How to Win: Uma Análise Exploratória de Dados de Partidas Vitoriosas da La Liga usando o FP-Growth

Lucca C. Augusto[1], **André L. M. Dutra**[1], **Carlos M. Silva**[1]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{andre, luccaaugusto, carlosmagalhaes}@dcc.ufmg.br

Resumo. *O futebol é um dos esportes mais relevantes do mundo, e a aplicação de análises orientadas a dados neste âmbito se mostra cada vez mais frequente. Este estudo se propõe a investigar padrões que influenciam vitórias na La Liga, visando auxiliar treinamentos das equipes e validar ou refutar ideias do senso comum sobre características associadas à vitória. Utilizando dados detalhados das temporadas de 2014-2015 a 2020-2021, foi aplicado o algoritmo FP-Growth para identificar características determinantes como posse de bola, precisão de passes e disciplina em campo. Os experimentos revelaram 43 padrões iniciais, refinados para 27 após análise de aleatoriedade. Os resultados apontam que chutes, chutes a gol e posse de bola são fundamentais para aumentar as chances de vitória, indo de acordo com a análise quantitativa e com percepções tradicionais sobre o futebol.*

1. Introdução

O futebol é um dos esportes mais relevantes do mundo, fomentando múltiplos campeonatos de escala continental e mundial, como a conhecida Copa do Mundo [Fifa 2024]. Neste contexto, a La Liga Santander, o principal campeonato espanhol de futebol, se destaca como uma das cinco grandes competições europeias, tendo como participantes times de relevância mundial, como o Real Madrid e o Barcelona [Fox 2024], e valendo cerca de 5 bilhões de euros em 2021 [Faria 2021]. Com a consolidada relevância do esporte, técnicas baseadas em ciência de dados têm sido cada vez mais buscadas pelos times como método de análise de desempenho e aporte à elaboração de estratégias, sendo um fator determinante em times de relevância mundial, como o Liverpool [StartSe 2023].

Com isso, este estudo se propõe a realizar uma análise exploratória de dados de partidas da La Liga Santander, com o objetivo de encontrar, dentre as partidas vencedoras, padrões entre as características mais comumente associadas no senso comum à vitória. Subsequentemente, esta análise também se propõe a validar quais destas características do senso comum de fato possuem ocorrência frequente entre as partidas vitoriosas, bem como quais ocorrem de maneira conjunta e quais as regras de associação entre estas características.

Partindo de uma base de dados contendo múltiplas informações sobre cada partida da La Liga ao longo dos anos de 2014-2015 a 2020-2021, criamos uma nova base de dados de itemsets, onde cada itemset representa uma partida e cada item corresponde a uma característica associada no senso comum à vitória, desde que o time vencedor da partida possua esta característica. Dentre as características definidas, temos como exemplo ”Time vencedor tem menos chutes para fora que o adversário”e ”Time vencedor tem mais posse

de bola que o adversário", por exemplo, e todas as características foram calculadas a partir de informações contidas na base de dados. Por fim, aplicamos o FP-Growth, uma abordagem de mineração de itemsets frequentes, ao conjunto de itemsets, encontrando itemsets frequentes e regras de associação entre os dados.

Este trabalho está dividido em cinco seções. A seção seguinte, a seção 2, discute trabalhos relacionados à proposta apresentada neste artigo. Em seguida, na seção 3 descrevemos a metodologia aplicada no artigo, como o tratamento dos dados utilizados e a técnica de mineração aplicada, seguido da seção 4, onde apresentamos e discutimos os resultados obtidos. Por fim, na seção 5, finalizamos a discussão da seção anterior e apresentamos perspectivas futuras de pesquisa.

2. Trabalhos Relacionados

Dada a relevância do esporte, diversas abordagens orientadas a dados já foram aplicadas no contexto do futebol. Rein e Memmert descrevem múltiplas abordagens de análise de dados aplicadas ao futebol e apontam que, atualmente, as abordagens mais populares se baseiam em aprendizado de máquina [Rein and Memmert 2016]. Decroos et al., por sua vez, propõem a mineração de itemsets como forma de realizar descoberta de táticas em dados espaço-temporais de partidas [Decroos et al. 2018]. Por fim, Klivansky e El-Hajj propõem uma abordagem baseada em mineração de itemsets para a análise de dados de partidas da La Liga [Klivansky and El-Hajj 2018], embora, ao contrário deste artigo, se resumam a usar apenas o Apriori como abordagem de descoberta de itemsets.

3. Metodologia

Nesta seção, apresentamos o processo de preparação dos dados para o estudar características que influenciem os resultados de jogos de futebol na La Liga. Utilizamos um dataset público que contém dados detalhados das temporadas de 2014-2015 a 2020-2021. Com base nesse dataset, selecionamos características consideradas determinantes para a vitória de uma equipe, como posse de bola, precisão de passes e disciplina em campo. Essas características foram utilizadas para construir um novo dataset focado em partidas com vitórias, excluindo empates e instâncias sem relevância estatística. Para a análise dos padrões frequentes, aplicamos o algoritmo FP-Growth, com implementação em Python.

3.1. Dataset utilizado

O dataset base foi obtido na plataforma Kaggle [Naik 2022], onde temos os arquivos separados dos anos de 2014 a 2020, e o arquivo *combined_data_laliga.csv* que nada mais é que a concatenação de todos os arquivos dos anos.

Neste dataset, cada instância é uma partida de futebol disputada na La Liga Santander pelas temporadas 2014-2015 a 2020-2021, ambas inclusos, totalizando 2660 linhas, já que cada uma das 7 temporadas é composta de 38 rodadas com 10 partidas em cada rodada.

As informações disponíveis no dataset são basicamente dados sobre ambos os times no pós-jogo, suas colunas são, respectivamente:

- Home Team $\rightarrow$ Nome do time jogando em casa
- Away Team $\rightarrow$ Nome do time jogando fora de casa

- Score → Placar final
- Half Time Score → Placar no intervalo
- Match Excitement → Nível de empolgação do jogo (0-10)
- Home Team Rating → Avaliação do time da casa (0-10)
- Away Team Rating → Avaliação do time de fora (0-10)
- Home Team Possession % → Percentual de posse de bola do time da casa
- Away Team Possession % → Percentual de posse de bola do time de fora
- Home Team Off Target Shots → Chutes fora do alvo do time da casa
- Home Team On Target Shots → Chutes no alvo do time da casa
- Home Team Total Shots → Total de chutes do time da casa
- Home Team Blocked Shots → Chutes bloqueados do time da casa
- Home Team Corners → Escanteios a favor do time da casa
- Home Team Throw Ins → Laterais cobrados pelo time da casa
- Home Team Pass Success % → Percentual de passes certos do time da casa
- Home Team Aerials Won → Disputas aéreas ganhas pelo time da casa
- Home Team Clearances → Desarmes do time da casa
- Home Team Fouls → Faltas cometidas pelo time da casa
- Home Team Yellow Cards → Cartões amarelos recebidos pelo time da casa
- Home Team Second Yellow Cards → Segundo cartão amarelo recebido pelo time da casa
- Home Team Red Cards → Cartões vermelhos recebidos pelo time da casa
- Away Team Off Target Shots → Chutes fora do alvo do time de fora
- Away Team On Target Shots → Chutes no alvo do time de fora
- Away Team Total Shots → Total de chutes do time de fora
- Away Team Blocked Shots → Chutes bloqueados do time de fora
- Away Team Corners → Escanteios a favor do time de fora
- Away Team Throw Ins → Laterais cobrados pelo time de fora
- Away Team Pass Success % → Percentual de passes certos do time de fora
- Away Team Aerials Won → Disputas aéreas ganhas pelo time de fora
- Away Team Clearances → Desarmes do time de fora
- Away Team Fouls → Faltas cometidas pelo time de fora
- Away Team Yellow Cards → Cartões amarelos recebidos pelo time de fora
- Away Team Second Yellow Cards → Segundo cartão amarelo recebido pelo time de fora
- Away Team Red Cards → Cartões vermelhos recebidos pelo time de fora
- Home Team Goals Scored → Gols marcados pelo time da casa
- Away Team Goals Scored → Gols marcados pelo time de fora
- Home Team Goals Conceded → Gols sofridos pelo time da casa
- Away Team Goals Conceded → Gols sofridos pelo time de fora
- year → Ano da temporada do jogo (ex.: 2014 → Temporada 2014/2015)

3.2. Montagem dos Itemsets

Com isso foram separadas as principais características que, no senso comum, levam um time à vitória, as características foram:

- Time vencedor tem alto índice de passes certos (pelo menos 85%)
- Time vencedor tem maior posse de bola que o adversário

- Time vencedor tem menos chutes para fora que o adversário
- Time vencedor tem mais chutes ao gol que o adversário
- Time vencedor tem mais chutes que o adversário
- Time vencedor ganha mais bolas aéreas que o adversário
- Time vencedor tem menos faltas que o adversário (no mínimo 3 a menos)
- Time vencedor toma menos cartões que o adversário (no mínimo 2 a menos)
- Time vencedor tem mais jogadores que o adversário (cartões vermelhos a menos)

Após a separação das principais características, foi construído um dataset próprio em que cada linha representa uma partida com vitória, e as características (itens citados acima) compõem cada partida, aquelas que não possuíam nenhuma destas características foram descartadas com base em dois fatos:

- Sem relevância estatística $\rightarrow$ não representavam nem 1% do dataset
- Os métodos utilizados no estudo não conseguem lidar com itemsets vazios

O notebook *MakeDataset.ipynb* utilizado para a criação do dataset está disponível no repositório do estudo no GitHub [Augusto et al. a] [Augusto et al. b].

3.3. Abordagem de Mineração Aplicada

A metodologia escolhida foi a FP-Growth por três principais motivos: alta eficácia, alta eficiência e alta praticidade. O cálculo feito no algoritmo FP-Growth não é probabilístico, portanto é garantido o melhor resultado, em comparação com outros problemas computacionais que temos que usar heurísticas, neste caso a explosão combinatória é evitada sem aproximações, garantindo o resultado ótimo [Han and Pei 2000].

Em comparação com os outros algoritmos para mineração de padrões frequentes vistos, temos uma boa eficiência, sendo um algoritmo mais rápido que o algoritmo de Força Bruta, que o algoritmo Apriori e que as duas variações do Eclat (com ou sem os Diffsets). A praticidade vem por meio da biblioteca pyfpgrowth do python [fpg], que implementa o algoritmo com grande eficiência, fornecendo funções com parâmetros simples e explicativos.

4. Calibragem do Modelo e Resultados Obtidos

Todos os dados e números citados a partir daqui também podem ser verificados no notebook criado no ambiente do Google Colaboratory [Augusto et al. c].

Inicialmente para conhecer melhor o dataset foram testados vários valores de suporte, parâmetro de entrada do algoritmo FP-Growth, após análise qualitativa foi definido o valor de 12.5% (2-3) de modo que foram encontrados 43 padrões. Essa porém seria uma análise muito simplista e generalizada, sem dar importância ao dado utilizado e o que ele significa, devido à forma de construção deste dataset os itens significam basicamente a ocorrência ou não daquela característica dentre aquelas definidas na seção 3.2.

Alguns destes dados possuem, estatisticamente, 50% de chance de ocorrer, outros possuem uma menor chance, que não cabe o cálculo neste estudo (como "Qual a chance de um time ter pelo menos 2 cartões a menos que o outro time?") já que eles por si só podem gerar um estudo sobre. Como a ideia hipótese é que todas essas características realmente contribuem para a vitória, foi considerado que todos os itens possuem 50% de chance de ocorrer.

Com isso em mente, foi feita uma outra análise que se baseia no seguinte questionamento: “Quais destes 43 padrões não são aleatórios?”, ou seja, qual tem índice de ocorrência maior do que a probabilidade aleatória do padrão encontrado, com base na premissa anterior de que todos os itens têm 50% de chance de ocorrer.

Assim definimos esta tabela com um limiar inferior de 110% do valor aleatório, para garantir uma margem de segurança, aqueles padrões que possuírem um índice de ocorrência menor serão descartados.

Quantidade de Itens no *Itemset*	Quantidade de *Itemsets*	Valor mínimo de ocorrências
1	2	1076
2	18	538
3	20	269
4	3	135

Tabela 1. Tabela de Frequência de *Itemsets*

Depois do procedimento descrito temos uma base de dados com a distribuição mostrada na Figura 1, totalizando 27 conjuntos.

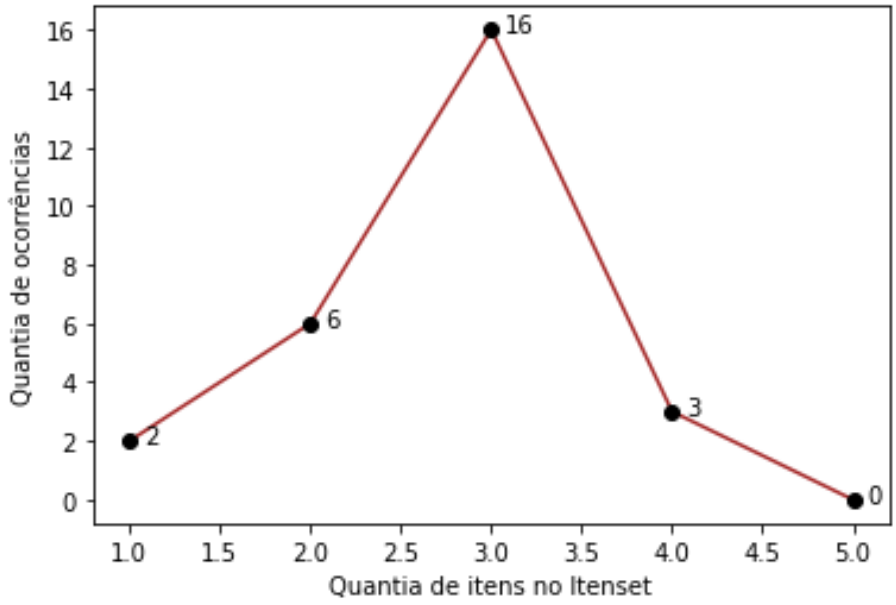

Figura 1. Distribuição da quantia de itemsets por tamanho de itemset

Agora, apenas com os dados considerados não aleatórios, partimos para uma exploração mais profunda, uma tentativa de correlação entre os padrões encontrados pelo algoritmo FP-Growth, que nos propicia a criação de regras, que são nada mais que uma correlação matemática de um padrão para outro.

Ao gerar as regras temos que escolher um limiar inferior de confiança, após análises qualitativas foi escolhido o valor de 60%, pela característica intrínseca deste tipo de análise, a correlação é como se os itens fossem independentes, gerando possíveis conclusões que não fazem muito sentido. Inicialmente temos 16 regras, precedidas pelo valor de confiança:

- 95.91: ('HighAccuracyPass', 'MoreOnTargetShots', 'MoreShots') → ('Possession')
- 93.52: ('HighAccuracyPass', 'MoreShots', 'Possession') → ('MoreOnTargetShots')
- 90.58: ('MoreAerials', 'MoreShots', 'Possession') → ('MoreOnTargetShots')
- 90.45: ('MoreShots', 'Possession') → ('MoreOnTargetShots')
- 90.14: ('HighAccuracyPass', 'MoreOnTargetShots', 'Possession') → ('MoreShots')
- 89.84: ('MoreShots',) → ('MoreOnTargetShots')
- 89.65: ('LessFouls', 'MoreShots', 'Possession') → ('MoreOnTargetShots')
- 89.30: ('MoreAerials', 'MoreShots') → ('MoreOnTargetShots')
- 85.96: ('MoreAerials', 'MoreOnTargetShots', 'Possession') → ('MoreShots')
- 84.58: ('LessFouls', 'MoreOnTargetShots', 'Possession') → ('MoreShots')
- 84.25: ('MoreOnTargetShots', 'Possession') → ('MoreShots')
- 81.23: ('LessFouls', 'MoreOnTargetShots', 'MoreShots') → ('Possession')
- 74.58: ('MoreAerials', 'MoreOnTargetShots') → ('MoreShots')
- 72.18: ('MoreOnTargetShots',) → ('MoreShots')
- 69.18: ('MoreAerials', 'MoreOnTargetShots', 'MoreShots') → ('Possession')
- 67.21: ('MoreOnTargetShots', 'MoreShots') → ('Possession')

Por fim, também geramos os padrões de itens frequentes, totalizando 27 padrões após a retirada daqueles considerados aleatórios, precedidos pelo índice de ocorrência:

- 1445 - ('MoreOnTargetShots')
- 1161 - ('MoreShots')
- 1043 - ('MoreOnTargetShots', 'MoreShots')
- 832 - ('MoreOnTargetShots', 'Possession')
- 783 - ('MoreAerials', 'MoreOnTargetShots')
- 775 - ('MoreShots', 'Possession')
- 701 - ('MoreOnTargetShots', 'MoreShots', 'Possession')
- 654 - ('MoreAerials', 'MoreShots')
- 584 - ('MoreAerials', 'MoreOnTargetShots', 'MoreShots')
- 557 - ('LessOffTargetShots', 'MoreOnTargetShots')
- 470 - ('MoreAerials', 'MoreOnTargetShots', 'Possession')
- 446 - ('MoreAerials', 'MoreShots', 'Possession')
- 416 - ('HighAccuracyPass', 'MoreOnTargetShots', 'Possession')
- 405 - ('LessFouls', 'MoreOnTargetShots', 'MoreShots')
- 404 - ('MoreAerials', 'MoreOnTargetShots', 'MoreShots', 'Possession')
- 401 - ('HighAccuracyPass', 'MoreShots', 'Possession')
- 391 - ('HighAccuracyPass', 'MoreOnTargetShots', 'MoreShots')
- 389 - ('LessFouls', 'MoreOnTargetShots', 'Possession')
- 375 - ('HighAccuracyPass', 'MoreOnTargetShots', 'MoreShots', 'Possession')
- 367 - ('LessFouls', 'MoreShots', 'Possession')
- 329 - ('LessFouls', 'MoreOnTargetShots', 'MoreShots', 'Possession')
- 313 - ('LessCards', 'MoreOnTargetShots', 'MoreShots')
- 305 - ('LessFouls', 'MoreAerials', 'MoreOnTargetShots')
- 285 - ('LessOffTargetShots', 'MoreAerials', 'MoreOnTargetShots')
- 276 - ('LessCards', 'MoreOnTargetShots', 'Possession')
- 274 - ('LessFouls', 'MoreAerials', 'Possession')
- 272 - ('LessFouls', 'MoreAerials', 'MoreShots')

5. Conclusões e Perspectivas Futuras

A partir das análises feitas na Seção 4 podemos chegar a algumas conclusões, tanto do porquê estes padrões são mais frequentes em uma vitória como das regras geradas.

A partir dos padrões frequentes coletados, podemos perceber que as características mais efetivas são os chutes, chute a gol e posse de bola, já que configuram o top 7 de forma praticamente independente, tendo apenas uma ocorrência com uma característica que não uma delas, o padrão ('MoreAerials', 'MoreOnTargetShots').

Tal conclusão, além de calcada nos dados, bate com a expectativa e o senso, validando que são importantes características para a equipe alcançar os 3 pontos. Já que, partindo do ponto que a chance de um chute ser gol é sempre igual, quanto mais chutes mais gols e consequentemente se o Time A chuta mais que o B tem mais chance de ganhar a partida.

Podemos afirmar ainda que, principalmente entre estas características mais impactantes, observamos um crescimento no multiplicador de chance de vitória grande quando em conjunto, como mostrado na Figura 2:

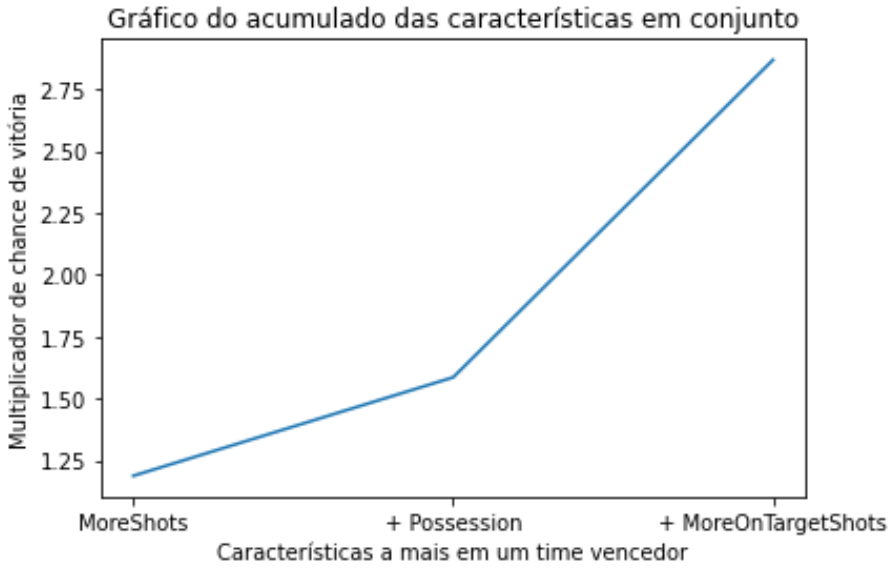

Figura 2. Multiplicador da chance de vitória no acumulado de cada característica

Além disso podemos perceber que algumas características, apesar de não terem uma relevância quando sozinhas, se mostram boas valências em um jogo quando em conjunto, como por exemplo chutar menos pra fora que seu adversário pode ser que apenas o time não chutou, porém quando em combinação com mais chutes em direção ao gol, mostra que a eficácia do chute que é maior, ou seja, o time vencedor perdeu menos chances de fazer gol que o adversário.

Alguns padrões ainda, aqueles com mais itens, normalmente significam uma qualidade destoante entre o time vencedor e o perdedor, de modo que o primeiro foi melhor em várias características do jogo, ou seja, uma dominância significativa na partida, como ('MoreAerials', 'MoreOnTargetShots', 'MoreShots', 'Possession').

Analisando também as regras, vemos que a mais frequente é levemente ilógica,

mostrando que os números sem análise não tem um real valor. já a implicação de mais chutes ao gol partindo de uma condição de alto índice de acertos nos passes, mais chutes e maior posse de bola. Esta regra combinada com o fato do padrão ('MoreOnTargetShots') ser o mais frequente nos indica que um time que treina passes e arrisca jogadas mais agressivas criando oportunidades de chute, tem um maior índice de acertos na direção ao gol e consequentemente tem maior chance de vencer a partida. Seguindo nesta linha vemos também que mesmo equipes com menor taxa de acerto de passe ainda assim conseguem chegar com mais perigo ao gol adversário, com uma confiança de 90.45%, apenas três pontos percentuais a menos que a análise anterior.

Podemos validar também uma das ideias mais difundidas no senso comum futebolístico, o famoso "Quem não arrisca não petisca", com base nas regras geradas vemos que o item comum presente com maior confiabilidade é a que quem chuta mais acaba acertando mais chutes na direção do gol, e como já visto, é a característica mais impactante para aumentar a chance de vitória de um time.

Por fim, como perspectivas futuras, temos a proposta de dar continuidade à pesquisa analisando a descoberta de padrões também entre os times perdedores, comparando os padrões e regras de associação entre os times perdedores e vencedores de maneira a validar os padrões exclusivos a cada um, fornecendo uma análise mais completa dos fatores relacionados à vitória. Além disso, tem-se também como perspectiva o uso de outras abordagens, como a descoberta de subgrupos, de maneira a observar características determinantes na proporção de vitórias e derrotas entre os times da La Liga.

Referências

Welcome to fp-growth's documentation! — fp-growth 1.0 documentation. https://fp-growth.readthedocs.io/en/latest/.

Augusto, L., Dutra, A., and Silva, C. https://github.com/LuccaAug/how-to-win.

Augusto, L., Dutra, A., and Silva, C. https://colab.research.google.com/drive/1Vj8llTKWjdJKklMdL4Aqk5MJ9xcHyABQ.

Augusto, L., Dutra, A., and Silva, C. https://colab.research.google.com/drive/1JL2iQhVThmHVXejnzKNSNsyuFa5cf5eI.

Decroos, T., Van Haaren, J., and Davis, J. (2018). Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 223–232, New York, NY, USA. Association for Computing Machinery.

Faria, G. (2021). O valor da laliga está próximo dos 5 mil milhões de euros. *Transfermarkt*.

Fifa (2024). Fifa world cup 26. https://www.fifa.com/en/tournaments/mens/worldcup/. Online; acesso em 20 de Julho de 2024.

Fox (2024). La liga teams. https://www.foxsports.com/soccer/la-liga/teams.

Han, J. and Pei, J. (2000). Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD explorations newsletter*, 2(2):14–20.

Klivansky, D. and El-Hajj, M. (2018). Causes of success in the la liga and how to predict them. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 1–8.

Naik, S. S. (2022). `https://www.kaggle.com/datasets/sanjeetsinghnaik/la-liga-match-data`.

Rein, R. and Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5:1–13.

StartSe, R. (2023). Como a ciência de dados fez do liverpool o melhor time de futebol do mundo. *StartSe Platform*.