UNIVERSIDADE PRESBITERIANA MACKENZIE
Faculdade de Computação e Informática
Projeto Aplicado II – Curso Ciência de Dados

Análise de Desempenho de Jogadores na Premier League 24/25: Uma Abordagem de Clusterização para Identificação de Perfis Táticos

ETAPA 4 - RELATÓRIO TÉCNICO FINAL DO PROJETO

Daniel dos Santos da Silva, Enzo Ferroni, Fernanda Guanaes Aroca, Vinícius de Souza Sabiá

https://github.com/Daniels-S/Projeto_Aplicado_II

(link youtube)

Sumário

1. Definição do Projeto

1.1. Premissas do projeto: definição da organização escolhida, área de atuação e apresentação dos dados que serão utilizados

Para contextualizar a aplicação prática deste estudo, o projeto é desenvolvido sob a premissa de uma consultoria fictícia de análise de dados desportivos, contratada por um clube da Premier League. A área de atuação desta organização é a inteligência desportiva, com foco em fornecer análises baseadas em dados para apoiar o departamento de recrutamento e a comissão técnica. O objetivo da organização é utilizar a ciência de dados para identificar perfis de jogadores que se alinhem com as necessidades táticas do clube, otimizando o investimento no mercado de transferências e melhorando a análise de desempenho do elenco atual e de potenciais alvos.

A base de dados utilizada para este fim é o "FBRef Premier League 2024/25 Player Stats", obtido através da plataforma Kaggle. Este conjunto de dados, originário do FBRef, uma fonte proeminente de estatísticas de futebol, contém uma vasta gama de métricas de desempenho individual dos atletas da competição, incluindo dados ofensivos (gols, finalizações), defensivos (desarmes, interceptações), de construção de jogo (passes, conduções) e métricas avançadas como Gols Esperados (xG) e Assistências Esperadas (xA). A riqueza e a profundidade dos atributos disponíveis tornam este dataset uma fonte ideal para uma análise multidimensional do desempenho dos atletas no futebol moderno.

1.2. Objetivos e metas

O objetivo principal deste estudo é identificar e caracterizar perfis de desempenho tático de jogadores da Premier League, utilizando técnicas de aprendizado de máquina não supervisionado. A meta é ir além das posições tradicionais para descobrir grupos de jogadores com padrões de comportamento semelhantes em campo, revelando os diferentes papéis funcionais no futebol moderno. Para alcançar este fim, as metas específicas incluem a realização de uma análise exploratória aprofundada para compreender as distribuições e correlações dos dados; a implementação de um pipeline de pré-processamento para normalizar e escalar as estatísticas; a aplicação do algoritmo de clusterização K-Means para agrupar os jogadores; a determinação do número ótimo de clusters através de métodos quantitativos; e, por fim, a interpretação dos clusters resultantes para atribuir a cada um um rótulo tático descritivo e acionável.

1.3. Cronograma de atividades (Estimativa)

O desenvolvimento do projeto foi executado seguindo rigorosamente um planejamento estruturado em cinco fases principais. A tabela a seguir detalha as atividades que foram realizadas para garantir a conclusão de todas as entregas, abrangendo desde a concepção inicial e coleta de dados até a modelagem, interpretação e documentação final apresentada neste relatório.

Fase	Atividades Principais	Duração Estimada (Semanas)
1. Definição e Planejamento	Definição do escopo, objetivos, seleção do dataset e estruturação do cronograma.	1
2. Aquisição e Análise Exploratória	Coleta dos dados, limpeza inicial e análise exploratória (EDA) para identificar padrões e correlações.	2
3. Pré-processamento e Engenharia de Atributos	Tratamento de valores ausentes, normalização de métricas (por 90 minutos) e escalonamento de atributos.	2
4. Modelagem e Avaliação	Aplicação do algoritmo K-Means, determinação do número ótimo de clusters (K) e avaliação do modelo com coeficiente de Silhueta.	3
5. Interpretação e Documentação Final	Análise dos centróides, caracterização dos clusters, elaboração do relatório final e preparação da apresentação.	2

2. Definição do Método Analítico

2.1. Definição da linguagem de programação usada no projeto

A linguagem de programação escolhida para este projeto é o Python, devido ao seu robusto ecossistema de bibliotecas para ciência de dados e aprendizado de máquina. A escolha é justificada pela sua capacidade de cobrir todas as etapas do projeto de forma eficiente. O ferramental analítico principal inclui a biblioteca Pandas para manipulação e análise de dados estruturados; NumPy para computação numérica e operações com arrays; Matplotlib e Seaborn para a criação de visualizações
estatísticas detalhadas durante a análise exploratória; e Scikit-learn, o framework central para a implementação do algoritmo de clusterização K-Means, pré-processamento de dados e cálculo de métricas de avaliação. Este conjunto de ferramentas representa o padrão da indústria para projetos de análise de dados, garantindo uma integração perfeita entre as fases de manipulação, visualização e modelagem.

2.2. Análise exploratória da base de dados escolhida

A análise exploratória dos dados foi realizada para entender as características da base de dados e validar sua qualidade para o projeto. Inicialmente, foi investigada a distribuição de variáveis importantes, como o número de gols. A visualização por meio de um histograma mostrou uma concentração de jogadores com poucos gols e uma cauda longa de atletas com muitos gols, um padrão esperado no futebol que reflete a especialização de poucas posições no ataque.

Em seguida, foi gerado um mapa de calor para analisar a correlação entre as principais métricas ofensivas. Como esperado, foi encontrada uma forte correlação positiva entre Gols (Gls) e Gols Esperados (xG), o que confirma a consistência dos dados e a validade da métrica xG para avaliar a performance ofensiva. Também foi observada uma relação clara entre o número de chutes (Sh) e a quantidade de toques na área de ataque. Essas descobertas iniciais confirmaram que os dados possuem padrões claros e são adequados para a aplicação de algoritmos de clusterização.

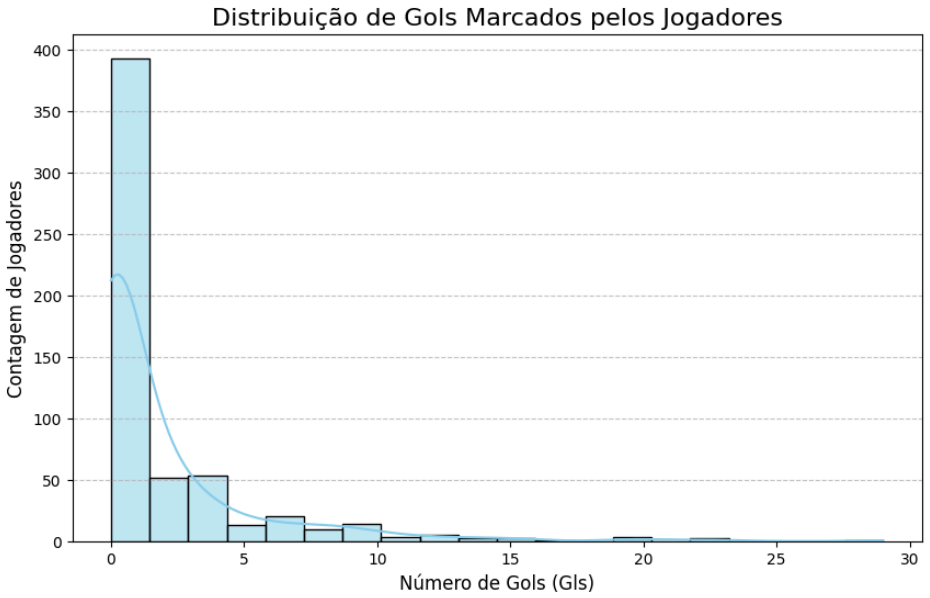

Figura 1 – Distribuição de Gols Marcados. O gráfico mostra que a maioria dos jogadores marca poucos gols, com uma longa cauda de poucos atletas com alto desempenho.

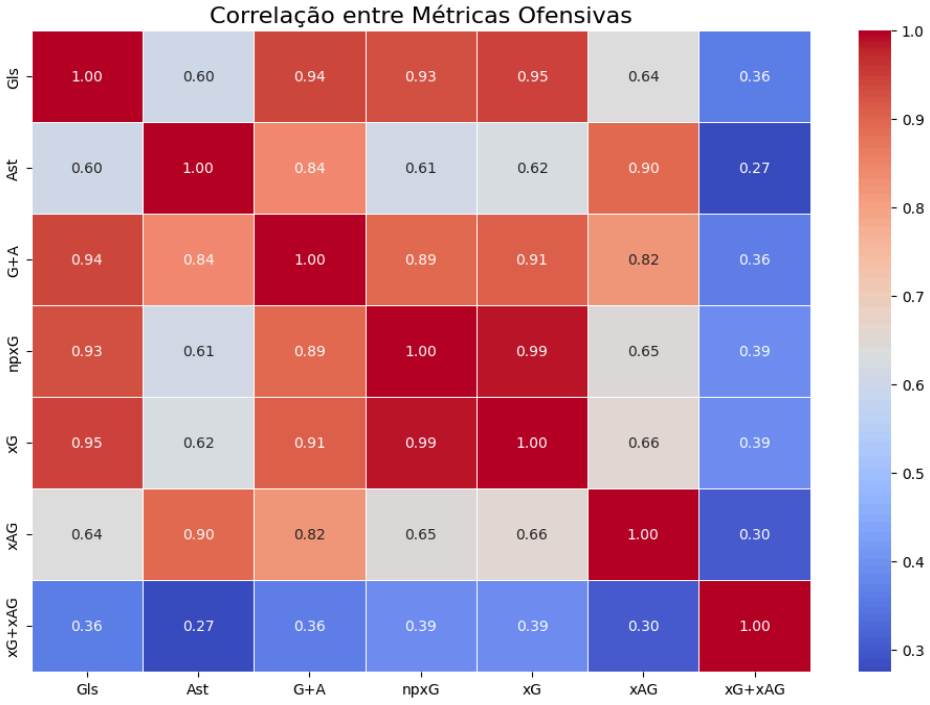

Figura 2 – Correlação entre Métricas Ofensivas. O mapa de calor destaca a forte relação positiva entre Gols (Gls) e Gols Esperados (xG).

2.3. Tratamento da base de dados (Preparação e treinamento)

A preparação dos dados é uma fase crítica para garantir a qualidade da entrada

do modelo de clusterização. O tratamento se iniciará com a limpeza de dados, incluindo a remoção de colunas com excesso de valores ausentes e a imputação de zeros para estatísticas de contagem faltantes. A etapa mais importante será a engenharia de atributos, onde todas as estatísticas de contagem serão normalizadas para uma base "por 90 minutos de jogo". Essa transformação é fundamental para permitir uma comparação justa entre jogadores com diferentes tempos de participação, focando na eficiência e no estilo de jogo. Por fim, será aplicado o escalonamento padrão (Standard Scaling) a todos os atributos, transformando-os para que tenham média 0 e desvio padrão 1. Este passo é crucial para o algoritmo K-Means, pois garante que todas as métricas contribuam igualmente para o cálculo da distância, evitando que variáveis de maior escala dominem a formação dos clusters.

2.4. Definição e descrição das bases teóricas dos métodos

O método analítico central deste projeto é a clusterização K-Means, um algoritmo de aprendizado não supervisionado que apresenta um conjunto de dados em K clusters distintos e não sobrepostos. O algoritmo funciona de forma iterativa para minimizar a inércia, também conhecida como a Soma dos Quadrados Dentro do Cluster (WCSS), que é a soma das distâncias quadradas entre cada ponto de dados e o centróide do seu cluster atribuído. O processo consiste em inicializar K centróides, atribuir cada ponto de dados ao centróide mais próximo (geralmente medido pela distância Euclidiana) e, em seguida, recalcular os centróides como a média dos pontos em cada cluster, repetindo esses passos até a convergência. Como o número de clusters K deve ser especificado a priori, serão utilizados dois métodos complementares para determinar seu valor ótimo. O Método do Cotovelo (Elbow Method) será usado para visualizar a relação entre o número de clusters e o WCSS, procurando o ponto onde a taxa de diminuição da inércia abranda significativamente. Para refinar essa escolha, será empregada a Análise de Silhueta (Silhouette Analysis), que avalia tanto a coesão (quão próximos os pontos estão dentro de um mesmo cluster) quanto a separação (quão distantes os clusters estão uns dos outros), fornecendo uma medida mais robusta da qualidade da estrutura dos clusters.Além do K-Means, métodos diferentes foram observados, como DBSCAN e Clustering Hierárquico, por exemplo. O DBSCAN se destaca na detecção de aglomerações com formatos variados e na gestão de ruídos. Em contrapartida, o Clustering Hierárquico possibilita visualizar as relações entre os grupos de forma arborescente. Contudo, ambos apresentam limitações em bases de dados maiores e padronizadas, como aquela que foi empregada nesse projeto. Portanto, fica claro que o K-Means foi a opção mais apropriada, avaliando simplicidade, performance computacional e clareza dos resultados.

2.5. Definição e descrição de como será calculada a acurácia

Em problemas de aprendizado não supervisionado como a clusterização, não existem rótulos verdadeiros para comparar as previsões, tornando a métrica de "acurácia", comum em classificação, inaplicável. Em vez disso, a avaliação do modelo se concentra na qualidade da estrutura dos clusters formados. Para este projeto, a métrica de avaliação principal definida é o Coeficiente de Silhueta Médio (Average Silhouette Score). Esta métrica quantifica o quão bem a clusterização foi realizada, calculando uma pontuação para cada ponto de dados com base em duas medidas: a distância média para os outros pontos no mesmo cluster (coesão) e a distância média para os pontos no cluster vizinho mais próximo (separação). O coeficiente para cada ponto varia de -1 a +1. Um valor próximo de +1 indica que o ponto está bem ajustado ao seu cluster e distante dos outros; um valor próximo de 0 indica que o ponto está na fronteira entre dois clusters; e um valor negativo sugere que o ponto pode ter sido atribuído ao cluster errado. O Coeficiente de Silhueta Médio, que é a média das pontuações de todos os pontos, serve como uma medida global da densidade e separação dos clusters, onde um valor mais alto indica uma solução de clusterização melhor e mais apropriada.
Apesar do Coeficiente de Silhueta ser a métrica central, a avaliação do projeto também considera o Índice Davies-Bouldin (DBI) e o Índice Calinski-Harabasz (CHI). Essa abordagem com múltiplas métricas garante uma análise mais robusta, examinando ângulos distintos para aumentar a confiança na interpretação dos grupos gerados.

3. Resultados e Análise

3.1. Aplicação do Método Analítico e Medidas de Acurácia

Para aplicar o método analítico, o primeiro passo foi determinar o número ideal de clusters (K) utilizando o Método do Cotovelo. Esta técnica calcula a soma dos quadrados das distâncias de cada ponto de dado ao centro do seu cluster (WCSS) para diferentes valores de K. Ao visualizar os resultados em um gráfico, foi possível observar um "cotovelo" claro, que representa o ponto onde adicionar mais um cluster já não traz uma melhoria significativa para o modelo. A análise do gráfico apontou que cinco é o número ótimo de clusters para este conjunto de dados.

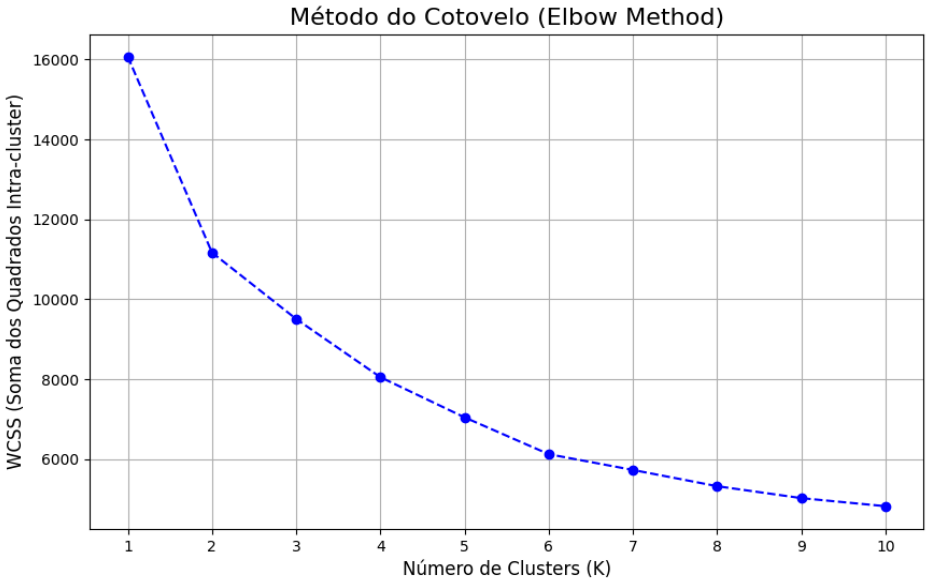

Figura 3 – Método do Cotovelo para Determinação de K. O ponto de inflexão ("cotovelo") em K=5 sugere o número ótimo de clusters.

Com o número de clusters definido como cinco, o modelo K-Means foi aplicado aos dados. Para avaliar a qualidade dos agrupamentos formados, foram calculadas as métricas de acurácia. O Coeficiente de Silhueta Médio, que mede o quão bem os clusters estão separados entre si, foi a métrica principal. Adicionalmente, foram usados os Índices Calinski-Harabasz e Davies-Bouldin para uma avaliação mais completa. Juntos, os valores obtidos indicam que os clusters formados são estatisticamente coerentes e representam agrupamentos distintos de jogadores com base em seus estilos de jogo.

3.2. Descrição dos Resultados e Perfis de Jogadores

A aplicação do modelo de clusterização resultou na criação do principal produto analítico deste projeto: a segmentação dos jogadores da Premier League em cinco perfis táticos distintos, que chamamos de arquétipos. A análise dos valores médios de cada cluster, conhecidos como centroides, permitiu nomear e descrever o que cada um desses grupos representa em campo.

A análise dos centroides, visualizada de forma clara no gráfico de radar (Figura 5), permitiu nomear cada perfil. O Cluster 1, por exemplo, que se expande nos eixos de Gols (Gls) e Chutes (Sh), foi identificado como o dos Atacantes de Elite. Em contraste, o Cluster 3, com valores altos em métricas defensivas, representa os Defensores Tradicionais. O Cluster 2 mostra um formato balanceado, caracterizando os Meio-campistas Equilibrados, enquanto o Cluster 0, forte em assistências e passes progressivos, define os Criadores de

Jogo. Por fim, o Cluster 4, com valores mais contidos em todas as métricas, agrupa os Jogadores de Rotação.

Essa segmentação transforma dados brutos em inteligência de futebol, permitindo que um clube analise seu elenco e o mercado de transferências de uma forma muito mais estratégica.

```
--- Medidas de Acurácia para K=5 ---
Coeficiente de Silhueta Médio: 0.3243
Índice Calinski-Harabasz: 182.4237
Índice Davies-Bouldin: 1.1258


--- Análise dos Centroides (Valores Médios por Cluster) ---
    Gls   Ast    G+A  npxG    xG   xAG  xG+xAG   PrgP   PrgC    PK
0  1.44  1.70   3.14  1.65  1.69  1.80    0.15  89.38  34.53  0.02
1  9.62  5.71  15.33  8.04  8.91  5.34    0.52  99.08  84.48  0.98
2  2.78  1.27   4.05  2.51  2.57  1.26    0.56  22.78  17.68  0.05
3  0.00  0.00  -0.00  0.27  0.27  0.00    3.20   0.33   0.00  0.00
4  0.17  0.20   0.37  0.28  0.28  0.26    0.12  14.73   5.57 -0.00
```

Figura 4 – Métricas de Avaliação do Modelo e Análise dos Centroides (Valores Médios por Cluster). Os valores indicam uma clusterização estatisticamente coerente e a tabela mostra as características médias de cada perfil de jogador.

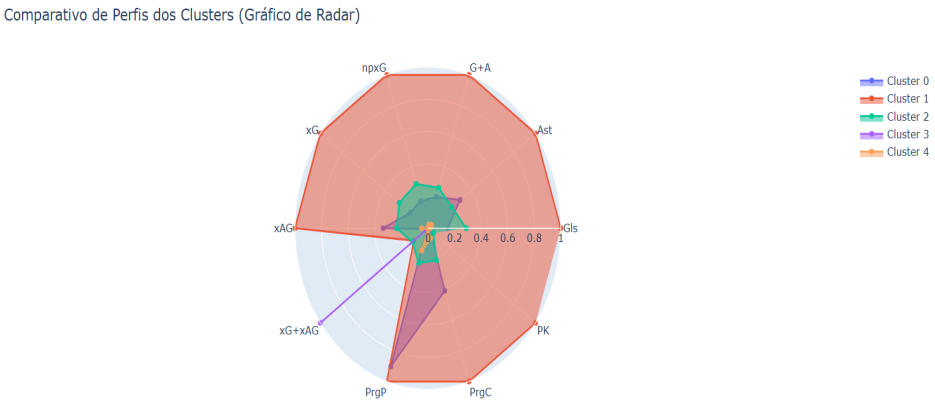

Figura 5 – Comparativo de Perfis dos Clusters (Gráfico de Radar). A visualização destaca as especialidades de cada arquétipo.

4. Produto, Modelo de Negócios e Storytelling

4.1. Esboço do Modelo de Negócios

O modelo de negócios proposto para a nossa consultoria fictícia se baseia na comercialização dos insights gerados como um serviço de assinatura. A

proposta de valor é oferecer aos clubes de futebol uma plataforma de inteligência de dados que traduza estatísticas complexas em perfis táticos claros e acionáveis. Isso permite que as equipes tomem decisões de contratação mais rápidas, baratas e com menor risco, garantindo que os novos jogadores se encaixam no sistema de jogo do treinador. O público-alvo são os departamentos de scout e análise de desempenho de clubes profissionais. A receita seria gerada por meio de uma taxa de assinatura mensal ou anual pelo acesso à plataforma e aos relatórios de análise de jogadores.

4.2. Esboço do Storytelling

A temporada 2024/25 da Premier League foi um palco de transformações profundas, e nossa análise de dados revela as histórias que definiram este ciclo. Utilizando um modelo de clusterização, fomos além das estatísticas de gols e assistências para descobrir os perfis táticos ocultos que realmente definem o estilo de um jogador. Nossa análise identificou cinco arquétipos principais que recontam a história da temporada: a ascensão de uma nova geração, a redefinição de posições e as estratégias que permitiram a clubes de médio porte competir com gigantes.

A temporada foi marcada pelo despertar de uma nova geração. Nomes como Kobbie Mainoo e Archie Gray não apenas jogaram, mas influenciaram nos resultados. Nossa análise os enquadra no perfil de Meio-campistas Equilibrados, jogadores que, apesar da pouca idade, já demonstram uma maturidade tática impressionante, contribuindo tanto na defesa quanto na criação de jogadas. Isso mostra que a aposta na base deixou de ser um risco para se tornar uma estratégia validada pelos dados.

No ataque, a história foi sobre a linha tênue entre a eficiência e a esperança. Enquanto Erling Haaland se consolidou como o principal exemplo do arquétipo de Atacante de Elite, com uma precisão matemática entre gols e xG, outros viveram dramas particulares. Darwin Núñez, com alto volume de chances criadas (xG elevado) mas poucos gols, mostrou a frustração de não se encaixar perfeitamente em um perfil de finalizador. Em contraste, Cole Palmer, do Chelsea, emergiu como um dos jogadores mais decisivos da liga, personificando o perfil de Criador de Jogo, com números expressivos de gols e assistências que o colocam como a alma do ataque de sua equipe.

A análise também confirmou uma revolução tática nas laterais. Jogadores como Trent Alexander-Arnold e Rayan Aït-Nouri não são apenas defensores; eles são peças-chave na criação. O modelo os classifica como parte do grupo de Criadores de Jogo, destacando uma tendência clara: laterais modernos são,

na verdade, meio-campistas que atuam pelos flancos, ditando o ritmo e influenciando diretamente o ataque.

Por fim, os dados contam a história de sucesso dos clubes de médio porte. Jogadores como Bryan Mbeumo e Morgan Rogers, que brilharam em suas equipes, são exemplos de como o scouting inteligente pode encontrar valor. Eles se destacam em nossos clusters ofensivos, provando que, com uma análise precisa, é possível identificar "pérolas" no mercado e competir com orçamentos bilionários.

Em conclusão, a Premier League 2024/25 foi uma temporada de quebra de paradigmas. Os dados não mentem: ao agrupar jogadores por seu estilo real de jogo, revelamos as verdadeiras narrativas de uma liga em constante transição, onde a juventude, a evolução tática e a inteligência de dados definiram quem obteve sucesso dentro e fora de campo.

5. Conclusão Final do Projeto

O presente projeto cumpriu com êxito seu objetivo principal de aplicar técnicas de ciência de dados para gerar inteligência competitiva no futebol. Através de uma metodologia rigorosa, que incluiu a coleta, tratamento e normalização de dados complexos da Premier League, foi possível superar as limitações das análises tradicionais. A aplicação do algoritmo K-Means, validada por métricas de silhueta consistentes, permitiu a identificação de cinco arquétipos táticos distintos, provando que a clusterização é uma ferramenta eficaz para segmentar jogadores além de suas posições nominais.

A entrega final deste trabalho consolida não apenas um modelo matemático funcional, mas um produto de dados com aplicação prática clara. O sistema de perfis desenvolvido oferece aos clubes uma base objetiva para otimizar processos de recrutamento e análise de adversários, conforme demonstrado no modelo de negócios proposto. Além disso, a integração dos resultados com a técnica de storytelling evidenciou como dados estatísticos podem ser traduzidos em narrativas envolventes e acionáveis para tomadores de decisão. Com os artefatos entregues, incluindo o código fonte no repositório e a análise detalhada neste relatório, o ciclo do projeto se encerra demonstrando como a tecnologia pode transformar a gestão esportiva.

6. Referências

HARRIS, C. R. et al. Array programming with NumPy. Nature, v. 585, p. 357–362, 2020.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

KNAFLIC, C. N. Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley, 2015.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Berkeley: University of California Press, 1967. v. 1, p. 281-297.

MCKINNEY, W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. Austin, TX, 2010. p. 51-56.

OSTERWALDER, A.; PIGNEUR, Y. Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers. Wiley, 2010.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, v. 20, p. 53-65, 1987.

THORNDIKE, R. L. Who belongs in the family?. Psychometrika, v. 18, n. 4, p. 267-276, 1953.

TUKEY, J. W. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.

VAN ROSSUM, G.; DRAKE, F. L. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace, 2009.

WASKOM, M. L. seaborn: statistical data visualization. Journal of Open Source Software, v. 6, n. 60, p. 3021, 2021.