

Análise Preditiva e fatores determinantes de sucesso para o voleibol - Estudo de Caso VNL Feminina

Predictive Analysis and determining factors of success in volleyball - Women's VNL Case Study

João Ferrari [Universidade Federal da Fronteira Sul | joao.ferrari@estudante.uffs.edu.br]
Denio Duarte [Universidade Federal da Fronteira Sul | duarte@uffs.edu.br]

Universidade Federal da Fronteira Sul, Rodovia SC 484, km 02, Bairro Fronteira Sul, Chapecó, SC CEP 89815-899
Tel. (49) 2049-2600 CNPJ 11.234.780/0007-46, Brasil.

Resumo. O aumento substancial no volume de dados esportivos tem impulsionado o uso de técnicas de Inteligência Artificial, tornando o Aprendizado de Máquina uma ferramenta essencial para a análise de desempenho e tomada de decisão no esporte moderno. Inserido neste contexto, este trabalho teve como objetivo desenvolver e avaliar modelos de Aprendizado de Máquina supervisionados capazes de prever os resultados (vitória ou derrota) de partidas de voleibol feminino de seleções. O estudo utilizou dados da Liga das Nações de Voleibol (VNL), cobrindo o recorte temporal de 2021 a 2025, extraídos via *Web Scraping* da plataforma oficial *Volleyball World*. A metodologia envolveu as etapas de coleta, limpeza e engenharia de *features* com foco em métricas coletivas e diferenciais técnicos, seguidas pela aplicação e comparação dos algoritmos *Random Forest* e *XGBoost*. Para a validação, foram utilizadas métricas como acurácia, precisão, *recall* e F_1 -score, nas quais os modelos alcançaram acurácia entre 77,17% e 79,35%. Os resultados permitiram não apenas classificar os desfechos dos confrontos com consistência, mas também identificar que as variáveis de *ranking* e o fundamento do bloqueio exercem maior influência como fatores determinantes para o sucesso das equipes.

Abstract. The substantial increase in the volume of sports data has driven the use of Artificial Intelligence techniques, making Machine Learning an essential tool for performance analysis and decision-making in modern sports. In this context, this work aimed to develop and evaluate supervised Machine Learning models capable of predicting the results (win or loss) of women's national volleyball team matches. The study used data from the Volleyball Nations League (VNL), covering the seasons from 2021 to 2025, extracted via *web scraping* from the official *Volleyball World* platform. The methodology involved the stages of data collection, cleaning, and feature engineering focusing on collective metrics and technical differentials, followed by the application and comparison of the *Random Forest* and *XGBoost* algorithms. For validation, metrics such as accuracy, precision, *recall*, and F_1 -score were used, with the models achieving accuracy between 77.17% and 79.35%. The results allowed not only predicting match outcomes consistently but also identifying that ranking variables and the blocking fundamental exert the greatest influence as determining factors for the teams' success.

Palavras-chave: Voleibol, Aprendizado de Máquina, Análise Preditiva, Aprendizado Supervisionado, Análítica de Esportes

Keywords: Volleyball, Machine Learning, Predictive Analysis, Supervised Learning, Sports Analytics

1 Introdução

O voleibol é um esporte de ampla prática em território brasileiro, tendo recebido enorme aceitação popular, seja na disciplina escolar de Educação Física, como atividade de lazer, ou até mesmo no cenário competitivo [Durão, 2025]. O esporte é jogado em um melhor de cinco sets, sendo necessário vencer três sets para ganhar a partida, com o objetivo de chegar aos vinte e cinco pontos em cada um dos sets. Caso seja jogado o quinto set, a pontuação diminui para quinze pontos. É obrigatório abrir uma vantagem de dois pontos para sair como vencedor de cada set, em situações nas quais tal vantagem não é concretizada, a pontuação pode ultrapassar os vinte e cinco pontos previamente estabelecidos [Federação Internacional de Voleibol (FIVB), 2021].

Além de seus objetivos competitivos, possui grande responsabilidade de integração, o que proporciona determinada aproximação entre seus atletas e pode garantir certa melhora no convívio, muitas vezes utilizado como escape para os empecilhos da rotina [Nunes *et al.*, 2014]. Ainda segundo Nunes *et al.* [2014], possui papel na composição da disciplina em seus praticantes, em virtude de ser um ambiente que apre-

senta regras e objetivos, em especial na vida de crianças e adolescentes, que utilizam a iniciação esportiva para um aprimoramento da socialização e regulamento de suas emoções. Todos esses aspectos tornam esse jogo um fator relevante para a vida de seus apoiadores.

No âmbito profissional, a coleta de estatísticas e a análise de dados vêm ganhando relevância contínua, configurando-se como ferramentas importantes para o planejamento estratégico de temporadas e o estabelecimento de metas competitivas [Sanjaykumar *et al.*, 2025; Lalwani *et al.*, 2022]. Alinhado com essa ideia, o Aprendizado de Máquina surge como um forte aliado para a análise esportiva, com uma utilização que transcende o ambiente acadêmico, sendo aplicado para desenvolvimento de atletas e ganho desportivo, como pode ser observado em Sanjaykumar *et al.* [2025] e Lalwani *et al.* [2022]. É válido constatar que a utilização de recursos estatísticos desponta como uma ferramenta de apoio, não sendo indicada para apontar verdades absolutas sobre determinado nicho esportivo.

O aumento substancial no volume de dados estruturados e não estruturados, que vai além dos eventos esportivos, im-

pulsionou o desenvolvimento e o uso de técnicas capazes de extrair informações e empregar algoritmos de Aprendizado de Máquina para a previsão de resultados. No contexto do cenário esportivo, a previsão de resultados tornou-se atraente não apenas para profissionais do esporte e gestores de equipe, mas também para o público mais amplo, destacando-se especialmente a área de apostas esportivas. Assim, os algoritmos de Aprendizado de Máquina têm o objetivo de auxiliar a comissão técnica, bem como dar suporte à tomada de decisões em apostas [Horvat and Job, 2020].

Sob a perspectiva da inteligência esportiva, a capacidade preditiva viabiliza uma preparação direcionada, permitindo antecipar cenários de sucesso ou falha contra adversários específicos. Ademais, ao isolar e avaliar os fatores determinantes que influenciaram um determinado resultado, apresenta-se a possibilidade de aprimorar lacunas técnicas específicas. Esse nível de análise dificilmente é alcançado por meio de métodos estatísticos tradicionais ou exames puramente visuais pós-jogo, o que justifica a introdução de modelos inteligentes no esporte.

A análise esportiva tem como objetivo proporcionar à comissão técnica informações que possam contribuir para o progresso de desempenho durante a temporada. É possível analisar em diversos estudos a utilização do Aprendizado de Máquina aliado a análise estatística, em torno disso, sua utilização possui grande variabilidade, em que modelos preditivos são utilizados em diferentes esportes, como no Basquete [Wang et al., 2022] e no Handebol [Felice and Ley, 2025], demonstrando que a aplicação do Aprendizado de Máquina não se limita a um escopo específico, adaptando-se às singularidades de cada esporte.

Este trabalho tem como objetivo desenvolver e avaliar um modelo preditivo baseado em Aprendizado de Máquina, sendo capaz de classificar os resultados de vitória ou derrota em partidas de voleibol feminino de alto rendimento. Para tanto, o modelo processará variáveis de desempenho acumulado, tais como eficiência de ataque, volume de bloqueios, qualidade do passe, consistência defensiva, totalização de pontos, histórico de confrontos diretos e o posicionamento das equipes no *ranking* da competição.

O site <https://br.volleyballworld.com/> foi utilizado para a extração dos dados via *Web Scraping*, tendo como foco a Liga das Nações de Voleibol (VNL), naipes feminino, cobrindo o recorte temporal das temporadas de 2021 até 2025. Na sequência, o conjunto de dados foi submetido às etapas de limpeza, tratamento e engenharia de atributos (*features*).

Experimentos foram realizados e apontaram para a viabilidade do modelo proposto, no qual o algoritmo *Random Forest* obteve 77,17% de acurácia e índices estáveis de 77% para precisão, *recall* e F_1 -score. Em contrapartida, o modelo *XGBoost* alcançou o desempenho superior deste estudo, registrando 79,35% de acurácia, 84% de precisão, 79% de *recall* e um F_1 -score de 80%. Ao analisar os fatores determinantes para o desfecho das partidas, as variáveis de *ranking* obtiveram maior importância em ambos os modelos. Como principal fundamento, o bloqueio demonstrou-se a ação técnica mais relevante para consolidar a vitória.

A estrutura deste trabalho está dividida em seções subsequentes. A Seção 2 revisa os Trabalhos Relacionados re-

levantes para o projeto. A fundamentação teórica de Aprendizado de Máquina, os algoritmos e a metodologia de implementação são apresentados na Seção 3. Os resultados empíricos obtidos e as suas respectivas interpretações analíticas encontram-se dispostos nas Seções 4 e 5. Por fim, a conclusão deste estudo é apresentada na Seção 6.

2 Trabalhos Relacionados

Esta seção apresenta alguns dos trabalhos relacionados à esta proposta. Após pesquisa em mecanismo de busca de artigos científicos, quatro trabalhos foram selecionados. Os mesmos são descritos a seguir.

No artigo Lalwani et al. [2022], foi utilizado como base 1.289 partidas da Liga Brasileira de Voleibol (Superliga), cobrindo as temporadas de 2010-11 até 2018-19, obtendo os dados por meio da plataforma *FlashScore.in* via *web scraping*. Os algoritmos utilizados foram divididos em dois grupos para previsão (*white-box* e *black-box*).

Entre os modelos de *white-box* estão: Regressão logística (LogReg) e *Boolean Rule Column Generation* (BRCG) e nos modelos de *black-box* estão *Support Vector Machine* (SVM), Redes Neurais (*Neural Networks / DNN*) e Análise Discriminante Linear (LDA). Para os modelos de explicação foram utilizados SHAP (*SHapley Additive exPlanations*) e *ProtoDash*.

Para a avaliação de desempenhos foram utilizadas as métricas de Acurácia, F_1 -score e AUC-ROC. Os resultados obtidos foram divididos em duas categorias: desempenho dos modelos de previsão e resultados de explicabilidade. No desempenho dos modelos de previsão, os modelos *black-box* superaram o desempenho dos modelos *white-box*, com o modelo SVM obtendo a maior precisão entre todos os testados, com acurácia de 0.7790. Em geral, modelos *black-box* tiveram pontuações mais altas em todas as métricas em comparação com os modelos *white box*.

Por outro lado, Sanjaykumar et al. [2025] abordam uma perspectiva diferente em relação ao uso do Aprendizado de Máquina no esporte. O objetivo do estudo foi investigar as relações entre os atributos demográficos e físicos das jogadoras de voleibol e seu desempenho em quadra, além de fornecer *insights* valiosos para estratégias de identificação de talentos e desenvolvimento de jogadoras. Nesse contexto, obteve foco também, para investigar a eficácia das técnicas de Aprendizado de Máquina para prever o desempenho futuro de jogadoras de voleibol universitárias na Índia, adquirindo novos conhecimentos sobre a aplicação do aprendizado na identificação de talentos e na previsão de desempenho no voleibol. Foram utilizados atributos de 174 jogadoras de voleibol feminino de nível universitário. Os dados foram coletados de várias faculdades e instituições que participam de torneios de voleibol em diferentes regiões da Índia. Os algoritmos selecionados foram *Linear Regression*, *Random Forest Regression* e *XGBoost Regression*.

O desempenho de cada modelo foi avaliado utilizando as métricas de desempenho padrão: Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE), R-squared (R^2) (ou Coeficiente de Determinação). Os resultados do estudo são divididos em duas áreas principais: o desempenho dos modelos de aprendizado

de máquina e a importância dos atributos físicos das jogadoras.

O algoritmo de *Random Forest* superou os demais, obtendo o melhor desempenho, com uma precisão de 94,18%, além de apresentar o menor erro com um MSE de 7.1505 e um RMSE de 2.6740. O *XGBoost* obteve o segundo melhor desempenho, tendo a precisão avaliada em 92,76%, além de um MSE de 8.8876 e um RMSE de 2,9812%. Com o menor desempenho entre os algoritmos o *Linear Regression*, apresentou uma previsão de 75,31% e com o maior erro, MSE de 30.3207 e RMSE de 5.5064. A análise de correlação de *Pearson* identificou que a altura, massa muscular e massa óssea, destacam-se como atributos físicos que estão mais fortemente ligados à previsão de desempenho. Idade e IMC não apresentaram uma correlação relevante.

Outra contribuição relevante nessa área é a de Wang et al. [2022], em que foram utilizados dados da NBA extraídos via *web scraping*¹, abrangendo as temporadas de 1980 a 2019. Para a análise descritiva, foram aplicados os algoritmos *K-means*, *DBSCAN* e Análise de Componentes Principais (PCA) sobre dados em nível de temporada. Na previsão de sucesso, definida pela taxa de vitórias mensal (Regressão) e pela classificação para os *playoffs* (Classificação), os modelos escolhidos foram *Random Forest* e *Feed-forward Neural Network*. Para a interpretabilidade das previsões, utilizou-se o LIME (*Local Interpretable Model-agnostic Explanations*).

As métricas utilizadas para a análise descritiva foram o *Silhouette Score* e a Proporção da Variância Explicada. Já para a tarefa de regressão, aplicou-se o R^2 e o *Mean Squared Error* (MSE), enquanto o F_1 -score foi empregado para a classificação dos *playoffs*.

O algoritmo *K-means* (com $N=3$ clusters) agrupou com sucesso temporadas de períodos diferentes, indicando estilos de jogo distintos por era. Na previsão da taxa de vitórias, a Rede Neural obteve desempenho superior ao *Random Forest*, alcançando um R^2 de 0.77 e MSE de 0.009, contra um R^2 de 0.65 e MSE de 0.014 do modelo de *Random Forest*. Para a previsão de *playoffs*, ambos os modelos alcançaram um F_1 -score de 0.87. Por fim, os resultados de interpretabilidade com o LIME validaram a famosa teoria do basquete de que a defesa ganha campeonatos.

Na mesma linha de pesquisa, o estudo de Felice and Ley [2025] trabalhou com partidas de clubes femininos de handebol, fazendo o uso de *SportScore API* (via *RapidAPI*), extraíndo dados históricos dos jogos, como placar, competição, local e horário. Além do uso do site *Handball Base*², que forneceu informações sobre as equipes e jogadoras, usadas para gerar variáveis como altura, peso, idade e nacionalidade.

A pesquisa tem como objetivo propor um modelo para prever os resultados de jogos de handebol. O estudo treinou modelos para tarefas de Regressão, com foco em prever o número exato de gols marcados por cada equipe e tarefas de Classificação, para prever o resultado final da partida (vitória em casa, empate ou vitória fora). Os algoritmos utilizados foram *Random Forest*, *XGBoost*, *CatBoost*, *Multi-Layered Perceptron*, além do *SHAP* para análise de explicabilidade. Para avaliação, foram utilizadas as métricas de Acurácia, F_1 -score

e *Brier score* para classificação, e *Root Mean Squared Error* (RMSE) e *Mean Absolute Percentage Error* (MAPE) para regressão.

Os resultados obtidos destacaram o impacto positivo das novas *features* estatísticas de forças (ataque e defesa). Nos modelos de classificação, o *Random Forest*, combinado com as variáveis de força, apresentou o melhor desempenho geral, atingindo uma acurácia de 81,32%. Para os modelos de regressão, o *CatBoost* com os mesmos atributos de força obteve o menor erro na previsão dos placares. A análise de explicabilidade, usando *SHAP*, confirmou que as *features* de força estavam entre as mais importantes para o modelo.

A proposta deste trabalho difere dessas apresentadas por integrar, de forma simultânea, a previsão de resultados via Aprendizado de Máquina de Classificação e a análise dos fatores determinantes de sucesso. Os modelos utilizaram indicadores de desempenho consolidados como pontuação, histórico de confrontos, posição no *ranking* e eficiência nos fundamentos.

3 Métodos

Esta seção apresenta a metodologia utilizada para implementar o modelo proposto neste trabalho bem como o referencial teórico que a apoia. Inicialmente, a competição Liga das Nações de Voleibol (VNL) é apresentada, seguida dos passos utilizados para obter o conjunto de dados. Depois são apresentados como os dados foram tratados e o desenvolvimento dos modelos de Aprendizado de Máquina.

3.1 Liga das Nações de Voleibol

A Liga das Nações de Voleibol (VNL) é uma competição que reúne as principais seleções do mundo, tanto femininas, quanto masculinas. É um dos eventos mais populares do voleibol moderno, representando a forte globalização no esporte [Volleyball World, 2025].

A competição foi criada em 2018 pela Federação Internacional de Voleibol (FIVB), tornando-se a principal competição anual do vôlei. Foi desenvolvida para substituir a antiga Liga Mundial, sendo desenhada em um formato mais competitivo e direto. As seleções se confrontam na fase de grupos, em seguida os melhores colocados avançam para a fase eliminatória [Volleyball World, 2025].

3.1.1 Fórmula da Competição

Atualmente, o torneio é disputado com 18 seleções em cada naipes, contando com 116 jogos entre a Fase de Classificação e a Fase Final e cada seleção irá disputar 12 partidas, no mínimo. Na primeira fase, os 12 jogos são distribuídos ao longo de três semanas, a cada semana, 3 grupos de 6 equipes serão formados em diferentes cidades e cada time entrará em quadra 4 vezes.

As 8 seleções bem mais colocadas na tabela ao término na fase de Classificação avançam para a fase Final. O país-sede da Fase Final tem sua vaga garantida entre os oito classificados. A fase Final é disputada em formato de eliminatória simples. Nas quartas de final o primeiro colocado joga contra o oitavo, segundo contra o sétimo e assim por diante. As seleções eliminadas são classificadas levando em consideração o posicionamento na fase de Classificação, já as demais disputam as semifinais, tendo em sequência a disputa de ter-

¹basketball.realgm.com/

²www.handball-base.com

ceiro lugar e a grande final [Volleyball World, 2025].

3.2 Visão Geral

Este projeto foi implementado em quatro etapas principais. A primeira consistiu na construção da base de dados via *web scraping*, utilizando a biblioteca *Selenium* para a extração de dados do site <https://br.volleyballworld.com/>, utilizando a Liga das Nações de Voleibol de 2021 a 2025 (VNL), naipes feminino, como competição para a elaboração do modelo (conforme apresenta a Figura 1). Na figura, é possível ver algumas estatísticas das maiores pontuadoras da temporada 2025. Por exemplo, a atleta Wu M. J. da China, até o momento da consulta das estatísticas, tinha feito 255 pontos, sendo 227 de ataque, 15 de bloqueio e 13 de saque.

A segunda etapa compreendeu o tratamento, isto é, pré-processamento, limpeza e a estruturação dos dados para garantir o melhor modelo. A terceira etapa focou no desenvolvimento e treinamento dos modelos (*Random Forest*, *XGBoost*, *LightGBM* e *SVM*). Por fim, a quarta etapa foi dedicada à avaliação e otimização do modelo, integrando a análise de importância dos atributos para identificar os fatores determinantes nas previsões.

3.3 Construção da base de dados

A elaboração do conjunto de dados foi estruturada em duas subetapas distintas e complementares: a extração e consolidação dos dados estatísticos de desempenho e a estruturação do histórico de resultados e classificações.

3.3.1 Base Estatística

Esta etapa consistiu na extração de dados brutos referentes aos fundamentos técnicos das partidas. As tabelas apresentavam *rankings* individuais das melhores atletas em sete categorias (pontuadoras, atacantes, bloqueadoras, sacadoras, levantadoras, defensoras e receptoras (passadoras)), sendo necessário uma etapa de engenharia de dados para converter métricas individuais em métricas coletivas.

Por meio da biblioteca *Pandas*, os dados foram agrupados por seleção nacional. Para cada fundamento, foram aplicadas as funções de agregação *média*, *soma* e *mediana* sobre as colunas estatísticas, permitindo quantificar o desempenho global de cada país.

Dessa forma, o processo gerou tabelas específicas para cada fundamento, nas quais cada coluna original foi transformada nas métricas de agregação citadas anteriormente. Posteriormente, estas tabelas foram unificadas por meio de uma operação de junção, resultando em uma base de dados consolidada para cada temporada (2021 a 2025).

3.3.2 Base Resultados

Esta subetapa focou na obtenção dos desfechos das partidas e do posicionamento na classificação final das equipes. Por meio da aba *Classificação*, foi extraído o posicionamento final de cada seleção em cada temporada. Já da aba *Programação*, os resultados de jogos, desde a fase preliminar até a eliminatória.

A partir dessas fontes, desenvolveu-se um algoritmo em *Python* para estruturar um conjunto de dados unificado contendo o resultado de cada confronto, o vencedor, a temporada vigente e o *ranking* da equipe na edição anterior. Para a temporada de 2021, chamada de *marco zero*, os dados

de classificação prévia e os atributos estatísticos das equipes para o ano de 2021, foram preenchidos inicialmente com valores nulos (*NaN*), dada a inexistência de dados históricos anteriores no formato padronizado. Para garantir a integridade da matriz de atributos sem descartar os registros dessa temporada, esses valores foram tratados posteriormente por meio de imputação utilizando a média global calculada estritamente sobre o conjunto de treinamento.

3.4 Tratamento

Com os dados estatísticos e os resultados estabelecidos, iniciou-se à etapa de seleção de atributos (*feature selection*). Neste processo consistiu na criação de um dicionário de configuração denominado *CONFIG*, que permitiu a experimentação/teste de diferentes estratégias de combinação dos atributos disponíveis.

Foram testadas diversas combinações de métricas agregadas (soma, média e mediana), tendo como objetivo identificar o cenário que melhor representa as dinâmicas do voleibol moderno para o modelo preditivo. As variáveis selecionadas abrangem as principais concepções do jogo. O Código 1 apresenta os atributos estatísticos propostos.

Código 1: Configuração do array de *features* selecionadas.

```
FEATURES = [
    'pon_pontos_sum',
    'pon_pontos_ataque_sum',
    'pon_pontos_bloqueio_sum',
    'blo_rebotes_sum',
    'pon_pontos_saque_sum',
    'lev_bem_sucedido_sum',
    'def_defesas_sum',
    'rec_bem_sucedido_sum',
    'ata_erros_sum',
    'blo_erros_sum',
    'sac_erros_sum',
    'lev_erros_sum',
    'def_erros_sum',
    'rec_erros_sum'
]
```

Foi escolhida a função de agregação soma (*_sum*), com o objetivo de mensurar o volume de jogo e a consistência das seleções ao longo da competição. Uma vez que as seleções que avançam para as fases finais disputam um número maior de jogos, a métrica atua como um *fator de premiação* para gratificar a capacidade técnica, resiliência física e o desempenho superior necessário para estender a trajetória dentro do campeonato. Além disso, essa abordagem contribuiu diretamente para reduzir vieses nos dados, deixando o modelo mais balanceado e elevando o seu desempenho preditivo final.

Dessa maneira, o modelo passa a apresentar um sinal preditivo que associa esse volume de jogo bem-sucedido com a solidez competitiva de cada elenco.

A nomenclatura adotada segue o seguinte raciocínio: o prefixo identifica a tabela de origem dos dados brutos, a seção intermediária aponta a coluna específica da tabela sobre a qual a função foi aplicada, e o sufixo *_sum* representa qual a função de agregação utilizada. Este padrão é visto no atributo *pon_pontos_sum*, no qual a operação de soma foi executada

Classificação	Nome Do Atleta	Team	Pontos	Pontos De Ataque	Pontos De Bloqueio	Pontos De Saque
1	Wu M.J.	CHN	255	227	15	13
2	Ndiaye	FRA	233	190	20	23
3	Zhuang Y.S.	CHN	225	190	25	10
4	Wade	JPN	223	208	5	10
5	Egonu	ITA	220	188	22	10
6	Ishikawa	JPN	216	192	5	19
6	B. Martinez	DOM	216	181	32	3

Figura 1. Disposição das tabelas, utilizando a temporada de 2025 como exemplo

sobre a coluna *pontos*, contida na tabela de melhores pontuadoras. Adicionalmente, este fluxo de engenharia de recursos foi replicado de forma independente para ambos os times, assim, a instância do atributo para o time A e para o time B gerou, respectivamente, as variáveis *a_pon_pontos_sum* e *b_pon_pontos_sum*.

As variáveis apresentam suas respectivas função táticas dentro de quadra. Para as ações de pontuação ativa o atributo *pon_pontos_sum* quantifica o total geral de pontos acumulados pela equipe, sendo este dividido em subcategorias que identificam a eficiência ofensiva por fundamento específico, os pontos convertidos por ações de ataque (*pon_pontos_ataque_sum*), o desempenho obtido nos bloqueios (*pon_pontos_bloqueio_sum*) e os pontos diretos obtidos a partir do saque (*pon_pontos_saque_sum*).

Em relação às variáveis de suporte e transição foram selecionados os indicadores que medem a sustentação do volume de jogo de cada elenco. O atributo *blo_rebotes_sum* representa os amortecimentos gerados pelo bloqueio que permitiram a continuidade do ponto, enquanto as variáveis *lev_bem_sucedido_sum*, *def_defesas_sum* e *rec_bem_sucedido_sum* avaliam, respectivamente, a precisão e qualidade técnica do levantamento para a construção ofensiva, a eficiência do sistema de defesa em segurar os ataques adversários e o índice de recepções e passes ideais na zona de controle técnico.

O bloco das variáveis de erro, responsável por demonstrar o desperdício de posse de bola e falhas operacionais das equipes. Esse grupo contabiliza as falhas diretas cometidas nos fundamentos de ataque (*ata_erro_sum*), infrações ou toques na rede no bloqueio (*blo_erro_sum*), saques desperdiçados na rede ou fora das delimitações da quadra (*sac_erro_sum*), além das perdas de precisão ou infrações de dois toques ocorridas durante os levantamentos (*lev_erro_sum*), defesas falhas (*def_erro_sum*) e recepções imperfeitas (*rec_erro_sum*).

3.4.1 Consolidação da Base de Dados

Durante a construção do conjunto de dados final, foi desenvolvida uma estrutura que unificou o histórico de confrontos aos dados de desempenho técnico. O processo iniciou-se com o alinhamento temporal dos dados, no qual cada partida da temporada T_i foi combinada com as estatísticas consolida-

das da temporada anterior (T_{i-1}). Essa estratégia, de atraso temporal, garante que o modelo realize as previsões baseadas em informações disponíveis antes do início do evento, mitigando o risco de vazamento de dados (*data leakage*). Para os casos de equipes sem histórico em T_{i-1} , as variáveis faltantes foram resolvidas via imputação pela média obtida a partir do conjunto de treino. Durante essa integração, realizou-se a padronização dos nomes das equipes por meio de mapeamentos predefinidos, garantindo apenas um formato de nomeação para os países. Com o intuito de proporcionar robustez estatística ao modelo, aplicou-se um filtro de permanência que manteve no conjunto de dados apenas as seleções com participação em pelo menos três temporadas. Caso um confronto programado envolva uma equipe desconsiderada por esse critério, a predição para aquela partida não será realizada, restringindo-se o escopo preditivo do estudo estritamente aos embates entre as dezesseis seleções validadas pelo filtro.

A distribuição histórica de participações das seleções no conjunto de dados, mapeada ao longo das edições da VNL, encontra-se na Tabela 1. A listagem apresenta o identificador numérico, gerado via *Label Encoding* no pré-processamento, o nome da respectiva seleção e o número total de temporadas disputadas.

As linhas, destacadas em verde claro, identificam os 16 times que atingiram o critério de consistência tática mínima de três ou mais temporadas estabelecido, garantindo uma base de dados robusta e representativa da elite do voleibol. Por outro lado, as linhas sem preenchimento denotam as seleções desconsideradas por apresentarem um volume histórico de dados insuficiente na série temporal analisada.

Além da utilização das estatísticas de ambas as equipes, o processo de engenharia de atributos introduziu o cálculo dos diferenciais de desempenho, denominados Deltas (Δ). A vantagem competitiva relativa é calculada conforme a Equação 1:

$$\Delta_{atributo} = Atributo_A - Atributo_B \quad (1)$$

Onde $Atributo_A$ e $Atributo_B$ representam os valores técnicos das equipes A e B, respectivamente. Por exemplo, o cálculo de *delta_pon_pontos_sum* foi definido a partir de $a_pon_pontos_sum - b_pon_pontos_sum$. O Código 2 apre-

Tabela 1. Mapeamento de participações das seleções e critério de consistência tática.

ID	Seleção	Temporadas que participaram
0	Bélgica	3
1	Brasil	5
2	Bulgária	4
3	Canadá	5
4	China	5
5	República Dominicana	5
6	Alemanha	5
7	Itália	5
8	Japão	5
9	Coreia	5
10	Países Baixos	5
11	Polônia	5
12	Sérvia	5
13	Tailândia	5
14	Turquia	5
15	Estados Unidos	5
-	França	2
-	Tchêquia	1
-	Croácia	1
-	Rússia	1

senta os atributos gerados utilizando a Equação 1.

Código 2: Configuração do *array* de *features* com a utilização dos Deltas.

```
DELTA_FEATURES = [
'delta_pon_pontos_sum',
'delta_pon_pontos_ataque_sum',
'delta_pon_pontos_bloqueio_sum',
'delta_blo_rebotes_sum',
'delta_pon_pontos_saque_sum',
'delta_lev_bem_sucedido_sum',
'delta_def_defesas_sum',
'delta_rec_bem_sucedido_sum',
'delta_ata_erros_sum',
'delta_blo_erros_sum',
'delta_sac_erros_sum',
'delta_lev_erros_sum',
'delta_def_erros_sum',
'delta_rec_erros_sum'
]
```

A utilização desse cálculo possibilita ao modelo identificar a superioridade técnica entre os competidores em uma escala linear. Um valor de Δ positivo indica que a equipe A possui um desempenho quantitativamente superior à equipe B no fundamento analisado, sugerindo uma vantagem competitiva. De modo inverso, valores negativos sinalizam que a equipe B dispõe de indicadores mais favoráveis, invertendo o favoritismo naquele atributo específico. Nos casos em que o valor do Δ é nulo ou próximo de zero, entende-se um cenário de equilíbrio técnico entre os competidores no fundamento em questão, ou a inexistência de dados históricos prévios para ambas as equipes. É importante destacar que a matriz final de atributos fornecida para o treinamento dos modelos não se limitou aos diferenciais técnicos. O conjunto completo de variáveis preditoras foi composto pela junção de todas as métricas de Deltas apresentadas no Código 2, combinadas aos identificadores das equipes envolvidas no confronto (*team_a* e *team_b*) e às suas respectivas classificações na temporada anterior (*rank_a_prev* e *rank_b_prev*). Adici-

onalmente, a variável alvo (*target*) foi mapeada para indicar o resultado real de vitória ou derrota da partida.

Para o tratamento de dados ausentes foram utilizadas informações provenientes do conjunto de treinamento, evitando a contaminação por dados futuros. As lacunas nos *rankings* foram preenchidas com a média histórica da própria seleção ou, na ausência desta, pela média global do treinamento. Da mesma forma, atributos técnicos faltantes foram imputados com a média aritmética das respectivas variáveis no conjunto de treino. Por fim, a variável alvo (*target*) foi estruturada de forma binária, representando o resultado da partida em relação à equipe A, ou seja, 1 representando vitória do time A e 0 vitória do time B.

3.5 Modelos Selecionados

É possível definir o Aprendizado de Máquina como um campo da Ciência da Computação com o objetivo de fazer computadores aprenderem (Duarte and Ståhl [2019]). Baseado na criação de modelos que, em diferentes modos, reconhecem padrões de aprendizagem para retornar um resultado de um objetivo proposto. No processo de Aprendizado de máquina, a saída é fornecida ao computador junto com os dados de entrada, resultando em um modelo preditivo que, ao receber novas entradas, é capaz de gerar as respectivas predições (Duarte and Ståhl [2019]). Quando um rótulo está presente no conjunto de dados, a tarefa de aprendizado é denominada supervisionada, caso contrário, é chamada de não supervisionada³. Na aprendizagem supervisionada, os rótulos dos conjuntos podem ser discretos (classes) ou contínuos (valores numéricos), o que torna viável aplicar uma abordagem de Classificação ou Regressão.

Nesse contexto, para este projeto será trabalhado o Aprendizado Supervisionado de Classificação, em virtude da utilização de classes para a construção do modelo, e para proporcionar uma avaliação mais completa e alinhada com os objetivos propostos. Foram escolhidos os algoritmos *Random Forest*, *XGBoost*, *LightGBM* e *SVM* (*Support Vector Machine*) para a realização dos testes empíricos. Em seguida, apenas os dois primeiros foram mantidos no escopo final por apresentarem desempenho superior e maior consistência nas métricas avaliadas.

Para a construção do modelo foram definidas quatro temporadas como o conjunto de treinamento (2021 – 2024) e uma temporada para o conjunto de teste (2025).

3.5.1 Random Forest

Random Forest combina um número de árvores de decisão que formam um comitê para decidir qual é a classe mais adequada para um determinado exemplo de entrada [Salman et al., 2024]. Emprega uma seleção aleatória de amostras e características. Como passo inicial, é obtida uma quantidade de dados escolhida de forma aleatória da amostra de treinamento nos dados da amostragem original, além disso, um subconjunto de características é escolhido para construir a árvore de decisão. Com a utilização dessas duas maneiras de aleatorização, resulta uma diminuição da correlação entre cada árvore, mitigando o potencial erro causado por *overfit*-

³Existem outras classes de algoritmos de Aprendizado de Máquina, como o semi-supervisionado, por reforço e sistemas de recomendação. Essas não serão discutidas neste trabalho

ting, ou seja, o modelo tem um bom desempenho no conjunto de treinamento, mas decai no conjunto de teste [Salman *et al.*, 2024]. Para a definição da classe para um dado exemplo, é escolhida a classe majoritária calculada por cada árvore da floresta.

Possui como vantagem a versatilidade, podendo ser aplicado em problemas de Regressão ou de Classificação. O *Random Forest*, geralmente, produz resultados superiores mesmo na ausência de ajuste de hiperparâmetros, isto é, fazendo o uso de valores padrão. Como todos os modelos baseados em árvores, o modelo gerado pode ser compreendido de maneira simples. Além disso, consegue superar a dificuldade de *overfitting*, em virtude de fazer o uso da agregação de múltiplas árvores de decisão, o que resulta na redução de viés e volatilidade. Possui como principal restrição o grande número de árvores, o que proporciona um tempo de treinamento prolongado, tornando-o lento e com baixa eficiência para previsões em tempo real. Em geral, apresenta rápida velocidade de treinamento, entretanto, com uma velocidade de previsão mais lenta após o treinamento.

3.5.2 XGBoost

Método baseado na implementação de árvores de decisão com *gradient-boosted*, geralmente utilizado em cenários supervisionados, tanto em problemas de Regressão quanto de Classificação. O algoritmo opera combinando aprendizes fracos, como árvores de decisão, de modo sequencial, em que cada novo aprendiz corrige efetivamente os erros cometidos pelos anteriores [Wiens *et al.*, 2025].

Nesse contexto, tem como objetivo minimizar uma função de perda, construindo árvores sequencialmente para ajustar o gradiente negativo dessa função em relação às previsões. Além disso, apresenta-se vantajoso devido à sua capacidade de lidar com grandes conjuntos de dados, sua robustez contra *overfitting* e sua flexibilidade de ajuste através de diversos hiperparâmetros. Inclui recursos como validação cruzada integrada, além de se beneficiar pela capacidade de ser executado em múltiplos núcleos durante o treinamento, tornando o tempo de computação muito mais acelerado, o que possibilita ser escalável para conjuntos de dados maiores Wiens *et al.* [2025].

3.6 Otimização do Modelo

Para a otimização dos modelos e a escolha dos melhores hiperparâmetros, foi definida como estratégia a utilização do *framework Optuna*⁴, um software de código aberto para otimização de hiperparâmetros [Akiba *et al.*, 2019]. O *framework* opera tendo a busca de hiperparâmetros como um processo de minimizar ou maximizar uma função objetivo, a qual recebe um conjunto de parâmetros como entrada e retorna uma pontuação baseada em métricas de validação.

A utilização deste *framework* permitiu uma busca de modo automatizado e eficiente, superando as limitações de métodos manuais e/ou exaustivos. O processo de otimização foi conduzido para encontrar o equilíbrio ideal entre viés e variância, garantindo que os modelos *XGBoost* e *Random Forest* atingissem sua melhor performance tendo como base o conjunto de dados. Os hiperparâmetros resultantes dessa

busca, que refletem as configurações mais bem avaliadas, estão na Tabela 2.

Tabela 2. Melhores Hiperparâmetros por modelo.

Modelo	Hiperparâmetros
XGBoost	n_estimators: 362 max_depth: 6 learning_rate: 0,0018 subsample: 0,8545 colsample_bytree: 0,9167 gamma: 2,6916 scale_pos_weight: 1,0159
Random Forest	n_estimators: 735 max_depth: 30 min_samples_split: 7 min_samples_leaf: 5 max_features: None criterion: 'gini'

Para o modelo *XGBoost*, destaca-se o uso de uma taxa de aprendizado (*learning_rate*) reduzida (0,0018), combinada com um número moderado de estimadores (362). Esses valores forçam o modelo a aprender de forma mais lenta e robusta, corrigindo erros residuais de maneira incremental. O parâmetro *gamma* (2,6916) atua como um controlador, tendo como instrumento a partição das árvores para evitar a criação de nós que não contribuam para a redução da perda. Ademais, o ajuste do *scale_pos_weight* foi escolhido para equilibrar a importância das classes no conjunto de dados.

Para o *Random Forest*, a profundidade máxima das árvores (*max_depth*) foi estabelecida em 30, permitindo que o modelo capture interações complexas entre os fundamentos técnicos. Entretanto, para manter a generalização, foram estabelecidos limites mínimos de amostras por folha (*min_samples_leaf*) e para divisão de nós (*min_samples_split*). O uso do critério de impureza de *Gini* auxiliou na seleção das melhores divisões estatísticas para a classificação dos resultados de vitória ou derrota.

Para garantir a assertividade e a confiança do modelo, é necessário avaliá-lo [Duarte and Ståhl, 2019]. As métricas de avaliação de desempenho, para modelos de classificação, são baseadas na contagem de quantas classes previstas são iguais às que foram observadas. Para o cenário em que as classes estão balanceadas, no qual o número de classes está distribuído homogeneamente pelo conjunto de dados, a acurácia desponta como a melhor métrica. Nessa métrica, as classes preditas são comparadas com as reais, sendo o resultado definido pelo percentual de acerto.

Na avaliação de modelos não balanceados, diversas métricas foram propostas, com três ganhando mais relevância: precisão, *recall* e *F1-score*. A seguir suas definições matemáticas, junto com a apresentação da acurácia.

$$Precisão = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1\text{-score} = \frac{2 \times precisão \times recall}{precisão + recall}$$

⁴optuna.org/

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN}$$

A precisão é utilizada quando o classificador está cobrindo as classes positivas com confiança, calculando a exatidão das previsões positivas, o recall (completude), possui o papel de indicar quantos positivos reais foram encontrados, já o F_1 -score fornece a média harmônica de precisão e recall.

4 Resultados

Nesta seção será apresentado os desempenhos que foram alcançados pelos modelos *Random Forest* e *XGBoost*, após o processo de otimização. O objetivo final foi verificar qual dos modelos melhor correspondia aos dados relacionados ao voleibol moderno. Na Tabela 3 é possível visualizar o resultado das principais métricas obtidas.

Tabela 3. Comparativo de desempenho entre os modelos finalistas.

Modelo	Acurácia	Precisão	Recall	F_1 -score
Random Forest	77,17%	77%	77%	77%
XGBoost	79,35%	84%	79%	80%

Os dois modelos apresentaram resultados satisfatórios e semelhantes, com o *XGBoost* levando uma pequena vantagem em todas as métricas. Enquanto o *Random Forest* atingiu uma acurácia de 77,17%, o *XGBoost* obteve 79,35%. De modo geral, é possível dizer que o segundo modelo consegue prever corretamente 8 a cada 10 vencedores.

A Precisão do *XGBoost* (84%) desponta como um indicador de elevada confiabilidade preditiva, visto que o valor indica alta confiança quando é apontado vitória de uma equipe, isto é, quando o modelo aponta a vitória de uma determinada seleção, a chance de o evento acontecer é expressiva, minimizando a ocorrência de falsos positivos. O F_1 -score (80%) reflete o equilíbrio, confirmando que o aprendizado é sólido sobre os diferenciais técnicos das seleções.

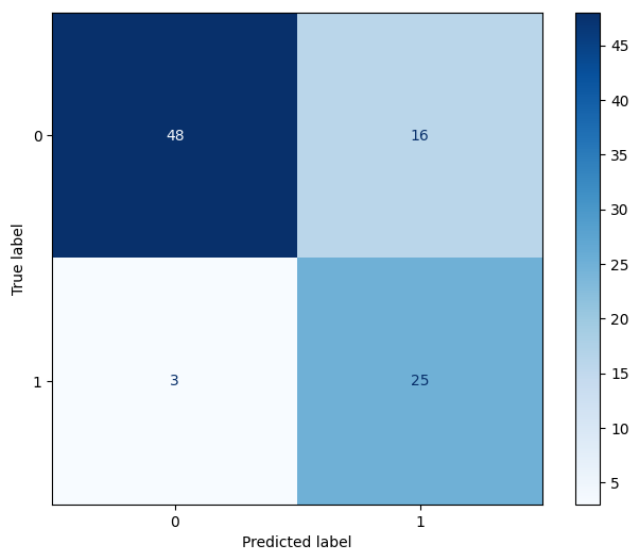


Figura 2. Matriz de Confusão do modelo *XGBoost*.

Observando a Figura 2, é destacado que dos 51 resultados previstos como verdadeiros para a vitória do time B (classe 0), o modelo conseguiu acertar 48, em contrapartida

dos 41 em que foi previsto vitória do time A (classe 1) acertou 25.

A concentração de valores na diagonal principal da matriz de confusão detalha o comportamento preditivo do modelo, revelando uma assimetria entre as classes mapeadas. Nesse contexto, a Classe 0 apresentou menor complexidade de mapeamento pelo modelo, obtendo uma taxa de acerto superior em relação à Classe 1. É possível identificar apenas 3 erros em relação a vitória do time B, indicando um *recall* elevado para a Classe 1, o que evidenciou a capacidade do modelo de identificar padrões associados a vitória do time B. Os 16 erros para vitória do time A, proporcionam a interpretação que o modelo está otimista demais para a Classe 1. No contexto esportivo, pode representar as seleções que são favorecidas em relação as estatísticas, mas fatores como erros ou pressão, ocasionam a derrota.

A seguir, na Figura 3, é possível observar a matriz do modelo utilizando *Random Forest*. Destaca-se uma divergência de comportamento preditivo quando comparado ao modelo *XGBoost*. Das 63 previsões realizadas, o modelo obteve 53 acertos, além de contar com 10 erros para vitória do time B (Classe 0). Em contrapartida, para a vitória do time A é adquirido uma quantidade mais enxuta de acertos, somando apenas 18 em 29 possíveis. Nesse sentido, constata-se que o valor da acurácia é sustentado pelo desempenho na Classe 0, mascarando uma baixa confiabilidade preditiva quando o modelo sinaliza a vitória do time A.

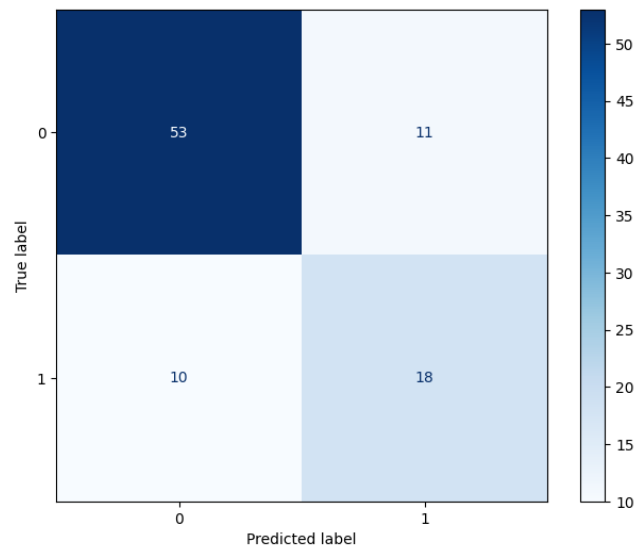


Figura 3. Matriz de Confusão do modelo *Random Forest*.

5 Discussão

Um dos principais objetivos deste trabalho é constatar, por meio de evidências estatísticas, quais fundamentos do voleibol moderno apresentam maior relevância para a predição da vitória.

Conforme ilustrado na Figura 4, observa-se que a classificação das equipes no ranking da temporada anterior (*rank_a_prev* e *rank_b_prev*) destaca-se como o atributo de maior peso para o modelo. Este atributo indica que a consistência histórica e a tradição das seleções ainda exercem forte influência no resultado esperado. Entretanto, ao des-

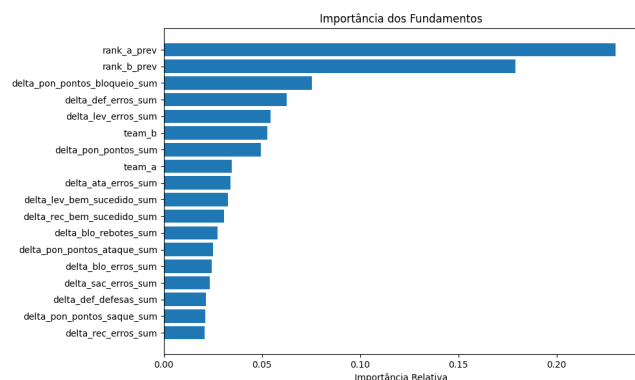


Figura 4. Importância dos atributos para o modelo *Random Forest*.

locar o foco para as ações técnicas executadas em quadra, a variável *delta_pon_pontos_bloqueio* surge como o principal indicador esportivo, apontando o bloqueio como fator determinante para garantir a vitória.

Na sequência, desconsiderando-se os atributos de erro, o levantamento e a recepção despontam como as ações principais no topo da hierarquia do modelo, representados pelas variáveis *delta_lev_bem_sucedido* e *delta_rec_bem_sucedido*, respectivamente. Vale a pena notar que as variáveis que mensuram falhas no jogo demonstram que o desfecho da partida é sensível à minimização de erros, especificamente nos fundamentos de defesa, levantamento e ataque, nesta ordem de importância relativa para o algoritmo.

Um ponto relevante diz respeito ao fundamento do saque. Embora, frequentemente citado pelo dito popular como uma das ações mais cruciais e determinantes do voleibol, ele ocupou apenas a penúltima colocação no quesito de importância para o modelo de *Random Forest*. Isso pode indicar que, embora o saque seja um facilitador para o ponto, a estabilidade defensiva e a eficiência do sistema de bloqueio possuem uma correlação estatística mais direta com o resultado final da partida na VNL.

Para o modelo *XGBoost*, os resultados apresentam determinada semelhança, como esperado, conforme indicado na Figura 5. Observa-se que, no *ranking* de importância, a classificação das seleções no ano anterior manteve-se como o atributo dominante. Isso consolida a premissa de que, a trajetória e a tradição de uma equipe permanecem como fatores cruciais para a previsibilidade de resultados na VNL. O fundamento do bloqueio (*delta_pon_pontos_bloqueio_sum*) reafirmou-se como a variável mais determinante para a predição de vitória.

Contudo, o topo das ações técnicas mais bem avaliadas pelo *XGBoost* apresenta atributos que o diferem do modelo anterior. Destacam-se as variáveis de rebotes de bloqueio (*delta_blo_rebotes_sum*) e a pontuação total de ataque (*delta_pon_pontos_sum*), sugerindo que este algoritmo prioriza a capacidade de transição e a eficiência ofensiva. Ademais, o modelo atribui alta relevância aos erros cometidos nos fundamentos de defesa e levantamento, indicando que a instabilidade nesses setores possui um impacto direto na probabilidade de derrota.

Por fim, o fundamento do saque mantém sua baixa posição entre os atributos, confirmando a tendência observada de que sua influência estatística é inferior à consistência do

sistema de rede e controle de erros.

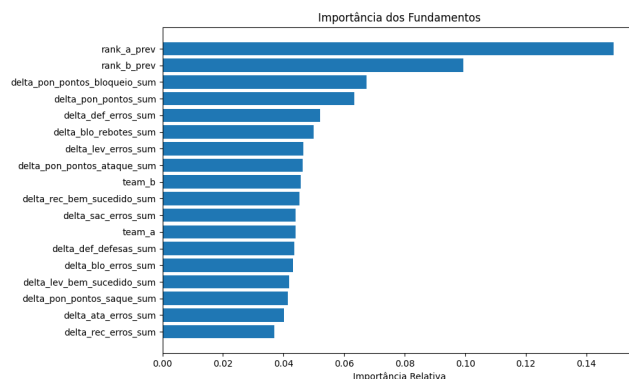


Figura 5. Importância dos atributos para o modelo *XGBoost*.

É possível realizar uma comparação entre os modelos, destacando a convergência em relação aos fatores determinantes de sucesso no voleibol. O atributo de *ranking* presente no topo de ambos os modelos confirma que o histórico de desempenho das seleções possui papel de uma importante variável preditora, abrangendo fatores como qualidade técnica do elenco e experiência em jogos de alto calibre.

Em relação às ações técnicas, o atributo de rebote de bloqueio (*delta_blo_rebotes_sum*) e o da pontuação total de ataque (*delta_pon_pontos_sum*) no *XGBoost* demonstra mais segurança nas predições, focando na capacidade de uma equipe em construir novas oportunidades de ataque e converter as ações técnicas em pontuação. O *Random Forest* evidenciou a importância das ações de levantamento e recepção, já o *XGBoost* aponta para um equilíbrio entre a defesa e o ataque.

Observando a baixa relevância do saque nos modelos, constata-se que embora esse fundamento possa gerar vantagem momentânea, a vitória é consolidada por meio da resiliência do sistema de bloqueio e da minimização de erros. Ademais, a transição do *Random Forest* para o *XGBoost* apresentou melhora nas métricas de avaliação e proporcionou uma visão mais dinâmica e abrangente sobre esse esporte tão popular, em que seu desfecho é obtido pela eficiência na rede e pela estabilidade técnica sob pressão.

Para melhor compreensão dos resultados obtidos pelos modelos, é utilizada a ferramenta SHAP (*SHapley Additive exPlanations*). Essa abordagem é definida por Lundberg and Lee [2017] como um *framework* metodológico que define a atribuição de importância dos atributos como um problema de modelagem aditiva de recursos lineares. Tem como premissa principal calcular a contribuição de cada variável para a previsão final do modelo, assegurando propriedades matemáticas de consistência e completude. O SHAP mensura o quanto o valor de um fundamento específico afastou a predição da partida em relação à base de previsão média esperada para o conjunto de dados.

O gráfico do tipo *Beeswarm*, apresentado na Figura 6, dispõe as variáveis no eixo vertical ordenadas de forma decrescente com base em sua importância para o modelo. No eixo horizontal, representa-se o valor SHAP, onde pontos posicionados à direita da linha de origem (valores positivos) indicam uma contribuição direcionada para a previsão de vitória do time A (Classe 1), enquanto pontos à esquerda (valores

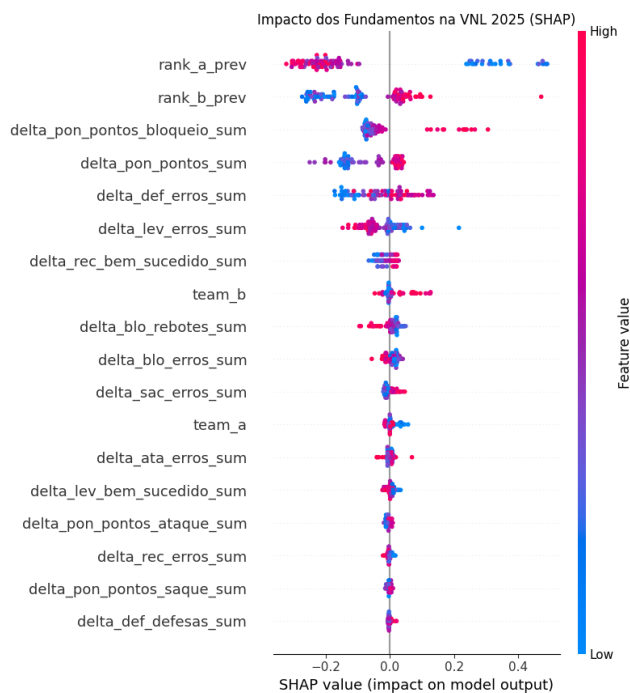


Figura 6. Impacto dos Fundamentos (SHAP) no modelo XGBoost.

negativos) empurram o modelo a prever a vitória do time B (Classe 0). A gradação de cores indica o valor do atributo, variando do azul (valores baixos) ao vermelho (valores altos).

É possível notar que os atributos de maior impacto preditivo mantêm-se associados ao *ranking* prévio das seleções (*rank_a_prev* e *rank_b_prev*). No caso de *rank_a_prev*, os pontos azuis estão concentrados à direita (SHAP positivo). Como o *ranking* é inversamente proporcional à qualidade técnica da equipe, isto é, valores baixos como 1, 2 e 3 indicam as melhores seleções do mundo, essa dispersão evidencia que um *ranking* numericamente baixo para o time A atua como um forte indutor para a previsão de sua vitória. De modo inverso, os pontos vermelhos que estão no lado esquerdo, os quais são valores numéricos altos, representam posições inferiores no *ranking*, o que diminui as chances de vitória do Time A. Para a variável *rank_b_prev*, ocorre o inverso, os valores baixos (pontos azuis à esquerda) empurram a predição em direção à derrota do Time A, enquanto valores altos (pontos vermelhos à direita) favorecem o Time A.

No quesito das ações técnicas, o saldo acumulado de pontos de bloqueio (*delta_pon_pontos_bloqueio_sum*) estabeleceu-se como o fator mais determinante para desequilibrar o comportamento do modelo. Há uma dispersão de pontos vermelhos espalhando-se de forma alongada à direita do ponto central, indicando que quando o Time A possui um volume de bloqueios superior ao time B ao longo da competição, é obtido um maior grau de confiança estatística para a vitória do time A.

O gráfico do tipo *Waterfall* complementa o anterior, demonstrando foco na interpretabilidade local, isto é, explica exatamente o comportamento do XGBoost para prever o resultado de uma única partida. Como é possível visualizar na Figura 7, são apresentados os detalhes da predição da partida final da VNL 2025, disputada entre Itália e Brasil em 27 de julho de 2025. Na ocasião, o resultado foi de 3x1 para a seleção italiana com parciais de 22x25, 25x18, 25x22 e 25x22,

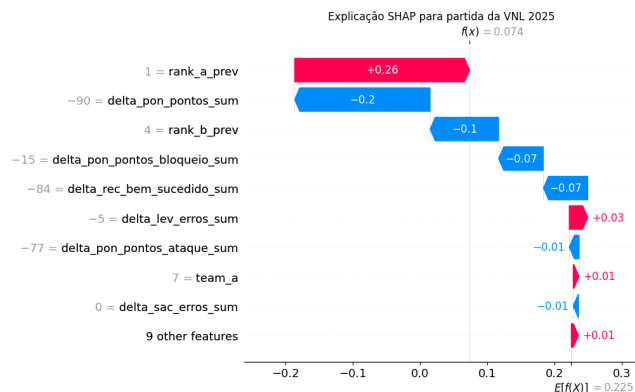


Figura 7. Explicação SHAP para partida da VNL no modelo XGBoost.

sendo o primeiro set vencido pela equipe brasileira e os demais pela Itália. O gráfico indica como o modelo ponderou os atributos para tomar a decisão nessa partida isolada.

É apresentado o ponto de partida na base inferior direita, indicado pelo valor de $E[f(X)] = 0.225$, o que representa a expectativa média de previsão do modelo antes de olhar para as características específicas da partida, indicando leve tendência histórica a favor da vitória do time A (classe 1). Nesse sentido, o valor de $f(x) = 0.074$ (topo central) é o retorno final do modelo após analisar todas as variáveis. Embora o resultado permaneça positivo, mantendo a predição da vitória do time A, a queda em relação ao valor base indica que as estatísticas de quadra reduziram significativamente a convicção inicial do algoritmo.

É válido destacar a contribuição de cada variável no desenvolvimento do resultado. Observa-se que o time A ocupando a primeira colocação do *ranking* prévio (*rank_a_prev* = 1), aumentou a sua pontuação preditiva em 0.26, enquanto o fato de enfrentar o quarto colocado (*rank_b_prev* = 4) penalizou o Time A em apenas -0.1, evidenciando o equilíbrio no confronto e do peso da tradição institucional na tomada de decisão do modelo.

Nas funções técnicas a variável *delta_pon_pontos_bloqueio_sum* apresentou um valor de -15, indicando que o time A tem um saldo acumulado de pontos marcados a menos do que o time B ao longo da competição, o que gerou uma penalização para o time A em -0.07. A variável *delta_rec_bem_sucedido_sum*, com valor de -84, custou uma penalização de -0.07 na confiança do modelo devido ao saldo inferior de pontos de recepções (passes) bem-sucedidos do time A. Por fim, a instabilidade do levantamento teve impacto contra a equipe mandante, em que a variável *delta_lev_erros_sum* registrou o valor de -5, adicionando 0.03 ao valor final da predição e consolidando o resultado do modelo orientado à vitória do time A.

Na Figura 8, em que é apresentado o impactos dos fundamentos para o modelo *Random Forest*. Observa-se a manutenção da concordância analítica em relação aos fatores principais identificados pelo XGBoost, ocorrendo divergências sutis apenas na atribuição dos pesos preditivos.

As três variáveis principais se mantiveram (*rank_a_prev*, *rank_b_prev* e *delta_pon_pontos_bloqueio_sum*). A dinâmica das cores também preservou o comportamento padrão de interpretabilidade.

Nas ações técnicas, o saldo acumulado de pontos de blo-

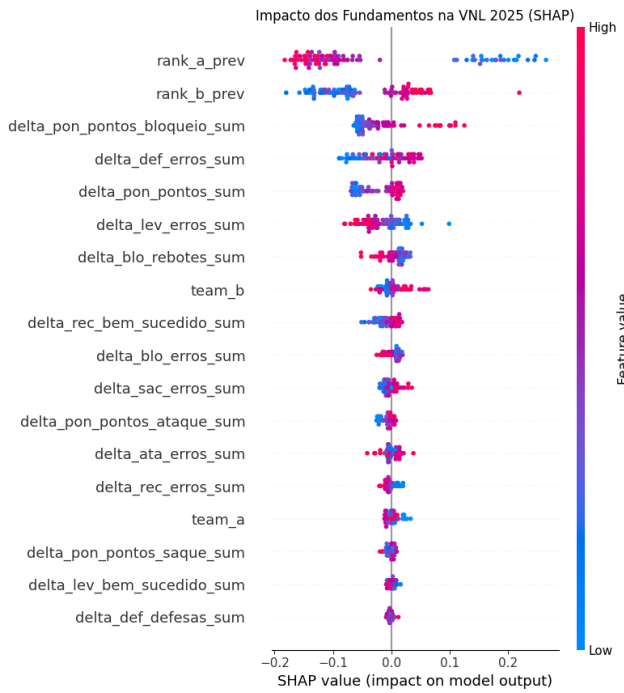


Figura 8. Importância dos atributos para o modelo *Random Forest*.

queio consolidou-se como o fundamento técnico de maior impacto positivo para a probabilidade de vitória do time A (Classe 1), evidenciado pela concentração de valores altos (pontos vermelhos) à direita do gráfico. Entretanto, uma análise comparativa revela alterações na hierarquia das variáveis secundárias. Houve uma inversão de posições entre os atributos *delta_def_erros_sum* e *delta_pon_pontos_sum*, em relação ao mesmo tipo de gráfico do modelo *XGBoost*. A variável de erros defensivos subiu para o quarto lugar e o saldo de pontos marcados decaiu para a quinta posição.

Adicionalmente, destaca-se o comportamento do eixo horizontal (*SHAP value*). Enquanto o *XGBoost* apresenta limites de impacto levemente mais espaçados, o *Random Forest* exibe uma escala de dispersão mais contida, variando entre -0.2 e $+0.2$. A variação do *XGBoost* é explicada pelo fato de o modelo ser baseado em árvores sequenciais, que corrigem os erros das anteriores, tornando-o mais agressivo, ao distribuir pesos maiores para as variáveis. No *Random Forest*, o funcionamento ocorre por meio da votação paralela das árvores independentes, o que gera um efeito de suavização e o torna mais conservador na atribuição dos pesos.

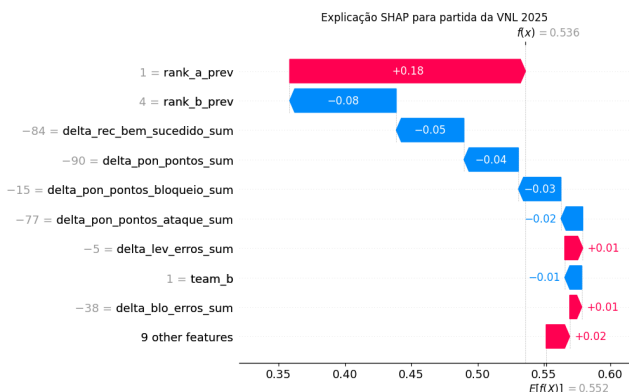


Figura 9. Explicação SHAP para partida da VNL no modelo *Random Forest*.

O gráfico da Figura 9 (*Waterfall*) complementa a análise global ao focar na interpretabilidade local do modelo *Random Forest*, detalhando como o algoritmo ponderou os atributos para prever o resultado da mesma partida detalhada pelo modelo anterior. Como grande trunfo do time A, destaca-se a variável *rank_a_prev*, visto que a seleção terminou a temporada passada como número 1 do mundo, dando um grande empurrão ($+0.18$) na probabilidade de vitória. O time B, no entanto (atual quarto colocado) reduziu a vantagem adversária em 0.08 por sua posição no ranking.

Esse resultado consolida a perspectiva de um confronto extremamente equilibrado, no qual o time A sustenta o favoritismo por uma margem mínima. Como fator relevante para a manutenção dessa condição, destaca-se a variável *rank_a_prev* = 1. O fato de o Time A ter encerrado a temporada anterior como líder do *ranking* mundial adicionou um acréscimo de 0.18 na probabilidade de vitória da equipe. Em contrapartida, a posição histórica do adversário (*rank_b_prev* = 4), atual quarto colocado do mundo, atuou ativamente para mitigar essa vantagem, aplicando uma penalização de -0.08 sobre o time A.

No âmbito das ações técnicas, o time B apresentou superioridade estatística, exercendo forte pressão sobre o favoritismo do líder. O volume superior de passes eficientes do oponente (*delta_rec_bem_sucedido_sum* = -84) subtraiu -0.05 da confiança no time A, dado que a precisão na recepção facilita as opções de armação do levantador adversário e equilibra o confronto. Ademais, o modelo aplicou penalizações de -0.04 devido ao saldo de pontos marcados (*delta_pon_pontos_sum* = -90), -0.03 pelo desempenho superior no bloqueio (*delta_pon_pontos_bloqueio_sum* = -15) e -0.02 pela dominância ofensiva do time B (*delta_pon_pontos_ataque_sum* = -77).

Diante desse cenário de pressão técnica, o time A conseguiu sustentar sua vantagem ao compensar os deslizos por meio da disciplina tática e do controle de erros. A maior precisão na distribuição de jogo (*delta_lev_erros_sum* = -5) e a redução de falhas ofensivas, que resultariam em pontos diretos para o adversário ao utilizar o bloqueio (*delta_blo_erros_sum* = -38), adicionaram, cada uma, $+0.01$ à equação final do modelo, evitando a perda de pontos por falhas táticas e assegurando a predição final voltada à vitória do time A.

6 Conclusão

O presente trabalho avaliou a capacidade dos modelos *Random Forest* e *XGBoost* na predição de resultados para partidas da VNL feminina, a partir de desempenho individuais das atletas, as quais foram convertidas em desempenho da equipe. Os testes mostraram que os modelos foram capazes de identificar o vencedor da partida produzindo resultados consistentes, apresentando um F_1 -score de 77% e Acurácia de 77,17% para o *Random Forest* e F_1 -score de 80% e Acurácia de 79,35% para o *XGBoost*. Contudo, a análise demonstrou desequilíbrio na capacidade de generalização inter-classe, com o algoritmo *XGBoost* demonstrando maior equilíbrio preditivo. Além disso, os resultados corroboraram a análise dos fatores determinantes para o sucesso do voleibol moderno, o qual apresenta-se como um dos principais obje-

tivos propostos.

A partir dos *insights* obtidos é válido destacar as ações técnicas evidenciadas pelos modelos. O bloqueio apresentou-se como o principal fundamento para ambos os modelos, isto é, em virtude do consenso, é possível confirmar que este trabalho apresenta o bloqueio como o principal ação técnica para a vitória de uma seleção feminina na VNL.

Para estabelecer os três principais fundamentos, foi exposto certas divergências entre os modelos, excluindo para a análise os atributos de erro. O *Random Forest* considerou as ações de levantamento e recepção como mais importantes, sendo possível notar a importância oferecida em relação ao jogo tático de consistência defensiva e de transição, enfatizando um bom passe para a realização do levantamento, além de focar no ponto defensivo citado, o bloqueio. O modelo *XGBoost*, por outro lado, considerou os fundamentos com as ações de rebotes e os clássicos pontos de ataque, o que sugere um estilo jogo agressivo e de insistência ofensiva.

Apesar de que exista a premissa popular de que indicadores estatísticos não *entram em quadra* e nem definem o sucesso de confrontos, o Voleibol, assim como o Basquete e o Handebol, consolidou-se como um esporte no qual o embasamento analítico está cada vez mais presente, como visto nos trabalhos mencionados ao longo dessa leitura. O propósito deste trabalho consistiu em evidenciar esse panorama, demonstrando que, embora os resultados preditivos não sejam deterministas ou soberanos perante a imprevisibilidade do esporte, os dados estatísticos configuram-se como uma ferramenta de suporte e aprimoramento tático para as equipes, não sendo utilizados de maneira isolada.

O estudo conclui que os modelos são capazes de prever o resultados das partidas propostas e possibilitar a análise de fatores determinantes. Entretanto, seu uso prático necessita planejamento, com ações de checagem, limpeza e tratamento dos dados, conforme o objetivo de cada aplicação.

Para trabalhos futuros, propõe-se a combinação de novos atributos e funções de agregação, além da incorporação da biblioteca *sklearn.inspection* para possibilitar uma nova perspectiva aprofundada sobre a interpretabilidade dos modelos. Esses enfoques possuem potencial para enriquecer a análise do conjunto de dados obtido, incrementando de modo significativo a explicabilidade das decisões algorítmicas e a robustez das predições. Ademais, destaca-se a viabilidade de replicar a metodologia proposta para outras competições expressivas do cenário internacional e nacional, validando a generalização dos modelos em diferentes contextos competitivos.

Referências

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*, pages 2623–2631, New York, NY, USA. ACM. DOI: 10.1145/3292500.3330701.
- Duarte, D. and Ståhl, N. (2019). Machine Learning: A Concise Overview. In Said, A. and Torra, V., editors, *Data Science in Practice*, volume 46 of *Studies in Big Data*, pages 27–58. Springer International Publishing.
- Durão, D. N. (2025). Resumo histórico do voleibol: da sua criação à ascensão esportiva. Trabalho de conclusão de curso, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Câmpus de Rio Claro, Rio Claro, SP. Orientador: Paulo Henrique de Araújo Guerra.
- Federação Internacional de Voleibol (FIVB) (2021). *Regras Oficiais de Voleibol: 2021–2024*. Lausanne. Aprovado pelo 37º Congresso Mundial da FIVB. Tradução: Fernanda Barbosa. Formatação: Flavio Fukuda. Revisão: Pedro Casanova e COBRAV.
- Felice, F. and Ley, C. (2025). Predicting handball matches with machine learning and statistically estimated team strengths. *Journal of Sports Analytics*, 11:1–12. DOI: 10.1177/22150218251313937.
- Horvat, T. and Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIRES Data Mining and Knowledge Discovery*, 10(5):e1380. DOI: 10.1002/widm.1380.
- Lalwani, A., Saraiya, A., Singh, A., Jain, A., and Dash, T. (2022). Machine Learning in Sports: A Case Study on Using Explainable Models for Predicting Outcomes of Volleyball Matches. arXiv preprint arXiv:2206.09258.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 4765–4774.
- Nunes, B. G., Nunes, J., Novikoff, C., Dias Neto, J. M. M., and Triani, F. d. S. (2014). Voleibol, iniciação desportiva e seus benefícios dentro do PRODUC. *FIEP Bulletin*, 84. Edição Especial, Artigo I.
- Salman, H. A., Kalakech, A., and Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024:69–79. DOI: 10.58496/BJML/2024/007.
- Sanjaykumar, S., Lakshmi, P. Y., Natarajan, S., Kalmykova, Y., Lobo, J., Pavlović, R., and Setiawan, E. (2025). Machine learning analysis for predicting performance in female volleyball players in India: Implications for talent identification and player development strategies. *Journal of Human Sport and Exercise*, 20(1):207–215. DOI: 10.55860/cn2vdj44.
- Volleyball World (2025). Volleyball nations league. Acessado em: 18 mai. 2026.
- Wang, Y., Liu, W., and Liu, X. (2022). Explainable AI techniques with application to NBA game-play prediction. *Neurocomputing*, 483:59–71. DOI: 10.1016/j.neucom.2022.01.098.
- Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., and Burton, J. (2025). A tutorial and use case example of the eXtreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clinical and Translational Science*, 18:e70172. DOI: 10.1111/cts.70172.