscou denotar a dependência entre as características geométricas dos chutes e a probabilidade de gol.

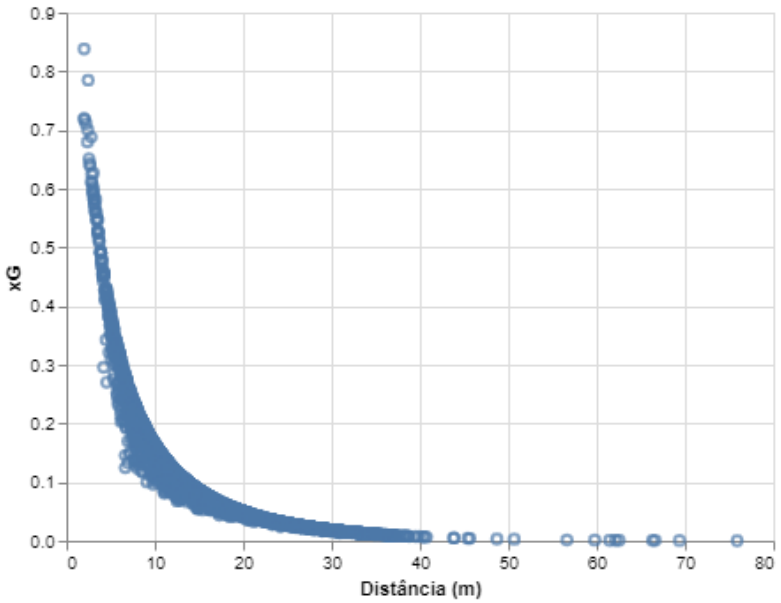
Figura 22 – Gráfico de dispersão da relação entre xG e ângulo



Fonte: do autor (2024)

Na Figura 22, a probabilidade de gol é plotada contra o ângulo de finalização. Observa-se uma relação positiva em que, conforme o ângulo aumenta, a probabilidade de gol também tende a crescer. Isso indica que chutes realizados em ângulos mais favoráveis (ou seja, com uma maior abertura em relação ao gol) têm uma maior chance de sucesso, suposição feita anteriormente e assim confirmada.

Figura 23 – Gráfico de dispersão da relação entre xG a distância



Fonte: do autor (2024)

A Figura 23 mostra a relação entre a probabilidade de gol e a distância da finalização. Percebe-se uma forte relação negativa: à medida que a distância aumenta, a probabilidade de gol diminui significativamente. Isso é esperado, pois chutes realizados de distâncias maiores tendem a ter menos precisão e uma menor chance de vencer o goleiro. Esse comportamento reflete a dificuldade aumentada de finalizações de longa distância em comparação com chutes mais próximos do gol. De maneira a quantificar os resultados observados, a Tabela 6, de correlação, mostra a relação entre o valor de xG e cada uma das variáveis.

Tabela 6 – Matriz de correlação entre as variáveis quantitativas e xG do modelo básico

	xG Logístico
Distância	-0.781253
Ângulo	0.983550
xG Logístico	1.000000

Fonte: do autor (2024)

Para a distância (-0.781253), existe uma forte correlação negativa com xG, confirmando que quanto maior a distância do chute, menor a chance de sucesso.

Mas para o ângulo (0.983550), por outro lado, percebe-se uma correlação positiva maior ainda em módulo, o que reforça que ângulos mais favoráveis (abertos) estão diretamente associados a maiores probabilidades de gol.

Ainda, foi feita a matriz de correlação entre as variáveis do modelo V2. Nesse caso, as variáveis qualitativas também foram incluídas e apresentaram um resultado diferente do primeiro. Os resultados podem ser observados na Tabela 7.

Tabela 7 – Matriz de correlação entre as variáveis quantitativas e xG do modelo V2

	xg_v2
distance_to_goal	-0.572078
situation_assisted	-0.134333
situation_free-kick	-0.079412
situation_corner	-0.039540
body_part_left-foot	-0.033291
situation_set-piece	-0.025829
situation_throw-in-set-piece	-0.005964
body_part_right-foot	-0.005594
is_home	0.032244
body_part_head	0.036057
body_part_other	0.051600
situation_regular	0.059228
situation_fast-break	0.109058
situation_penalty	0.596110
angle_degrees	0.721074
xg_v2	1.000000

Fonte: do autor (2024)

Para o segundo caso, além das correlações esperadas das duas variáveis vistas anteriormente no modelo básico, observou-se uma dependência de variáveis qualitativas relacionadas ao tipo de situação de jogo e a parte do corpo utilizada para realizar a finalização.

As correlações negativas mais fortes observadas foram, novamente, a distância, mas também a situação de jogo “assistido” e “falta direta” – resultados, no mínimo, curiosos. Como visto na Figura 15, Gráfico do percentual de gols marcados

por situação de jogo, aproximadamente 45% dos gols do campeonato foram derivados de situações de um passe direto de companheiro com bola rolando, ou seja, uma assistência. Mas a correlação negativa deve fazer sentido, porque mais de 50% das finalizações do campeonato (5697 finalizações) resultam desse tipo de situação – proporção superior ao valor visto nos gols. E quanto à situação da falta direta, com correlação de -0.079, o valor segue a linha do que vem sendo observado ao longo dos últimos anos. Em uma análise feita pelo Ciência da Bola (2024), entre 2011 e 2020, os gols de falta diminuíram no Campeonato Brasileiro. “Com exceção da edição de 2021, são 62% de gols a menos nos últimos 8 anos” (Ciência da Bola, 2024).

Já as correlações positivas mostraram uma influência notável de pênaltis e contra-ataques. Há uma correlação de quase 0.6 entre uma situação de pênalti e a probabilidade de gol calculada pelo xG. Também, um valor de correlação de aproximadamente 0.1 para situações que envolvam contra-ataques, evidenciando como esses são diretamente relacionados a situações oportunas de gol.

5.2 AVALIAÇÃO DOS MODELOS

A avaliação dos modelos envolveu o cálculo das métricas definidas na seção 2.7, as quais Eggels (2016) e VanderPlas (2016) definem como as mais adequadas para problemas de classificação, como os modelos aqui abordados.

A partir das definições dessas métricas, descritas nas equações 2.3, 2.4, 2.5 e 2.6, foi possível quantificar o desempenho do modelo básico para prever gols. Essas métricas concluem, de forma numérica, a capacidade do modelo de identificar corretamente oportunidades de gol, refletindo tanto a precisão quanto a sensibilidade do modelo em situações de previsão do evento nas partidas. Um compilado da Tabela 8 explica os resultados.

Tabela 8 – Resultado da avaliação do modelo básico pelas métricas de desempenho

	Modelo Básico de xG
Precision	0.6707
Recall	0.0608
F1-score	0.1114
AUC-ROC	0.7701

Fonte: do autor (2024).

O Modelo Básico apresenta uma precisão de 0.6707 e um *recall* de 0.0608. Isso indica que este consegue identificar corretamente uma parte das oportunidades de gol, mas tem uma baixa capacidade de recuperação dos VP (verdadeiros positivos, os gols reais). O *F1-score* de 0.1114 refletiu essa limitação, mostrando que o modelo tem dificuldades para equilibrar precisão e recuperação. A *AUC-ROC* de 0.7701 indica uma capacidade moderada do modelo de classificar corretamente as oportunidades de gol em comparação a tentativas mal sucedidas.

Tabela 9 – Comparativo de desempenho entre os Modelos Básico e V2 pelas métricas de desempenho

Modelo	Precision	Recall	F1-score	AUC-ROC
Modelo Básico	0.6707	0.0608	0.1114	0.7701
Modelo V2	0.7104	0.2033	0.3162	0.8160

Fonte: do autor (2024)

Em contrapartida, o Modelo V2 apresenta uma melhora significativa em todas as métricas:

a) A precisão aumenta para 0.7104, indicando que o modelo aprimorado é mais preciso na previsão de gols;
b) O Recall quase quadruplica para 0.2033, sugerindo que o Modelo V2 tem uma melhor recuperação de verdadeiros positivos;
c) O F1-score também aumenta para 0.3162, refletindo um equilíbrio muito melhor entre precisão e recuperação, o que representa uma evolução substancial em termos de eficácia;

d) A área sob a curva ROC sobe para 0.8160, indicando que o Modelo V2 tem uma capacidade mais robusta de distinguir entre chutes bem-sucedidos e mal sucedidos.

Essas diferenças demonstram que a inclusão de variáveis adicionais no Modelo V2 (situação do chute e parte do corpo utilizada para execução) proporcionam resultados mais precisos nas previsões.

5.3 VALIDAÇÃO COM MODELO COMERCIAL

Ainda, de forma a realizar um comparativo com um modelo já presente e estabelecido no mercado mundial esportivo, comparou-se o Modelo V2 com os valores obtidos pelo modelo do SofaScore, disponíveis nos dados brutos obtidos na etapa de extração.

Tabela 10 – Comparativo de desempenho entre os três modelos pelas métricas de desempenho

Modelo	Precision	Recall	F1-score	AUC-ROC
Modelo Básico	0.6707	0.0608	0.1114	0.7701
Modelo V2	0.7104	0.2033	0.3162	0.8160
Modelo SofaScore	0.7211	0.2343	0.3536	0.8159

Fonte: do autor (2024).

Ao comparar o Modelo V2 com o Modelo SofaScore, percebe-se que o modelo da SofaScore apresenta um desempenho superior, mesmo que de forma breve, na maior parte das métricas:

a) A precisão do Modelo SofaScore é ligeiramente maior (0.7211) em comparação com o Modelo V2 (0.7104);
b) A recuperação do Modelo SofaScore (0.2343) também é superior ao do Modelo V2 (0.2033), sendo mais eficaz em capturar as oportunidades de gol verdadeiras;
c) O *F1-score* do SofaScore é 0.3536, comparado ao 0.3162 do Modelo V2, o que sugere que o Modelo SofaScore consegue equilibrar melhor a precisão e a recuperação;

d) A AUC-ROC é mais alta no Modelo V2 (0.8160) do que no SofaScore (0.8159), o que indica uma capacidade de classificação geral equivalente entre os dois modelos.

Em resumo, o Modelo V2 representa uma diferença em relação ao Modelo Básico, denotando que a inclusão de variáveis contextuais aumenta substancialmente a capacidade de previsão de gols. Embora o Modelo do SofaScore possua vantagem em três das quatro métricas, o desempenho do Modelo V2 se mostrou próximo, com uma *AUC-ROC* ligeiramente superior. Isso demonstra que o Modelo V2 se aproxima da precisão de um modelo comercial como o da SofaScore, mas vale ressaltar que os dados de ambos os modelos, básico e V2, estiveram limitados ao Campeonato Brasileiro de 2022, diferentemente do SofaScore, que possui uma base de dados superior à esta. O alto valor encontrado pela área sob a curva ROC pode ser decorrente de *overfitting*, como mencionado na seção 2.4.4.

Esses resultados sugerem que o Modelo V2 oferece uma alternativa para a previsão de gols, com uma estrutura de variáveis adaptada para capturar nuances que o Modelo Básico não contemplava. O modelo V2 permite uma análise mais aplicável em avaliações de desempenho de jogadores e equipes.

5.4 AVALIAÇÃO DE JOGADORES

Esta é a etapa chave do trabalho, momento em que se agrega valor a todo o desenvolvimento do estudo de caso, ponto em que se justifica a realização deste trabalho. Para a avaliação de desempenho dos jogadores finalizadores da Série A do Brasileirão 2022, três métricas foram utilizadas: taxa de gols por finalizações totais, taxa de gols por finalização no gol e taxa de gols por gols esperados. Ao longo desta seção, todas serão abordadas de maneira mais aprofundada, buscando tirar conclusões a partir da interpretação dos dados obtidos.

5.4.1 Taxa de Conversão Geral

Considerando os 10 jogadores que mais finalizaram no campeonato, a Tabela 11 apresenta suas estatísticas de finalização, incluindo o número total de

finalizações, gols marcados, a taxa de conversão (gols por finalização) e a quantidade média de finalizações necessárias para cada gol marcado.

Tabela 11 – Os 10 jogadores que mais finalizaram no campeonato.

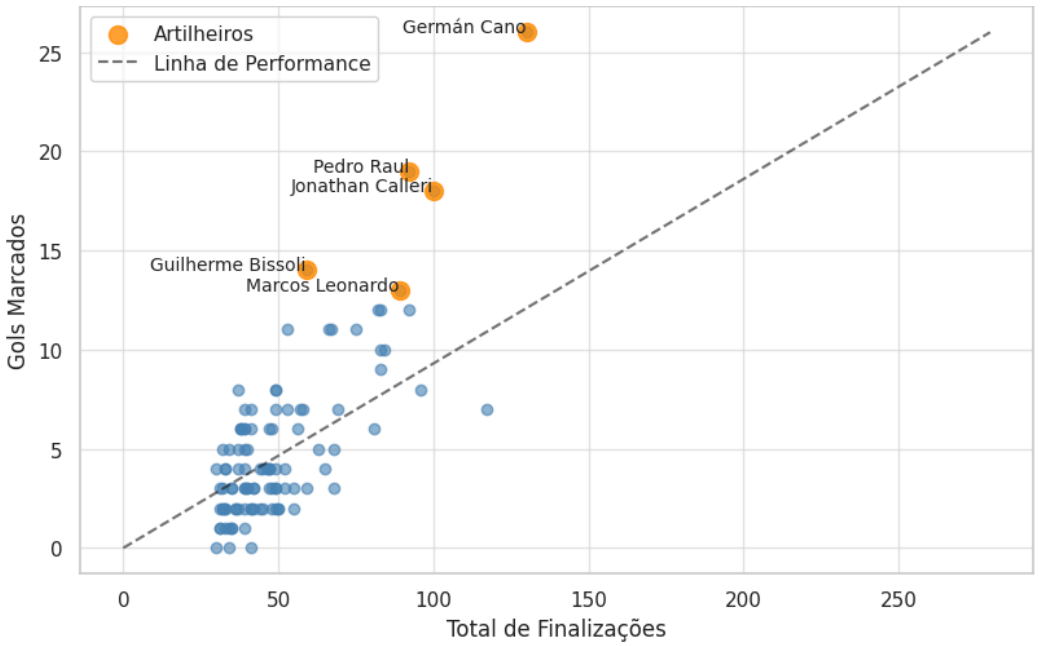
	Jogador	Posição	Nº de Finalizações	Gols Marcados	Gol/Finalização	Finalização/Gol
0	Germán Cano	F	130	26	0.200000	5.000000
1	Gustavo Scarpa	M	117	7	0.059829	16.714286
2	Jonathan Calleri	F	100	18	0.180000	5.555556
3	Wellington Rato	M	96	8	0.083333	12.000000
4	Pedro Raul	F	92	19	0.206522	4.842105
5	Hulk	F	92	12	0.130435	7.666667
6	Marcos Leonardo	F	89	13	0.146067	6.846154
7	Róger Guedes	F	84	10	0.119048	8.400000
8	Alef Manga	M	83	9	0.108434	9.222222
9	Steven Mendoza	M	83	10	0.120482	8.300000

Fonte: do autor (2024)

A taxa de gols por finalização indica o quão eficiente é cada atleta em converter suas finalizações em gols. Quanto maior este valor, melhor é o desempenho do jogador em termos de precisão nas finalizações. A taxa de finalização por gol representa o número médio de finalizações que o jogador precisa realizar para marcar um gol. Neste caso, um valor menor indica um desempenho superior, pois o jogador necessita de menos tentativas para atingir o sucesso. Dessa forma, a segunda métrica é complementar à primeira, proporcionando uma visão sobre a consistência do jogador em transformar oportunidades em gols.

Após a análise da taxa de conversão, foi feito um comparativo visual de gols em relação ao total de finalizações através de um gráfico de dispersão. A representação gráfica facilita a visualização do desempenho relativo entre diferentes jogadores, permitindo identificar aqueles que se destacaram no torneio.

Figura 24 – Comparativo de Gols por Total de Finalizações

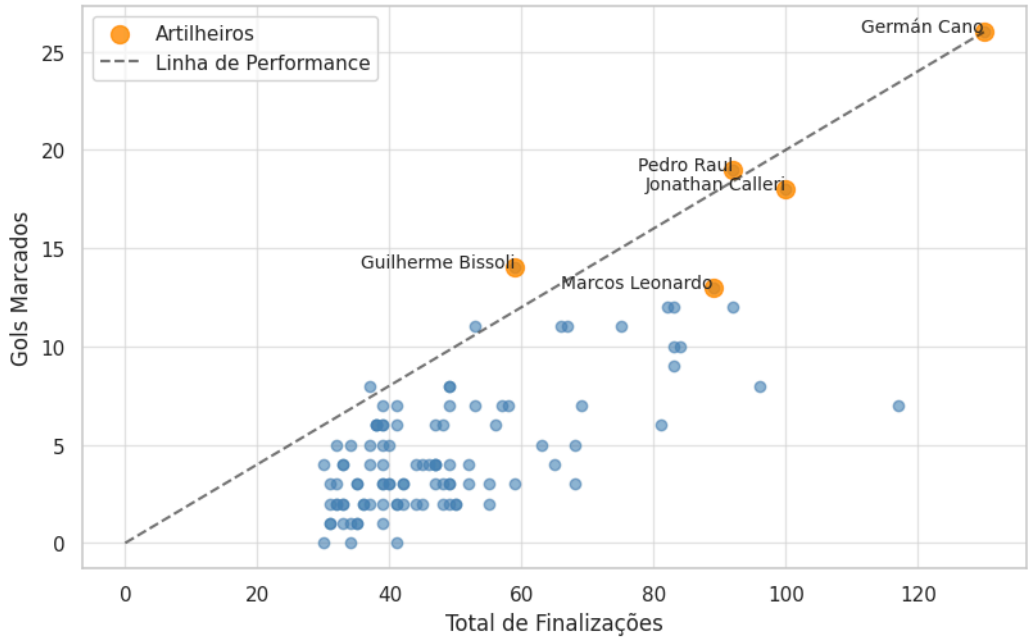


Fonte: do autor (2024)

Para o gráfico representado na Figura 24, foram considerados somente os jogadores com mais de 30 finalizações ao longo de toda a temporada da liga. Esse filtro foi aplicado para focar nos atletas com uma frequência de finalizações mais relevante, tornando a análise mais direcionada para jogadores com maior presença ofensiva. A linha tracejada representa a taxa de conversão média do campeonato. Jogadores que se posicionam acima desta linha apresentam uma eficiência superior, convertendo suas finalizações em gols com uma taxa mais elevada. Em contrapartida, os atletas abaixo da linha indicam uma eficiência de finalização inferior. É interessante observar que todos os maiores artilheiros se posicionam acima da linha.

Neste contexto, Germán Cano, do Fluminense Football Club se destacou, tanto no número de finalizações quanto na eficiência, estando acima da linha de performance e demonstrando sua capacidade de converter chances em gols. Na Figura 25, Cano foi considerado o balizador para a linha de performance. Sua taxa de conversão foi muito acima da média, mais que o dobro, mas ainda assim outros dois atletas, Pedro Raul e Guilherme Bissoli, conseguiram a façanha de superar este valor.

Figura 25 – Comparativo de Gols por Total de Finalizações.



Fonte: do autor (2024)

5.4.2 Taxa de Gols por Finalização no Gol

Nesta seção, a análise busca observar a capacidade de um jogador em converter finalizações que foram diretamente no gol, ou seja, aquelas que resultaram em uma defesa, trave ou gol.

Tabela 12 – As 10 melhores taxas de gols por chutes no alvo do campeonato.

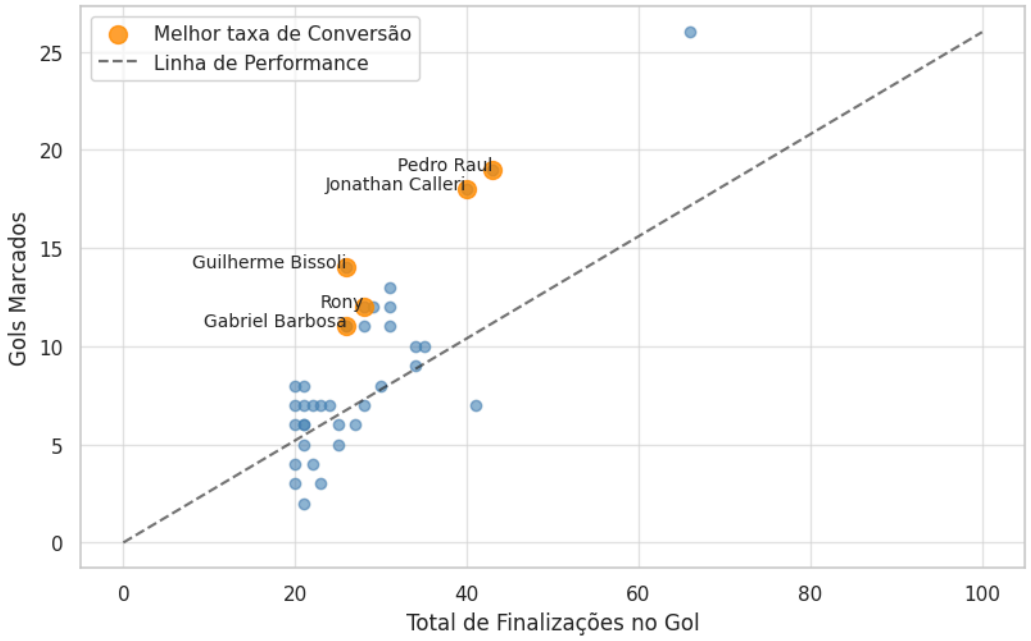
	Jogador	Posição	Nº de Chutes no Alvo	Gols Marcados	Gol/Chute no Alvo	Chute no Alvo/Gol
0	Guilherme Bissoli	F	26	14	0.54	1.86
1	Jonathan Calleri	F	40	18	0.45	2.22
2	Pedro Raul	F	43	19	0.44	2.26
3	Rony	F	28	12	0.43	2.33
4	Gabriel Barbosa	F	26	11	0.42	2.36
5	Marcos Leonardo	F	31	13	0.42	2.38
6	Pedro	F	26	11	0.42	2.36
7	Hulk	F	29	12	0.41	2.42
8	Yuri Alberto	F	20	8	0.40	2.50
9	Germán Cano	F	66	26	0.39	2.54

Fonte: do autor (2024)

De forma análoga à análise primária, a Tabela 12 apresenta somente jogadores com mais de 20 finalizações no campeonato, isto é, os mais envolvidos no ataque de maneira objetiva. Os resultados foram ordenados considerando as melhores taxas de conversão de chutes no alvo em gol.

Deve-se entender que quanto maior o valor da taxa de “Gol/Chute no Alvo”, melhor o desempenho do jogador em converter chances. Por outro lado, a coluna “Chute no Alvo/Gol” é o inverso da anterior e representa a quantidade média de finalizações no gol necessárias para marcar, em que valores menores indicam desempenho superior. Nesse caso, Cano, que foi o artilheiro do campeonato, apresentou somente a décima melhor conversão entre os atletas avaliados no recorte.

Figura 26 – Comparativo de Gols por Finalização no Gol.



Fonte: do autor (2024)

Os mesmos cinco atletas posicionados no topo da Tabela 12 são os mesmos destacados na Figura 26, sendo os atletas com melhor taxa de conversão de gols por chutes no alvo: Guilherme Bissoli, Jonathan Calleri, Pedro Raul, Rony e Gabriel Barbosa. É interessante ressaltar que, dentre os cinco, Pedro Raul foi o que mais marcou gols, mas não foi o atleta com melhor taxa de conversão. Também, Bissoli

foi o que menos finalizou (dentre os cinco), juntamente com Gabriel, mas foi o atacante mais certeiro.

5.4.3 Taxa de Gols por xG

Para a análise final, na primeira tabela, dentre todos os atletas do campeonato, foram selecionados os que tiveram o maior valor acumulado de gols esperados. Em outras palavras, a coluna “xG V2” representa o valor correspondente ao somatório total dos valores de gols esperados, considerando cada finalização no alvo realizada pelos jogadores durante a temporada. O valor de xG acumulado deve ser comparado ao número real de gols marcados, uma vez que tenta prever seu valor. Quanto mais próximo xG estiver do número de gols reais, melhor foi a previsão. Para os casos de Marcos Leonardo, Rony, Pedro e Steven Mendoza, o modelo performou entregando um resultado muito próximo da realidade.

Além de comparar o desempenho do Modelo V2 com os gols reais marcados no campeonato, foram incluídos os valores acumulados do Modelo do SofaScore para comparação entre modelos.

Tabela 13 – Os 10 maiores gols esperados acumulados do campeonato.

	Jogador	Posição	Gols Marcados	xG V2	xG SofaScore
0	Germán Cano	F	26	16.81	18.46
1	Jonathan Calleri	F	18	14.13	13.34
2	Pedro Raul	F	19	13.36	12.60
3	Marcos Leonardo	F	13	12.35	13.41
4	Rony	F	12	12.11	11.37
5	Guilherme Bissoli	F	14	11.24	11.60
6	Pedro	F	11	10.62	12.09
7	Steven Mendoza	M	10	10.45	11.34
8	Gabriel Barbosa	F	11	9.94	11.08
9	David Terans	M	12	8.84	9.81

Fonte: do autor (2024)

Em seguida, na Tabela 14, é apresentada a taxa de gols por gols esperados, que mede o desempenho real dos jogadores. Para isso, foram considerados somente os atletas com mais de 10 gols marcados no campeonato, novamente para direcionar as análises para o campo ofensivo e evitar *outliers*. Pela primeira vez, dentre as análises realizadas, surge um defensor. Luan Cândido, lateral esquerdo do RB Bragantino, performou muito além do previsto e teve a segunda melhor taxa de Gols/xG do campeonato, ficando atrás apenas de Hulk, atacante do Atlético Mineiro.

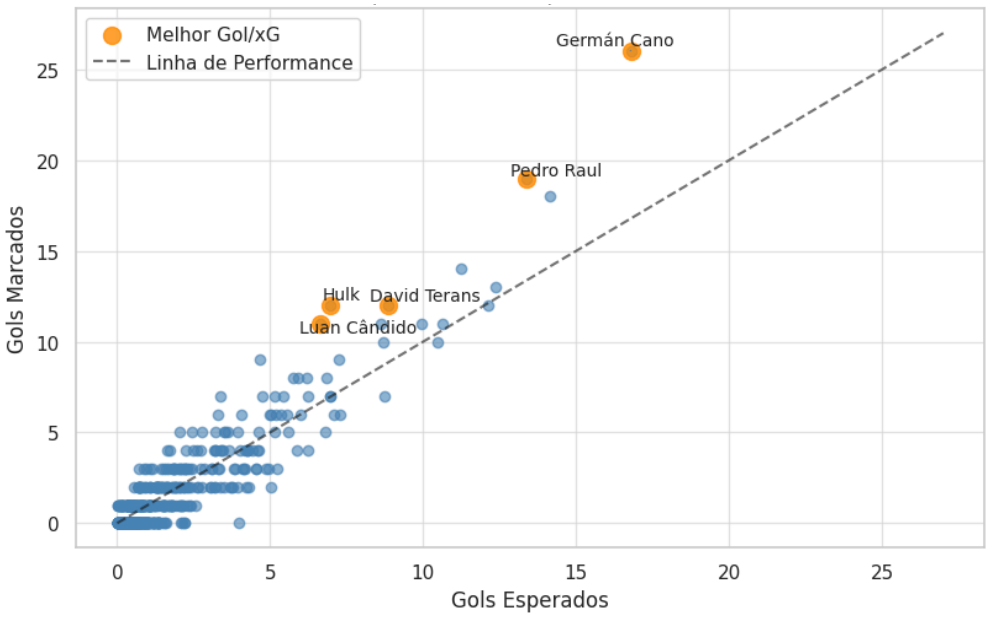
Tabela 14 – Os 10 jogadores com maior taxa de Gols por xG no campeonato.

	Jogador	Posição	Gols Marcados	xG V2	Gols/xG
0	Hulk	F	12	6.98	1.72
1	Luan Cândido	D	11	6.62	1.66
2	Germán Cano	F	26	16.81	1.55
3	Pedro Raul	F	19	13.36	1.42
4	David Terans	M	12	8.84	1.36
5	Jonathan Calleri	F	18	14.13	1.27
6	Luciano	M	11	8.63	1.27
7	Guilherme Bissoli	F	14	11.24	1.25
8	Róger Guedes	F	10	8.69	1.15
9	Gabriel Barbosa	F	11	9.94	1.11

Fonte: do autor (2024)

O gráfico de dispersão da Figura 27 apresenta a relação entre os gols marcados e os valores de xG acumulados para todos os jogadores do campeonato. A linha de performance, nesse caso, representa uma função linear onde os gols marcados possuem o mesmo valor dos gols esperados. Os cinco primeiros da Tabela 14 foram destacados dentre os demais, de forma a mostrar (visualmente) como as altas taxas de conversão se diferenciam entre os demais. Como era de se esperar, os pontos seguem aproximadamente a diagonal, o que significa que os gols esperados dos jogadores estão próximos dos gols reais (Eggels, 2016).

Figura 27 – Comparativo de gols por xG acumulado.



Fonte: do autor (2024)

5.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Com isso, concluíram-se as análises de performance realizadas sobre os atletas. Utilizando as métricas propostas na fundamentação teórica, para avaliação de jogadores, foi possível perceber quem foram os que melhor desempenharam em suas funções, sendo os principais destaques do campeonato:

a) German Cano: o artilheiro e jogador com maior xG acumulado;
b) Guilherme Bissoli: jogador com a melhor taxa de gols por chute no gol;
c) Pedro Raul: jogador com a melhor taxa global – gol por finalização;
d) Hulk: dono da melhor taxa de gols por xG.

Sendo o xG o objeto de estudo principal deste trabalho, o que pode ser concluído é que Hulk, do Atlético Mineiro, foi o jogador do Campeonato Brasileiro de 2022 que mais criou gols de situações improváveis de gol. Hulk, por assim dizer, contribuiu muito mais para sua equipe do que se era esperado dele.

6 CONCLUSÃO

6.1 CONCLUSÕES

O objetivo geral deste trabalho era desenvolver um modelo de gols esperados para avaliar o desempenho de jogadores da Série A do Campeonato Brasileiro de Futebol de 2022. Durante o desenvolvimento, foram construídos dois modelos – um básico e um aprimorado – que possibilitaram a criação de uma ferramenta capaz de estimar a probabilidade de uma finalização resultar em gol, oferecendo uma análise quantitativa da eficiência dos jogadores em diferentes contextos de jogo. Esse desenvolvimento foi a chave para alcançar o objetivo geral proposto.

Além do objetivo geral, foram delineados os objetivos específicos, os quais também foram atendidos ao longo do trabalho. O primeiro, que visava avaliar a eficácia dos jogadores na criação e conversão de oportunidades de gol, foi atingido por meio da análise dos resultados produzidos pelos modelos. As métricas de desempenho revelaram como cada jogador se destacou em suas habilidades de finalização.

O segundo objetivo específico, voltado para a identificação dos jogadores com melhor desempenho na conversão de oportunidades em gol, foi alcançado quando feita a comparação das taxas de conversão dos principais artilheiros do campeonato, em que métricas específicas definidas na fundamentação teórica foram utilizadas para identificar os finalizadores mais eficientes.

Em relação ao terceiro objetivo, que buscava validar os resultados do modelo proposto em comparação a um modelo comercial de referência, o trabalho atingiu esse ponto ao analisar o desempenho do modelo avançado (V2) em relação ao modelo de xG fornecido pela plataforma SofaScore. Essa comparação proporcionou uma validação externa para o modelo desenvolvido, confirmando sua qualidade e capacidade preditiva. Entretanto, aqui, o modelo não performou tão bem quanto o modelo comercial, o que era esperado, uma vez que empresas privadas que atuam no setor possuem mais recursos para construir um modelo mais robusto.

O quarto e último objetivo, que consistia em aprimorar as previsões de gols utilizando métricas específicas do campeonato de 2022, foi atendido pela inclusão de variáveis adicionais no modelo avançado, permitindo previsões mais precisas e adequadas ao contexto específico do campeonato brasileiro. Dessa forma, pode-se

afirmar que tanto o objetivo geral quanto os objetivos específicos foram atendidos, consolidando o propósito estabelecido para o presente estudo de caso.

6.2 TRABALHOS FUTUROS

Conforme mencionado anteriormente na seção 4.5.2, uma possibilidade de desenvolvimento que pode enriquecer o campo de análise é a criação de um modelo de Expected Goals on Target (xGOT). Esse modelo permitiria estimar a probabilidade de uma finalização no alvo resultar em gol após o chute. Esse modelo permite incorporar variáveis adicionais além dos aspectos já considerados. Assim, o desenvolvimento de um modelo xGOT proporcionaria uma análise ainda mais detalhada do desempenho ofensivo e, também, defensivo das equipes, expandindo a compreensão dos fatores que influenciam o sucesso de uma finalização.

A inclusão de um modelo xGOT é vista como uma oportunidade de avanço pois possibilitaria não apenas avaliar a eficácia dos jogadores em acertar o gol, mas também medir a eficiência dos goleiros em impedir a entrada da bola. Um modelo de xGOT considera a posição do goleiro no gol, qualidade do goleiro e coordenadas finais da bola na boca do gol. O presente estudo não abarcou o desenvolvimento desse modelo mais avançado, pois essa abordagem não se inclui no escopo original do trabalho, porém fica a possibilidade de posteriormente trabalhar-se em uma proposta dessa natureza.

REFERÊNCIAS

ANZER, G.; BAUER, P. **A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer)**. 2021. Disponível em: https://www.frontiersin.org/articles/10.3389/fspor.2024.1348983/full. Acesso em: 29 set. 2024.

BERNSTEIN, Peter L. **Desafio aos Deuses: A Fascinante História do Risco**. 1. ed. Rio de Janeiro: Alta Books, 2018.

CBF ACADEMY. **Ciência de Dados no Futebol**. Disponível em: https://www.cbfacademy.com.br/pt-br/cursos/89-ciencia-de-dados-no-futebol. Acesso em: 13 nov. 2024.

CIÊNCIA DA BOLA. **Scouting no Futebol**. 2024. Disponível em: https://www.cienciadabola.com.br/scouting/. Acesso em: 1 nov. 2024.

CIÊNCIA DA BOLA. **Por que os gols de falta no futebol estão diminuindo?**. Disponível em: https://www.cienciadabola.com.br/por-que-os-gols-de-falta-no-futebol-estao-diminuindo/#:~:text=Com%20exce%C3%A7%C3%A3o%20da%20edi%C3%A7%C3%A3o%20de,os%20culpados%20por%20essa%20redu%C3%A7%C3%A3o%3F. Acesso em: 21 nov. 2024.

CNN BRASIL. **Brasileirão é a segunda liga mais imprevisível do mundo, aponta levantamento**. 2024. Disponível em: https://www.cnnbrasil.com.br/esportes/brasileirao/brasileirao-e-a-segunda-liga-mais-imprevisivel-do-mundo-aponta-levantamento/#:~:text=A%20S%C3%A9rie%20A%20do%20Campeonato,acordo%20com%20levantamento%20do%20BolaVip. Acesso em: 13 nov. 2024.

CORREIO DO ESTADO, 2020. **Futebol brasileiro tenta salvar 156 mil empregos durante a crise do coronavírus.** Disponível em: https://correiodoestado.com.br/esportes/futebol-brasileiro-tenta-salvar-156-mil-empregos-durante-a-crise-do-co/370169/. Acesso em: 09 de abril de 2024.

DATA SCIENCE PROCESS ALLIANCE. **What is CRISP DM?** 2024. Disponível em: https://www.datascience-pm.com/crisp-dm-2/. Acesso em: 14 mai. 2024.

EGGELS, H. P. **Expected Goals in Soccer: Explaining Match Results using Predictive Analytics**. 2016. Disponível em: https://pure.tue.nl/ws/files/46945853/855660-1.pdf. Acesso em: 09 abr. 2024.

ÉPOCA NEGÓCIOS. **Como o Liverpool usou dados e tecnologia para chegar à final da Champions League**. 2019. Disponível em: https://epocanegocios.globo.com/Empresa/noticia/2019/06/como-o-liverpool-usou-dados-e-tecnologia-para-chegar-final-da-champions-league.html. Acesso em: 13 nov. 2024.

FAIRCHILD, Alexander; KOKKODIS, Marios; PELECHRINIS, Konstantinos. **Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality**. Journal of Sports Analytics, v. 4, n. 3, p. 1-10, 2018. Disponível em:https://www.researchgate.net/publication/324738240_Spatial_analysis_of_shots_in_MLS_A_model_for_expected_goals_and_fractal_dimensionality. Acesso em: 13 jul. 2024.

FLUMINENSE. **Fluminense fecha parceria com o SofaScore**. Disponível em: https://www.fluminense.com.br/noticia/fluminense-fecha-parceria-com-o-sofascore. Acesso em: 13 nov. 2024.

GLOBO ESPORTE. **Tudo igual dentro das 4 linhas: CBF padroniza gramados das séries A e B**. 2016. Disponível em: https://ge.globo.com/futebol/noticia/2016/01/tudo-igual-dentro-das-4-linhas-cbf-padroniza-gramados-das-series-e-b.html. Acesso em: 1 nov. 2024.

GLOBO ESPORTE. **Como a tecnologia ajuda Palmeiras a melhorar jogo, contratar reforços e até buscar um novo Endrick**. 2024. Disponível em: https://ge.globo.com/futebol/times/palmeiras/noticia/2024/03/20/como-a-tecnologia-ajuda-palmeiras-a-melhorar-jogo-contratar-reforcos-e-ate-buscar-um-novo-endrick.ghtml. Acesso em: 1 nov. 2024.

GONÇALVES, Lucas de Camargo; MENEZES, Mário Olímpio de. **O impacto da tecnologia no futebol brasileiro na Era Digital: uma análise das ferramentas de análise de desempenho e sua influência na gestão esportiva**. 2023. Trabalho de Conclusão de Curso (Graduação em Gestão Esportiva) – Universidade Presbiteriana Mackenzie, São Paulo, 2023. Disponível em: https://adelpha-api.mackenzie.br/server/api/core/bitstreams/8f21410e-9269-4ebc-a341-df4e1f90e72f/content. Acesso em: 13 nov. 2024.

GOOGLE COLAB. **Colab Research Platform**. 2024. Disponível em: https://colab.research.google.com/#scrollTo=UdRyKR44dcNI. Acesso em: 16 jun. 2024.

HAKES, Jah K.; SAUER, Raymond D. **An Economic Evaluation of the Moneyball Hypothesis.** The Journal of Economic Perspectives Vol. 20, No. 3, p. 173-186. Disponível em: https://www.jstor.org/stable/30033672. Acesso em: 14 de maio de 2024.

HAMIL, S.; WALTERS, G.; WATSON, L. **The Business of Football: A Game of Two Halves?** In: **Football in the Digital Age: Whose Game Is It Anyway?** Routledge, 2010. p. 19-30. Disponível em: https://library.olympics.com/Default/doc/SYRACUSE/43201/a-game-of-two-halves-the-business-of-football-ed-by-sean-hamil-jonathan-michie-and-christine-oughton?_lg=en-GB. Acesso em: 13 jul. 2024.

HOSMER, David W., LEMESHOW, Stanley. **Applied Logistic Regression**. 2000. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/0471722146. Acesso em: 1 nov. 2024.

IFFHS. **IFFHS MEN'S STRONGEST NATIONAL LEAGUE IN THE WORLD - THE TOP 100**. 2024. Disponível em: https://iffhs.com/posts/3336#. Acesso em: 18 nov. 2024.

IFFHS. **IFFHS MEN'S STRONGEST NATIONAL LEAGUE IN THE WORLD - THE TOP 100**. 2023. Disponível em: https://www.iffhs.com/posts/2483. Acesso em: 18 nov. 2024.

IFFHS. **IFFHS MEN'S STRONGEST NATIONAL LEAGUE IN THE WORLD - THE TOP 100**. 2022. Disponível em: https://www.iffhs.com/posts/1607. Acesso em: 18 nov. 2024.

IFFHS. **IFFHS MEN'S STRONGEST NATIONAL LEAGUE IN THE WORLD - THE TOP 100**. 2021. Disponível em: https://www.iffhs.com/posts/911. Acesso em: 18 nov. 2024.

INSPER. **Saiba Como Funciona a Mineração de Dados (Ou Data Mining)**. 2022. Disponível em: https://www.insper.edu.br/noticias/mineracao-de-dados-ou-data-mining/. Acesso em: 14 maio 2024.

DUTTA, Nabanita *et al*. **Centrifugal Pump Cavitation Detection Using Machine Learning Algorithm Technique**. Disponível em: https://www.researchgate.net/publication/343792287_Centrifugal_Pump_Cavitation_Detection_Using_Machine_Learning_Algorithm_Technique. Acesso em: 13 nov. 2024.

KHDER, M. A. **Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application**. 2021. Disponível em: https://www.i-csrs.org/Volumes/ijasca/2021.3.11.pdf. Acesso em: 14 out. 2024.

KLEPPMANN, M. **Designing Data-Intensive Applications**. Disponível em: https://archive.org/details/designing-data-intensive-applications-th/page/n9/mode/2u. Acesso em: 13 nov. 2024.

LANCE!. **Premiação do Brasileirão 2024: entenda como os valores são calculados**. 2024. Disponível em: https://www.lance.com.br/lancebiz/financas/premiacao-do-brasileirao-2024-entenda-como-os-valores-sao-calculados.html. Acesso em: 2 nov. 2024.

LIPSCHUTZ, S. **Álgebra Linear**. 10. ed. 2019. Disponível em: https://www.academia.edu/43438879/%C3%81lgebra_Linear_D%C3%89CIMA_EDI%C3%87%C3%83O. Acesso em: 1 nov. 2024.

LOPES, P. **Geometry of Shooting**. In: **Soccermatics**. Disponível em: https://soccermatics.readthedocs.io/en/latest/lesson2/GeometryOfShooting.html. Acesso em: 1 nov. 2024.

MACKENZIE. **Análise de Dados no Futebol**. Disponível em: https://dspace.mackenzie.br/items/31060f26-880f-46ec-980f-f423ce214d56. Acesso em: 2 nov. 2024.

MEAD, J.; O'HARE, A.; McMENEMY, P. **Expected goals in football: Improving model performance and demonstrating value**. PLoS ONE, v. 18, n. 4, p. e0282295, 2023. Disponível em: https://doi.org/10.1371/journal.pone.0282295. Acesso em: 13 nov. 2024.

MIRROR. **Kevin De Bruyne uses data to boost football performance**. Disponível em: https://www.mirror.co.uk/sport/football/news/kevin-de-bruyne-uses-data-23870686. Acesso em: 1 nov. 2024.

MITCHELL, R. **Web Scraping with Python: Collecting More Data from the Modern Web**. O'Reilly Media, 2018. Disponível em: https://www.oreilly.com/library/view/web-scraping-with/9781491985564/. Acesso em: 13 jul. 2024.

MCKINNEY, W. **Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython**. O'Reilly Media, 2017. Disponível em: https://www.oreilly.com/library/view/python-for-data/9781491957653/. Acesso em: 13 jul. 2024.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. Wiley, 2018. Disponível em: https://kolegite.com/EE_library/books_and_lectures/Douglas%20C.%20Montgomery,%20George%20C.%20Runger%20-%20Applied%20Statistics%20and%20Probability%20for%20Engineers-Wiley%20(2018).pdf. Acesso em: 13 nov. 2024.

REEP, C. BENJAMIN, B. **Skill and Chance in Association Football.** Journal of the Royal Statistical Society, 1968, Series A (General) Vol. 131, No. 4, p. 581-585. Disponível em: https://www.jstor.org/stable/2343726. Acesso em: 14 de maio de 2024.

SCHOLTES, A.; KARAKUŞ, O. **Bayes-xG: Player and Position Correction on Expected Goals (xG) using Bayesian Hierarchical Models**. Frontiers in Sports and Active Living, 2024. Disponível em: https://www.frontiersin.org/journals/sports-and-active-living/articles/10.3389/fspor.2024.1348983/full. Acesso em: 1 nov. 2024.

SHEARER, C. **The CRISP-DM Model: The New Blueprint for Data Mining**. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000. Disponível em: https://www.academia.edu/42079490/CRISP_DM_The_New_Blueprint_for_Data_Mining_Colin_Shearer_Fall_2000_. Acesso em: 13 jul. 2024.

SIRISURIYA, D. S. **A Comparative Study on Web Scraping**. In: **Proceedings of 8th International Research Conference, KDU**. 2015. Disponível em: http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y. Acesso em: 14 out. 2024.

SOFASCORE. **About Us**. 2024. Disponível em: https://corporate.sofascore.com/about. Acesso em: 16 jul. 2024.

SOFASCORE. **Match: Internacional vs Palmeiras**. Disponível em: https://www.sofascore.com/football/match/internacional-palmeiras/nOsqO#id:10114142. Acesso em: 1 nov. 2024.

SOFASCORE. **Match: Coritiba vs São Paulo**. Disponível em: https://www.sofascore.com/football/match/coritiba-sao-paulo/GOsHO#id:10114000. Acesso em: 1 nov. 2024.

SOFASCORE. **Match: Goiás vs Corinthians**. Disponível em: https://www.sofascore.com/football/match/goias-corinthians/hOskO#id:10113966. Acesso em: 1 nov. 2024.

STATS PERFORM. **Introducing Expected Goals on Target (xGOT)**. 2024. Disponível em: https://www.statsperform.com/resource/introducing-expected-goals-on-target-xgot/. Acesso em: 13 nov. 2024.

STATS PERFORM. **Opta Event Definitions**. 2024. Disponível em: https://www.statsperform.com/opta-event-definitions/. Acesso em: 21 nov. 2024.

THE IFAB. **Laws of the Game 2024-25**. Disponível em: https://downloads.theifab.com/downloads/laws-of-the-game-2024-25-brazilian-portuguese?l=en. Acesso em: 1 nov. 2024.

TRANSFERMARKT. **Campeonato Brasileiro Série A**. 2024. Disponível em: https://www.transfermarkt.com.br/campeonato-brasileiro-serie-a/transfers/wettbewerb/BRA1. Acesso em: 16 jun. 2024.

VANDERPLAS, Jake. **Python Data Science Handbook: Essential Tools for Working with Data**. O'Reilly Media, 2016. Disponível em: https://www.oreilly.com/library/view/python-data-science/9781491912126/. Acesso em: 13 jul. 2024.

WITHOEFT, Jordy R. **Além do número de finalizações: criação e aplicação de um modelo de estimação de gols esperados (xG)**. 2020. Trabalho de Conclusão de Curso (Graduação em Matemática Industrial) – Universidade Federal do Paraná, Curitiba, 2020. Disponível em: https://acervodigital.ufpr.br/handle/1884/71067. Acesso em: 13 nov. 2024.

YEN, Show-Jane; LEE, Yue-Shi. **Cluster-based under-sampling approaches for imbalanced data distributions**. Expert Systems with Applications, v. 36, n. 3, p. 5718–5727, 2009. Disponível em: https://sci2s.ugr.es/keel/pdf/specific/articulo/yen_cluster_2009.pdf. Acesso em: 24 de outubro de 2024.