UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

IGOR RODRIGUES LIMA

APLICAÇÃO DE MODELO DE APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO NO JOGO FIFA 2022

Revisado pelo Orientador

Assinatura do Orientador
Data ____/ ____ / 2024

PRESIDENTE PRUDENTE
2024

IGOR RODRIGUES LIMA

APLICAÇÃO DE MODELO DE APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO NO JOGO FIFA

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientador: Prof. Dr. Klaus Schlunzen Junior.

PRESIDENTE PRUDENTE
2024

L732a

Lima, Igor Rodrigues
Aplicação de modelo de aprendizado de máquina não supervisionado no jogo FIFA 2022 / Igor Rodrigues Lima. -- Presidente Prudente, 2024
68 p. : tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Klaus Schlunzen Junior

1. Estatística. 2. Análise estatística multivariada. 3. Aprendizado de máquinas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

Igor Rodrigues Lima

APLICAÇÃO DE MODELO DE APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO NO JOGO FIFA

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Drº. Klaus Schlunzen Junior.
Departamento de Estatística (Dest)

_____

Prof. Drº. Sérgio Minoru Oikawa.
Departamento de Estatística (Dest)

Presidente Prudente, 7 de Dezembro de 2024.

RESUMO

Todo entusiasta de jogos eletrônicos busca alcançar o melhor desempenho possível em sua jogabilidade, e no contexto do FIFA, não poderia ser diferente. A possibilidade de gerenciar uma equipe, realizar contratações e traçar estratégias táticas torna a experiência ainda mais envolvente e imersiva, com um único objetivo em mente: alcançar a glória. No entanto, esse objetivo exige mais do que habilidade individual; é necessário planejamento estratégico, especialmente na identificação e seleção de atletas que se alinhem ao orçamento e às necessidades da equipe. Por meio da aplicação de aprendizado de máquina, utilizando o algoritmo K-Médias, atletas foram agrupados em clusters distintos com base em suas características, como idade, valor de mercado e potencial de desempenho. Esses grupos permitem que os jogadores avaliem e escolham atletas de maneira mais eficiente, considerando fatores como custo-benefício e estratégia de longo prazo. A análise revelou diversas oportunidades valiosas de contratação. Entre elas, destacam-se atletas jovens com excelente potencial futuro e preços acessíveis, que representam ótimos investimentos para equipes que buscam equilíbrio entre desempenho e orçamento. Ao explorar as possibilidades que os dados oferecem, o trabalho busca aprimorar a experiência dos jogadores no FIFA, transformando informações complexas em conhecimentos práticos e acessíveis. Para os entusiastas que desejam elevar sua jogabilidade e tomar decisões estratégicas fundamentadas, este projeto oferece uma ótima ferramenta para alcançar o sucesso no universo dos jogos eletrônicos.

Palavras-chave: Futebol; FIFA; K-Médias; Clusterização; Aprendizado de Máquina não supervisionado; Método do Cotovelo; Método da Silhueta; Contratação de atletas no FIFA.

ABSTRACT

Every electronic game enthusiast seeks to achieve the best possible performance in their gameplay, and in the context of FIFA, it is no different. The ability to manage a team, make signings, and devise tactical strategies makes the experience even more engaging and immersive, with a single objective in mind: to achieve glory. However, this goal requires more than individual skill; it demands strategic planning, especially in identifying and selecting athletes that align with the budget and needs of the team. By applying machine learning using the K-Means algorithm, athletes were grouped into distinct clusters based on characteristics such as age, market value, and performance potential. These groups allow players to evaluate and choose athletes more efficiently, considering factors such as cost-effectiveness and long-term strategy. The analysis revealed several valuable signing opportunities. Among them, standout players include young athletes with excellent future potential and affordable prices, representing great investments for teams seeking a balance between performance and budget. By exploring the possibilities that data offers, this work aims to enhance the FIFA player experience by transforming complex information into practical and accessible knowledge. For enthusiasts who wish to elevate their gameplay and make informed strategic decisions, this project provides a great tool for achieving success in the world of electronic games.

Keywords: Football; FIFA; K-Means; Clustering; Unsupervised Machine Learning; Elbow Method; Silhouette Method; Athlete Recruitment in FIFA.

LISTA DE FIGURAS

Figura 1 - Exemplo modelo de classificação 12
Figura 2 - Exemplo de modelo de regressão 13
Figura 3 - Dados originais x Clusterização. 14
Figura 4 - Distância euclidiana 16
Figura 5 - Representação de métodos hierárquicos através de dendograma. 17
Figura 6 - Passos K-Médias 21
Figura 7 - Exemplo Método Cotovelo 24
Figura 8 - Método cotovelo representação algébrica 25
Figura 9 - Exemplo coeficiente de silhueta 26
Figura 10 - Representação visual silhueta 27
Figura 11 - Estatística das variáveis 29
Figura 12 - Gráfico das variáveis 31
Figura 13 - Média do valor dos atletas por posição 34
Figura 14 - Média do valor de goleiros 35
Figura 15 – Pontuação geral médio dos atletas por país. 36
Figura 16 - Pontuação Geral para países com mais de 30 atletas 36
Figura 17 - Correlação Pontuação Geral vs Valor. 37
Figura 18 - Gráfico de dispersão destacando Kylian Mbappé. 38
Figura 19 - Gráfico de dispersão destacando Lionel Messi. 38
Figura 20 - Gráfico de dispersão destacando atletas bons e “baratos”. 39
Figura 21 - Jogadores selecionados manualmente 40
Figura 22 – Método do cotovelo. 43
Figura 23 – Coeficiente de Silhueta. 46
Figura 24 – Gráfico dos clusters. 47
Figura 25 - Cluster Zero: Gráfico. 48
Figura 26 - Cluster Um: Gráfico. 51
Figura 27 - Cluster Dois: Gráfico. 53
Figura 28 - Cluster Três: Gráfico. 56
Figura 29 - Cluster Quatro: Gráfico 59

SUMÁRIO

1 Introdução 7
1.1 Objetivo geral e específicos 8
1.2 Justificativa 9
2 Referencial Teórico 10
3 Aprendizado de máquina 11
3.1 Aprendizado de máquina supervisionado 11
3.1.1 Modelos de classificação 11
3.1.2 Modelos de regressão 12
3.2 Aprendizado de máquina não supervisionado 13
4 Clusterização 14
4.1 Medidas de distância 15
4.1.1 Distância euclidiana 16
4.2 Tipos de cluster 17
4.2.1 Métodos hierárquicos 17
4.2.2 Métodos não-hierárquicos 19
5 K-Médias 19
5.1 Etapas da aplicação 20
5.2 Vantagens e desvantagens 22
5.3 Determinação do número de clusters 23
5.3.1 Método do Cotovelo 23
5.3.2 Método da Silhueta 26
6 Análise exploratória dos dados 28
6.1 Conhecimento das variáveis 28
6.2 Estatística das variáveis 29
6.2.1 Informações sobre as variáveis 29
6.2.2 Tratamento dos dados 31
6.2.3 Análises extras 34
7 Aplicação do modelo 41
7.1 Apresentação dos dados 41
7.2 Tratamento das variáveis 42
7.3 Escolha da quantidade de cluster 43
8 Resultados 48
8.1 Cluster zero 48
8.2 Cluster Um 51
8.3 Cluster Dois 53
8.4 Cluster Tres 56
8.5 Cluster Quatro 59
9 Conclusão 62
Referências 63

1 Introdução

Nos últimos anos, o cenário dos jogos de simulação esportiva passou por mudanças significativas, e a experiência está cada vez mais próxima da realidade, representada pelo famoso jogo de futebol Federação Internacional de Futebol Associado (FIFA). Estes jogos atraem não apenas fãs de futebol, mas também uma comunidade global de jogadores para competir no mundo virtual do futebol.

Desenvolvido em colaboração com a EA Sports, o FIFA 22 é um jogo eletrônico que simula partidas de futebol. No entanto, a parceria entre a EA Sports e a FIFA teve um ponto final após o FIFA 23, resultando no lançamento do novo jogo da EA Sports denominado EA FC 24.

Dentro do FIFA 22, os jogadores (pessoas que jogam FIFA) têm a liberdade de escolher qualquer time das principais divisões e ligas do mundo. Eles têm a capacidade de definir estratégias, como escalar atletas, atribuir funções específicas a eles e escolher o estilo de jogo da equipe, entre outras opções. Todas essas escolhas têm um impacto significativo no desempenho do jogador durante as partidas.

Além disso, os jogadores têm a opção de escolher o modo de jogo que preferem. Eles podem optar por controlar apenas um jogador em campo ou assumir o controle de toda a equipe, tomando decisões táticas e estratégicas em nível mais amplo. Isso torna o FIFA um jogo versátil, onde os jogadores podem experimentar diferentes estratégias e desafios, dependendo de suas preferências pessoais.

Um dos modos existentes no jogo, é o Modo Carreira Treinador, onde o jogador assume a posição de treinador em determinada equipe. Neste modo, é possível realizar transferências de jogadores, elaborar estratégias de jogo e controlar o time em partidas. Para muitos jogadores, o objetivo deste modo é ter sucesso como comandante, conquistando títulos e alcançando as metas estipuladas pelo clube.

O sucesso da equipe depende diretamente das decisões tomadas e da habilidade do jogador. Portanto, considerando como as negociações de atletas afetam diretamente o desempenho da equipe, a ideia de agrupamento de jogadores de acordo com suas características para auxiliar a tomada de decisão sobre qual jogador contratar levando em consideração seu custo e potencial.

O modelo pode ser utilizado para classificar atletas (personagens dentro do jogo) em vários grupos, ajudando o usuário a identificar bons atletas para a estratégia

do time. O modelo selecionado foi o K-Médias, um algoritmo de aprendizado de máquina não supervisionado que visa agrupar os dados em um número específico de clusters (grupos). Este modelo usa a distância euclidiana entre os pontos para determinar a classificação em cada grupo.

Sendo o jogo uma simulação real do universo do futebol, as técnicas utilizadas podem ser úteis no contexto de gerenciamento de equipes reais. Pois a composição de bons atletas em um elenco é importante para qualquer time que almeja o topo. Em resumo, o objetivo deste projeto é demonstrar como a análise de dados e a aplicação de agrupamento podem proporcionar uma ótima ferramenta para análises de dados auxiliando na extração de conhecimento dos dados. As potenciais contribuições não se restringem apenas aos jogadores e à comunidade do jogo, mas se estendem ao mundo real. Isso ocorre porque, à medida que o jogo se torna cada vez mais realista, os conhecimentos obtidos podem contribuir para o avanço das instituições de futebol.

1.1 Objetivo geral e específicos

O objetivo do projeto foi analisar o uso de modelo estatístico com foco em agrupamento por meio do algoritmo K-médias para agrupar atletas de acordo com as suas características visando melhorar o desempenho esportivo. Além disso, realizou-se uma análise exploratória dos dados para compreender, resumir e visualizar os principais aspectos do conjunto de dados. Desta forma, o propósito deste foi aplicar uma análise descritiva e aplicação do algoritmo para agrupar jogadores de acordo com suas características.

Os objetivos específicos a serem alcançados são:

- Aplicar técnicas de limpeza e processamento de dados no contexto de jogos eletrônicos.
- Realizar técnicas de análise exploratória de dados a fim de compreender a base de dados.
- Desenvolver habilidades em Python e R Studio.
- Adquirir conhecimentos em modelagem estatística utilizando K-Médias para agrupar atletas do jogo FIFA 22.
- Analisar o modelo.

1.2 Justificativa

Sabendo que os esportes eletrônicos tiveram um crescimento considerável nos últimos anos, contando com diversas equipes profissionais e competições intercontinentais. Seja o jogador profissional ou o jogador casual, ambos buscam aperfeiçoar seu desempenho o máximo possível, a fim de conquistar melhores resultados. Além da habilidade do jogador, a escolha de atletas para compor o seu time impactam diretamente seu desempenho. Neste caso, a aplicação de modelos de agrupamento oferece uma oportunidade para melhorar a compreensão e o desempenho no jogo.

Pensando em como as decisões sobre qual jogador utilizar impactam significativamente no desenvolvimento e desempenho da equipe, serão realizadas análises e agrupamento de jogadores que podem oferecer melhores atuações de acordo com a condição financeira do clube. Além disso, o projeto tem como objetivo incentivar o uso de técnicas estatísticas em contextos não explorados, e também promover o uso dos jogos eletrônicos como forma de entretenimento. Uma vez que, um estudo realizado pelo site Casino.org em parceria com a Universidade de Leeds, no Reino Unido, diz que, apesar do jogo ser um dos que mais provocam raiva em seus jogadores, este pode reduzir o nível de estresse e ansiedade.

Neste contexto, o projeto busca atingir o terceiro objetivo de desenvolvimento sustentável da Organização Das Nações Unidas (ONU) que busca promover a saúde mental e o bem-estar, também prevenir o suicídio. Visto que a ansiedade é uma das principais situações de vulnerabilidade para o suicídio, segundo o Manual de Bolso Prevenção do Suicídio e Promoção da vida.

Portanto, o projeto visa auxiliar a contratação de jogadores, enquanto destaca a importância crescente da análise de dados em um dos setores de entretenimento mais dinâmicos da atualidade. As lições aprendidas e as habilidades desenvolvidas podem ser aplicadas em diversos contexto do mundo real, sendo o modelo de agrupamento um dos mais utilizados para segmentação de clientes e recomendação de conteúdo.

2 Referencial Teórico

Atualmente, vivemos na Era do Conhecimento, onde uma vasta quantidade de dados são coletados diariamente e precisam ser analisados e estudados. Com isso, técnicas mais aprofundadas foram criadas para tirar conhecimentos destes dados, como a Mineração de Dados (Han, Kamber, Pei, 2012). A Mineração de dados é utilizada para a tomada de decisão, extraindo informações de um conjunto de dados sem um conhecimento prévio (Diniz, Louzada Neto, 2000), utilizando técnicas e ferramentas computacionais como classificação, regressão, análise de cluster, etc. (Doni, 2004)

Uma das técnicas mais utilizadas no processo de mineração de dados é a análise de agrupamento que consiste em criar grupos dentro do conjunto de dados e descobrir possíveis padrões. A análise de agrupamento consiste em uma técnica de aprendizado de máquina não supervisionado, onde o algoritmo “aprende” a criar os grupos sozinho, sem a necessidade de um supervisor. (Haldiki, Batistakis, Vazirgiannis, 2001). Estes grupos são divididos de forma que dentro dos grupos sejam o mais parecido possível e que haja uma distinção entre eles.

Alguns estudos já exploraram a aplicação de técnicas estatísticas no jogo FIFA. Entre eles, destaca-se o trabalho realizado por Gerolamo, que utilizou técnicas de clusterização para formar uma equipe competitiva com o menor orçamento possível, comparando-a à equipe vencedora do Barcelona. Por meio do agrupamento de jogadores, analisou e otimizou a composição do elenco. Outro exemplo relevante é o estudo de Lima, que empregou técnicas de regressão para prever o salário de atletas com base em suas características individuais.

Este projeto busca se diferenciar dos trabalhos anteriores ao utilizar técnicas multivariadas, com enfoque no agrupamento de atletas, a fim de auxiliar os jogadores na tomada de decisão durante o processo de contratação dentro do jogo.

3 Aprendizado de máquina

O aprendizado de máquina (em inglês, machine learning) é uma área específica dentro da inteligência artificial (IA) que usa modelos estatísticos treinados em conjuntos de dados para criar sistemas que aprendem sem necessariamente serem programados. Trata-se de uma técnica que aperfeiçoa a performance aprendendo por meio de métodos de computação baseados na experiência (ZHOU; LIU, 2016). Dessa forma, é capaz de prever resultados com base nos dados

Atualmente, o aprendizado de máquina está presente em muitos aspectos. Seja na concessão de crédito em bancos, na detecção de fraudes em sistemas de pagamento ou recomendação de conteúdo em redes sociais, os algoritmos de aprendizado de máquina atuam para tornar essas interações mais eficientes, fluidas e seguras.

Os principais algoritmos de aprendizado de máquina são os supervisionados e não supervisionados.

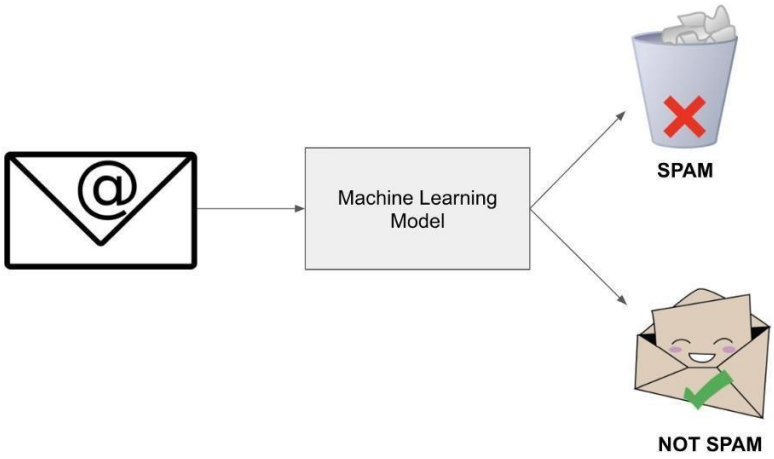
3.1 Aprendizado de máquina supervisionado

O aprendizado de máquina supervisionado é uma das abordagens mais populares e amplamente utilizadas no campo da ciência de dados. Trata-se de uma técnica em que um algoritmo é treinado a partir de um conjunto de dados rotulados para fazer previsões. Essa metodologia se baseia na ideia de que o modelo pode identificar padrões e relações nos dados que podem ser generalizados para novos exemplos. Existem dois tipos principais de modelos supervisionados: classificação e regressão.

3.1.1 Modelos de classificação

Os modelos de classificação geralmente são utilizados para prever resultados categóricos. Isso pode envolver a distinção entre duas classes (classificação binária) ou entre mais de duas classes (classificação multiclasse). É possível pensar na classificação binária como resposta alvo sendo 0 ou 1, por exemplo, aplicando o modelo com o objetivo de descobrir se um email é spam (1) ou não (0) assim como mostra a Figura 1.

Figura 1 - Exemplo modelo de classificação



Fonte: Dharani, Mandla. Disponível em: https://www.naukri.com/code360/library/binary-classification. Acesso em 28/05/2024.

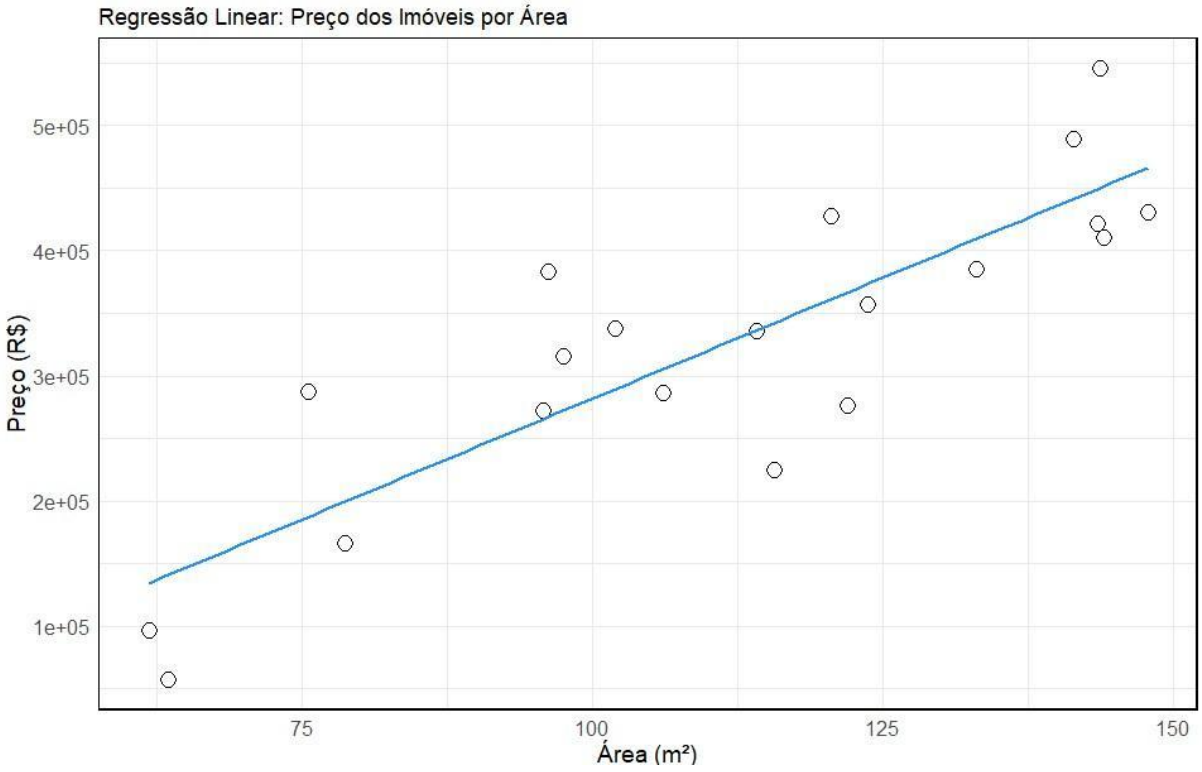
Entretanto, o modelo de classificação é capaz de realizar previsões envolvendo mais de duas classes, por exemplo, na previsão se um cliente será classificado como: Ouro, Prata ou Bronze.

3.1.2 Modelos de regressão

Os modelos de regressão, por outro lado, têm como objetivo principal retornar valores numéricos. Eles modelam a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras). A regressão busca prever ou explicar o comportamento da variável dependente com base nos valores das variáveis independentes.

Um exemplo clássico de regressão é prever o valor de venda de casas de acordo com a sua área. Nesse caso, como mostra a Figura 2, o modelo utiliza a área da casa como variável independente para estimar o preço de venda, a variável dependente.

Figura 2 - Exemplo de modelo de regressão



Fonte: Elaborado pelo autor (2024).

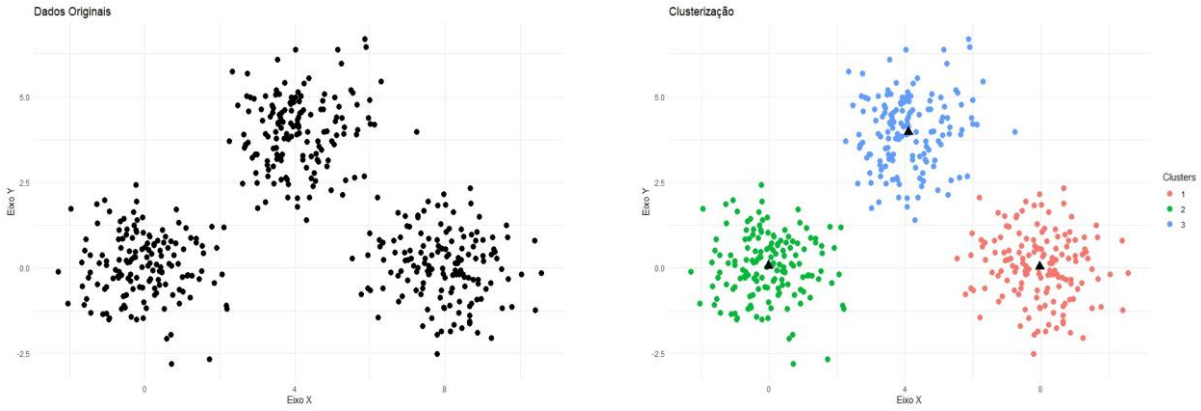
3.2 Aprendizado de máquina não supervisionado

Ao contrário do aprendizado supervisionado, que utiliza dados rotulados para treinar algoritmos, o aprendizado não supervisionado trabalha com dados que não possuem rótulos ou respostas predefinidas. O objetivo principal dessa abordagem é descobrir padrões ocultos, estruturas subjacentes ou relações nos dados sem a necessidade de orientação explícita. Entre as diversas técnicas disponíveis no aprendizado não supervisionado, a clusterização se destaca como uma das mais importantes e será a técnica principal utilizada neste projeto.

4 Clusterização

A clusterização é um tipo de classificação, trata-se de uma técnica de classificação não supervisionada que busca dividir a base de dados em grupos chamados clusters. O objetivo principal é agrupar os dados de modo que os elementos dentro de um cluster sejam o mais semelhantes possível, enquanto os elementos de clusters diferentes sejam distintos. A Os algoritmos de clusterização atribuem cada ponto de dados a um cluster específico, de maneira semelhante aos algoritmos de classificação como mostra a Figura 3.

Figura 3 - Dados originais x Clusterização.



Fonte: Elaborado pelo autor (2024).

A clusterização pode ser utilizada em diversos contextos. Algumas das situações:

a) Redução de dados: Em contextos envolvendo grande volume de dados onde processamento é muito alto, a redução de dados pode ser uma alternativa. A técnica consiste em diminuir o volume de dados mantendo o máximo de informações relevantes possível. A técnica divide os dados em clusters que representam grupos de dados semelhantes, tornando mais fácil o processamento e a análise. Ao invés de processar todos os dados, a técnica adota apenas os representantes de cada cluster. (Haldiki, Batistakis, Vazirgiannis, 2001).
b) Segmentação de Clientes: Empresas utilizam a clusterização para segmentar seus clientes em grupos com base em padrões de compra, preferências de produto, comportamento de navegação online, entre outros. Permitindo à

empresa personalizar campanhas de marketing, recomendações de produtos e estratégias de atendimento ao cliente para cada segmento.

c) Diagnóstico Médico: Na área da saúde, geralmente a clusterização é empregada para identificar subgrupos de pacientes com base em características clínicas, biomarcadores ou histórico médico. Isso pode ajudar no diagnóstico precoce de doenças, personalização de tratamentos e identificação de fatores de risco.

4.1 Medidas de distância

As medidas de distância são cruciais na clusterização. Pois, fornecem uma forma quantitativa de avaliar quão semelhantes ou diferentes são dois objetos, com base na proximidade ou distância entre objetos em um conjunto de dados. Existem várias medidas de similaridade que são utilizadas, cada uma com suas próprias características e áreas de aplicação. Cada uma dessas medidas oferece uma perspectiva única sobre como os dados podem ser comparados e agrupados, influenciando diretamente os resultados da clusterização.

A definição de distância segundo Hunter (2014) é definida da seguinte forma:

Definição: uma distância d em um conjunto X é uma função

$$d: X \times X \to R$$

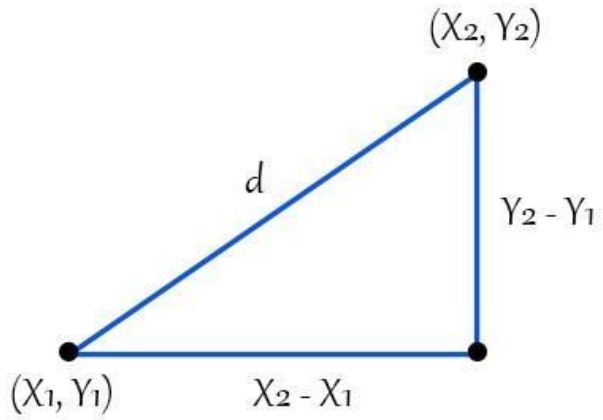
tal que para todo x, y, z função deve satisfazer às seguintes condições para todo x, y, $z \in X$:

1) $d(x, y) \geq 0$;
2) $d(x, y) = 0 \Leftrightarrow x = y$;
3) $d(x, y) \leq d(x, z) + d(z, y)$+ $d(x, y) = d(y, x)$;

4.1.1 Distância euclidiana

A distância euclidiana é uma das medidas de similaridade mais comumente utilizadas. Ela é especialmente aplicável quando lidamos com dados numéricos em um espaço euclidiano. Essa medida calcula a distância entre dois pontos no espaço, utilizando a geometria euclidiana clássica. A Figura 4 mostra como é calculada a distância d entre dois pontos.

Figura 4 - Distância euclidiana



Fonte: Mai, Mike.Disponível: https://maideveloper.com/blog/how-to-calculate-euclidean - distance-how-to-measure-the-distance-beteen-two-points. Acesso em: 27/05/2024.

Sendo $X = [X_1, X_2, \ldots, X_n]$ e $Y = [Y_1, Y_2, \ldots, Y_n]$. A distância euclidiana entre estes dois conjuntos é definida da seguinte forma:

$$d(x, y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \ldots + (X_n - Y_n)^2}$$

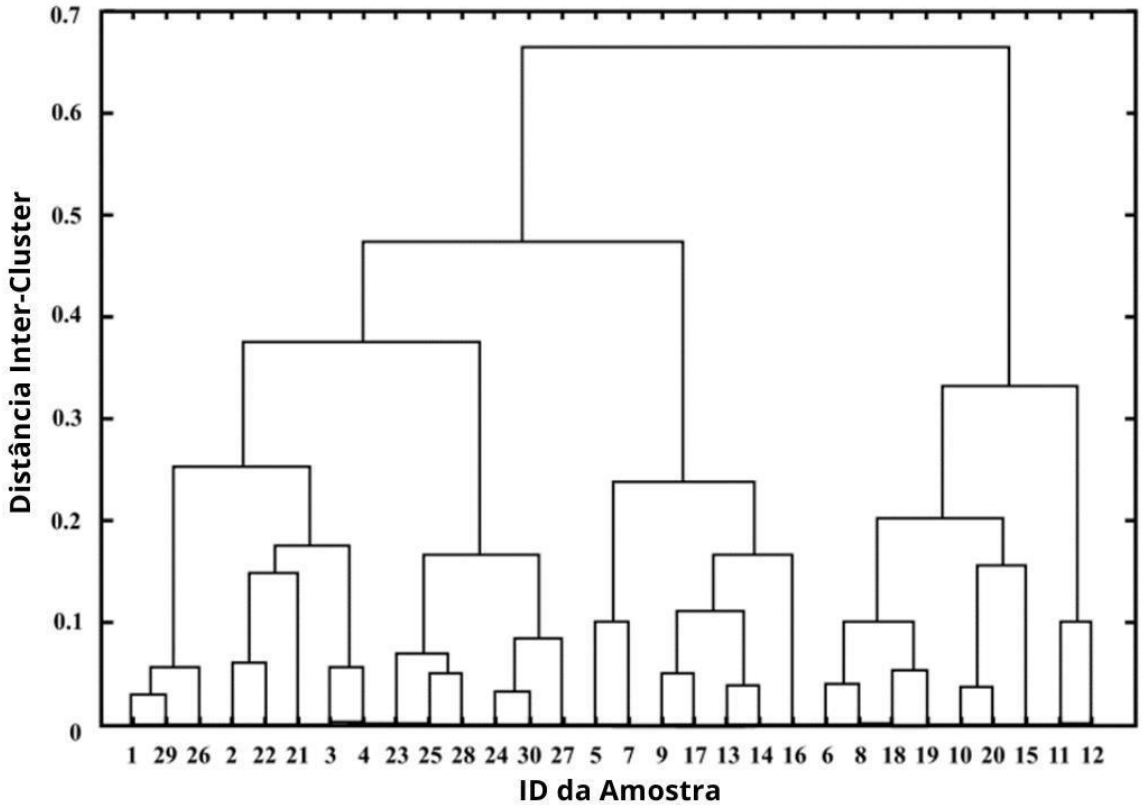
4.2 Tipos de cluster

Neste momento, os tipos de clusterização serão apresentados. Existem dois tipos principais: métodos hierárquicos e métodos não hierárquicos ou por particionamento.

4.2.1 Métodos hierárquicos

Os métodos hierárquicos são técnicas de clusterização que organizam os dados em uma estrutura hierárquica, na qual os clusters são agrupados em subgrupos sucessivos. Essa abordagem é intuitiva e permite uma visualização fácil da estrutura dos dados, frequentemente representada por meio de dendrogramas (Doni, 2004), como ilustra a Figura 5 .

Figura 5 - Representação de métodos hierárquicos através de dendograma.



Fonte: Zhou, Liu (2016, p. 231).

Segundo Halkidi, Batistakis e Vazirgiannis, os métodos hierárquicos podem ser classificados como aglomerativos ou divisivos. Os métodos hierárquicos aglomerativos têm como objetivo formar clusters com a menor distância possível, cada

ponto de dados é considerado um cluster e ao decorrer das etapas todos os elementos se agrupam sucessivamente até formar apenas um agrupamento. Por outro lado, os métodos divisivos começam com um único cluster e, em cada etapa, com base na similaridade, se dividem em clusters menores até que reste apenas um elemento por cluster.

4.2.2 Métodos não-hierárquicos

Os métodos de clusterização não hierárquicos, também conhecidos como métodos de particionamento, são técnicas de agrupamento de dados onde os clusters são formados de maneira independente, sem a formação de uma estrutura hierárquica de clusters. Este método atribui os elementos a um determinado grupo desde que um valor K de clusters tenha sido definido inicialmente.

A maior parte desses métodos atribui inicialmente uma partição e posteriormente, realiza mudanças desses elementos com a finalidade de obter a melhor partição.

Seja um conjunto de dados, denominado D, com n elementos e K o número de clusters definidos para serem formados. Sendo (k<=n), os algoritmos organizam esses elementos em k clusters, onde cada cluster representa uma partição. O objetivo é que haja semelhança entre os elementos de cada cluster e diferença entre os clusters (Han, Kamber, Pei, 2012).

Contudo, os principais algoritmos são o K-Médias e o K-Medóides. Para a realização da análise de agrupamento deste trabalho o algoritmo escolhido foi o K-Médias.

5 K-Médias

Como mencionado anteriormente, o K-Médias é uma técnica popular de agrupamento. Esta técnica consiste em agrupar dados não rotulados em grupos distintos, chamados de clusters, baseados em características similares. O objetivo principal é encontrar centros de clusters (centróides) que representem áreas distintas dos dados (Muller, Guido, 2016), onde cada ponto de dados pertence ao cluster com a distância mais próxima. Desta forma, busca minimizar a variância dentro de cada cluster e maximizar a variância entre os clusters. A qualidade dos clusters é medida pela variação dentro do cluster, que consiste na soma do erro quadrático entre todos os elementos e o centróide, definido por:

$$E = \sum_{i=1}^{k} \quad \sum_{p \in Ci} \quad dist(p, ci)^2$$

onde E é a soma do erro quadrático de todos os objetos no conjunto de dados; p é o ponto em espaço representando um determinado objeto; e ci é o centróide do cluster Ci (Han, Kamber, Pei, 2012).

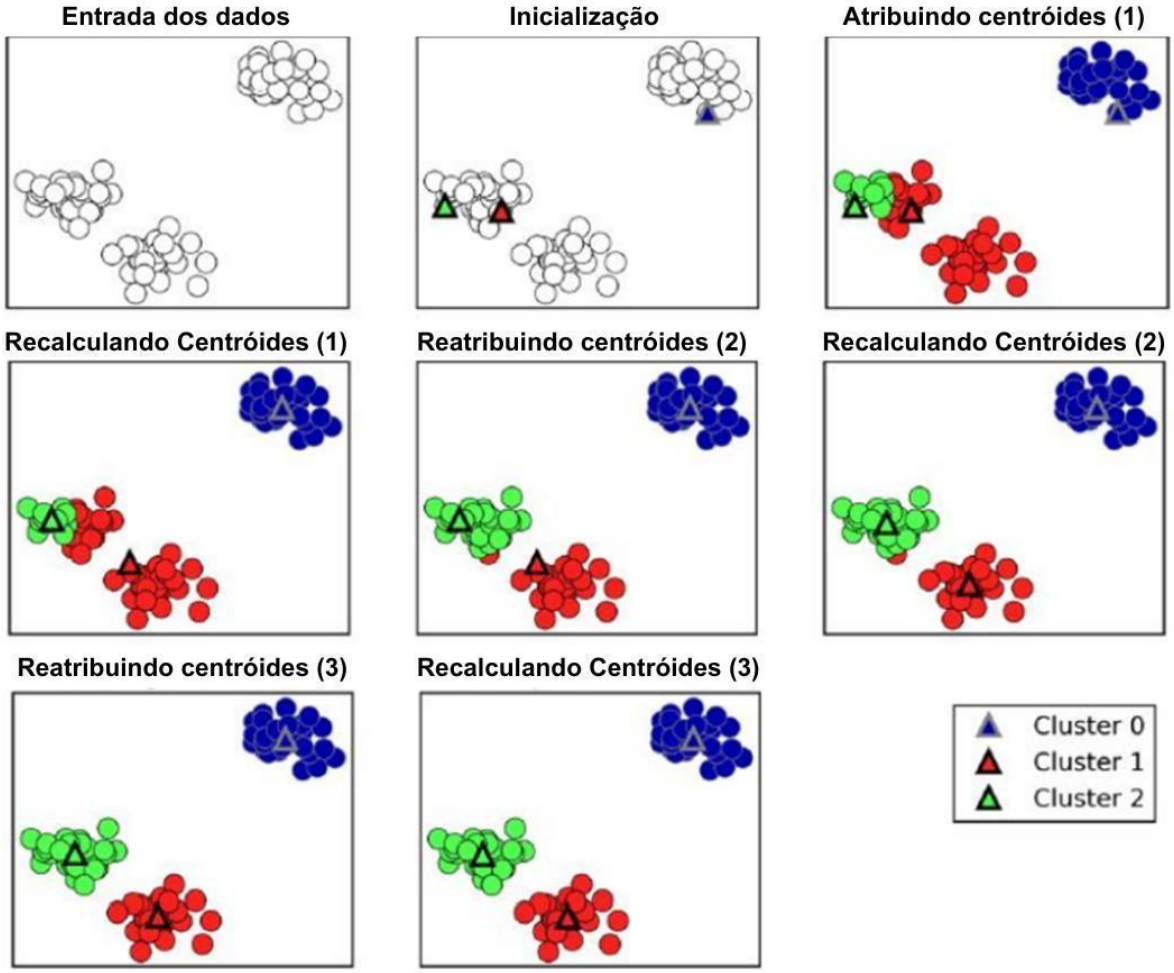
5.1 Etapas da aplicação

Johnson & Wichern (2007) descrevem os seguintes passos para o funcionamento do algoritmo:

1) Inicialmente, é necessário estipular uma quantidade k de clusters, cada um desses clusters possuíra um centróide. A quantidade e o centróide pode ser atribuída aleatoriamente ou devido a algum critério específico.
2) A partir do centróide, o algoritmo calcula a distância entre o centróide e os pontos de dados. Cada elemento será atribuído ao cluster que possuir a menor distância.
3) Após a finalização do passo 2, calcula-se a média da distância de cada ponto ao seu centróide. Então, esta média representará um novo centróide.
4) Com a definição de um novo centróide, os passos 2) e 3) serão repetidos até que a cada nova definição de centróide não haja mais modificações. Ademais, caso seja definido um número de iterações pré-definido o processo pode ser encerrado precocemente (Gerolamo, 2020).

A Figura 6 evidencia os passos desde a entrada dos dados até a definição definitiva dos clusters.

Figura 6 - Passos K-Médias



Fonte: Muller, Guido (2016, pg 169)

Onde cada triângulo representa o centróide de cada cluster.

5.2 Vantagens e desvantagens

Apesar de ser uma ferramenta amplamente utilizada na área de análise de dados, os algoritmos apresentam tanto vantagens quanto limitações significativas. Sua simplicidade e eficácia são inegáveis, pois possui métodos robustos para o agrupamento de dados. Contudo, é importante reconhecer que algumas desvantagens podem limitar sua aplicação em determinados contextos Algumas das vantagens são:

a) A simplicidade do algoritmo, este possui uma abordagem fácil de entender e implementar, mesmo para iniciantes em aprendizado de máquina.
b) A escalabilidade para conjuntos de dados de tamanhos variados, tornando-o adequado para uma ampla gama de aplicações.
c) Os clusters gerados são facilmente interpretáveis e podem fornecer conhecimentos úteis sobre a estrutura dos dados.

Porém, as desvantagens que este apresenta serão citadas abaixo:

a) Dependendo da escolha inicial dos centróides, o resultado pode variar, realizando agrupamentos ruins.
b) A necessidade do usuário fornecer o número de clusters específico, tornando desafiador em algumas situações onde não há um conhecimento prévio sobre a estrutura dos dados.
c) O algoritmo é sensível a outliers, o que pode distorcer os resultados se as variáveis possuírem escalas muito diferentes.

Embora tenha suas limitações, com a devida compreensão de suas características e técnicas adequadas, o algoritmo pode ser uma ótima ferramenta para a análise de dados e a descoberta de padrões em conjuntos de dados de variados domínios.

5.3 Determinação do número de clusters

Uma das etapas mais importantes do processo de aplicação do algoritmo é a determinação do número de clusters, pois essa escolha definirá como os dados serão segmentados. Entretanto, não existe uma resposta definitiva para a escolha desse número; a decisão depende da estrutura dos dados e do problema a ser resolvido. Algumas técnicas auxiliam nessa decisão, sendo o Método do Cotovelo e o Método da Silhueta os mais utilizados.

5.3.1 Método do Cotovelo

O método do cotovelo é um método muito utilizado para a escolha da quantidade de clusters, consiste na ideia de executar o algoritmo diversas vezes com quantidades diferentes de clusters e observar qual a melhor. Este método se utiliza de um índice que calcula a distância dentro dos clusters, sendo esse um valor decrescente. Pois, ao aumentar a quantidade de clusters, os clusters acabam se tornando mais parecidos entre eles e diminuindo a semelhança entre os elementos dentro do cluster. O índice é calculado da seguinte maneira:

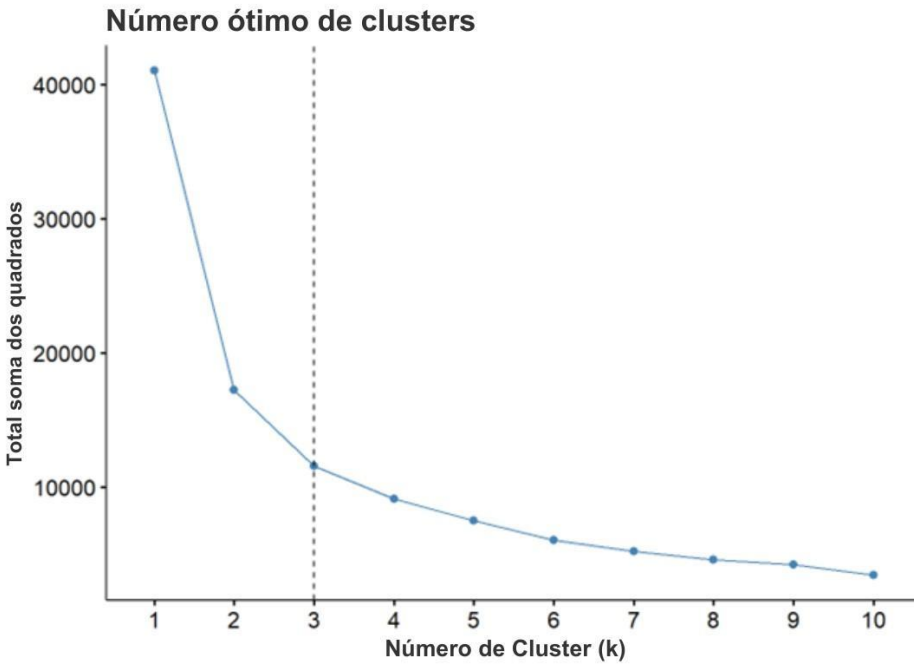
$$W_j = \sum_{i=1}^{j} \quad (\frac{1}{N_j} * D_j)$$

onde j representa a quantidade de clusters, Nj o número de elementos em cada cluster, Dj a distância de cada ponto com seu centróide (Gerolamo, 2020).

O objetivo é encontrar um ponto onde ocorra a menor variação na quantidade deste índice, mantendo assim os elementos intra-clusters o mais homogêneos possível e garantindo que os clusters possuam a maior distinção possível.

Através da Figura 7, é possível observar que a linha tracejada está no número 3 de clusters. Isso significa que este número representa a melhor quantidade, pois se aumentarmos para 4 clusters, não haverá uma diferença significativa no erro, o que forma o "cotovelo”.

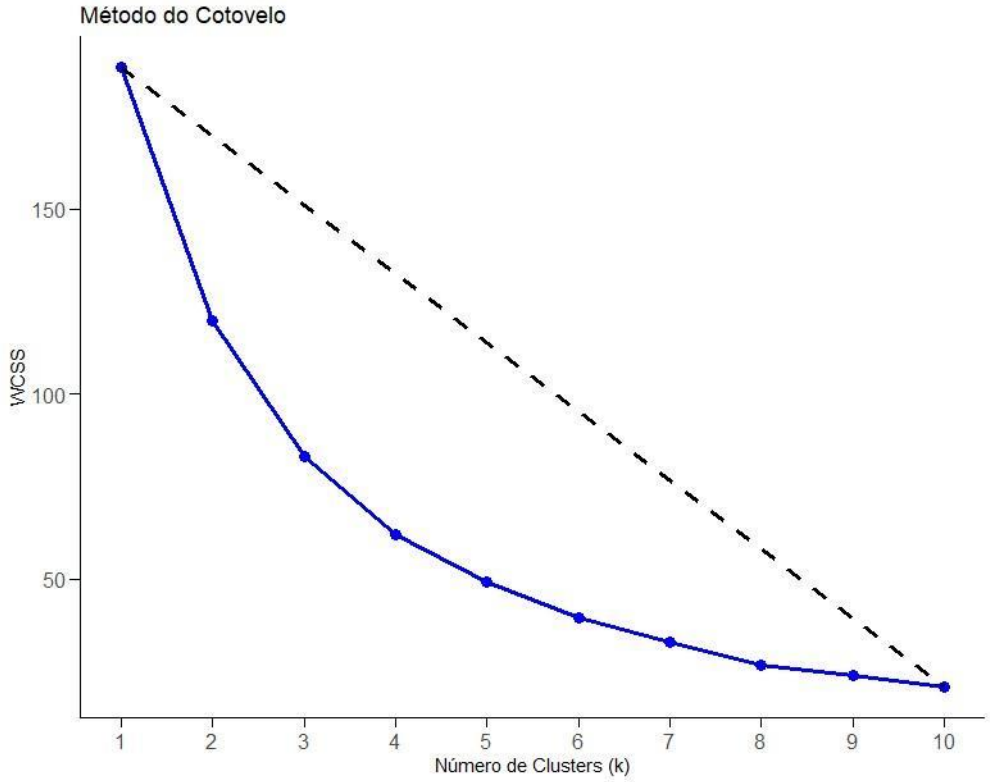
Figura 7 - Exemplo Método Cotovelo



Fonte: Gabriel Dias. Disponível em: https://rpubs.com/diascodes/770518. Acesso em: 03/06/2024.

Entretanto, em certos casos, pode ser visualmente complicado encontrar a quantidade via gráfico. Pensando nisso, é possível recorrer ao método algébrico. Esse método envolve traçar uma reta entre o primeiro e o último ponto no eixo x e, em seguida, calcular a distância do ponto desejado até esta reta.

Figura 8 - Método cotovelo representação algébrica



Fonte: Elaborado pelo autor

A distância entre o ponto e a reta é calculado matematicamente da seguinte forma:

$$Dist\ (P_1 P_n, (x,y)) = \frac{|(y_n - y_1)x - (x_n - x_1)y + x_n * y_1 - y_n * x_1|}{\sqrt{(y_n - y_1)^2 + (x_n - x_1)^2}}$$

Sendo x_1, y_1 coordenadas com 1 cluster; x_n, y_n da coordenadas com 10 clusters e x, y as coordenadas do ponto que desejamos calcular. O ponto escolhido será aquele que possui a maior distância em relação à reta tracejada.

5.3.2 Método da Silhueta

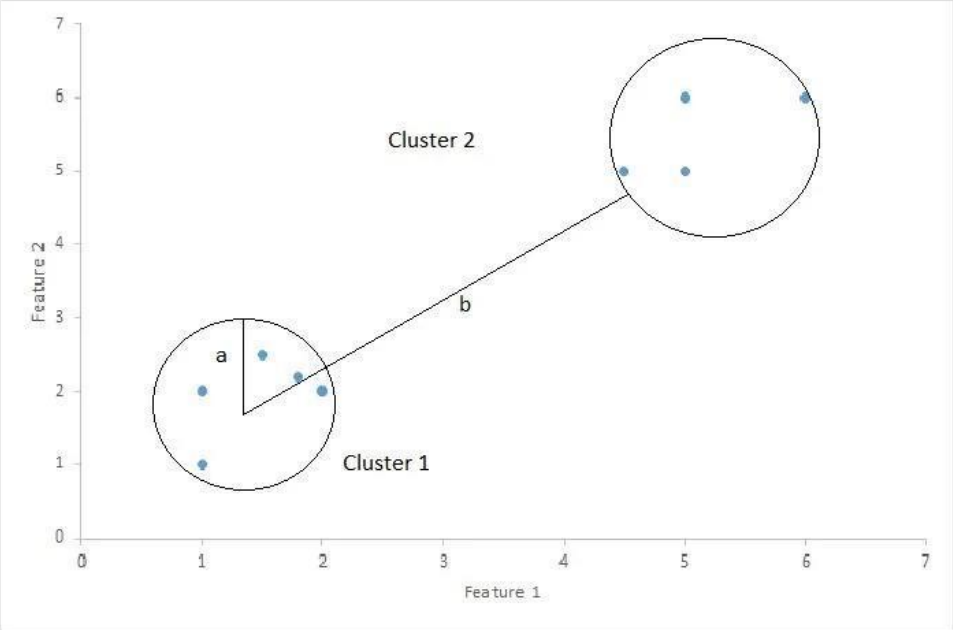
O método da silhueta é uma técnica de avaliação de qualidade do algoritmo. Ele mede o quão bem cada objeto foi classificado em seu cluster em comparação com outros clusters. Em outras palavras, ajuda a determinar quão bem os clusters estão separados.

O coeficiente de silhueta varia de -1 a 1. Um valor próximo de 1 indica que o ponto está bem dentro de seu próprio cluster e distante dos outros clusters. Um valor próximo de 0 indica que o ponto está próximo da fronteira entre dois clusters. Um valor próximo de -1 indica que o ponto pode ter sido atribuído ao cluster errado. A fórmula deste coeficiente é:

$$s = \frac{b-a}{max(a,b)},$$

onde, s é o coeficiente, a é a distância entre um elemento do grupo e o restante do mesmo grupo, b é a distância entre um elemento e todos os elementos do grupo próximo e $max(a, b)$ representa o valor máximo entre a e b. A Figura 9 mostra intuitivamente o funcionamento do método.

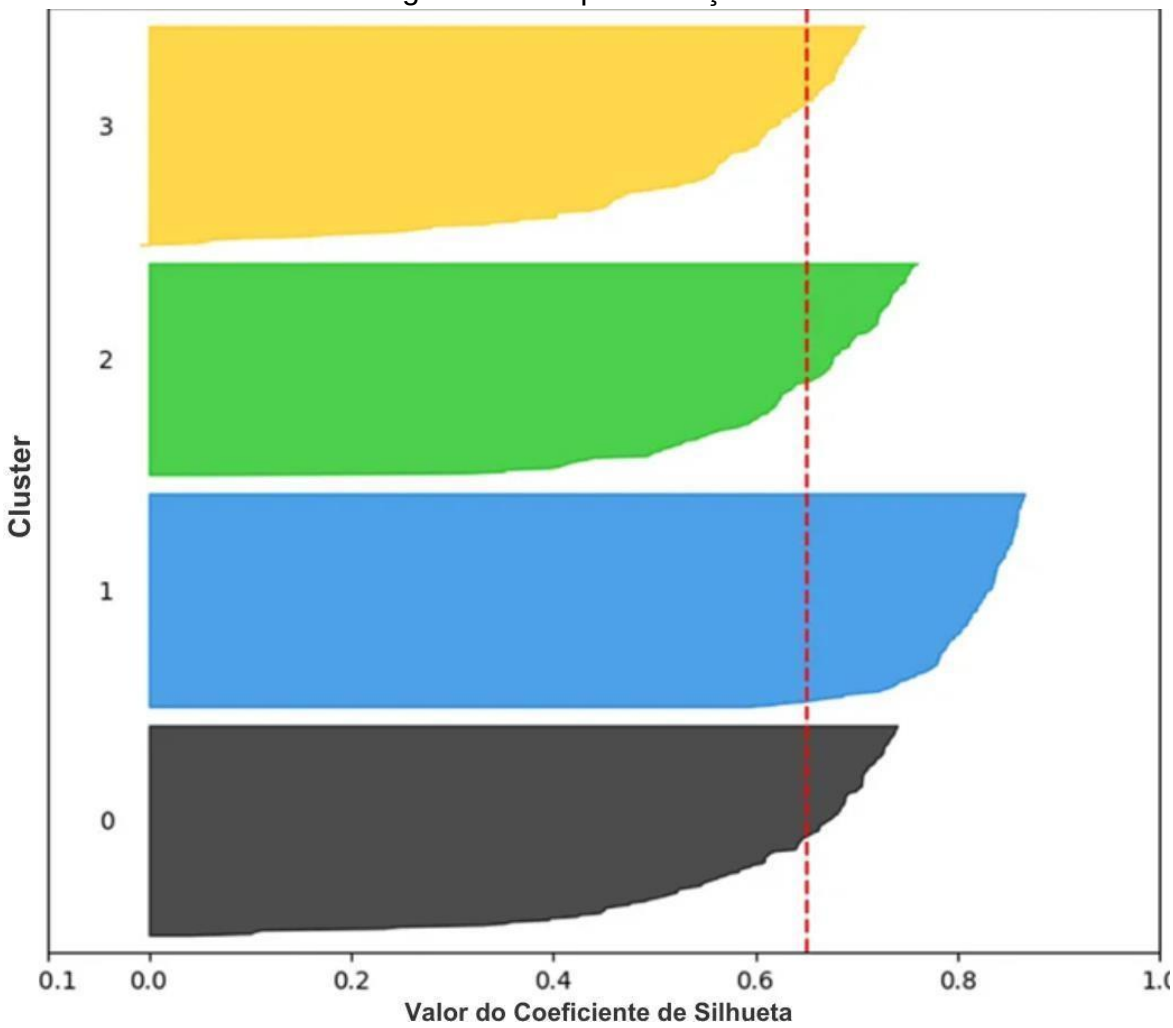
Figura 9 - Exemplo coeficiente de silhueta



Fonte: Bhardwaj, Ashutosh. Disponivel em: https://towardsdatascience.com/silhouettecoefficient-validating-clustering-techniques-e976bb81d10c. Acesso em: 04/06/2024.

O nome "silhueta" vem da analogia visual que o método proporciona como ilustra a Figura 10. Pois, visualmente, você pode pensar nas silhuetas de cada figura projetadas contra um fundo. Se uma figura se encaixa bem em seu cluster e está bem separada dos outros clusters, sua silhueta será nítida e bem definida. Se a figura está próxima dos limites do cluster ou sobreposta com outras figuras de clusters diferentes, sua silhueta será menos nítida.

Figura 10 - Representação visual silhueta



Fonte: Zuccarelli, Eugenio. Disponível em: https://towardsdatascience.com/performancemetrics-in-machine-learning-part-3-clusteringd69550662dc6. Acesso em: 04/06/2024.

6 Análise exploratória dos dados

A análise descritiva foi uma etapa fundamental na exploração e compreensão do conjunto de dados. O principal objetivo foi resumir e descrever as características essenciais dos dados de uma maneira que seja fácil de entender e interpretar. De maneira que auxilie na compreensão dos dados para o desenvolvimento do projeto.

6.1 Conhecimento das variáveis

O conjunto de dados original possui 19239 linhas e 23 colunas contendo informações sobre os atletas dentro do jogo. Porém, para o desenvolvimento do projeto não serão utilizadas todas as 23 colunas. Portanto, serão selecionadas apenas as que mais interessam para a análise descritiva.

As variáveis escolhidas para compor o novo conjunto de dados foram:

Quadro 1 - Tipo das variáveis

Variáveis	Tipo da variável
short_name (Nome)	Objeto
age (Idade)	Número inteiro
player_positions (Posição)	Objeto
overall (Pontuação geral)	Número inteiro
potential (Pontuação que pode ser alcançada)	Número inteiro
value_eur (Valor de mercado)	Número inteiro
nationality_name (Nacionalidade)	Objeto

Fonte: Elaborado pelo autor (2024).

No entanto, o FIFA 22 enfrenta um problema relacionado aos direitos autorais devido à utilização das imagens e nomes dos jogadores e clubes. Para garantir autenticidade, o FIFA requer o licenciamento desses elementos. Portanto, para aumentar a semelhança entre o jogo e a realidade, serão utilizados todos os clubes e atletas das principais ligas licenciadas presentes no jogo. As ligas selecionadas para compor o conjunto de dados são: “Ligue 1” (Campeonato Francês), “Bundesliga” (Campeonato Alemão), “La Liga” (Campeonato Espanhol), “Serie A” (Campeonato Italiano) e “Premier League” (Campeonato Ingles). Desse modo, o conjunto foi reduzido de 19239 para 2976 linhas.

6.2 Estatística das variáveis

Neste momento, as estatísticas de cada variável foram analisadas, incluindo média, mediana, desvio padrão e os quartis, conforme apresentado na Figura 11.

Figura 11 - Estatística das variáveis

	Quantidade	Média	Desvio padrão	Mínimo	25%	50%	75%	Máximo
Idade	2.976	25	5	16	21	25	28	40
Pontuação Geral	2.976	72	7	51	67	73	77	93
Potencial	2.976	78	5	60	74	78	81	95
Valor_EUR	2.976	10.330.225	16.504.702	60.000	1.600.000	3.900.000	12.000.000	194.000.000

Fonte: Elaborado pelo autor (2024).

6.2.1 Informações sobre as variáveis

Observando a Figura 11, diversas informações relevantes foram extraídas, oferecendo uma visão sobre as características dos jogadores:

a) Faixa Etária dos Atletas: O atleta mais jovem tem apenas 16 anos, enquanto o mais velho tem 40 anos. A idade média dos jogadores é de 24,5 anos, o que sugere uma ampla quantidade de atletas em fase de desenvolvimento e maturidade física e técnica.

b) Pontuação Geral: O atleta com a melhor pontuação geral possui 93 pontos, destacando-se como o mais completo entre os analisados. Além disso, o

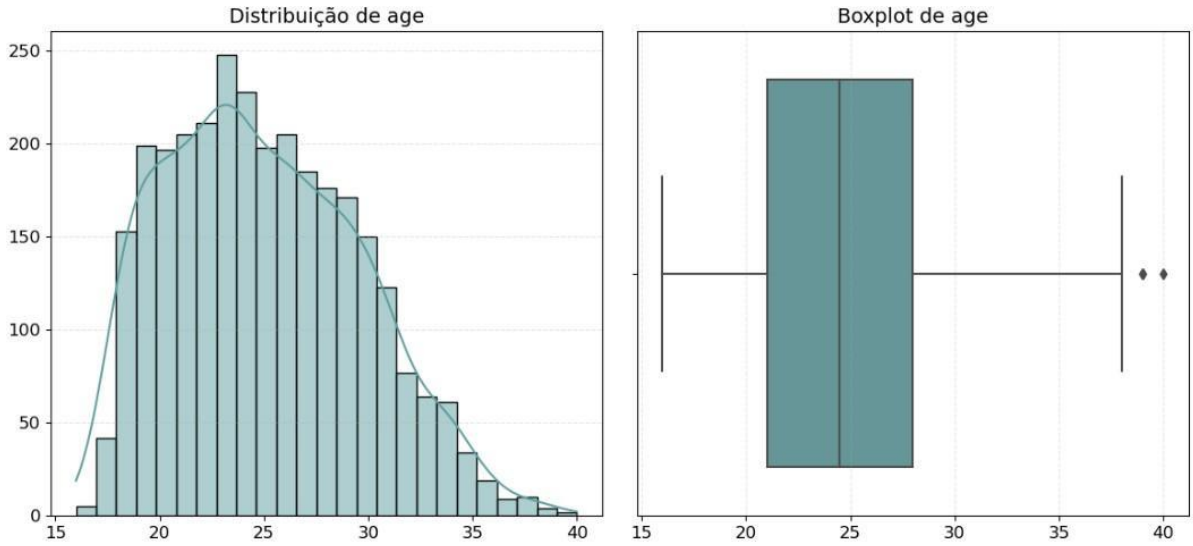
jogador mais promissor tem o potencial de alcançar 95 pontos, indicando um talento que pode se desenvolver.

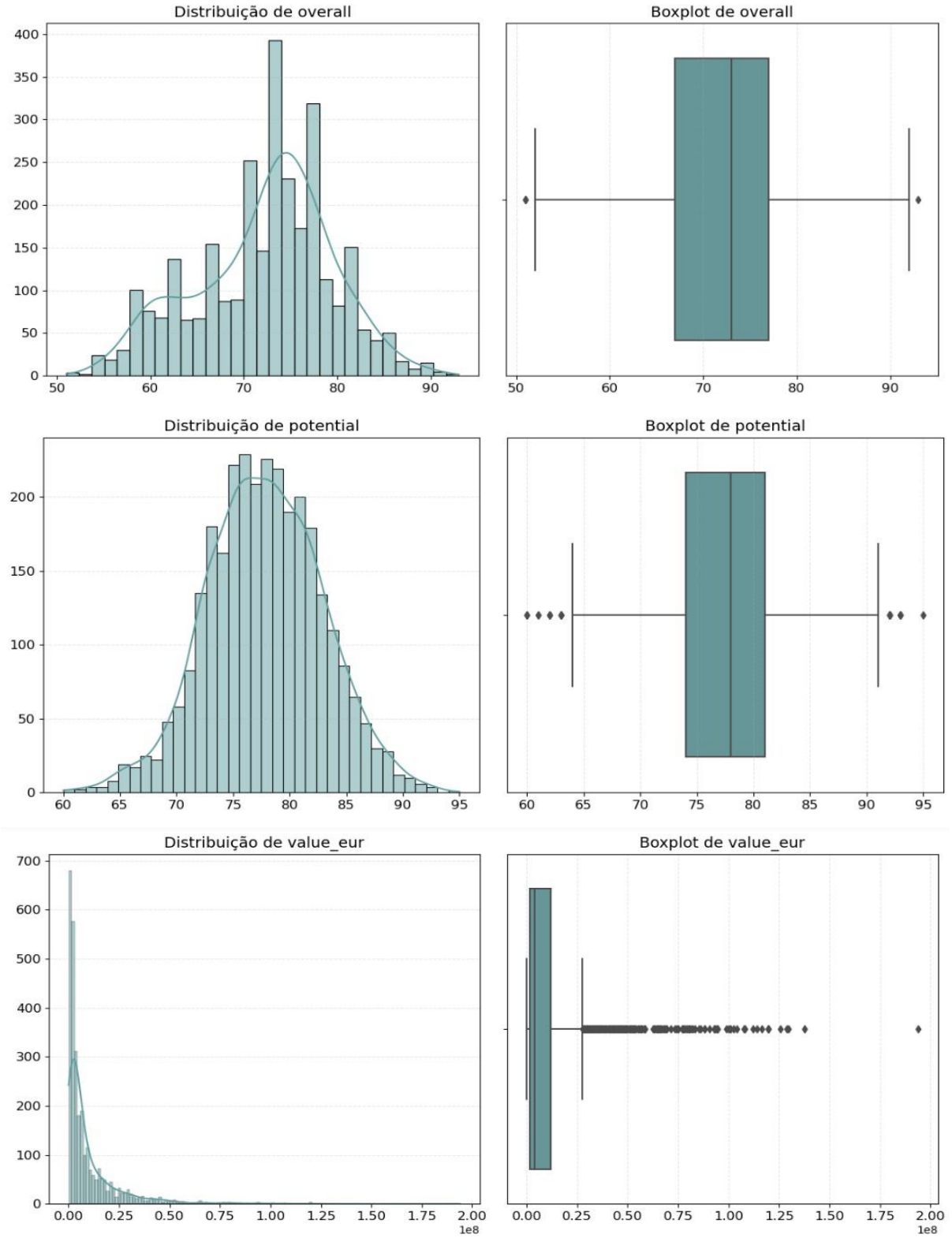
c) Valor de Mercado: O atleta mais caro tem um valor quase 19 vezes maior do que a mediana dos jogadores. Isso destaca a disparidade significativa nos valores atribuídos a alguns jogadores, possivelmente refletindo fatores como habilidade e popularidade.

6.2.2 Tratamento dos dados

O conjunto de dados não contém valores duplicados ou valores nulos. Neste instante, é crucial verificar os valores extremos (outliers) para a aplicação do algoritmo K-Médias, pois esses valores podem desregular os centróides, formando clusters irregulares. Também, verificar as distribuições das variáveis, conforme os gráficos da Figura 12.

Figura 12 - Gráfico das variáveis





Fonte: Elaborado pelo autor (2024).

A presença de outliers é evidenciada pelos boxplots. A principal decisão a ser tomada é como lidar com esses outliers. Devemos prestar especial atenção aos que estão relacionados ao valor dos atletas, como a variável “Valor”. Um dos valores extremos identificados corresponde a Kylian Mbappé, um promissor jogador francês de 23 anos[1]. Seu alto valor de mercado é justificado por suas habilidades e por ser considerado um dos melhores jogadores do jogo. No entanto, ao formar um bom elenco, é importante questionar: seria vantajoso contratar esse jogador ou seria possível encontrar alternativas mais acessíveis que também sejam promissoras? A resposta a essas questões poderá ser abordada posteriormente na aplicação do algoritmo. Por enquanto, a decisão também foi manter os outliers na base de dados, uma vez que esses valores refletem a realidade dos atletas e a autenticidade da base de dados.

Através dos gráficos de histogramas, é possível analisar as distribuições dos dados. Todas as variáveis são quantitativas contínuas, ou seja, variáveis que podem assumir qualquer valor dentro de um intervalo contínuo. Esses valores geralmente incluem números com frações ou casas decimais.

Nas variáveis que expressam os valores dos atletas, fica explícito que alguns são considerados exceções. No entanto, ao observar as variáveis que medem o desempenho dos atletas, essa distinção não é muito significativa. Apesar de as variáveis de desempenho estarem limitadas entre 0 e 100, os jogadores com maiores valores deveriam ter maior impacto no quesito desempenho.

[1] No FIFA 22, o jogador possuía 23 anos de idade, atualmente este tem 25 anos.

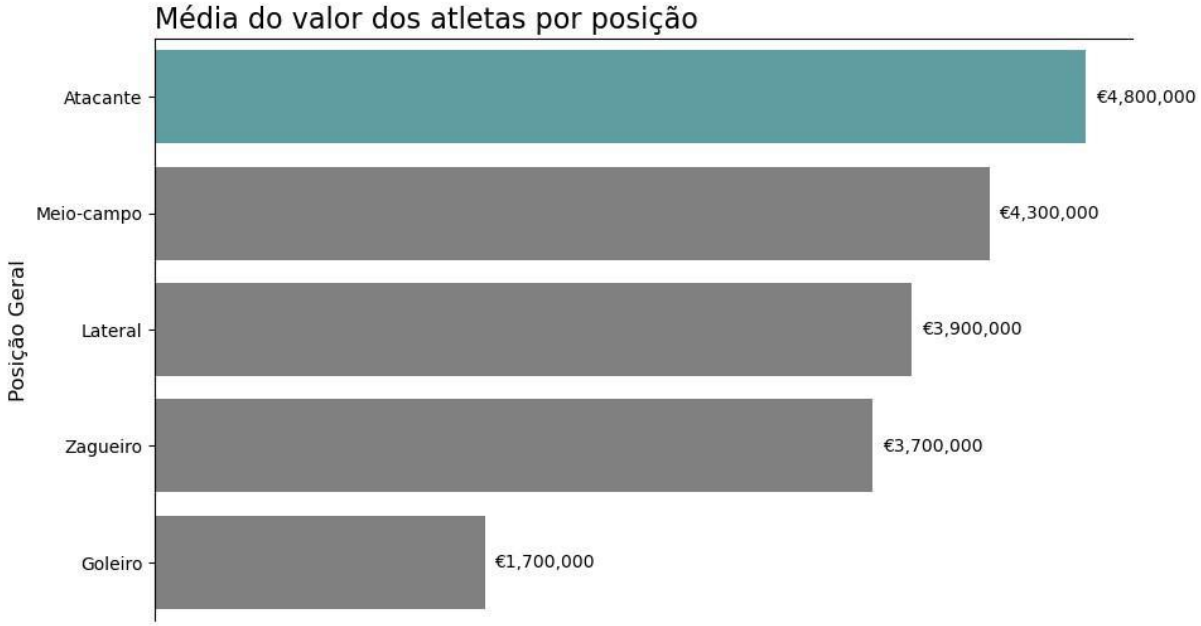
6.2.3 Análises extras

Neste momento, o estudo foi detalhando as variáveis com o objetivo de entender suas características e comportamentos na base de dados. Cada variável foi analisada para identificar padrões e possíveis correlações. Este processo foi essencial para adquirir uma compreensão profunda dos dados, o que permitirá realizar análises e tirar conclusões mais precisas. Além disso, a compreensão do comportamento das variáveis ajudou na melhoria da modelagem dos dados, contribuindo para a eficácia geral do estudo.

A primeira dúvida que surgiu foi: será que a mediana do valor dos atletas varia de acordo com sua posição?

Através da Figura 13, observa-se que os atletas que são atacantes possuem um valor de €4.800.000 de mediana, sendo a maior mediana de valor de mercado. Isso pode ser explicado pela sua importância em decidir jogos, tendo como principal objetivo a realização de gols, o que, supostamente, os torna mais valiosos em comparação com as demais posições.

Figura 13 - Média do valor dos atletas por posição

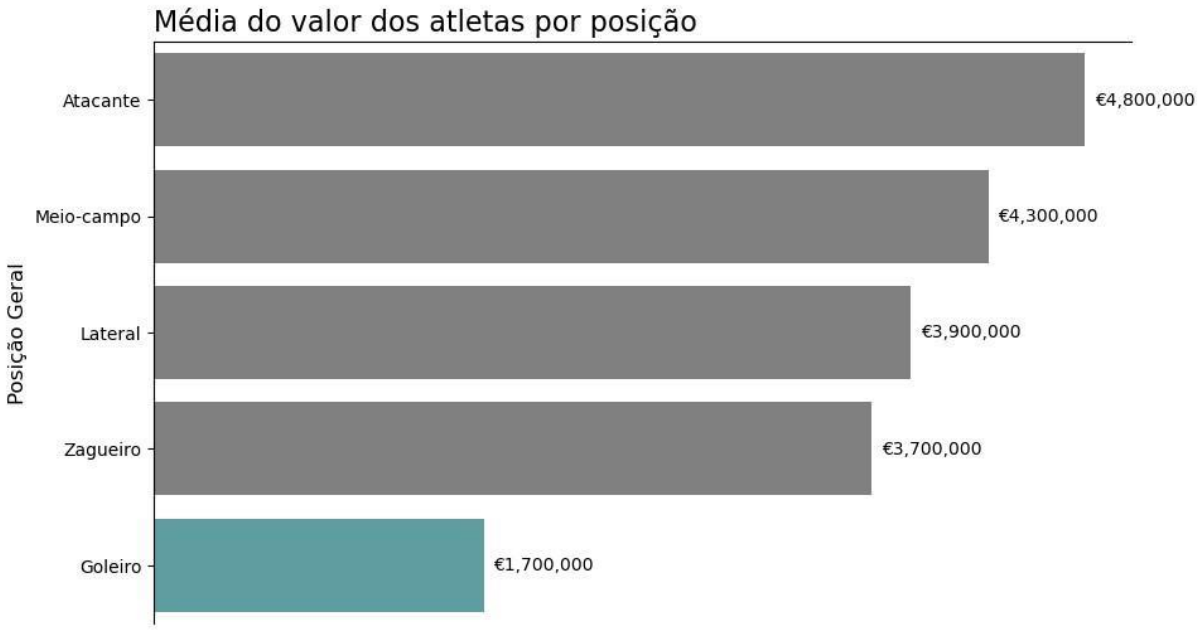


Fonte: Elaborado pelo autor (2024).

Entretanto, apesar de desempenharem um papel crucial para o time ao defender bolas difíceis que poderiam se transformar em gols, os goleiros têm um valor

€1.700.000 de mediana, o que representa cerca de 47,2% abaixo da mediana das outras posições. Conforme mostra a Figura 14.

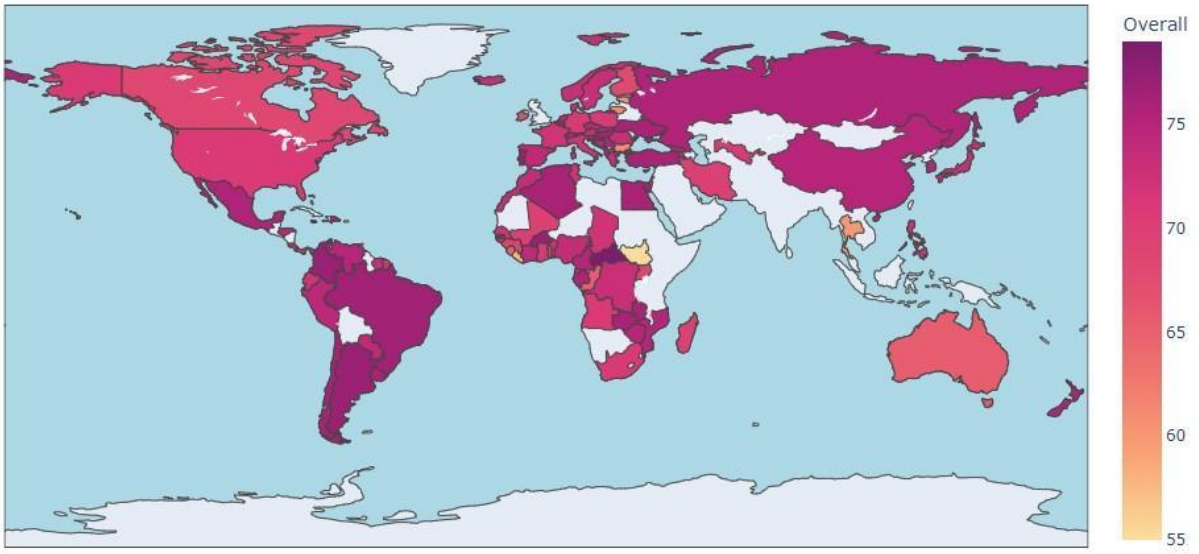
Figura 14 - Média do valor de goleiros



Fonte: Elaborado pelo autor (2024).

Por meio da Figura 15, verifica-se que o Sudão do sul (País em amarelo) possui o menor Pontuação Geral de mediana, sendo considerado o país com os piores jogadores da base de dados. Ao analisar a base, observou-se que alguns países possuem mediana que chama a atenção, como o exemplo da República Africana Central, averiguando percebeu-se que há apenas um jogador, o que elevou a mediana do país. Neste momento, serão selecionados apenas países com mais de 30 atletas na base de dados, a fim de evitar esta tendência.

Figura 15 – Pontuação geral médio dos atletas por país.

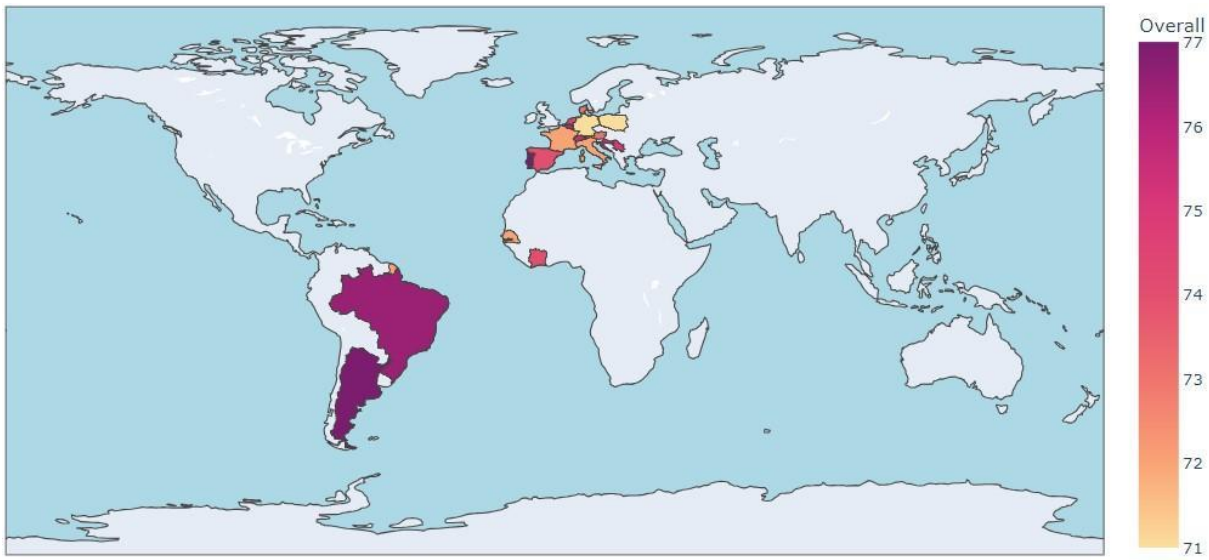


Fonte: Elaborado pelo autor (2024)

Como anteriormente apenas as ligas licenciadas foram selecionadas, estas são todas europeias. Portanto, apenas atletas que jogam na Europa foram incluídos na análise, resultando em poucos países com mais de 30 jogadores ativos na Europa. Analisando a Figura 16, um dos países com a maior mediana de Pontuação Geral é a Argentina, atual campeã da Copa do Mundo. Isso evidencia uma forte relação entre o desempenho no jogo e a realidade, considerando que o lançamento do FIFA 22 ocorreu antes do título da Argentina[2]. O Brasil ocupa a terceira posição entre os países com as maiores medianas de Pontuação Geral. Não foi possível avaliar as seleções em específico, apenas as nacionalidades dos atletas. Pois algumas seleções não são licenciadas pelo jogo, como o Brasil.

[2] O jogo FIFA 22 foi lançado em 27 de setembro de 2021, o título mundial da Argentina foi 12 de dezembro de 2022.

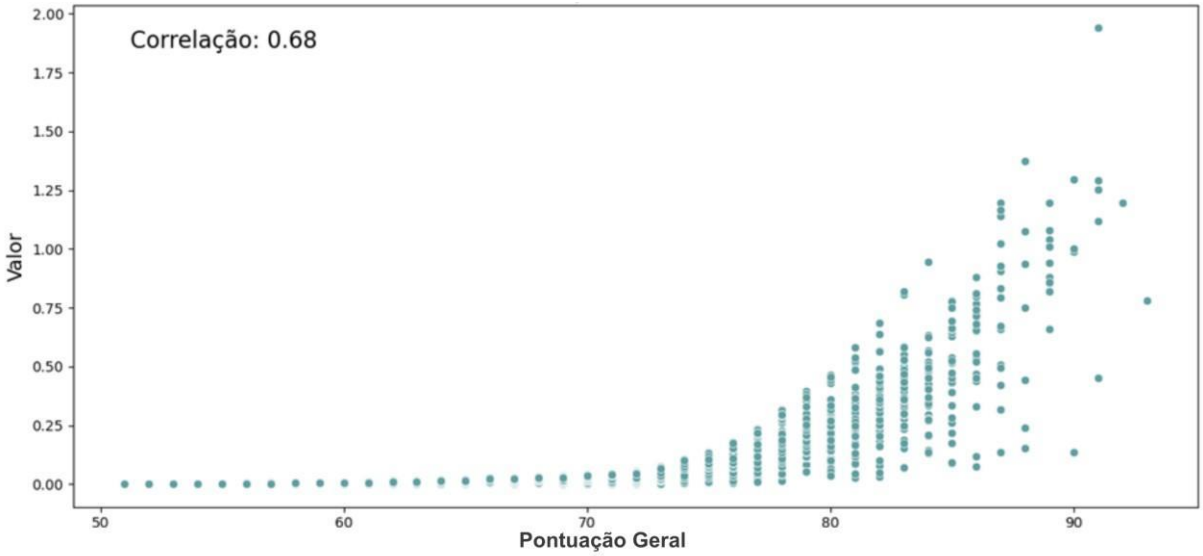
Figura 16 - Pontuação Geral para países com mais de 30 atletas



Fonte: Elaborado pelo autor (2024)

Segundo a Figura 17, destaca-se a correlação moderada entre o Pontuação Geral do atleta e seu valor. Após a pontuação 70 de Pontuação Geral a correlação fica mais explícita.

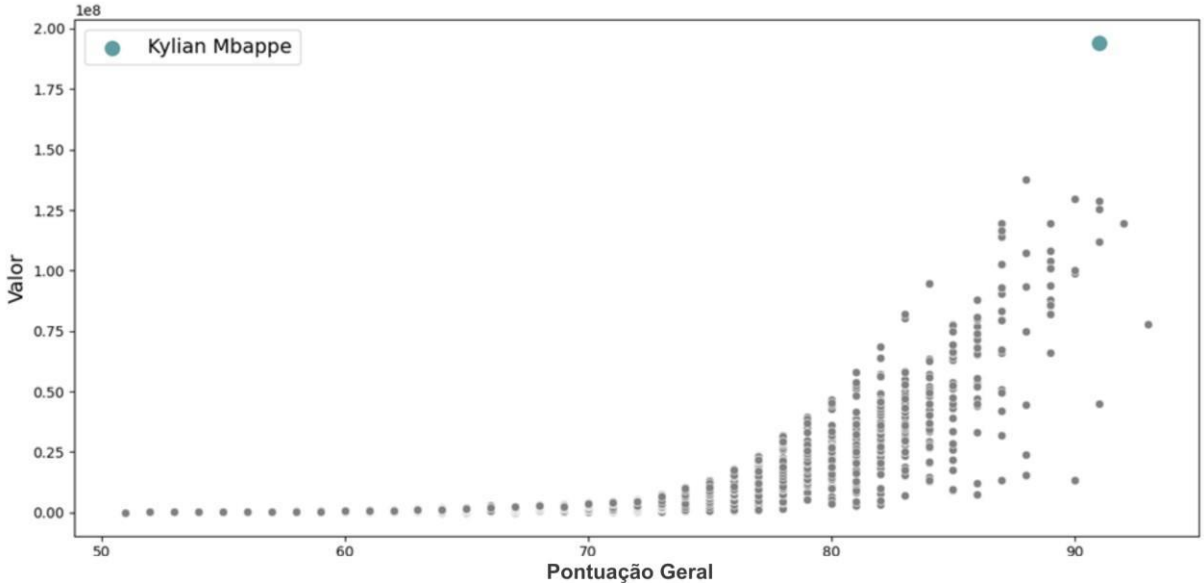
Figura 17 - Correlação Pontuação Geral vs Valor.



Fonte: Elaborado pelo Autor (2024)

Conforme mostra a Figura 18, o atleta Mbappé é o jogador mais valioso do conjunto de dados, porém ele não é o atleta com maior Pontuação Geral. Como explicado anteriormente, este valor pode ser explicado pela sua juventude, possibilidade de evolução e desempenho regular por maior quantidade de tempo.

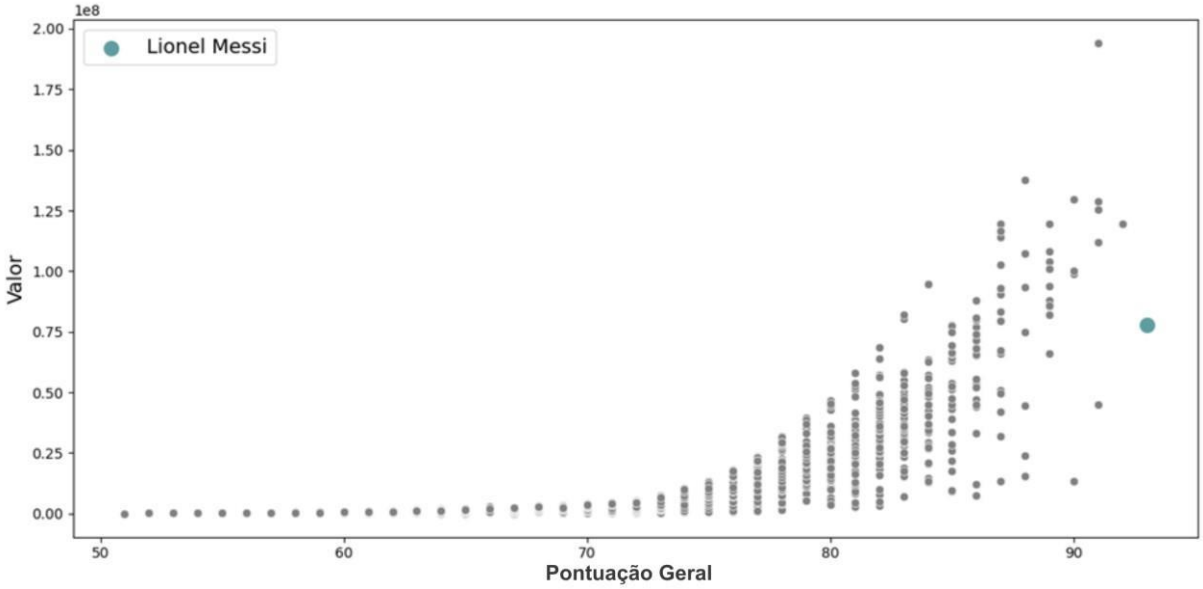
Figura 18 - Gráfico de dispersão destacando Kylian Mbappé.



Fonte: Elaborado pelo autor (2024).

De acordo com a Figura 19, Lionel Messi é o jogador com maior Pontuação Geral do conjunto de dados, apesar disso, ele é o vigésimo jogador mais valioso do jogo. O valor ainda é alto, entretanto evidencia que as vezes nem sempre é necessário desembolsar valores altos para conseguir bons jogadores.

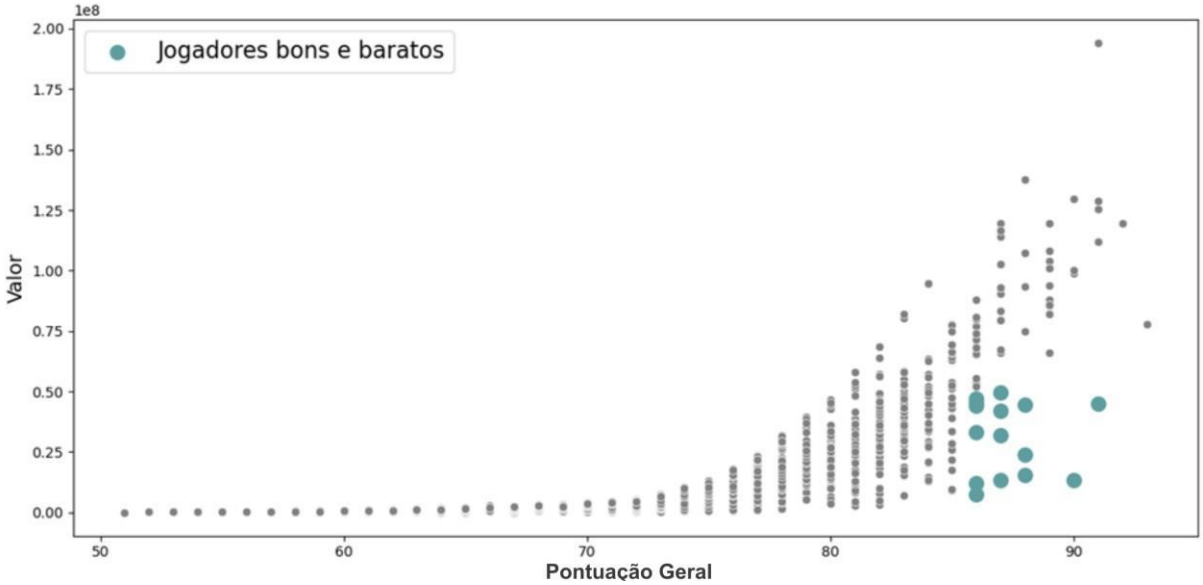
Figura 19 - Gráfico de dispersão destacando Lionel Messi.



Fonte: Elaborado pelo autor (2024).

Após reparar a existência de eventuais bons jogadores com menores preços. Foram selecionados manualmente atletas que podem ser considerados bons jogadores, mas são baratos em comparação à sua Pontuação Geral. Os jogadores selecionados manualmente estão representados na Figura 20.

Figura 20 - Gráfico de dispersão destacando atletas bons e “baratos”.



Fonte: Elaborado pelo autor (2024).

A Figura 21, evidencia a semelhança entre esses jogadores, devido que todos possuem idade superior a 30 anos, o que significa que dificilmente alcançarão seu potencial máximo por muito tempo. Desta forma, é importante considerar o momento atual do elenco e quais jogadores se adequam melhor ao time.

Figura 21 - Jogadores selecionados manualmente

Nome	Idade	Overall	Valor
Cristiano Ronaldo	36	91	45.000.000,00
M. Neuer	35	90	13.500.000,00
Sergio Ramos	35	88	24.000.000,00
L. Suárez	34	88	44.500.000,00
K. Navas	34	88	15.500.000,00
H. Lloris	34	87	13.500.000,00
L. Modrié	35	87	32.000.000,00
Á. Di María	33	87	49.500.000,00
W. Szczesny	31	87	42.000.000,00
G. Chiellini	36	86	12.000.000,00
S. Handanovit	36	86	7.500.000,00
M. Hummels	32	86	44.000.000,00
Jordi Alba	32	86	47.000.000,00
Sergio Busquets	32	86	45.000.000,00
J. Vardy	34	86	33.000.000,00

Fonte: Elaborado pelo autor (2024).

7 Aplicação do modelo

Após a análise exploratória dos dados, foram extraídas informações relevantes da base e compreendidas suas características principais. Neste momento, será aplicado o modelo de clusterização com o objetivo de identificar grupos de jogadores com características semelhantes.

7.1 Apresentação dos dados

Os dados utilizados neste momento foram coletados do site "Kaggle", disponibilizados pelo usuário Stefano Leone, contendo informações sobre jogadores em um banco de dados com 110 colunas referentes ao "FIFA 22". Dentre essas 110 colunas, apenas 3 serão utilizadas para o treinamento do modelo, sendo elas:

a) Potencial: Refere-se ao máximo de "Overall" (Pontuação Geral) que o atleta pode atingir, variando de 0 a 99. Quanto maior o potencial, melhor o atleta será no futuro. No entanto, à medida que a idade avança, o potencial tende a deixar de crescer.
b) Valor EUR: Representa o valor do atleta em Euros.
c) Idade: Indica a idade do atleta em 2022.

As estatísticas das variáveis podem ser observadas no Quadro 2.

Quadro 2 - Estatísticas das variáveis do modelo.

	Potencial	Valor EUR	Idade
Quantidade	2976	2976	2976
Média	78	10330225	25
Desvio Padrão	5	16504702	5
Mínimo	60	60000	16
25%	74	1600000	21
50%	78	3900000	25
75%	81	12000000	28
Máximo	95	194000000	40

Fonte: Elaborado pelo autor (2024).

7.2 Tratamento das variáveis

Ao analisar as estatísticas, é possível observar que as variáveis possuem escalas diferentes. Por exemplo, a variável “Valor EUR” tem uma escala muito maior do que as demais. Como o K-Médias depende da métrica de distância, ele é diretamente afetado pela escala das variáveis. Se uma variável tiver uma escala muito superior, ela pode "dominar" o cálculo das distâncias, distorcendo os resultados do modelo. Portanto, é necessário utilizar uma técnica que garanta que cada variável contribua de maneira proporcional para as distâncias, permitindo uma classificação ou agrupamento mais equilibrado. Para este projeto, a técnica utilizada será a padronização.

A padronização (ou normalização Z-score) é uma técnica estatística usada para transformar os dados de modo que eles tenham uma média igual a zero e um desvio padrão igual a um. Isso permite que variáveis em diferentes escalas possam ser comparadas ou analisadas conjuntamente, sem que as diferenças nas unidades de medida influenciem a análise. Esta técnica é muito utilizada em situações onde os modelos são baseados em distâncias.

Matematicamente, a padronização de uma variável $X_1,..,X_n$ com média μ e desvio padrão σ é realizada através da fórmula:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

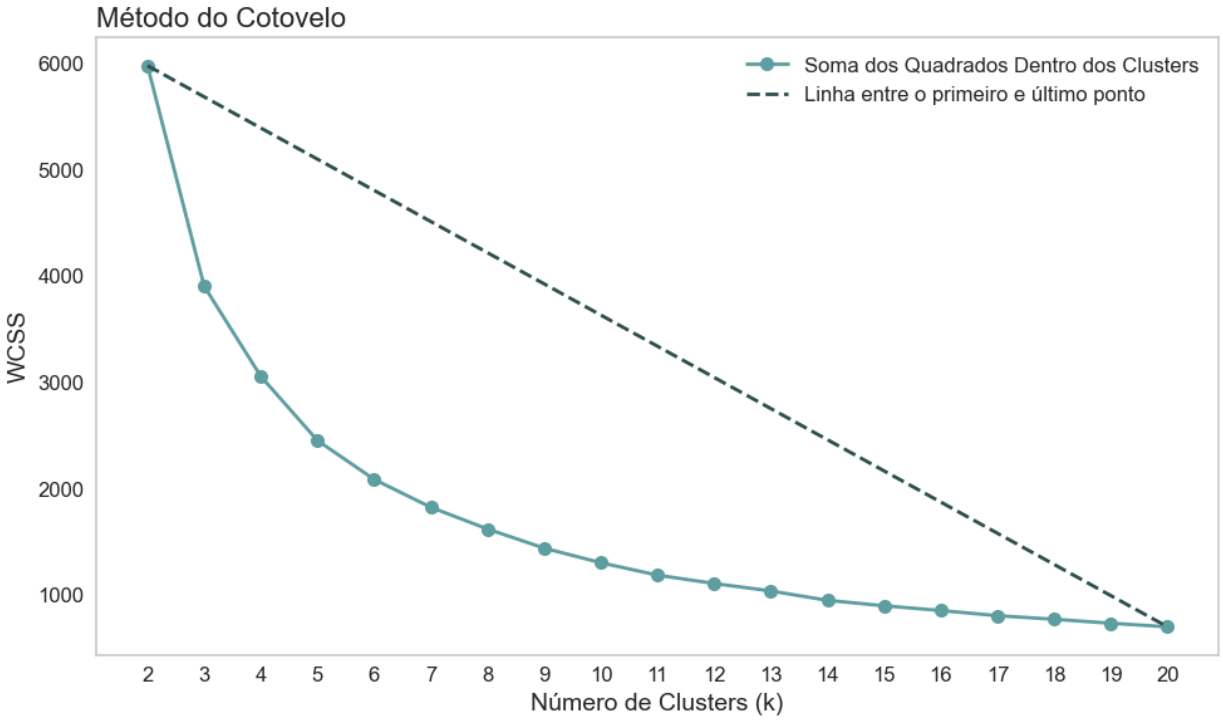
Após a padronização, as variáveis mantêm as informações originais, porém passam a ter uma distribuição normal padrão, facilitando a comparação entre elas.

7.3 Escolha da quantidade de cluster

Como mencionado anteriormente, a escolha da quantidade de cluster será feita a partir do método do cotovelo. Para este método, a base de dados foi treinada em 20 modelos diferentes, cada um com um número de clusters (k) variando em sequência. O resultado dessa análise é apresentado na Figura 22 e no Quadro 3.

Através desta figura, repara-se que a partir de cinco clusters, a soma dos quadrados dentro dos cluster (WCSS - Weighted Cumulative Sum of Squares) tem uma queda menos acentuada em comparação com os valores anteriores indicando que pode não ser necessário adicionar mais cluster, pois a diferença entre eles se torna cada vez menos significativa.

Figura 22 – Método do cotovelo.



Fonte: Elaborado pelo autor (2024).

Quadro 3 – Distâncias método do cotovelo.

Diferença até a linha tracejada
k=2: 0.00
k=3: 1782.62
k=4: 2337.35
k=5: 2647.58
k=6: 2720.08
k=7: 2689.01
k=8: 2599.97
k=9: 2486.97
k=10: 2330.71
k=11: 2154.46
k=12: 1939.41
k=13: 1716.48
k=14: 1511.34
k=15: 1269.23
k=16: 1020.34
k=17: 775.61
k=18: 515.87
k=19: 259.76
k=20: 0.00

Fonte: Elaborado pelo autor (2024).

Para confirmar essa escolha, será utilizado o método da silhueta. O número de clusters será definido com base na melhor distribuição dos atletas, buscando a métrica mais próxima de 1. Isso ocorre porque quanto mais próximo de 1, mais adequado o atleta está ao seu respectivo cluster. Por outro lado, se o valor estiver próximo de -1, indica que o atleta não está bem alocado naquele cluster.

Ao observar o Quadro 4, nota-se que os coeficientes de silhueta para 11 a 15 clusters são melhores comparados aos observados para 6 a 10 clusters. Porém, aumentar significativamente o número de clusters comprometeria o objetivo do estudo, pois resultaria em clusters com menor distinção entre si. Portanto, com objetivo de equilibrar um bom coeficiente de silhueta e uma quantidade pequena de clusters, foi determinado o número cinco de clusters.

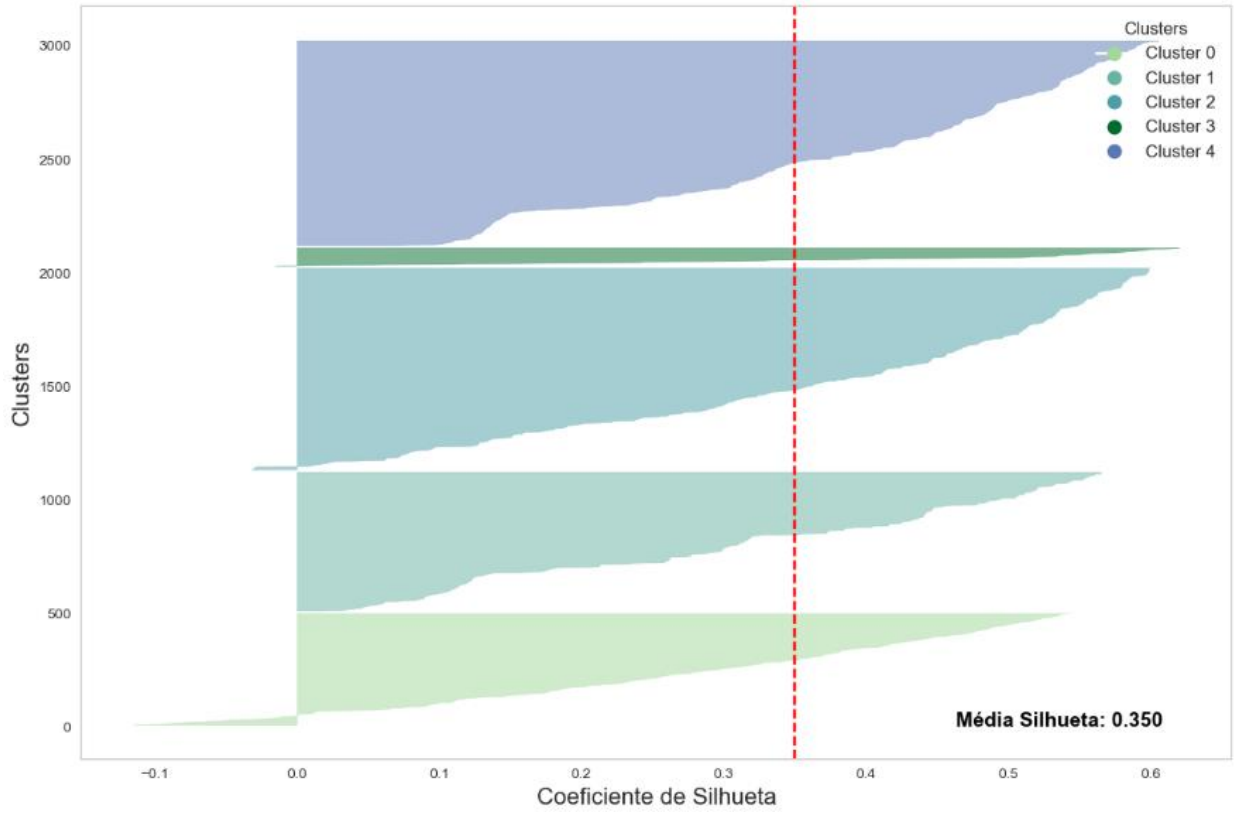
Quadro 4 – Coeficiente de Silhueta para diferentes quantidades de clusters.

Coeficiente de Silhueta
2 clusters: 0.420
3 clusters: 0.383
4 clusters: 0.372
5 clusters: 0.350
6 clusters: 0.333
7 clusters: 0.330
8 clusters: 0.328
9 clusters: 0.333
10 clusters: 0.332
11 clusters: 0.337
12 clusters: 0.337
13 clusters: 0.338
14 clusters: 0.343
15 clusters: 0.341
16 clusters: 0.330
17 clusters: 0.329
18 clusters: 0.343
19 clusters: 0.320
20 clusters: 0.324

Fonte: Elaborado pelo autor (2024).

A Figura 23 mostra como os dados estão distribuídos em cada clusters. Os clusters estão proporcionalmente distribuídos, exceto o cluster três. Após análise, concluiu-se que o cluster reúne os melhores jogadores do jogo (mais caros, maior potencial e idade variada), como será mostrado posteriormente. Também é possível observar que o primeiro, o quarto e o último cluster contém dados que foram atribuídos de forma incorreta.

Figura 23 – Coeficiente de Silhueta.



Fonte: Elaborado pelo autor (2024).

Além disso, conforme mostra o Quadro 5, todos os coeficientes médios estão mais próximos de 0 do que de 1, o que indica que, para certos atletas, pode ser necessário atribuí-los a mais de um cluster devido ao baixo valor do coeficiente de silhueta. Dessa forma, qualquer atleta com coeficiente de silhueta inferior a 0,350 (média dos coeficientes) será reatribuído ao cluster e ao cluster vizinho mais próximo.

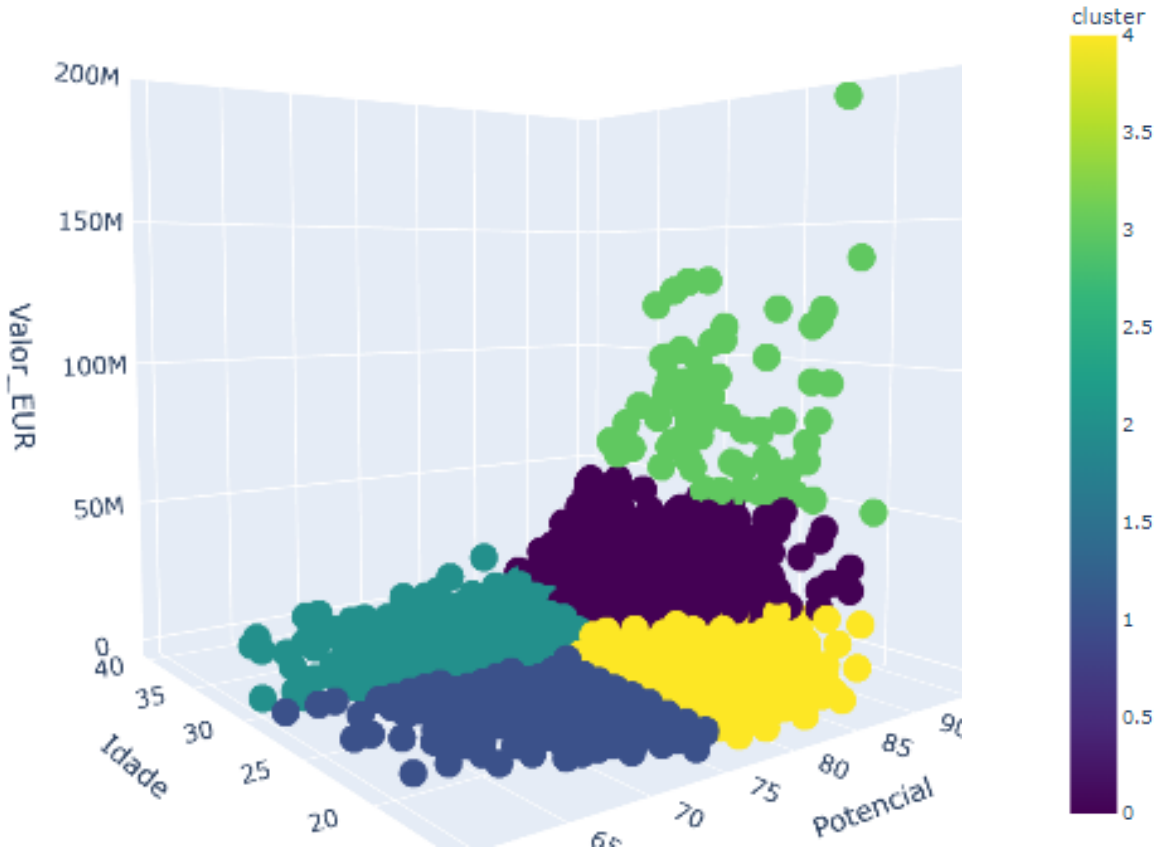
Quadro 5 - Coeficientes médios.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Coeficiente Médio de Silhueta	0.276	0.313	0.372	0.426	0.386

Fonte: Elaborado pelo autor (2024).

Por fim, a distribuição dos clusters está representada na Figura 24, onde nossas variáveis Potencial, Idade e Valor_EUR estão representadas como x, y e z, respectivamente, no gráfico de dispersão tridimensional. Neste momento, é possível ver que os atletas com maiores valores e alto potencial foram atribuídos ao cluster três.

Figura 24 – Gráfico dos clusters.



Fonte: Elaborado pelo autor (2024).

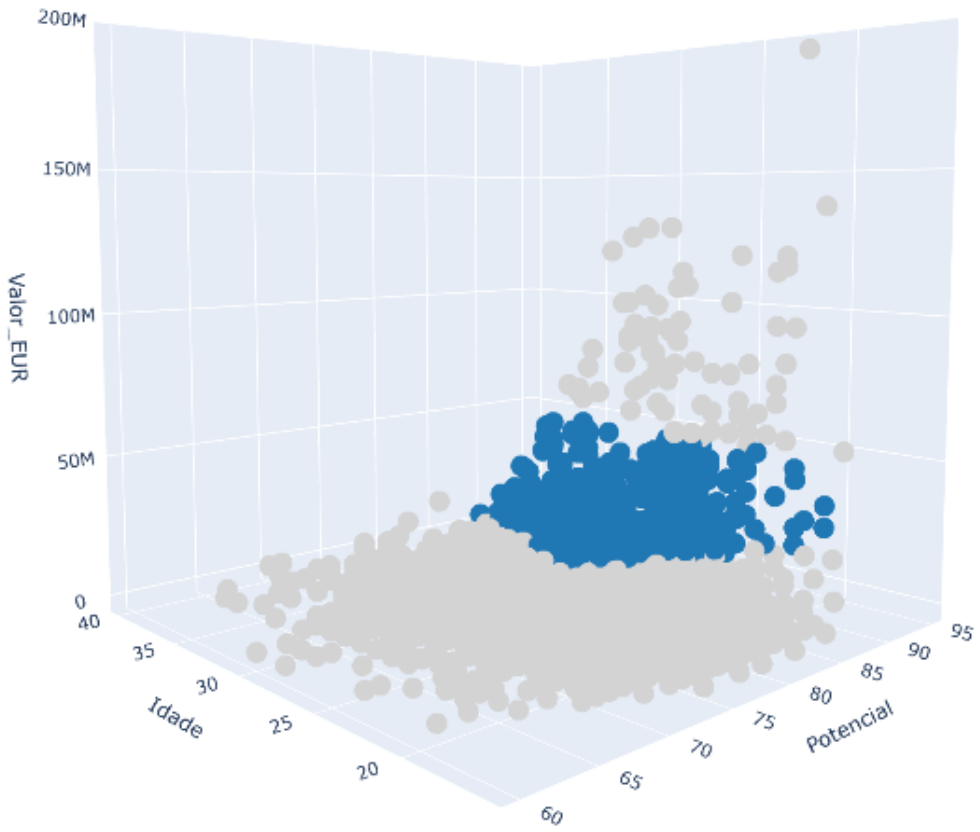
8 Resultados

Após realizar a clusterização, é necessário compreender cada um dos clusters, como suas características, identificar os jogadores que os compõem e avaliar se serão benéficos para o objetivo do projeto. Relembrando, o objetivo central do projeto é auxiliar o jogador na contratação de atletas

8.1 Cluster zero

Este cluster é composto por atletas de idades variadas, com alto potencial e com um valor de mercado mediano. Ele pode ser útil caso o jogador busque um atleta com alto potencial, mas sem a necessidade de investir valores tão altos quanto os exigidos nos próximos clusters, conforme ilustrado na Figura 25.

Figura 25 - Cluster Zero: Gráfico.



Fonte: Elaborado pelo autor (2024).

O Quadro 6 mostra as estatísticas revelam que o cluster contém 493 atletas, com idade entre 18 e 36 anos, com potencial entre 79 e 91 e valor de mercado variando entre 7,5 milhões e 55,5 milhões. Embora esses valores ainda sejam elevados, este cluster oferece boas oportunidades de negócio para aqueles que buscam talento promissor a um custo relativamente acessível.

Quadro 6 - Cluster Zero: Estatísticas.

	Potencial	**Valor_EUR**	**Idade**
Quantidade	**493**	**493**	**493**
Média	**83**	**28026369**	**26**
Desvio Padrão	**2**	**10431996**	**3**
Mínimo	**79**	**7500000**	**18**
25%	**82**	**20000000**	**24**
50%	**83**	**25500000**	**26**
75%	**85**	**35000000**	**29**
Máximo	**91**	**55500000**	**36**

Fonte: Elaborado pelo autor (2024).

Uma possível aplicação deste cluster seria um jogador que busca bons atletas experientes, que já alcançaram seu potencial máximo, mas que não possuem valores exorbitantes. Atualmente, muitos clubes brasileiros utilizam essa abordagem. Alguns exemplos de clubes que contrataram atletas deste cluster incluem: Alex Sandro, recentemente contratado pelo Flamengo na reta final da temporada, e que conquistou a Copa do Brasil 2024, uma das competições mais importantes do cenário nacional; Thiago Silva, recentemente contratado pelo Fluminense, que tem sido peça-chave na

luta contra o rebaixamento para a segunda divisão do Campeonato Brasileiro; e Luiz Suárez, contratado pelo Grêmio em 2023, onde marcou 29 gols e 17 assistências em 54 partidas, contribuindo para o vice-campeonato do Campeonato Brasileiro 2023. No Quadro 7, estão listados outros possíveis jogadores com oportunidades semelhantes.

Quadro 7 - Cluster Zero: Oportunidades.

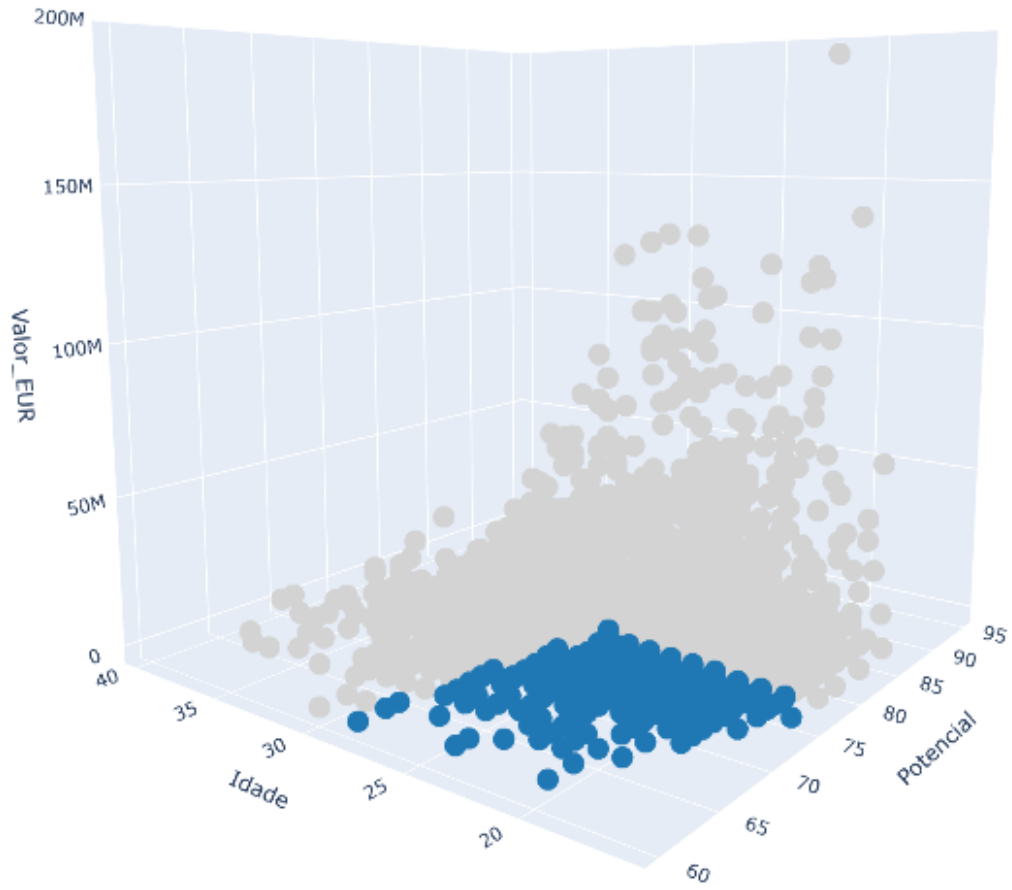
Nome	**Potencial**	**Valor_EUR**	**Idade**
M. Neuer	90	13500000.00	35
Sergio Ramos	88	24000000.00	35
K. Navas	88	15500000.00	34
G. Chiellini	86	12000000.00	36
Thiago Silva	85	9500000.00	36
David Silva	85	22000000.00	35
I. Rakitić	82	18500000.00	33
G. Bale	82	25000000.00	31

Fonte: Elaborado pelo autor (2024).

8.2 Cluster Um

Este cluster é formado por atletas jovens, com baixo potencial e com um valor de mercado reduzido, como mostra a Figura 26. Não parece haver uma aplicação estratégica útil para este cluster, uma vez que esses atletas têm baixo potencial de desenvolvimento. Com o mesmo valor, é possível buscar no mercado jogadores com melhor desempenho e perspectivas de evolução.

Figura 26 - Cluster Um: Gráfico.



Fonte: Elaborado pelo autor (2024).

De acordo com as estatísticas apresentadas no Quadro 8, o cluster contém 612 atletas, com idade entre 16 e 28 anos, potencial entre 60 e 76 e valores entre 899 mil e 6,5 milhões. Dado o baixo potencial e o valor de mercado, optou-se por não avaliar os atletas deste cluster, pois não se mostram vantajosos para o objetivo do projeto.

Quadro 8 - Cluster Um: Estatísticas.

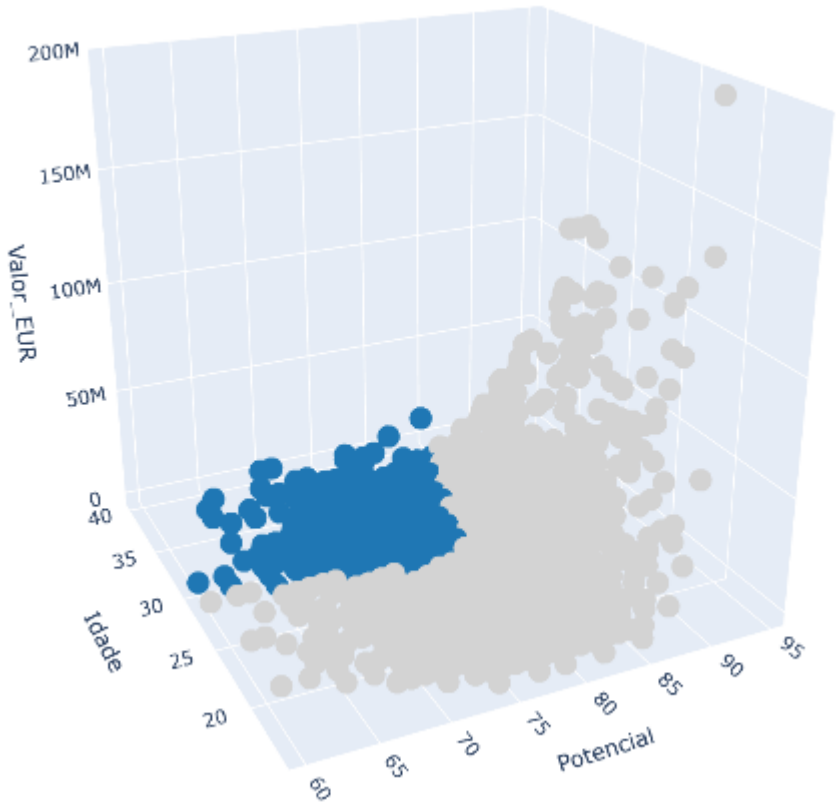
	Potencial	**Valor_EUR**	**Idade**
Quantidade	**612**	**612**	**612**
Média	**73**	**1258064**	**22**
Desvio Padrão	**3**	**899249**	**3**
Mínimo	**60**	**130000**	**16**
25%	**71**	**525000**	**20**
50%	**73**	**950000**	**22**
75%	**75**	**1800000**	**24**
Máximo	**76**	**6500000**	**28**

Fonte: Elaborado pelo autor (2024).

8.3 Cluster Dois

Este cluster é composto por atletas experientes, com potencial mediano e baixo valor de mercado ilustrado pela Figura 27. Ele pode ser útil para jogadores que desejam contratar um atleta acessível, experiente e com um bom nível de desempenho. Embora os atletas deste grupo não sejam tão talentosos quanto os do Cluster Zero, eles oferecem uma opção mais econômica.

Figura 27 - Cluster Dois: Gráfico.



Fonte: Elaborado pelo autor (2024).

A partir das estatísticas apresentadas no Quadro 9, observa-se que este cluster é composto por 889 atletas, com idades entre 26 e 40 anos, potencial variando de 60 a 84, e valor de mercado entre 60 mil e 17,5 milhões.

Quadro 9 - Cluster Dois: Estatísticas

	Potencial	Valor_EUR	Idade
Quantidade	**889**	**889**	**889**
Média	**75**	**5139021**	**30**
Desvio Padrão	**3**	**4013222**	**3**
Mínimo	**60**	**60000**	**26**
25%	**73**	**2000000**	**28**
50%	**75**	**3900000**	**30**
75%	**77**	**7000000**	**32**
Máximo	**84**	**17500000**	**40**

Fonte: Elaborado pelo autor (2024).

Assim como o Cluster Zero, os clubes brasileiros também utilizam amplamente a estratégia de buscar jogadores experientes com preços reduzidos. Um exemplo dessa abordagem é a contratação de Fernandinho pelo Athletico Paranaense em 2022, e de Arturo Vidal pelo Flamengo no mesmo ano. No entanto, diferentemente dos atletas do Cluster Zero, esses jogadores não tiveram participações tão expressivas. O Quadro 10 apresenta algumas oportunidades de atletas pertencentes a este cluster.

Quadro 10 - Cluster Dois: Oportunidades.

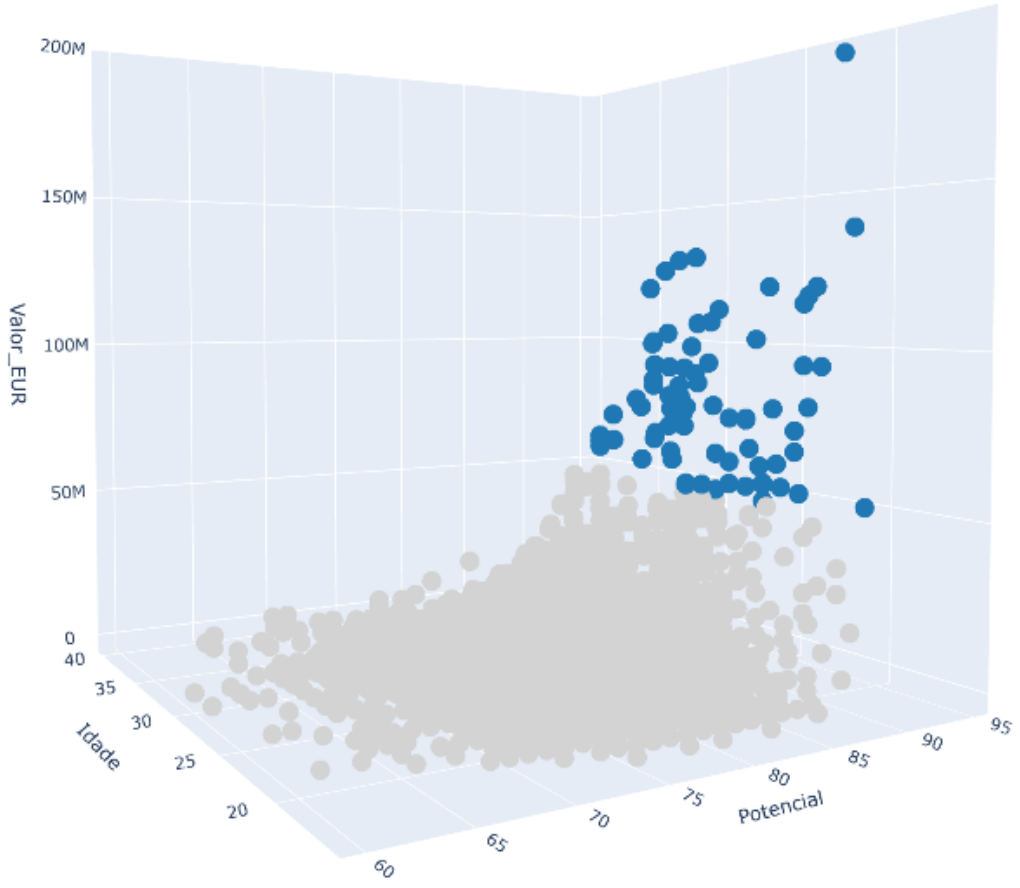
Nome	Potencial	Valor_EUR	Idade
Raúl Albiol	82	8000000	35
Ł. Fabiański	82	3400000	36
S, Sirigu	82	5000000	34
A, Consigli	81	4400000	34
S. Mandanda	81	2900000	36
José Fonte	81	4600000	37
Nacho Monreal	80	6500000	35
D. Godín	80	5500000	35
Guaita	80	3600000	34

Fonte: Elaborado pelo autor (2024).

8.4 Cluster Tres

Este cluster é composto pelos atletas mais caros da base de dados, representando os principais e melhores atletas disponíveis conforme a Figura 28. Ele pode ser útil para jogadores que tenham um orçamento considerável para investir em um atleta de alto nível. A faixa etária dos atletas é variada, porém os mais jovens são promessas avaliadas em milhões de dólares, o que torna difícil para clubes de médio porte contratá-los.

Figura 28 - Cluster Três: Gráfico.



Fonte: Elaborado pelo autor (2024).

Através das estatísticas apresentadas no Quadro 11, pode-se observar que a quantidade de atletas diminuiu consideravelmente em comparação aos clusters anteriores, com apenas 79 atletas. Esses atletas têm idades entre 18 e 34 anos, potencial variando de 86 a 95, e valores de mercado entre 53 milhões e 194 milhões.

Quadro 11 - Cluster Três: Estatísticas.

	Potencial	Valor_EUR	Idade
Quantidade	**79**	**79**	**79**
Média	**89**	**83917722**	**26**
Desvio Padrão	**2**	**24579989**	**3**
Mínimo	**86**	**53000000**	**18**
25%	**88**	**65750000**	**23**
50%	**89**	**79500000**	**26**
75%	**91**	**96750000**	**28**
Máximo	**95**	**194000000**	**34**

Fonte: Elaborado pelo autor (2024).

Neste cluster, encontram-se os jogadores no auge de suas carreiras, as principais celebridades do futebol, atletas vencedores que, mesmo jovens, já proporcionam atuações espetaculares para suas equipes. Entre eles está Lionel Messi, amplamente considerado por muitos como o melhor jogador do esporte. Messi, campeão mundial em 2022, é também o vencedor de oito prêmios de Melhor Jogador do Mundo, concedidos pela revista *France Football*. O Quadro 12 fornece os principais atletas deste cluster.

Quadro 12 - Cluster Três: Principais Jogadores.

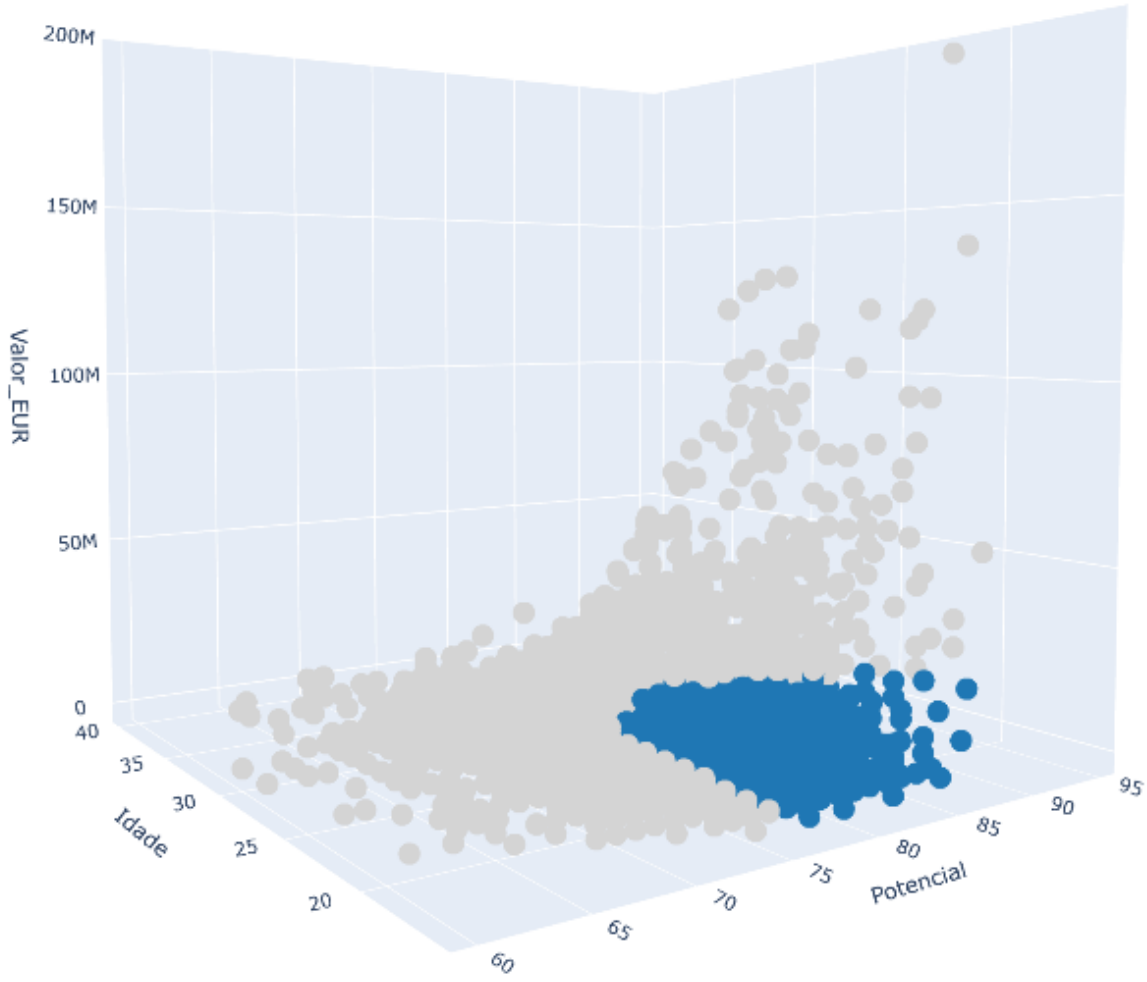
Nome	Potencial	Valor_EUR	Idade
K. Mbappé	95	194000000	22
L. Messi	93	78000000	34
J. Oblak	93	112000000	28
G. Donnarumma	93	119500000	22
E. Haaland	93	137500000	20
R. Lewandowski	92	119500000	32
M. ter Stegen	92	99000000	29
F. de Jong	92	119500000	24
T. Alexander-Arnold	92	114000000	22
K. Havertz	92	94500000	22

Fonte: Elaborado pelo autor (2024).

8.5 Cluster Quatro.

Talvez este seja o melhor cluster em termos financeiros, pois reúne jogadores jovens com valor de mercado médio e excelente potencial. Isso significa que ele representa um ótimo investimento, pois ao adquirir o atleta por um valor relativamente baixo, há a perspectiva de que ele evolua significativamente ao longo do tempo. Os atletas são indicados pelas bolinhas azuis na Figura 29.

Figura 29 - Cluster Quatro: Gráfico.



Fonte: Elaborado pelo autor (2024).

Por meio das estatísticas do Quadro 13, é possível ver que o cluster possui 903 atletas, com idades entre 16 e 26 anos, potencial variando de 77 a 90, e valores de mercado entre 2,75 milhões e 22 milhões.

Quadro 13 - Cluster Quatro: Estatísticas

	Potencial	Valor_EUR	Idade
Quantidade	**903**	**903**	**903**
Média	**80**	**5490282**	**21**
Desvio Padrão	**2**	**4377109**	**2**
Mínimo	**77**	**275000**	**16**
25%	**78**	**2100000**	**19**
50%	**80**	**4000000**	**21**
75%	**82**	**8000000**	**23**
Máximo	**90**	**22000000**	**26**

Fonte: Elaborado pelo autor (2024).

Neste cluster, encontramos jogadores com grande potencial de valorização por um baixo preço. Embora possam demorar para se consolidar, representam ótimas oportunidades de negócio para equipes em busca de um elenco vencedor.

Alguns exemplos incluem Jamal Musiala, de 18 anos em 2022, cujo valor no EA FC 24 é de 76,5 milhões, com uma valorização de 5,89 vezes em dois anos. Gavi, jogador espanhol, teve uma valorização de 17,85 vezes, passando de 2,1 milhões para 37,5 milhões em dois anos. Além deles, há muitos outros atletas com grande potencial para investimento que podem ser vistos através do Quadro 14.

Quadro 14 - Cluster Quatro: Principais Jogadores

Nome	Potencial	Valor_EUR	Idade
Ansu Fati	90	17500000	18
G. Raspadori	88	10500000	21
Gabriel Martinelli	88	18000000	20
T. Kubo	88	13500000	20
J. Musiala	88	13000000	18
R. Cherki	88	7000000	17
J. Gvardiol	87	12500000	19
C. Tzolis	87	10000000	19
N. Rovella	87	4100000	19
G. Reyna	87	22000000	18
H. Elliott	87	7000000	18
Kayky	87	2700000	18

Fonte: Elaborado pelo autor (2024).

9 Conclusão

O presente trabalho teve como objetivo principal agrupar atletas em diferentes grupos, auxiliando jogadores em situações estratégicas de contratação, com o intuito de promover o desenvolvimento e a aplicação de técnicas estatísticas no contexto dos jogos eletrônicos. Através da utilização de técnicas de clusterização, foi possível obter resultados expressivos, como a identificação de cinco grupos distintos de atletas, cada um apresentando características específicas que os tornam mais adequados para diferentes estratégias.

Os resultados demonstraram que, em alguns clusters, foram identificados atletas com ótimas oportunidades de compra, representando um diferencial significativo para a construção de uma equipe vencedora. Esses achados evidenciam que a aplicação de técnicas de clusterização pode gerar conhecimentos valiosos sobre o desempenho dos atletas, facilitando a tomada de decisões estratégicas por parte dos jogadores. Dessa forma, o trabalho reforça a importância da utilização de métodos estatísticos para extrair valor de dados em ambientes ainda pouco explorados, como os eSports.

Como sugestão para novos projetos, recomenda-se explorar mais a fundo as características específicas de cada atleta de acordo com sua posição, uma vez que a clusterização baseada em posições pode aprimorar a identificação de talentos com habilidades especiais que se alinhem às necessidades táticas de uma equipe. Além disso, o uso de dados provenientes de partidas reais, apesar do desafio de mensurar o potencial dos atletas de forma objetiva, pode contribuir para a criação de métricas que sejam aplicáveis em cenários reais, proporcionando benefícios práticos para equipes que buscam melhorar seus processos de recrutamento.

Em suma, o processo de desenvolver e implementar métodos de clusterização para análise de jogadores não só proporcionou uma compreensão mais profunda sobre a aplicação prática de métodos estatísticos, como também destacou a importância de um trabalho contínuo para refinar essas abordagens e torná-las ainda mais precisas e úteis para aplicação em contextos reais.

Referências

DINIZ, C. A. R; LOUZADA NETO, F. **Data mining:** uma introdução. São Paulo: Associação Brasileira de Estatística, 2000.

DONI, M. V. **Análise de cluster:** métodos hierárquicos e de particionamento. São Paulo, 2004.

GEROLAMO, Eduardo Henrique. Rumo a glória: Técnicas multivariadas para formação de equipe no game FIFA. Trabalho de conclusão de curso (Bacharelado em Estatística) - Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências e Tecnologia, Presidente Prudente, 2020. Orientador: Professor Dr. Klaus Schlunzen Junior.

HALDIKI, M. BATISTAKIS, Y. VAZIRGIANNIS, M. **On Clustering Validation Techniques.** Journal of Intelligent Information Systems, 107–145, 2001.

HAN, J. KAMBER, M. PEI, J. **Data Mining Concepts and Techniques.** Elsevier Inc**.** 2012.

HUNTER, J. K. An Introduction to Real Analysis. Disponível em: https://www.math.ucdavis.edu/~hunter/intro_analysis pdf/intro_analysis.pdf. Acesso em: 15/07/2024.

JOHNSON, RICHARD A.; WICHERN, DEAN W. **Applied multivariate statistical analysis.** New Jersey: Prentice Hall, ed. 6, 2007.

LIMA, Raphael Rodrigues. **Análise estatística e utilização de Machine Learning para previsão salarial de futebolistas do jogo FIFA. 2022**. Trabalho de conclusão de curso (Bacharelado em Estatística) - Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências e Tecnologia, Presidente Prudente, 2022. Orientadora: Miriam Rodrigues Silvestre.

MULLER, C. GUIDO, S. **Introduction to Machine Learning with Python: A GUIDE**

FOR DATA SCIENTISTS. O'Reilly Media, Inc. 2016

Orientações para Profissionais da Rede Local. Disponível em: <https://www.cevs.rs.gov.br/suicidio >. Acesso em: 31 Out. 2023.

PALMA, L. FARIAS. **AGRUPAMENTO DE DADOS: K- MÉDIAS**, Trabalho de conclusão de curso (Bacharel em Ciências Exatas e Tecnológicas) - Universidade Federal do Recôncavo da Bahia, Centro de Ciências Exatas e Tecnológicas. Cruz das Almas, 2018. Orientadora: Profa. Dra. Julianna Pinele Santos Porto.

SHIELDS, A. **Revealed:** What Really Happens To Your Body And Brain When You Play FIFA. Cassino.org**.** 2021.
Disponível em: <https://www.casino.org/blog/real-effects-of-playing-fifa/>.
Acesso em: 31 Out. 2023.

YOUGUO LI, HAIYAN WU. **A Clustering Method Based on K-Means Algorithm.** Physics Procedia, p.1104-1109, 2012.

ZHOU, Zhi-Hua; LIU, Shaowu. **Machine Learning.** Tradução da edição em inglês. Singapura: Springer Nature Singapore Pte Ltd., 2021. Publicado originalmente pela Tsinghua University Press, 2016.