CEUB | PIC

O USO DE MACHINE LEARNING PARA IDENTIFICAÇÃO PREVENTIVA DE LESÕES EM JOGADORES DE VÔLEI

Professor orientador: Fabio Oliveira Guimaraes

Aluno: Matheus Pereira Gomes Moraes

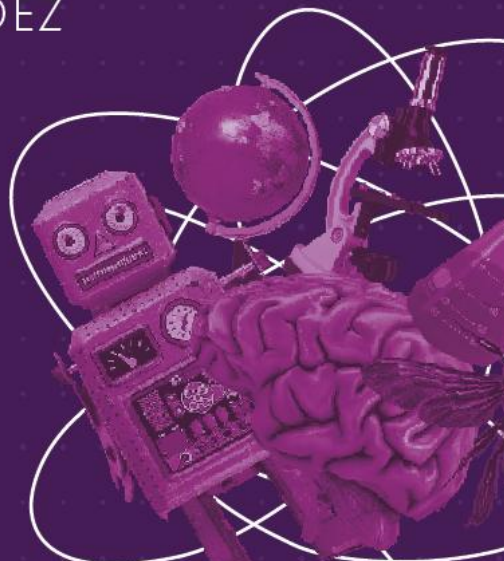
ISSN: 2595-4563 | **EDIÇÃO 2024** | VOLUME 10

PROGRAMA DE INICIAÇÃO CIENTÍFICA PIC/CEUB

RELATÓRIOS DE PESQUISA

VOLUME 10 Nº 1- JAN/DEZ

2024



fapdf CNPq CEUB

CENTRO UNIVERSITÁRIO DE BRASÍLIA - CEUB
PROGRAMA DE INICIAÇÃO CIENTÍFICA

MATHEUS PEREIRA GOMES MORAES

O USO DE MACHINE LEARNING PARA IDENTIFICAÇÃO PREVENTIVA DE LESÕES EM JOGADORES DE VÔLEI

Relatório final de pesquisa de Iniciação Científica apresentado à Assessoria de Pesquisa e Extensão.
Orientação: Fabio Oliveira Guimaraes

BRASÍLIA
2025

RESUMO

O voleibol é bastante praticado e seus atletas podem sofrer com lesões devido a natureza do esporte, seja em saltos e aterrisagens, contato com a bola ou até mesmo com outros jogadores. O presente estudo então buscou realizar uma análise exploratória das contusões sofridas por jogadores e aplicar um algoritmo de machine learning não supervisionado para clusterização dos registros a fim de identificar padrões e servir de insight para estratégias de prevenção. Os dados foram obtidos por meio do Sistema Nacional de Vigilância Eletrônica de Lesões (NEISS), um sistema de observação de lesões exercido pela Comissão de Segurança de Produtos de Consumo dos EUA (CPSC). Foram selecionados dados de 2004 a 2023, contemplando 28.201 casos, que reduziram a 17.332 após filtragens de idade e relevância das categorias de cada coluna. Foram selecionadas as colunas de idade, gênero, parte do corpo afetada, diagnóstico e narrativa (que descreve em poucas palavras a lesão). A variável idade foi filtrada para as idades entre 10 e 35 anos e padronizada utilizando o método *StandardScaler* do pacote scikit-learn. Já as variáveis categóricas foram convertidas para um formato numérico, aplicando a técnica de *one-hot enconding.* Por fim na coluna de narrativa foi aplicada a *Term Frequency–Inverse Document Frequency* (TF-IDF), uma técnica de processamento de linguagem natural (NLP), bem como a extração dos verbos mais presentes, que após análise de relevância para o contexto da pesquisa, foram transformados em colunas. A partir disso, com base nos métodos do cotovelo e de silhouette o número de clusters foi definido em k=5 e após foi aplicado o algoritmo KMeans. Uma função de similiaridade do cosseno foi aplicada entre os registros de um mesmo cluster na coluna narrativa para avaliar a consistência dos casos intraclusters. Os clusters foram analisados a partir de sua distribuição de variáveis, identificando padrões que podem servir para que estratégias de treino e fortalecimento de partes do corpo possam ser traçadas. Foi possível reconhecer padrões como por exemplo entorses nos dedos de meninas adolescentes em ações de torção por compressão e impacto/colisão, o que pode indicar que foram em ações do bloqueio ou ataque, comuns à prática do voleibol. A base de dados, por ser mais generalista carece de informações mais específica que poderiam trazer ainda mais informações aos clusters como por exemplo peso, altura, posição em que o atleta atua e se o fato ocorreu em jogo ou treino.

Palavras-chave: clustering; kmeans; voleibo; lesão esportiva

SUMÁRIO

1 Introdução

1.1 Contextualização da pesquisa

O voleibol é um esporte bastante popular no Brasil e no mundo. Em 2022 era esperado que a Federação Internacional de Vôlei (FIVB) tivesse 222 federações nacionais filiadas, com 800 milhões de participantes ao redor do mundo YOUNG (2023). A última edição da Superliga Feminina de Vôlei, principal competição nacional, foi a mais assistida desde a temporada 20/21, alcançando 8,7 milhões de pessoas (WEBVÔLEI, 2024). Segundo VERHAGEN et al. (2004) mesmo com alto número de praticantes profissionais e amadores a incidência de lesões pode ser considerada baixa com relação a outros esportes, tendo em vista por exemplo a separação dos times em lados opostos da quadra, separados pela rede. Ainda assim há o risco de lesões, que são causadas geralmente por saltos e voltas ao solo, bem como bloqueios e ataques (BERE et al., 2015).

O voleibol é um esporte que tem como característica um alto impacto com o solo. Segundo MESQUITA et al. (2008), há uma grande exigência muscular e da ação das forças gravitacionais agindo sobre o corpo em movimentos de parada, salto e bloqueio que são performados constantemente durante uma partida.

A partir disso, é possível notar, devido ao impacto e exigência muscular, algumas lesões no vôlei, sendo as mais comuns as que ocorrem no ombro, joelho e tornozelo de acordo com KILIC et al. (2017). Ao terem suas lesões, os atletas têm um tempo de recuperação diferente de acordo com o local do corpo onde ocorreu e com a gravidade. Alguns casos demoram dias enquanto outros podem durar meses.

Ao se afastarem das quadras o ritmo de jogo muitas vezes é perdido, sessões de fisioterapia e tratamento para recuperação são necessários e a volta às partidas se dá gradativamente. É importante então identificar causas e buscar por soluções para prevenir e minimizar tais desgastes e ocorrências a fim de ter um atleta saudável e que possa desempenhar suas funções durante as temporadas de jogos.

Um atleta é uma peça fundamental no planejamento de um clube, tanto pelo sistema tático para qual foi contratado quanto pelo orçamento disponível. Um time com seus jogadores à disposição, sem lesões é um time competitivo e que consegue desempenhar melhor seu potencial. Ao perder um atleta seja por dias ou meses, o clube precisa se adaptar e em alguns

casos até mesmo contratar mais alguém para compor o time, tendo então desfalques táticos e orçamentários.

Ao analisar a literatura em respeito às lesões no voleibol é possível identificar uma revisão sistemática que procura estabelecer as principais e mais comuns. Lesões no ombro, joelho e tornozelo figuram entre as mais comuns (KILIC et al., 2017). É importante notar tais características a fim de ter profissionais mais preparados para prevenir e tratar melhor tais danos aos jogadores.

A presente pesquisa então busca fazer uma análise exploratória de lesões para identificar se há algum padrão na ocorrência de contusões em jogadores, bem como a partir de algoritmos de *machine learning* não supervisionado analisar possíveis correlações entre fatores tais como idade, sexo, parte do corpo lesionada e o tipo de lesão.

1.2 Objetivos

Levando em consideração a saúde e um melhor desempenho dos atletas que praticam o voleibol, o presente projeto almeja investigar lesões sofridas pelos jogadores. E perceber, a partir de algoritmos de *machine learning* não supervisionado (K-means), características que possam estar diretamente ligadas às contusões sofridas, investigando correlações e buscando identificar cluster que tragam informações importantes para o auxiliona prevenção de contusões.

1.2.1 Gerais:

Contribuição para a prevenção de lesões na prática do voleibol utilizando modelos não supervisionados de *machine learning*.

1.2.2 Específicos:

Análises descritivas explorando as relações entre fatores de risco e contusões em atletas de vôlei.

Treinamento de um algoritmo de machine learning não supervisionados para identificação de padrões de lesões em atletas de voleibol.

2 Fundamentação teórica

A investigação de fatores de risco para prevenção de lesões em esportes é um assunto que permeia a literatura e está em voga atualmente segundo VAN EETVELDE et al. (2021), mesmo que haja certa complexidade em realizar modelos preditivos devido aos inúmeros fatores e situações que permeiam as interações do atleta com o esporte. Outro fator de dificuldade, indicado por AMENDOLARA (2023), reside no fato da falta de uniformidade nos dados esportivos, o que frustra novos testes e validação de tais dados.

Para mitigar efeitos de desconfortos e buscando uma maior qualidade de vida para os atletas, estudos que tentam desenvolver algoritmos de *machine learning* estão aumentando. Apesar da dificuldade em adaptar tais modelos para situações completas e cercadas de fatores que são as lesões em esportes, estudos como o de OLIVER (2020), em que jogadores de futebol foram analisados, algumas correlações, ainda que baixas, entre características individuais e risco de lesões conseguem ser identificadas. Isso é importante para o presente momento, de forma tal a ter um maior entendimento do risco, bem como para futuras pesquisas, instigando caminhos pelos quais as pesquisas devem se guiar. É mencionado por YOUNG (2023) que programas de prevenção de lesões, especialmente as que ocorrem no tornozelo, mostram grandes benefícios, o que aumenta a necessidade de se ter mais previsões para que cuidados prévios sejam tomados.

No voleibol é bastante comum a análise de lesões mais frequentes. Na revisão da literatura realizada por GUERZONI (2022), há muitos artigos publicados e analisados a fim de identificar contusões mais comuns em atletas profissionais e amadores, sendo mais comuns as que ocorrem no tornozelo. O estudo feito por SANTOS (2021), também realiza uma revisão da literatura no mesmo intuito e encontra resultados similares, além de apurar posições mais vulneráveis e prevalência durante treinos ou jogos.

Além de demonstrar apenas o que ocorre com mais frequência, é interessante notar também uma análise exploratória, de maneira a identificar se há alguma tendência ao longo dos anos e servir como base para estudos que venham a surgir futuramente.

Desta forma, aliando os estudos preditivos já feitos em outros esportes, é interessante começar a direcionar também o olhar para o voleibol mais especificamente, identificando suas características e especificidades. De acordo com VAN EETVELDE et al. (2021) os modelos preditivos mais utilizados atualmente na área são os baseados em árvores. Segundo KERN (2019), tais modelos são capazes de se adaptar a inter-relações complexas não lineares, o que é

beneficia assim as predições de contusões tendo em vista que se adaptam aos cenários do mundo real, que podem não ser explicados por modelos lineares tradicionais.

Entretanto, a clusterização como método de prevenção a lesão e identificação de padrões também vem ganhando espaço, segundo ZHAO; LI. (2023) a combinação de redes neurais profundas com clusterização funciona bem para predizer se atletas possuem risco de lesão. SUAREZ-DEL FUEYO et al. (2020) também se utiliza de clusters para risco de lesões, mas com acidentes de trânsito, mostrando a versatilidade e aplicabilidade do algoritmo.

3 Método

3.1 Obtenção dos dados

Para a realização da pesquisa, foram analisados dados dos últimos 10 anos de lesões relacionadas ao voleibol, a partir do Sistema Nacional de Vigilância Eletrônica de Lesões (NEISS), um sistema de observação de lesões exercido pela Comissão de Segurança de Produtos de Consumo dos EUA (CPSC). O NEISS acompanha lesões tratadas em departamentos de emergência selecionados em todo os Estados Unidos a partir de uma amostra de aproximadamente 100 hospitais com serviços de emergência 24 horas (CPSC, 2024).

Primeiramente foi feita a coleta dos dados a partir do site do NEISS, em que o download é feito individualmente para cada ano por meio de planilhas de Excel. Foram selecionados então os últimos 20 anos disponíveis, extraindo dados de 2004 a 2023.

3.2 Tratamento e filtragem dos dados

A partir disso todas as planilhas foram unidas em uma só e posteriormente filtradas a partir da coluna *Product_1* (Produto_1), que identifica a causa da lesão, para essa pesquisa foram filtradas as que continham o valor 1266, que se refere ao voleibol. Desta forma foram encontrados 28.201 registros de lesões com características como idade, gênero, parte do corpo afetada e descrição da fratura.

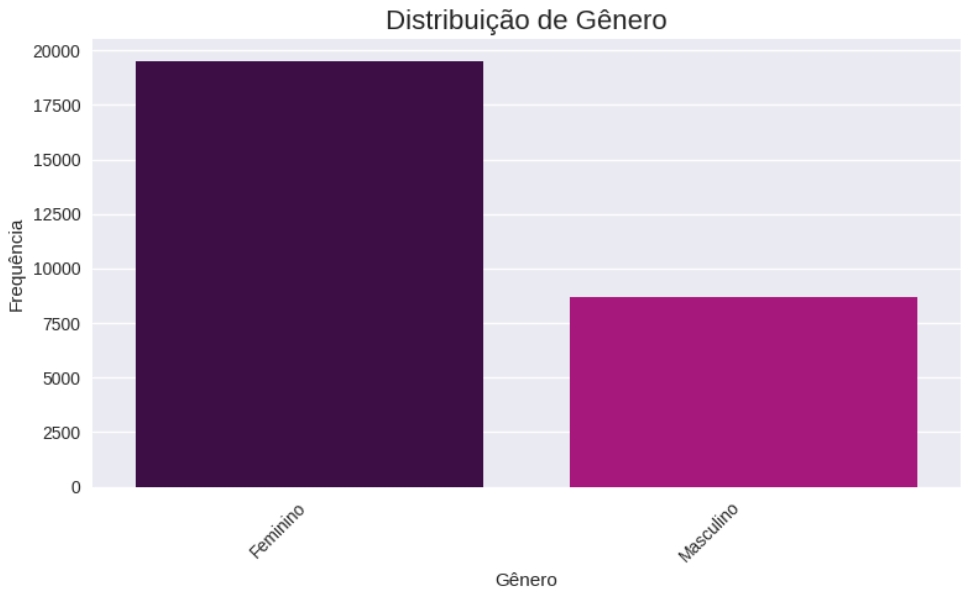
Ao estudar a base de dados, muitas colunas estavam presentes, porém foram consideradas aquelas que tinham relevância para o estudo, sendo elas *Age* (Idade), *Gender* (Gênero), *Body_Part* (Parte do Corpo), *Diagnosis* (Diagnóstico) e *Narrative* (Narrativa), esta última descreve em poucas palavras o que ocorreu no momento da lesão.

Os gráficos a seguir apresentam a distribuição dos dados antes dos tratamentos e seleções feitas.

O gráfico abaixo demonstra a distribuição de gênero entre as lesões filtradas, é interessante ver como o gênero feminino representa 69,3% dos indivíduos, com 19.532 casos, para a pesquisa foram considerados apenas os gêneros masculino e feminino, pois as demais categorias presentes na base de dados se resumiam em apenas 3 casos, o que foi considerado extremamente baixo para prover alguma informação relevante. A presença maior de indivíduos do gênero feminino nos registros de lesão pode ser explicada pela popularidade do esporte nos Estados Unidos, país da pesquisa, pois a participação feminina no vôlei no ensino médio

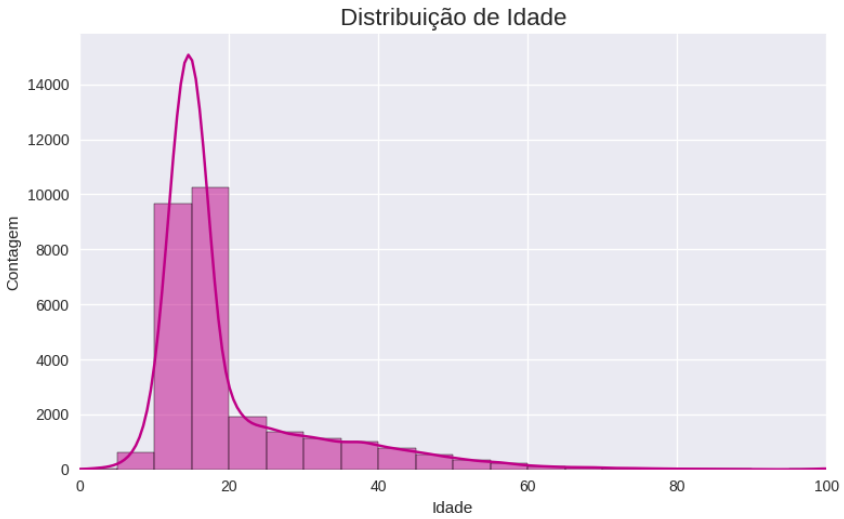
(479.125 atletas) supera em 4,8 vezes a masculina, sendo o segundo esporte mais praticado por mulheres (NATIONAL FEDERATION OF STATE HIGH SCHOOL ASSOCIATIONS, 2023).

Gráfico 1 – Distribuição de gênero



Fonte: Elaborado pelo autor

Gráfico 2 – Distribuição de Idade

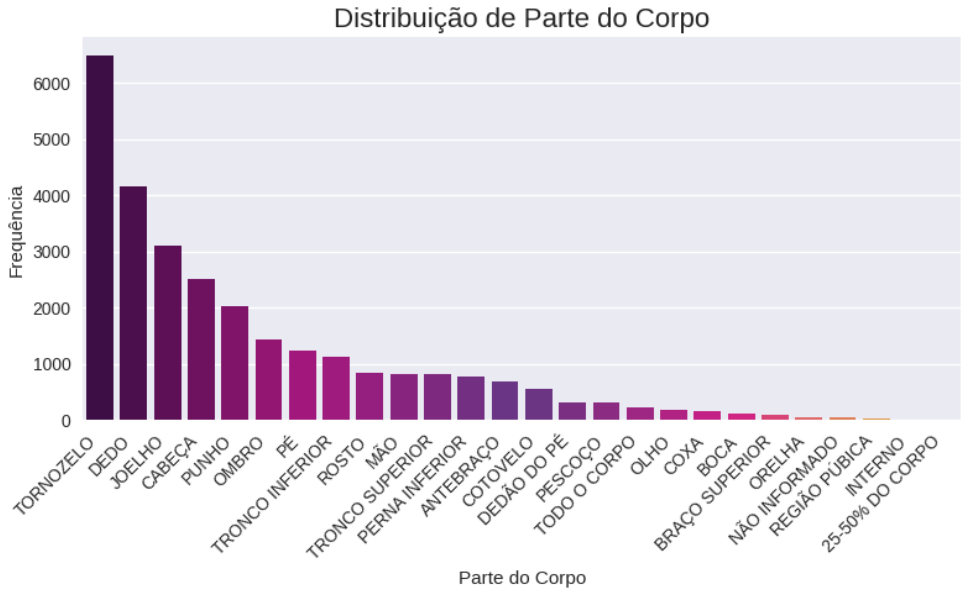


Fonte: Elaborado pelo autor

Já o segundo gráfico apresenta a distribuição de idade entre os lesionados, sendo que boa parte destes tem entre 10 e 20 anos. Estas idades concentram quase 70% do total, com 19.915 casos. Segundo Wylleman e Lavallee (2004) o jogador de vôlei tem sua fase da carreira esportiva iniciada aos 10 anos de idade e aposentadoria aos 35. Alguns atletas podem fugir a esse padrão, principalmente no que diz respeito a aposentadoria, em que temos exemplos de

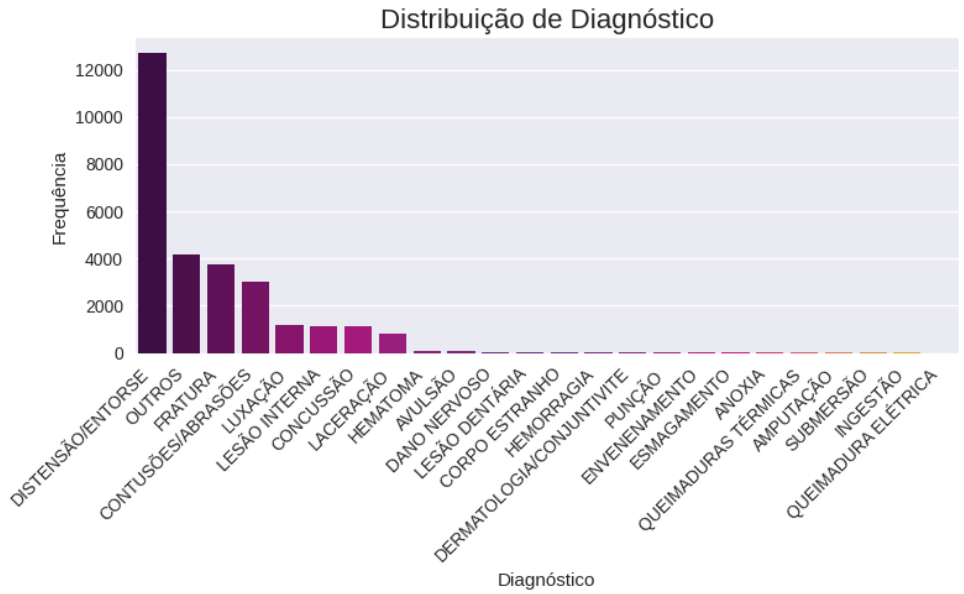
jogadores com mais de 40 anos jogando em alto nível. Porém, a fim de delimitar a pesquisa em uma idade que os jogadores estejam em plena forma física e baseados no estudo feito por Wylleman e Lavallee foram considerados aqueles que se lesionaram com a idade entre 10 e 35 anos.

Gráfico 3 – Distribuição de partes do c



Fonte: Elaborado pelo autor

Gráfico 4 – Distribuição de diagnóstico



Fonte: Elaborado pelo autor

Ao analisar variáveis como parte do corpo e diagnóstico de lesão, muitas categorias foram encontradas, sendo assim foram consideradas relevantes apenas as que apresentaram mais que 5% de representatividade na base de dados. Essa seleção é importante tanto para melhorar a qualidade dos algoritmos empregados, como para filtrar casos que façam mais sentido no contexto do vôlei e retirar casos incomuns que possam trazer ruídos para a análise. Após as seleções feitas restaram 17.332 registros.

As variáveis categóricas (*Gender*, *Body_Part* e *Diagnosis*) foram convertidas para um formato numérico, aplicando a técnica de *one-hot enconding*, que é importante para aplicação de algoritmos de *machine learning* que tenham como requisito entradas numéricas ao mesmo temo que mantém a condição não-ordinal das categorias presentes. O parâmetro *drop = "first"* foi empregado a fim de eliminar a primeira categoria de cada variável, evitando assim a multicolinearidade, desta maneira a coluna *Gender* como exemplo teria apenas uma coluna em que a presença do número 1 representa o gênero masculino e a presença do número 0 representa o gênero feminino, não sendo necessária uma nova coluna. Cada variável se torna então uma coluna binária, com 1 nos casos em que se correspondem a sua categoria e 0 nas demais.

No intuito de garantir comparabilidade entre as variáveis e adequação aos algoritmos de *machine learning*, foi feita a padronização da variável idade utilizando o método *StandardScaler* do pacote scikit-learn. Tal transformação converte os valores originais para uma distribuição com média zero e desvio padrão igual a um, seguindo a fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Onde:

x = valor original da idade

μ = média da distribuição original

σ = desvio padrão da distribuição original

A padronização foi aplicada apenas na coluna *Age* (Idade), pois foi a única variável contínua identificada no conjunto de dados que requeria normalização. Os valores padronizados foram então convertidos para o formato de matriz esparsa (csr_matrix) utilizando a biblioteca SciPy, permitindo armazenamento eficiente em memória, concatenação direta com as variáveis categóricas previamente codificadas via *one-hot encoding* e compatibilidade com algoritmos de clusterização que operam sobre matrizes esparsas.

3.3 Processamento de linguagem natural (NLP)

Para tratar a coluna *Narrative* (Narrativa) e extrair informações das descrições das lesões foi aplicada a *Term Frequency–Inverse Document Frequency* (TF-IDF), uma técnica de processamento de linguagem natural (NLP) que transforma textos em vetores de números. Foram considerados unigramas e bigramas (1 a 2 palavras) e as 300 expressões mais representativas foram escolhidas. As *stopwords*, palavras bastante comuns e que não agregam para o algoritmo, foram removidas. Assim, uma matriz esparsa foi obtida como resultado representando a importância relativa de cada termo para cada caso, utilizada como entrada para o modelo de *clusterização*.

A seguir foi feita uma extração dos verbos mais comuns presentes na coluna de narrativas, a fim de encontrar ações realizadas durante a lesão. Para tal, uma função foi feita para extrair os verbos mais comuns, retirando-se previamente aqueles genéricos como *be* (ser/estar), *have* (ter), *do* (fazer), *say* (dizer), *make* (fazer), *play* (jogar). Com os verbos mais comuns extraídos da função, uma análise manual foi feita a fim de identificar aqueles que tem relevância no contexto da pesquisa. Os verbos encontrados foram fall (cair), hit (impacto), land (aterrissar/pousar), jump (pular), twist (torção rotacional), roll (rolar), strain (distender), strike (golpear), come down (cair), run (correr), jam (torção por compressão), step (pisar), collide (colidir), trip (tropeçar), spike (cortar no vôlei).

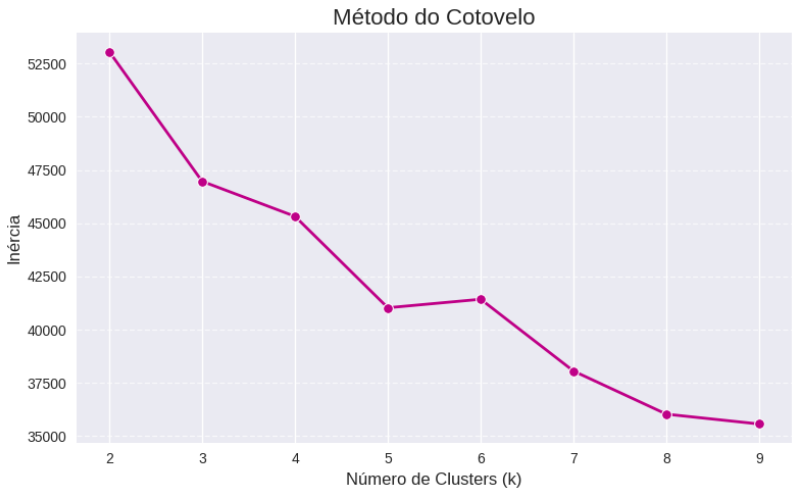
3.4 Aplicação da técnica de PCA

Para a etapa de modelagem, todas as variáveis tratadas acima foram combinadas em uma única matriz de atributos. Após essa união, foi aplicada a técnica de Análise de Componentes Principais (*PCA – Principal Component Analysis*), reduzindo a dimensionalidade do conjunto de dados. O número de componentes foi escolhido de modo a preservar 95% da variância original, garantindo assim a manutenção das informações mais relevantes ao mesmo tempo que se simplifica a estrutura dos dados para os algoritmos de agrupamento.

3.5 Escolha do número de clusters

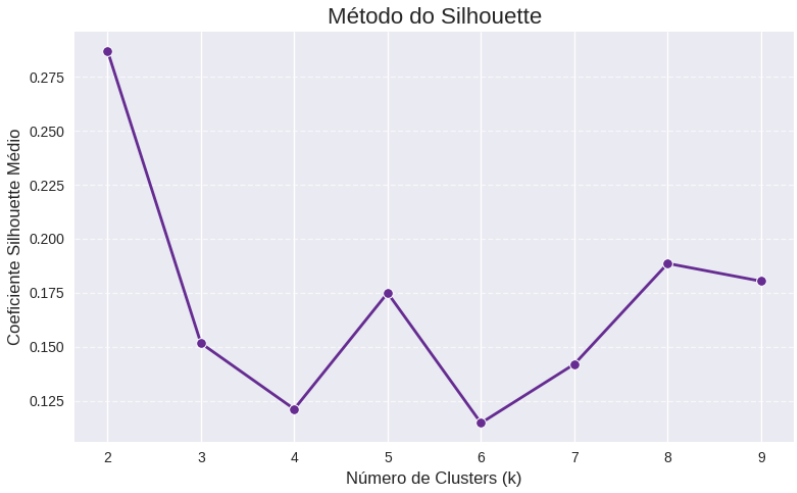
Para selecionar o número ideal de *clusters* foram utilizados tanto o Método do Cotovelo (*Elbow Method*), que avalia o número de agrupamentos e a inércia e identifica o ponto em que o ganho na redução da inércia começa a se estabilizar e o Coeficiente *Silhouette,* que mede a coesão intracluster, identificando a qualidade da separação para cada quantidade *k* de agrupamentos.

Gráfico 5 – Método do cotovelo



Fonte: Elaborado pelo autor

Gráfico 6 – Método silhouette



Fonte: Elaborado pelo autor

O Método do Cotovelo indicou um ponto de mudança em k = 5, insinuando que aumentos acima desse valor resultariam em ganhos marginais na explicação da variância. Por outro lado, o Coeficiente de *Silhouette* apresentou seu valor máximo em k = 2, porém com menor detalhamento das estruturas presentes nos dados.

Considerando a necessidade de maior granularidade para a análise e boa separação entre os agrupamentos, optou-se por k = 5, conciliando a indicação do Cotovelo com valores aceitáveis de Silhouette, pois apenas 2 clusters como sugerido pelo coeficiente de Silhouette seria muito pouco pela quantidade de casos disponíveis.

3.6 Aplicação do algoritmo KMeans

Para agrupamento dos dados em perfis similares, foi utilizado o algoritmo *KMeans*, que é um método de clusterização que divide os dados em grupos com base na proximidade das observações em um espaço de variáveis. O algoritmo atribui cada ponto ao cluster cujo centroide (média dos pontos do grupo) é o mais próximo, iterando até que a posição dos centroides se estabilize. Com a definição do número de clusters igual a 5, tal modelo foi aplicado com *random-state* = 42 para garantir a reprodutibilidade.

3.7 Similaridade intracluster

Por fim, com o objetivo de visualizar a similaridade dos registros de cada cluster foi realizada um cálculo de similaridade cosseno a partir da coluna de narrativas, utilizando a técnica de *Term Frequency–Inverse Document Frequency* (TF-IDF), com as 200 expressões mais representativas. Desta forma será possível perceber se as narrativas de cada cluster são parecidas e possuem características em comum para justificar seu agrupamento.

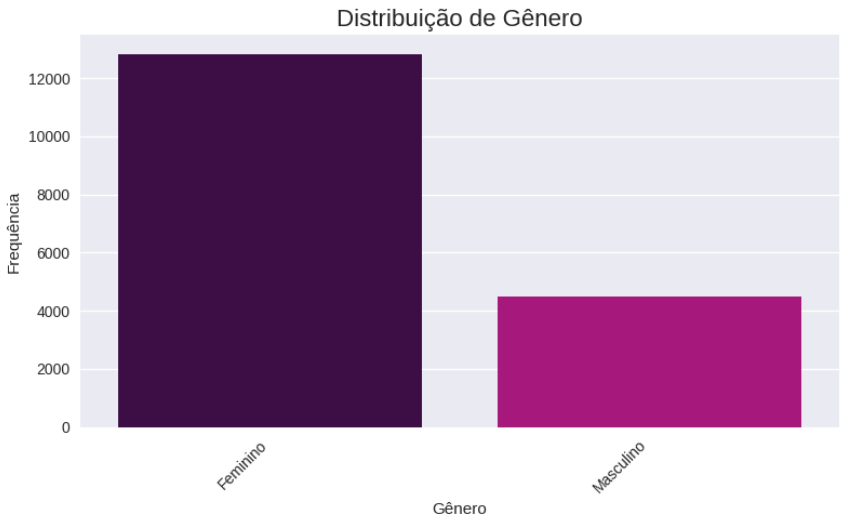
4 Resultados e discussão

4.1 Análise exploratória

A análise exploratória, realizada após a filtragem dos dados, permitiu uma compreensão melhor da distribuição dos dados, revelando previamente alguns padrões nas lesões, antes mesmo do agrupamento em clusters. Essa etapa foi essencial para avaliar a consistência dos dados e orientar as decisões subsequentes na escolha das técnicas de análise.

Nos gráficos apresentados a seguir, é possível observar a distribuição das categorias de cada variável, evidenciando diferenças de frequência entre gênero, faixas etárias, partes do corpo lesionadas e diagnóstico. Essa visualização inicial contribuiu para a formulação de hipóteses e para antecipar possíveis agrupamentos que seriam confirmados ou refutados na etapa de clusterização.

Gráfico 7 – Distribuição de gênero após filtragem dos dados

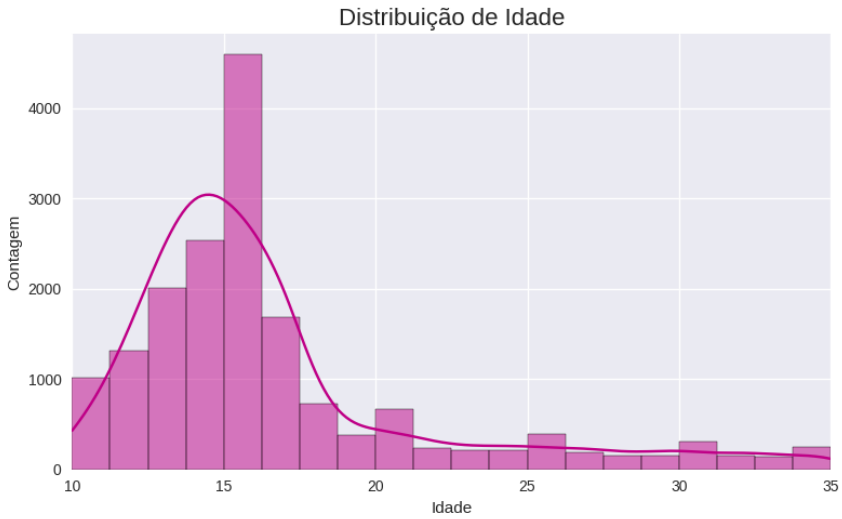


Fonte: Elaborado pelo autor

O gráfico acima demonstra a distribuição de gênero nos casos analisados. É possível perceber que a quantidade de indivíduos do sexo feminino continua maior que a do sexo masculino, inclusive aumentando para 74,1% de representatividade com 12.838 registros. Isso pode indicar que uma maior quantidade de mulheres também estará presente nos clusters. Tal fator pode ser entendido a partir da popularidade do esporte entre as mulheres no país de origem dos dados, pois conforme visto anteriormente na seção de metodologia, os dados da National Federation of State High School Associations (2023) mostram que a participação feminina no vôlei entre os alunos de ensino médio é significativamente maior que a masculina.

Na última pesquisa realizada a quantidade de atletas mulheres era de 479.125, sendo o segundo esporte mais praticado por elas, enquanto os atletas do gênero masculino eram apenas 85.255.

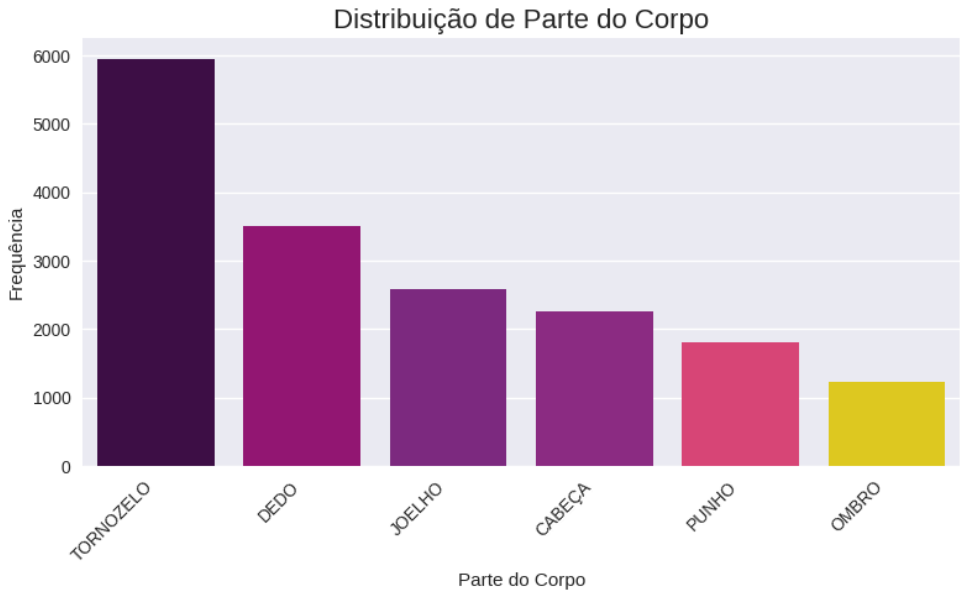
Gráfico 8 – Distribuição de idade após filtragem dos dados



Fonte: Elaborado pelo autor

A distribuição de idade mostra uma concentração maior entre as idades de 10 a 20 anos, onde pouco mais de 80% dos registros estão concentrados. Com essa informação é possível sugerir que os agrupamentos feitos terão tais idades mais presentes, e que idades maiores não terão uma maior representatividade.

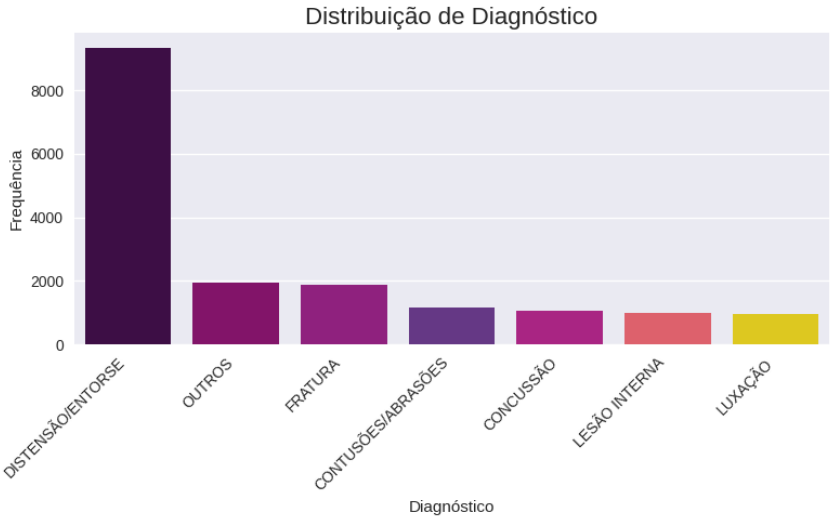
Gráfico 9 – Distribuição de parte do corpo após filtragem dos dados



Fonte: Elaborado pelo autor

Após a seleção de partes do corpo com um mínimo de 5% de representatividade, as partes do corpo mais afetadas pelas lesões foram tornozelo, dedos, joelho, cabeça, punho e ombro. Estudos como o de KILIC et al. (2017) corroboram tal achado e mostram que tais partes do corpo são mais afetadas pela natureza das ações realizadas em uma partida de voleibol.

Gráfico 10 – Distribuição de diagnóstico após filtragem dos dados



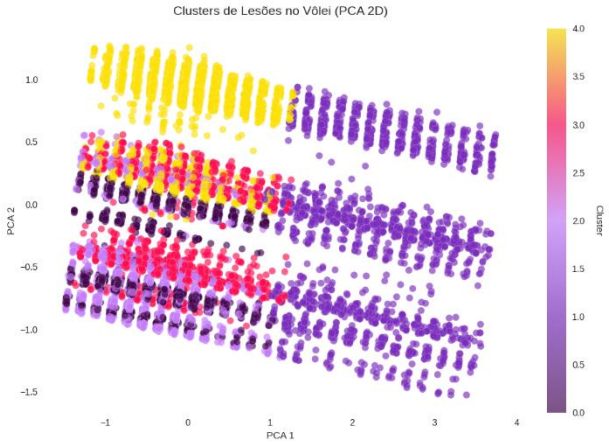
Fonte: Elaborado pelo autor

No gráfico acima é perceptível a diferença na quantidade de casos com distensão/entorse e os demais diagnósticos, o que se dá pelos movimentos praticados no esporte. Tendo essa informação, foi importante fazer uma extração dos verbos presentes na coluna narrativa para entender o que ocorreu para que tal lesão acontecesse. Desta forma combinando o diagnóstico com a parte do corpo afetada e o movimento realizado foi possível entender melhor o que ocorreu e desta forma sugerir prevenções para tal.

4.2 Divisão dos clusters

Ao aplicar o algoritmo *KMeans* com k=5, dividindo os dados em 5 clusters, foi possível identificar alguns padrões. A fim de facilitar a visualização e interpretação dos agrupamentos obtidos, os dados foram projetados nos dois primeiros componentes principais por meio da Análise de Componentes Principais (PCA). Tal técnica permitiu reduzir a dimensionalidade dos dados preservando a maior parte da variabilidade e possibilitando representar graficamente as observações em um espaço bidimensional. Cada ponto no gráfico representa um registro do conjunto de dados, sendo a coloração correspondente ao cluster atribuído pelo algoritmo *KMeans*. Essa projeção evidencia a distribuição dos grupos, sendo possível identifica grupos mais separados e nítidos e alguns mais sobrepostos.

Gráfico 11 – Clusters de lesões no vôlei (PCA 2D)

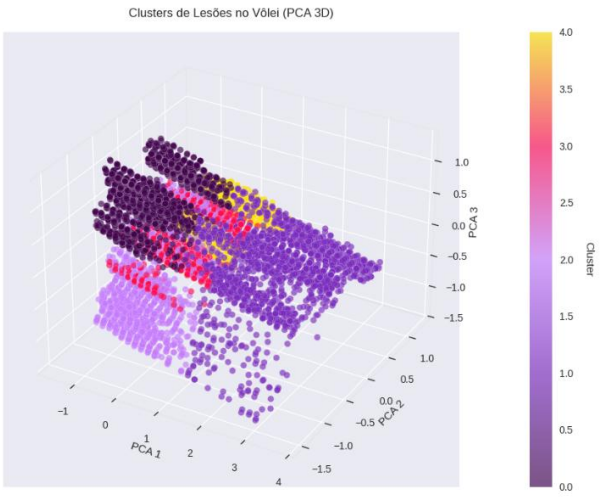


Fonte: Elaborado pelo autor

O gráfico acima mostra a divisão dos clusters em um espaço bidimensional formado pelas duas primeiras componentes principais (PC1 e PC2) obtidas pelo PCA. Essas componentes agrupam a maior parte da variabilidade dos dados e permitem uma visualização simplificada da distribuição dos registros e de seus respectivos clusters.

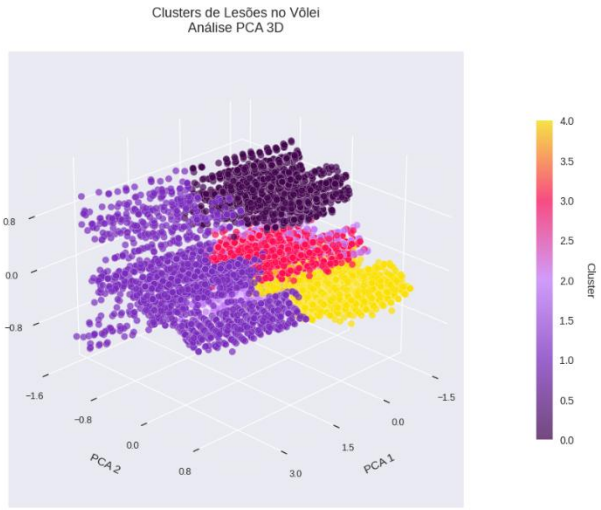
Já os gráficos abaixo demonstram a divisão dos clusters em um espaço tridimensional com as três primeiras componentes principais (PC1, PC2 e PC3) obtidas pelo PCA. Foram feitos dois gráficos com as mesmas características, porém em ângulos diferentes para melhor visualização e entendimento da distribuição dos dados em cada cluster.

Gráfico 12 – Clusters de lesões no vôlei (PCA 3D) – Ângulo 1



Fonte: Elaborado pelo autor

Gráfico 12 – Clusters de lesões no vôlei (PCA 3D) – Ângulo 2

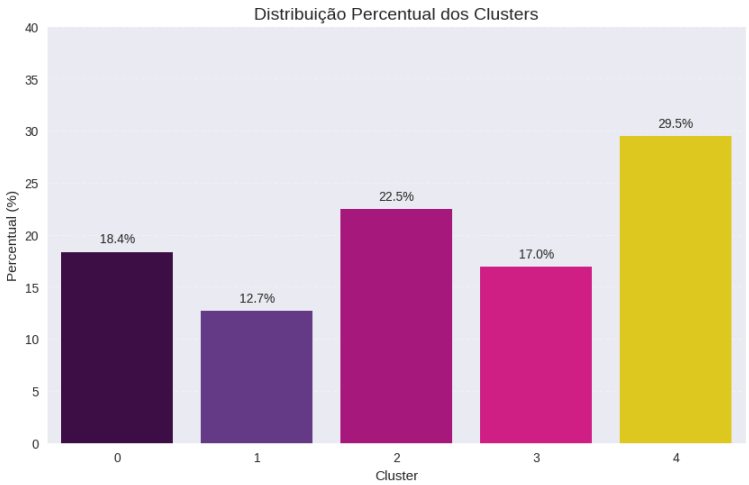


Fonte: Elaborado pelo autor

A análise dos *loadings* mostrou que, na Componente Principal 1 (PC1), a variável idade apresentou o maior peso (0,973), seguida por gênero feminino (0,150) e diagnóstico de entorse e distensão (0,078), indicando que este eixo está fortemente associado à idade dos atletas e, em menor grau, ao gênero feminino e tipo de diagnóstico. Já a Componente Principal 2 (PC2) teve mais influência do diagnóstico de entorse e distensão (0,595) e tornozelo (0,560), sugerindo que este eixo está relacionado principalmente a lesões por entorse no tornozelo. Por fim a Componente Principal 3 (PC3), presente nos gráficos tridimensionais, foi mais impactada pela parte do corpo referente aos dedos (0.602) cabeça (0.457) e pelo verbo hit (0.316), no sentido de bater ou impactar, indicando um eixo motivado por lesões de impacto nos dedos e na cabeça.

O gráfico acima apresenta a distribuição de casos entre os 5 clusters obtidos, mostrando uma distribuição quase uniforme, em que apenas o cluster 4 destoa levemente dos demais com uma maior quantidade de registros, representando 30% do total.

Gráfico 13 – Distribuição percentual dos clusters



Fonte: Elaborado pelo autor

4.3 Análise dos clusters

A tabela abaixo mostra a distribuição dos dados entre os clusters, o que permite uma análise dos padrões. Sendo assim, será possível identificar quais combinações de características ocorrem no intuito de prevenir que novas lesões ocorram.

Tabela 1 – Distribuição das variáveis entre os clusters

Cluster	*Nº Casos*	*Idade Média (anos)*	*% Feminino*	*% Masculino*	*Partes do Corpo Mais Afetadas*	*Diagnósticos Mais Frequentes*	*Ações Associadas*
0	3184	14.3	74.50%	25.50%	Dedo (3170) Ombro (14)	Distensão/Entorse (1255) Fratura (1005) Contusões/Abrasões (374)	Torção por compressão (673) Impacto/Colisão (506) Distender (198)
1	2203	28.2	43.35%	56.65%	Tornozelo (840) Joelho (481) Dedo (341)	Distensão/Entorse (1352) Fratura (234) Luxação (219)	Torção rotacional (323) Distender (288) Impacto/Colisão (138)
2	3896	14.5	84.52%	15.48%	Cabeça (2156) Punho (1622) Ombro (118)	Concussão (1025) Lesão Interna (951) Distensão/Entorse (944)	Impacto/Colisão (1713) Distender (206) Cair (103)
3	2938	15.5	78.39%	21.61%	Joelho (2098) Ombro (832) Punho (8)	Distensão/Entorse (1374) Outros (705 Luxação (515)	Distender (535) Torção rotacional (310) Aterrissar/Pousar (230)
4	5111	15.4	76.60%	23.40%	Tornozelo (5110) Ombro (1)	Distensão/Entorse (4413) Outros (332) Fratura (318)	Torção rotacional (1303) Rolar (733) Aterrissar/Pousar (624)

Fonte: Elaborado pelo autor

Conforme visto na tabela acima a maioria do cluster é composta majoritariamente por atletas do sexo feminino, tendo em vista a quantidade maior de mulheres nos casos registrados.

A idade também tem uma predominância entre os 14 e 15 anos, apenas o cluster 1 onde a maioria é composta por homens a média de idade está em 28.2 anos. Os diagnósticos mais presentes são distensão/entorse na maioria dos clusters, já as partes do corpo afetadas e os verbos de ação associados variam entre os clusters, demonstrando combinações diferentes.

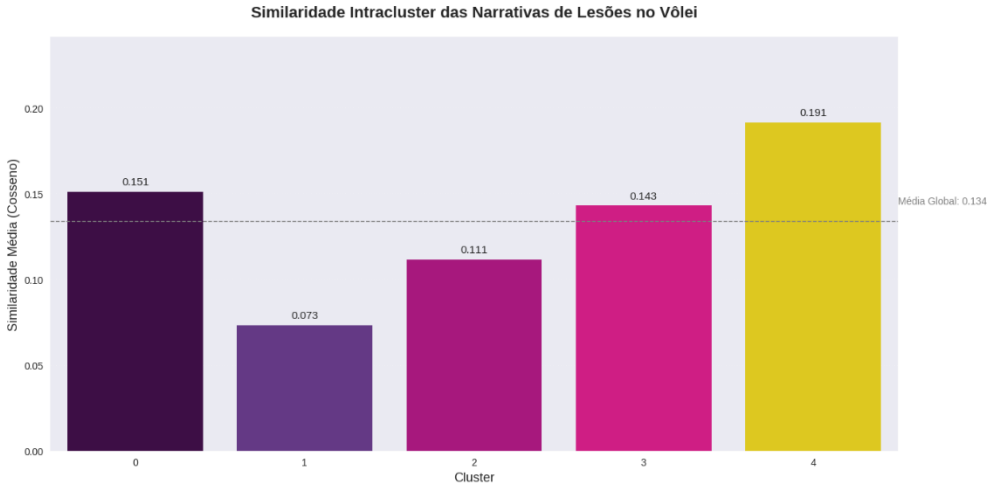
O cluster 0 é composto em sua maioria por meninas com uma média de idade de 14.3 anos com lesões principalmente nos dedos, com distensões e entorses em ações de torção por compressão e colisões. Tal agrupamento pode indicar contatos com a bola em ações como bloqueio, em que a atleta posiciona as mãos de forma a impedir que a bola atacada pelo time adversário passe para seu lado da quadra. Desta forma é interessante encontrar meios de proteger as mãos e fortalecer os músculos desta parte do corpo para que tal fato se minimize.

Já o cluster 1, é o único composto em sua maioria por homens, tendo estes uma média de idade de 28.2 30 anos, apresenta lesões principalmente nos membros inferiores como tornozelo e joelho, com diagnósticos de distensão e entorse em ações de torção e impacto/colisão. Isso pode indicar que tais danos ocorreram em ações de pulos, em que ao saltar e aterrissar novamente no chão o atleta tem um impacto grande e pode torcer o tornozelo ou machucar o joelho. Desta forma, é importante direcionar os treinamentos de forma a orientar a melhor forma de saltar com segurança. A adesão de acessórios como joelheira e tornozeleira também pode ser estudada para identificar se são necessárias para minimizar a ocorrência de lesões.

O cluster 2, assim como os demais que serão apresentados, é composto em sua maioria de meninas com uma média de idade de 14.5 anos, com lesões na cabeça em que algum tipo de impacto ou colisão aconteceu. Tal fato pode indicar queda, colisões com outra jogadora, com partes da quadra como o poste da rede e arquibancadas e colisões até mesmo com a bola. Pode ser indicado treinar a movimentação em quadra, para estar atento aos possíveis obstáculos em quadra e ter uma noção de como agir de forma a evitar o contato com objetos que possam atingir a cabeça. Ao fixar o olhar na bola é possível que a atleta não observe o que está ao redor, mas a colisão também pode ocorrer de forma inevitável.

Os clusters 3 e 4 são focados principalmente nos membros inferiores e associados a ações que indicam pulos e aterrisagens, desta forma assim como visto no cluster de maioria masculina os treinos de saltos são bastante importantes. Isso porque no voleibol em especial os pulos são bastante frequentes, seja em ações de bloqueio, saque, ataque ou até mesmo defesa. Desta forma é necessário ensinar a melhor forma de voltar ao chão após uma ação de pulo.

Gráfico 14 – Similaridade intracluster das narrativas de lesões no vôlei



Fonte: Elaborado pelo autor

Por último, a similaridade do cosseno intracluster, obtida a partir da coluna de narrativas que descreve como ocorreu a lesão, demonstra o quão coeso os clusters obtidos são. Os valores obtidos podem ser identificados no gráfico acima.

É interessante notar que o cluster 4, apesar de ser o maior em quantidade de casos, também é o que possui uma maior similaridade entre seus registros. Os demais casos também possuem uma boa similaridade, com exceção do cluster 1 que está um pouco mais diverso que os demais. Para demonstra exemplos das narrativas encontradas nos clusters, alguns registros da base de dados foram disponibilizados na tabela abaixo.

Tabela 2 – Exemplos de narrativas em cada cluster (idioma original)

Cluster	Exemplo 1	Exemplo 2	Exemplo 3
0	13YR F PLAYING VOLLEYBALL;DX FINGER FX	20YOM WHILE PLAYING VOLLEYBALL INJURED HIS LEFT THUMB;SPRAIN LEFT THUMB	RIGHT THUMB CONTUSION. 18 YOF.18 YOF HAD A THUMB CONTUSIONFROM VOLLEYBALL.
1	LEFT KNEE STR. PT STRAINED KNEE WHEN PLAYING VOLLEYBALL	SPRAIN RT.KNEE.PT.REFERS WHILE PLAYING VOLLEYBALL.	SHOULDER STR. 25 YOM STRAINED SHOULDER WHEN PLAYING VOLLEYBALL
2	18 YOF HAD FOREARM PX AFTER STRIKING VOLLEYBALL IN PE.DX: WRIST PX, FOREARM PX.	PT FELL HITTING HEAD ON GROUND WHILE PLAYING VOLLEYBALL @SCHOOL TODAY,CONCUSSION	13YOF C/O HEADACHE WAS HIT IN HEAD WITH VOLLEYBALL, DX CONCUSSION
3	KNEE STR. 16 YOF STRAINED KNEE WHEN PLAYING VOLLEYBALL	15YOM WAS AT SCHOOL PLAYING VOLLEYBALL WHEN HE WAS ABOUT TO "TAKE OFF" BUT RIGHT KNEE BENT INWARD & THEN OUTWARD AGAIN W/ PAIN DX: KNEE SPRAIN	TWIST KNEE, PLAYING VOLLEYBALL DX: KNEE SPRAIN

4	17YOF TWISTED LEFT ANKLE PLAYING VOLLEYBALL DX: STRAINED ANKLE	17 YO MALE HURT PLAYING VOLLEYBALL. DX ANKLE SPRAIN B	15YOF TWISTED R ANKLE WHILE PLAYING VOLLEYBALL / SPRAIN ANKLE

Fonte: Elaborado pelo autor

Tabela 3 – Exemplos de narrativas em cada cluster (traduzida)

Cluster	Exemplo 1	Exemplo 2	Exemplo 3
0	13 anos Feminino jogando vôlei; Diagnóstico: Fratura no dedo	20 anos Masculino lesionou o polegar esquerdo durante jogo de vôlei; Entorse do polegar esquerdo	Contusão no polegar direito. 18 anos Feminino. Lesão ocorreu jogando vôlei
1	Lesão no joelho esquerdo. Paciente tensionou o joelho jogando vôlei	Entorse no joelho direito. Ocorreu durante partida de vôlei	Lesão no ombro. 25 anos Masculino tensionou o ombro jogando vôlei
2	13 anos Feminino com queixa de dor de cabeça após ser atingida na cabeça por bola de vôlei. Diagnóstico: Concussão	Paciente caiu e bateu a cabeça no chão durante jogo de vôlei na escola hoje. Diagnóstico: Concussão	18 anos Feminino com fratura no antebraço após golpear bola de vôlei na aula de educação física. Diagnóstico: Fratura no pulso e antebraço
3	Lesão no joelho. 16 anos Feminino tensionou o joelho jogando vôlei	15 anos Masculino lesionou o joelho direito durante salto no vôlei escolar quando o joelho moveu-se para dentro e depois para fora com dor. Diagnóstico: Entorse no joelho	Torção no joelho durante jogo de vôlei. Diagnóstico: Entorse no joelho
4	17 anos Feminino torceu o tornozelo esquerdo jogando vôlei. Diagnóstico: Entorse no tornozelo	17 anos Masculino lesionou-se jogando vôlei. Diagnóstico: Entorse no tornozelo B	15 anos Feminino torceu o tornozelo direito durante jogo de vôlei. Diagnóstico: Entorse no tornozelo

Fonte: Elaborado pelo autor

A partir da tabela acima é possível perceber que as narrativas possuem certa semelhança entre si, indicando uma boa divisão de dados entre os clusters. Também é possível verificar a prevalência de lesões que corroboram com os resultados encontrados na tabela de características de cada cluster, como por exemplo os registros de lesões na cabeça no cluster 2 e as ocorrências no tornozelo no cluster 4. Desta forma é possível identificar um bom agrupamento dos registros apresentados, o que pode ser útil para traçar estratégias de prevenção para as contusões que podem ocorrer ao praticar o esporte.

5 Considerações finais

O presente estudo teve como objetivo realizar uma análise exploratória de lesões ocorridas no voleibol, bem como aplicar algoritmos de *machine learning* não supervisionados no intuito de identificar clusters, ou seja, agrupamentos, de casos para que padrões em tais lesões fossem encontrados. Tal pesquisa é importante para que estratégias adequadas de prevenção sejam tomadas para que os atletas e jogadores casuais possam diminuir o risco de contusões.

Com a análise exploratória foi possível identificar a presença majoritária de atletas do sexo feminino, dada a popularidade do esporte no país de origem dos dados, Estados unidos. As idades se concentram entre 10 e 20 anos, indicando que um cuidado maior tenha que ser tomado na prática desportiva entre jovens e adolescentes. Já as partes do corpo e diagnósticos encontrados vão ao encontro às mencionadas na literatura existente, dada a natureza das ações praticadas no esporte.

Os dados foram divididos em cinco clusters, apresentando uma boa separação e que pode servir de insights para profissionais da saúde e preparadores físicos que buscam diminuir as lesões de atletas. Foram identificados agrupamentos que mostram como cada ação praticada juntamente com a idade e sexo do atleta tem relação com a parte do corpo lesionada. As ações e verbos relacionadas as partes do corpo e diagnósticos de cada ocorrência foram importantes para mostrar como estão relacionadas. Treinos específicos focados em tais partes do corpo e fortalecimento dessas partes podem ser feitos, como por exemplo a ocorrência de entorse nos dedos associada à torção por compressão, o que pode indicar contato com a bola em ações de bloqueio típicas de uma partida de vôlei.

A base de dados, apesar de possuir bastante registros, não possui colunas que poderiam auxiliar ainda mais na divisão dos clusters. Colunas essas que poderiam ser mais gerais como altura e peso e algumas mais específicas do esporte como por exemplo posição em que o atleta atua e se a lesão ocorreu em treino ou durante jogo.

Desta forma, para estudos futuros, uma coleta de registros com mais informações pode levar a resultados mais claros e que auxiliem cada vez mais na identificação de padrões e definição de estratégias de prevenção.

REFERÊNCIAS

AMENDOLARA A, PFISTER D, SETTELMAYER M, SHAH M, WU V, DONNELLY S, JOHNSTON B, PETERSON R, SANT D, KRIAK J, BILLS K. An Overview of Machine Learning Applications in Sports Injury Prediction. **Cureus**. 2023 Sep 28;15(9):e46170. doi: 10.7759/cureus.46170. PMID: 37905265; PMCID: PMC10613321.

BERE, T. et al. Injury risk is low among world-class volleyball players: 4-year data from the FIVB Injury Surveillance System. **British journal of sports medicine**, v. 49, n. 17, p. 1132–1137, 2015.

Consumer Product Safety Commission. **National Electronic Injury Surveillance System 2004-2023 on NEISS Online Database**, released April, 2024. Generated at https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx. on: May 6, 2024 at 12:38:02

GUERZONI, G. T. **Diferenças na prevalência de lesões no voleibol amador e profissional: uma revisão de literatura**. 2022.

KERN, C.; KLAUSCH, T.; KREUTER, F. **Tree-based machine learning methods for survey research. Survey research methods**, v. 13, n. 1, p. 73–93, 2019.

KILIC, O. et al. Incidence, aetiology and prevention of musculoskeletal injuries in volleyball: A systematic review of the literature. **European journal of sport science: EJSS: official journal of the European College of Sport Science**, v. 17, n. 6, p. 765–793, 2017.

MESQUITA, W. G. DE .; FONSECA, R. M. C.; FRANÇA, N. M. DE .. Influência do voleibol na densidade mineral ossea de adolescentes do sexo feminino. **Revista Brasileira de Medicina do Esporte**, v. 14, n. 6, p. 500–503, nov. 2008.

NATIONAL FEDERATION OF STATE HIGH SCHOOL ASSOCIATIONS. **High school athletics participation survey 2022-23.** Indianápolis: NFHS, 2023. 25 p. Disponível em: <https://members.nfhs.org/participation_statistics>. Acesso em: 10 out. 2023.

OLIVER, J. L. et al. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. **Journal of science and medicine in sport**, v. 23, n. 11, p. 1044–1048, 2020.

SANTOS, B. **Caracterização das lesões em atletas de voleibol: revisão de literatura**. [s.l: s.n.]. 2021

Superliga feminina: aumento na audiência em 23/24. Veja o top 5! Disponível em: <https://webvolei.com.br/superliga-feminina-aumento-na-audiencia-em-23-24-veja-o-top-5/>. Acesso em: 1 maio. 2024.

VAN EETVELDE H, MENDONÇA LD, LEY C, SEIL R, TISCHER T. Machine learning methods in sport injury prediction and prevention: a systematic review**. J Exp Orthop**. 2021 Apr 14;8(1):27. doi: 10.1186/s40634-021-00346-x. PMID: 33855647; PMCID: PMC8046881.

VERHAGEN, E. A. L. M. et al. A one season prospective cohort study of volleyball injuries. **British journal of sports medicine**, [s. l.], v. 38, n. 4, p. 477–481, 2004. DOI 10.1136/bjsm.2003.005785. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=15273190&lang=pt-br&site=ehost-live. Acesso em: 1 maio. 2024.

YOUNG, W. K.; BRINER, W.; DINES, D. M. Epidemiology of common injuries in the volleyball athlete. **Current reviews in musculoskeletal medicine**, v. 16, n. 6, p. 229–234, 2023