Lorrayne Somerlatte dos Santos

Análise de Desempenho de Jogadores de Tênis Usando Dados Históricos

Belo Horizonte

2024

Lorrayne Somerlatte dos Santos

Análise de Desempenho de Jogadores de Tênis Usando Dados Históricos

Pesquisa Científica

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas – ICEx
Departamento de Ciência da Computação

Orientadora: Ana Paula Couto da Silva

Belo Horizonte

2024

Sumário

1 Introdução

Nos últimos anos, a evolução da Inteligência Artificial (IA) e sua capacidade de processar grandes volumes de dados com eficiência têm transformado profundamente diversas áreas, incluindo o esporte. No tênis, em particular, a análise de dados esportivos vai além de monitorar o desempenho dos atletas: ela fornece insights valiosos para treinadores e equipes técnicas, permitindo decisões estratégicas e embasadas. Cada jogador possui características únicas e um estilo de jogo próprio, e o uso de IA e modelos preditivos possibilita prever resultados de partidas, identificar padrões táticos e personalizar estratégias de treino de maneira mais assertiva (TAN, 2023).

A aplicação de modelos de aprendizado de máquina no tênis tem mostrado impacto significativo em áreas como a previsão de resultados, a prevenção de lesões e a análise tática, conforme destacado por (COHEN; SLOANE, 2015). Este estudo enfatiza a relevância de monitorar métricas como aces, quebras de serviço e percentuais de acerto de saque, variáveis diretamente associadas ao sucesso em quadra. A análise de dados históricos permite identificar padrões de desempenho e características individuais dos jogadores, fornecendo uma base sólida para o desenvolvimento de treinamentos adaptativos que maximizem o rendimento esportivo.

Explorando o potencial da análise de dados históricos no desempenho dos jogadores de tênis, esta pesquisa visa identificar quais fatores mais influenciam o resultado de uma partida e como essas variáveis podem ser utilizadas em modelos preditivos de alta precisão. A relevância deste estudo reside no potencial de aprimorar as metodologias de treino e a compreensão estratégica do tênis, contribuindo tanto para a ciência esportiva quanto para a prática do esporte.

Ao unir tecnologia com técnicas avançadas de análise de dados, esta abordagem não apenas proporciona uma vantagem competitiva aos jogadores, mas também estabelece um novo patamar para o treinamento esportivo, onde cada decisão é fundamentada em dados concretos e análises detalhadas. Além disso, ela promove a evolução contínua do esporte, ao permitir que treinadores e atletas entendam melhor os fatores que impactam diretamente o desempenho. Compreender essas variáveis, seja por meio de padrões táticos ou pela identificação de pontos fortes e fracos, pode ajudar a desenvolver estratégias mais eficazes e a otimizar o uso de recursos durante os treinamentos. Assim, o tênis se torna uma modalidade cada vez mais estratégica, desafiadora e acessível para inovações tecnológicas, criando um impacto significativo tanto no esporte profissional quanto em sua popularização e engajamento global (BODEMER, 2023).

1.1 Objetivo Geral

O objetivo geral deste estudo é desenvolver modelos preditivos baseados em dados históricos de tênis, capazes de analisar e prever o desempenho de jogadores. Para isso, serão utilizadas métricas de avaliação adequadas ao contexto esportivo, como precisão e recall, que são particularmente relevantes para problemas de classificação em que a identificação correta de vitórias e derrotas é crucial (BRANCO; TORGO; RIBEIRO, 2016). A precisão mede a proporção de previsões corretas entre todas as previsões feitas, enquanto o recall avalia a capacidade do modelo de identificar corretamente todos os casos positivos (vitórias, por exemplo). Essas métricas são essenciais para garantir que os modelos não apenas acertem as previsões, mas também minimizem falsos positivos e falsos negativos, que podem levar a decisões estratégicas equivocadas.

Os modelos serão testados e validados com o objetivo de identificar padrões estratégicos e variáveis-chave que influenciam o sucesso em quadra. Com isso, o estudo pretende contribuir para a ciência esportiva, fornecendo uma base sólida para que treinadores, atletas e equipes técnicas tomem decisões mais fundamentadas e estratégicas, baseadas em evidências quantitativas.

1.2 Objetivos Específicos

Os objetivos específicos deste trabalho incluem realizar uma análise detalhada de modelos de *Machine Learning*, com foco nos algoritmos *Random Forest* e *K-Nearest Neighbors* (KNN), para análise de desempenho no tênis, considerando suas características e adequação para a previsão de resultados e identificação de padrões de jogo. Implementar e conduzir testes dos modelos *Random Forest* e KNN, variando parâmetros como o número de árvores (*n_estimators*) e a profundidade máxima das árvores (*max_depth*) para o *Random Forest*, e o número de vizinhos (*n_neighbors*) para o KNN, a fim de identificar a configuração ideal para previsão de desempenho em partidas de tênis. Avaliar os modelos com base em métricas de desempenho, como precisão, recall e F1-score, comparando sua eficácia para o conjunto de dados históricos de tênis.

Além disso, preparar e pré-processar o conjunto de dados históricos de tênis, provenientes de fontes como Kaggle[1] [2], Tennis Abstract[3], garantindo a remoção de inconsistências e o balanceamento das classes para evitar vieses nos resultados preditivos. Treinar os modelos *Random Forest* e KNN usando o conjunto de dados preparado, explorando a influência de diferentes variáveis de desempenho, como *aces*, quebras de serviço, porcentagem de primeiro saque e erros não forçados, e analisando o impacto desses indicadores no sucesso dos jogadores. Realizar uma análise aprofundada do impacto dos hiperparâmetros no desempenho dos modelos, com especial foco na precisão das previsões e na capacidade dos modelos de identificar padrões

[1] <https://www.kaggle.com/datasets/dissfya/atp-tennis-2000-2023daily-pull>
[2] <https://www.kaggle.com/code/dissfya/wta-tennis-daily-update>
[3] <https://www.tennisabstract.com/>

relevantes no comportamento dos jogadores.

Finalmente, analisar e interpretar os resultados obtidos na experimentação, avaliando a eficácia e relevância dos modelos *Random Forest* e KNN para prever o desempenho dos jogadores de tênis e identificar os principais fatores de sucesso. Discutir as conclusões, destacando as variáveis mais relevantes para o desempenho, e propor diretrizes para a implementação prática dos modelos, visando auxiliar treinadores e equipes técnicas na elaboração de estratégias e planos de treinamento.

2 Referencial Teórico

Na literatura, a análise de dados esportivos tem se tornado cada vez mais relevante, impulsionada pelo avanço das técnicas de *Machine Learning* e *Deep Learning* (K.; K., 2023). Essas tecnologias possibilitam o desenvolvimento de modelos preditivos de alto desempenho para prever resultados e otimizar o desempenho dos atletas. A seguir, são apresentados estudos que fundamentam o uso dessas técnicas em contextos esportivos, com ênfase no tênis.

O estudo de (COHEN; SLOANE, 2015) explora a aplicação de modelagem preditiva e sistemas de recomendação em esportes como atletismo e basquete. Utilizando dados históricos, o estudo investiga o efeito do "momentum" e desenvolve modelos de previsão de desempenho. A análise desses dados revelou um erro médio de apenas 1,38 % nas previsões de atletismo, evidenciando a importância e o potencial dos modelos preditivos em outros esportes, como o tênis.

No artigo de (SAMPAIO et al., 2024), os autores revisam sistematicamente a aplicação de aprendizado de máquina para melhorar o desempenho no tênis. Eles destacam aplicações como o monitoramento psicológico dos jogadores, a previsão de resultados das partidas e a análise tática. O estudo sugere que o aprendizado de máquina pode ajudar treinadores e atletas a se adaptarem às demandas específicas do esporte, oferecendo insights valiosos para a otimização do desempenho e a prevenção de lesões.

Os autores em (CHMAIT; WESTERBEEK, 2021) discutem os desafios e as oportunidades que a inteligência artificial traz para a ciência esportiva e as equipes de medicina esportiva. Entre os benefícios, estão a automação de tarefas repetitivas e a realização de análises de sentimentos e padrões táticos, proporcionando uma visão mais completa da saúde e bem-estar dos atletas. O artigo também alerta para os desafios, como a complexidade dos sistemas e a possível substituição de funções humanas, que precisam ser cuidadosamente considerados ao aplicar IA em análises de tênis.

O trabalho apresentado em (TAN, 2023) analisa o papel do Big Data e da IA na predição esportiva, desenvolvendo um sistema preditivo para monitorar e prever o desempenho dos atletas. Ele destaca como a combinação de IA e Big Data pode apoiar planos de treinamento e monitoramento de saúde, facilitando a tomada de decisões e otimizando o desempenho. Esse estudo reforça a importância de considerar grandes volumes de dados para análises mais precisas e completas, aspecto essencial para a análise de desempenho no tênis.

Por fim, o estudo de (BRITO et al., 2024) investiga a influência da cinemática na velocidade do saque no tênis, utilizando tecnologia de medição inercial para capturar dados de movimento. A pesquisa demonstra como a biomecânica dos jogadores afeta seu desempenho, oferecendo insights valiosos para a otimização do treinamento de saque. Essa análise fornece uma base para que treinadores e atletas ajustem seus treinos com base em dados de movimento,

complementando a análise preditiva de desempenho em partidas.

Baseando-se nas conclusões e metodologias apresentadas nesses estudos, este trabalho propõe investigar a utilização de dados históricos para a análise de desempenho no tênis, buscando construir modelos preditivos que identifiquem padrões e variáveis-chave de sucesso. A pesquisa visa contribuir para a prática esportiva, oferecendo uma base sólida para decisões estratégicas e o aprimoramento no treinamento dos atletas.

3 Metodologia

Para atingir os objetivos propostos neste estudo, foi adotada uma abordagem metodológica composta por várias etapas distintas. O fluxo de trabalho mostrado a seguir descreve as principais fases da metodologia.

3.1 Aquisição e Armazenamento dos Dados

A aquisição dos dados foi realizada a partir de repositórios públicos disponíveis no GitHub, especificamente os repositórios relacionados ao tênis profissional (WTA e ATP) (SACKMANN, 2025b; SACKMANN, 2025a), bem como do site oficial da ATP (Association of Tennis Professionals, 2025) e WTA (Women's Tennis Association, 2025). Esses dados, referentes ao período de 2018 até 2024, foram transferidos para um bucket no Google Cloud Platform (GCP), denominado *tennis-data-lake*, visando centralizar e organizar as informações coletadas.

No GCP, os dados foram estruturados em categorias distintas, como jogadores, rankings e partidas, e organizados em pastas específicas de acordo com o ano e o tipo de dado. Essa organização facilitou o acesso e o gerenciamento das informações ao longo do processo de análise.

3.2 Exploração e Entendimento dos Dados

Para compreender a estrutura e o conteúdo dos dados, foi realizada uma análise inicial dos arquivos disponíveis. Essa etapa incluiu a visualização das colunas presentes nos arquivos de rankings, jogadores e partidas, tanto para a WTA quanto para a ATP. Os principais arquivos analisados foram os de rankings, como *wta_rankings_current.csv* e *atp_rankings_current.csv*, os de jogadores, como *wta_players.csv* e *atp_players.csv*, e os de partidas, que incluíam diversos arquivos organizados por ano, como *wta_matches_2018.csv*, *atp_matches_2018.csv*, e assim por diante, abrangendo o período de 2018 a 2024 (SACKMANN, 2025b) (SACKMANN, 2025a).

A inspeção inicial dos dados foi realizada por meio da exibição dos cabeçalhos e de amostras das primeiras linhas de cada arquivo, o que permitiu a identificação das estruturas e a compreensão das variáveis disponíveis.

3.3 Filtragem e Preparação dos Dados

3.3.1 Filtragem do Top 50

A filtragem do top 50 jogadores de cada categoria, feminina e masculina, foi realizada com base no atributo *rank* presente nos arquivos de rankings. Esse processo permitiu identificar

e selecionar os 50 melhores jogadores de acordo com suas classificações. Para enriquecer as informações, as tabelas de rankings e jogadores foram combinadas, resultando em um conjunto de dados detalhado que incluía nome, *rank* e pontuação de cada jogador. Adicionalmente, uma coluna de gênero foi adicionada para distinguir entre os jogadores da ATP e da WTA. Com isso, foi criada uma tabela consolidada contendo os 50 melhores jogadores de ambas as categorias. Para visualização e verificação dos dados filtrados, foram geradas tabelas estilizadas utilizando as bibliotecas *pandas* e *plotly*, além da plotagem de gráficos interativos para explorar possíveis correlações entre *rank* e pontuação.

3.3.2 Preparação dos Dados de Partidas

A preparação dos dados de partidas teve como objetivo garantir que essas informações fossem consistentes e adequadas para a modelagem. Inicialmente, os arquivos de partidas dos anos de 2018 a 2024 foram combinados em um único conjunto de dados. Em seguida, foi aplicada uma filtragem para manter apenas as partidas que envolviam jogadores do top 50. As *features* relevantes para a análise foram selecionadas, incluindo características do torneio, como *surface*, *round*, *draw_size*, *best_of* e *minutes*, além de métricas dos jogadores, como *winner_rank*, *loser_rank*, *winner_rank_points* e *loser_rank_points*. Para tratar colunas categóricas, como *surface* e *round*, foi utilizada a técnica de *one-hot encoding*, convertendo-as em variáveis numéricas.

Valores ausentes foram preenchidos utilizando a técnica de imputação por média (*mean imputation*), onde os valores ausentes são substituídos pela média dos valores presentes na coluna correspondente. Essa estratégia foi escolhida para minimizar o impacto dos dados ausentes e garantir a integridade do conjunto de dados para análise.

3.3.3 Verificação dos Dados Preparados

Após a aplicação das transformações, a estrutura final dos dados foi verificada para garantir sua adequação à etapa de modelagem. Essa verificação incluiu a conferência das primeiras linhas do *DataFrame* tratado, a inspeção das colunas resultantes após as transformações e a garantia de que o *DataFrame* estava consistente e pronto para a próxima fase do projeto. Essa etapa foi essencial para assegurar que os dados estivessem corretamente preparados e que nenhum erro ou inconsistência pudesse comprometer a análise subsequente.

3.4 Engenharia de Features

A engenharia de *features* foi uma etapa crucial nesta pesquisa, permitindo transformar os dados brutos em informações relevantes para a análise preditiva. Nessa fase, novas variáveis foram criadas para capturar diferenças e características essenciais entre os jogadores vencedores e perdedores, facilitando a identificação de padrões significativos que influenciam os resultados das partidas (VERDONCK et al., 2021).

Foram geradas variáveis que destacam as discrepâncias entre os jogadores em métricas importantes, como a diferença de *ranking*, que avalia o impacto da posição no *ranking* no resultado das partidas, e a diferença de pontos no *ranking*, que fornece uma medida mais granular do desempenho relativo entre os jogadores. Além disso, foram consideradas a diferença de idade, analisando se a experiência relacionada à idade influencia os resultados, e a diferença de altura, avaliando o impacto de características físicas no desempenho.

Variáveis relacionadas ao desempenho de saque foram incorporadas, como a diferença no número de *aces*, que avalia a habilidade dos jogadores em marcar pontos diretos com o saque, e a diferença nas duplas faltas, que mede a eficiência dos jogadores em minimizar erros nos serviços.

A capacidade dos jogadores em salvar *break points* foi modelada para capturar sua resiliência em momentos críticos. Foram criadas variáveis como a eficiência do vencedor ao salvar *break points*, a eficiência do perdedor ao salvar *break points* e a diferença na eficiência de *break points*, que compara a eficiência entre vencedor e perdedor, destacando sua importância em momentos decisivos da partida.

As variáveis derivadas foram criadas com o objetivo de enriquecer a base de dados, destacando aspectos críticos do desempenho dos jogadores. A análise subsequente demonstrou que variáveis como diferença de *ranking*, diferença de pontos no *ranking* e eficiência de *break points* estão entre as mais influentes na previsão dos resultados das partidas.

A engenharia de *features* não apenas melhorou a qualidade dos dados analisados, mas também contribuiu diretamente para a eficácia dos modelos preditivos. Ao capturar nuances específicas do jogo de tênis, como estatísticas de saque e diferenças físicas, esta etapa revelou-se fundamental para o entendimento dos fatores que impactam o desempenho dos jogadores.

3.5 Modelos de Aprendizado

3.5.1 Modelo Random Forest

O algoritmo de classificação Random Forest foi escolhido devido à sua robustez e capacidade de lidar com dados de alta dimensionalidade e complexidade. Ele opera criando múltiplas árvores de decisão durante o treinamento e combina suas saídas para melhorar a precisão preditiva e evitar problemas como *overfitting* (LOUPPE, 2015). Os dados foram preparados separando as variáveis preditoras (*features*) e a variável-alvo. A variável-alvo foi definida como a comparação direta entre os *rankings* dos jogadores, sendo o valor 1 quando o vencedor tinha um *ranking* melhor e 0 caso contrário. Para assegurar que as variáveis com escalas diferentes não influenciassem o treinamento do modelo, os dados foram normalizados utilizando o método *StandardScaler*, ajustando todas as *features* para uma média zero e desvio padrão unitário. Os dados foram divididos em 70% para treinamento e 30% para teste, utilizando uma amostragem

aleatória com uma semente fixa para garantir reprodutibilidade.

O modelo Random Forest foi configurado com 100 árvores de decisão (*n_estimators=100*) e ajustado utilizando os dados de treinamento. Após o treinamento, foi realizada uma análise de importância das *features* para identificar os fatores que mais influenciaram as decisões do modelo. Entre os principais fatores, destacaram-se a diferença de *ranking*, a diferença de pontos no *ranking* e os pontos do *ranking* do vencedor. Esses fatores mostraram-se cruciais para prever o resultado das partidas, evidenciando a importância dos *rankings* no desempenho dos jogadores. A relevância de cada variável foi representada em um gráfico de barras, destacando as 10 *features* mais influentes para a classificação. Essa visualização permitiu uma compreensão clara das variáveis que mais contribuíram para o desempenho do modelo, evidenciando a importância relativa de cada uma delas na previsão dos resultados das partidas. Entre as principais, destacaram-se a diferença de *ranking*, a diferença de pontos no *ranking* e os pontos do *ranking* do vencedor, que se mostraram determinantes para a precisão do modelo.

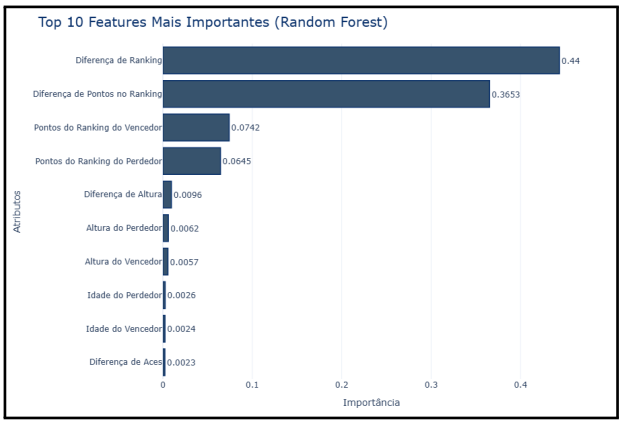

Figura 1 – Importância das Features no Modelo Random Forest

3.5.2 Modelo K-Nearest Neighbors (KNN)

O modelo K-Nearest Neighbors (KNN) foi implementado para explorar a eficácia de algoritmos baseados em proximidade na classificação de resultados de partidas de tênis. Este método classifica os exemplos com base em seus vizinhos mais próximos no espaço das *features*, o que é útil para identificar padrões em conjuntos de dados estruturados (HALDER et al., 2024). Assim como no modelo Random Forest, os dados foram preparados com a separação entre *features* preditivas e variável-alvo, normalização e divisão em conjuntos de treino e teste. A normalização foi particularmente importante no KNN, dado que este modelo é sensível às escalas das variáveis. O modelo KNN foi configurado para considerar os 5 vizinhos mais próximos (*n_neighbors=5*) com a métrica de distância padrão Euclidiana. Essa configuração foi escolhida para equilibrar a precisão do modelo e evitar *overfitting*.

O modelo foi treinado com o conjunto de treino e avaliado com o conjunto de teste. A importância das *features* foi avaliada utilizando a técnica de *Permutation Importance*, que mede

a redução de desempenho do modelo ao embaralhar os valores de uma *feature* específica. Os resultados mostraram que as variáveis mais importantes para o modelo foram a diferença de pontos no *ranking*, os pontos do *ranking* do perdedor e a diferença de *ranking*. Essas variáveis tiveram maior impacto na decisão do modelo, destacando a relevância de fatores diretamente relacionados ao desempenho histórico dos jogadores. As importâncias das *features* foram representadas em gráficos interativos, destacando as 10 variáveis mais relevantes para o modelo KNN, seguindo a mesma abordagem utilizada no modelo Random Forest. Essa visualização permitiu uma compreensão clara dos fatores que mais influenciaram as decisões do modelo, com destaque para a diferença de pontos no *ranking*, os pontos do *ranking* do perdedor e a diferença de *ranking*. A Figura X ilustra a importância relativa dessas *features*, evidenciando sua contribuição para a precisão do modelo.

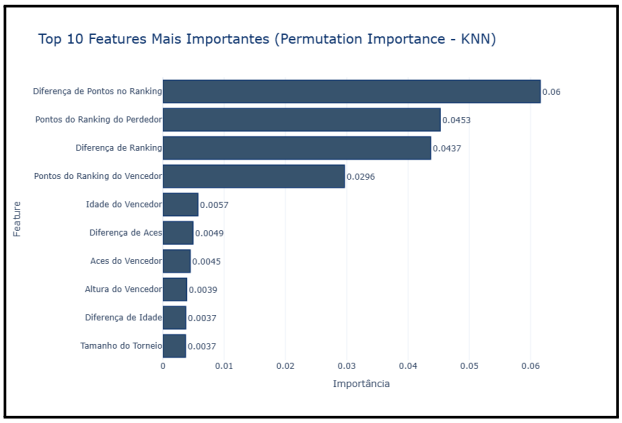

Figura 2 – Importância das Features no Modelo K-Nearest Neighbors

4 Análise dos Dados e Resultados

4.1 Avaliação dos Modelos: Random Forest e K-Nearest Neighbors (KNN)

Os modelos Random Forest e K-Nearest Neighbors (KNN) foram avaliados com base em várias métricas de desempenho, como acurácia, *precision*, *recall* e *F1-score*, para determinar sua eficácia na previsão dos resultados das partidas de tênis. O modelo Random Forest apresentou uma acurácia de 1.00 (100%), indicando que classificou todos os exemplos corretamente. O relatório de classificação mostrou *precision*, *recall* e *F1-score* de 1.00 para ambas as classes (0 e 1), com suporte de 922 exemplos para a classe 0 e 1701 exemplos para a classe 1. As médias macro e ponderada também foram de 1.00 para todas as métricas, refletindo um desempenho excelente na base de teste.

Para o modelo K-Nearest Neighbors (KNN), os resultados obtidos foram uma acurácia de 81.66%. O relatório de classificação mostrou *precision* de 0.80, *recall* de 0.64 e *F1-score* de 0.71 para a classe 0 (perdedor tinha melhor ranking), com suporte de 922 exemplos. Para a classe 1 (vencedor tinha melhor ranking), os valores foram *precision* de 0.82, *recall* de 0.91 e *F1-score* de 0.87, com suporte de 1701 exemplos. As médias macro e ponderada também foram calculadas, com *precision* de 0.81, *recall* de 0.78 e *F1-score* de 0.79 para a média macro, e *precision* de 0.81, *recall* de 0.82 e *F1-score* de 0.81 para a média ponderada. Esses resultados indicam que o modelo foi eficaz em classificar os exemplos, especialmente na identificação dos vencedores com melhor *ranking*.

A identificação de *outliers* e padrões específicos nos gráficos oferece oportunidades de exploração para otimização de treinamentos e estratégias no tênis. Esses insights podem ser úteis para técnicos e jogadores que buscam melhorar seu desempenho com base em dados. Além disso, o uso de múltiplos modelos de aprendizado de máquina, como o Random Forest e o KNN, demonstra a eficácia dessas técnicas na análise preditiva no esporte, destacando a importância de uma abordagem multifacetada para a compreensão e melhoria do desempenho dos jogadores.

4.2 Gráfico de Dispersão: Diferença de Ranking vs. Diferença de Pontos no Ranking

Os gráficos de dispersão foram utilizados para visualizar relações entre variáveis contínuas, permitindo identificar padrões, tendências e *outliers* nos dados. Essa técnica é particularmente útil para explorar a influência de múltiplas dimensões simultaneamente, como diferenças

de *ranking*, pontos, idade e altura, no contexto de resultados de partidas de tênis (BERGSTROM; WEST, 2018). O gráfico de dispersão Diferença de *Ranking* vs. Diferença de Pontos no *Ranking* apresenta a relação entre essas duas variáveis principais, além de considerar dimensões adicionais, como diferença de idade (representada pelas cores dos pontos) e diferença de altura (representada pelo tamanho dos pontos). Esse gráfico fornece insights importantes sobre os padrões de vitória no tênis com base em características dos jogadores. As principais observações incluem a diferença de *Ranking*, onde valores negativos no eixo X indicam que o vencedor possuía um *ranking* superior ao perdedor. A maioria dos pontos está concentrada em valores negativos, evidenciando que jogadores com melhores *rankings* tendem a vencer. A diferença de Pontos no *Ranking*, onde valores positivos no eixo Y mostram que o vencedor tinha mais pontos no *ranking* do que o perdedor. A concentração de pontos positivos reforça a relevância dos pontos no *ranking* como indicador de desempenho. A relação entre Diferença de *Ranking* e Diferença de Pontos mostra uma correlação positiva, onde vencedores com grandes diferenças de pontos no *ranking* geralmente têm *rankings* superiores. Alguns *outliers* indicam resultados inesperados, onde jogadores com *rankings* inferiores venceram. A diferença de Idade, representada pelas cores dos pontos, não mostra um padrão claro que sugira que a idade seja um fator determinante nos resultados. A diferença de Altura, representada pelo tamanho dos pontos, também não apresenta uma relação direta consistente com os resultados. As conclusões indicam que *ranking* e pontos são os principais indicadores de vitória, conforme esperado no contexto competitivo do tênis. *Outliers* sugerem situações de surpresa, que podem ser interessantes para investigações futuras, enquanto diferenças de idade e altura têm impacto marginal.

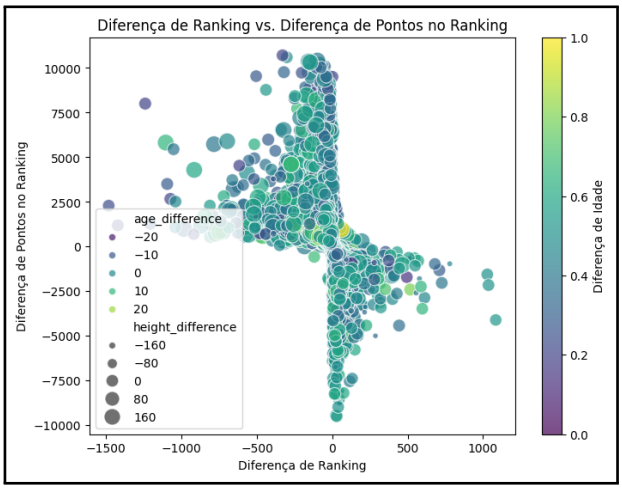

Figura 3 – Diferença de *Ranking* vs. Diferença de Pontos no *Ranking*

4.3 Gráfico de Dispersão: Diferença de Idade vs. Diferença de Altura

O segundo gráfico de dispersão analisa a relação entre diferenças de idade e altura, com dimensões adicionais, como diferença de *aces* (cores dos pontos) e diferença de duplas

faltas (tamanhos dos pontos). Esse gráfico foi escolhido para explorar como características físicas e estatísticas de jogo interagem e influenciam os resultados. As principais observações incluem a diferença de Idade, onde valores negativos no eixo X indicam vencedores mais jovens, enquanto valores positivos indicam vencedores mais velhos. A maioria dos pontos está próxima de zero, indicando partidas entre jogadores de idades similares. A diferença de Altura, onde valores positivos no eixo Y mostram vencedores mais altos, enquanto valores negativos indicam vencedores mais baixos. A concentração ao redor de zero sugere que diferenças extremas de altura são menos comuns. A diferença de *aces*, representada pelas cores dos pontos, mostra que vencedores que fizeram mais *aces* tendem a ter vantagem. A diferença de duplas faltas, representada pelo tamanho dos pontos, reflete maior diferença de erros nos serviços. Os padrões observados indicam que jogadores mais altos frequentemente apresentam vantagem em *aces*, mas diferenças extremas de idade ou altura não são predominantes. A altura é um fator mais relevante do que a idade no desempenho, especialmente em relação à geração de *aces*. No entanto, erros como duplas faltas podem anular essa vantagem. A análise gráfica revela padrões esperados no desempenho dos jogadores, destacando a relevância de variáveis como *ranking* e pontos. Entretanto, aspectos físicos, como idade e altura, mostram impacto limitado em cenários gerais, sugerindo que a combinação de habilidades técnicas, táticas e contexto do jogo desempenha um papel crucial nos resultados. A identificação de *outliers* e padrões específicos nos gráficos oferece oportunidades de exploração para otimização de treinamentos e estratégias no tênis. Esses insights podem ser úteis para técnicos e jogadores que buscam melhorar seu desempenho com base em dados.

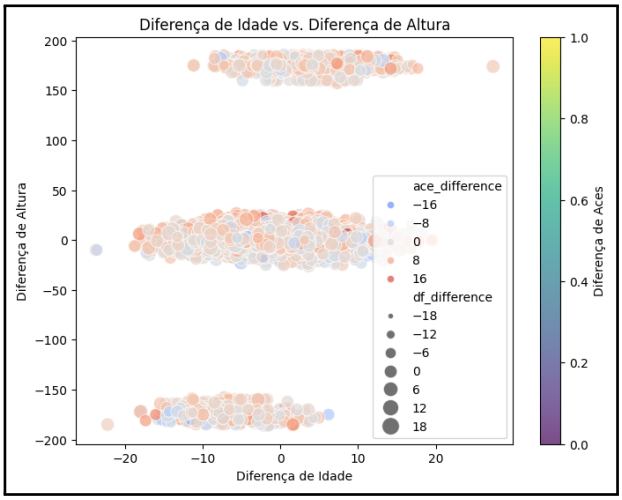

Figura 4 – Diferença de Idade vs. Diferença de Altura

4.4 Heatmap de Correlação

O *heatmap* de correlação oferece uma visão abrangente das relações entre as variáveis utilizadas na análise (DAVILA; PAZ; MOQUILLAZA, 2023). Ele utiliza a matriz de correlação para destacar a intensidade e o tipo de relação (positiva ou negativa) entre as diferentes *features*.

Essa análise é essencial para identificar os principais fatores que influenciam os resultados das partidas. O *heatmap* destaca como as variáveis estão relacionadas entre si, com cores quentes (vermelho) indicando correlação positiva e cores frias (azul) indicando correlação negativa. As variáveis com correlações mais fortes sugerem relações diretas ou indiretas que podem impactar os resultados das partidas. As variáveis que apresentam maior correlação com a Diferença de *Ranking*, que é uma das principais *features* da análise, são a Diferença de *Ranking*, que é correlacionada consigo mesma (valor de 1.0, esperado), o *Ranking* do Vencedor (0.508), que mostra alta correlação com a Diferença de *Ranking*, reforçando que o *ranking* é um dos indicadores mais fortes, e os Pontos do *Ranking* do Perdedor (0.386), que indicam que a pontuação do perdedor também contribui para a explicação da diferença. Outras variáveis incluem a Altura do Perdedor (0.234), que indica que características físicas, como altura, têm um impacto leve na diferença de *ranking*, Primeiros Saques em Quadra (Vencedor) (0.094), que sugere que a eficiência no primeiro saque pode influenciar o desempenho, Duração da Partida (min) (0.092), que indica que partidas mais longas podem estar associadas a diferenças de *ranking*, *Aces* do Perdedor (0.087), que mostra que a capacidade de marcar *aces* pode ter um impacto, mesmo para o perdedor, *Break Points* Enfrentados (Vencedor) (0.085), que reflete a resiliência do vencedor em momentos críticos, Pontos Ganhos no Primeiro Saque (Perdedor) (0.076), que indica a eficiência do perdedor no primeiro saque, e Pontos Ganhos no Primeiro Saque (Vencedor) (0.069), que mostram a importância do primeiro saque para o vencedor.

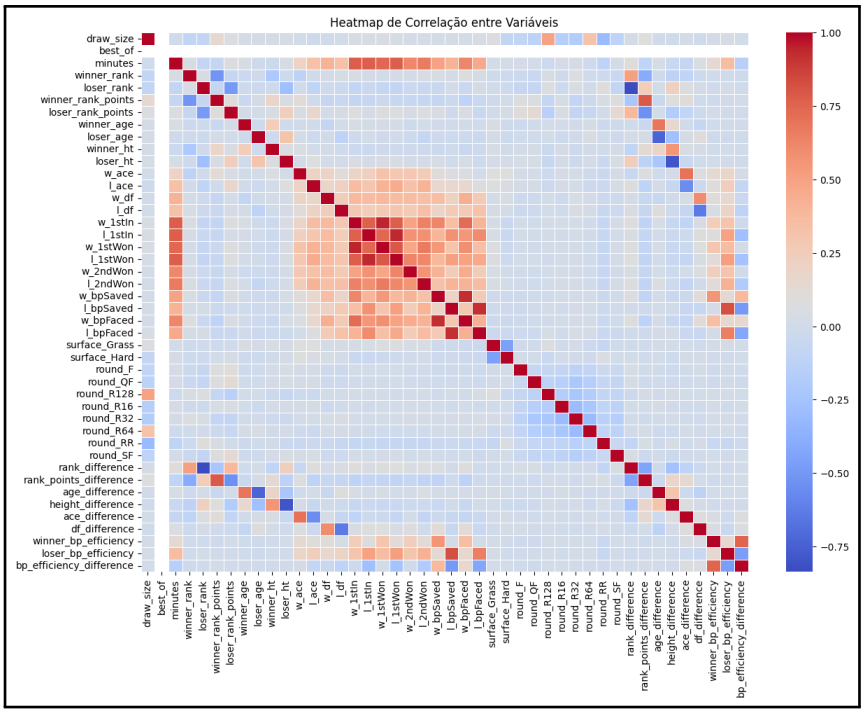

Figura 5 – *Heatmap* de correlação

Ranking e pontuação são os fatores mais correlacionados à Diferença de Ranking, reforçando sua relevância na análise de desempenho. Fatores como altura e estatísticas de saque, embora apresentem correlações menores, ainda contribuem para uma análise mais ampla do

jogo, oferecendo insights complementares. O heatmap de correlação se mostrou uma ferramenta visual poderosa, permitindo a identificação de padrões gerais e fatores específicos que podem ser explorados em modelos preditivos ou estratégias táticas. Essa análise reforça a importância de considerar múltiplas variáveis para uma compreensão mais completa dos fatores que influenciam os resultados das partidas de tênis.

5 Considerações Finais e Trabalhos Futuros

5.1 Considerações Finais

Este trabalho explorou o uso de técnicas de aprendizado de máquina e análise de dados no contexto do tênis, com foco em prever os resultados de partidas e identificar os fatores mais relevantes que influenciam o desempenho dos jogadores. A partir de bases de dados robustas provenientes de fontes confiáveis, como a ATP, WTA e repositórios públicos, foi possível realizar um *pipeline* completo de análise, que incluiu a exploração, filtragem, engenharia de *features*, treinamento de modelos e avaliação dos resultados.

Os modelos desenvolvidos, como *Random Forest* e *K-Nearest Neighbors* (KNN), demonstraram capacidade significativa de prever os resultados das partidas, com o *Random Forest* alcançando uma precisão impressionante de 100% nos dados de teste. Este desempenho pode ser atribuído à riqueza das *features* criadas e ao poder de generalização do modelo, que soube explorar a correlação entre variáveis como diferença de *ranking*, diferença de pontos no *ranking* e estatísticas de saque.

Apesar de alguns modelos apresentarem limitações, como o KNN, que mostrou um desempenho inferior ao *Random Forest* devido à sua sensibilidade a padrões específicos nos dados, os resultados obtidos comprovam que técnicas de aprendizado de máquina são ferramentas poderosas para análise preditiva no esporte. Além disso, as análises gráficas, como os gráficos de dispersão e o *heatmap* de correlação, forneceram insights adicionais sobre a relação entre variáveis, destacando que, embora *ranking* e pontuação no *ranking* sejam os fatores mais relevantes, outros elementos, como altura e estatísticas de saque, também desempenham papéis importantes em determinadas circunstâncias.

Este estudo contribui para a crescente literatura sobre a aplicação de análise de dados e aprendizado de máquina no esporte, reforçando como essas técnicas podem melhorar a tomada de decisão de treinadores e atletas, além de tornar o tênis ainda mais emocionante e estratégico.

5.2 Trabalhos Futuros

Embora este trabalho tenha atingido seus objetivos, ele abre caminho para diversas direções futuras que podem complementar e expandir os resultados obtidos. Uma dessas direções é o aprimoramento de modelos, experimentando técnicas mais avançadas, como *Gradient Boosting Machines* (GBMs), *XGBoost* ou redes neurais profundas, que podem capturar padrões mais complexos nos dados. Além disso, a incorporação de dados adicionais, como informações

mais recentes, históricos ampliados e variáveis contextuais (clima, tipo de torneio e condição física dos jogadores), pode enriquecer a análise.

Outra área promissora é a aplicação em cenários reais, desenvolvendo ferramentas para uso em tempo real que permitam a treinadores e analistas fazer previsões durante as partidas ou ajustar estratégias com base nos dados mais recentes. A integração com sistemas de visualização de dados interativos também pode facilitar a interpretação dos resultados por especialistas no esporte.

A prevenção de lesões é outro campo de interesse, onde a análise pode ser expandida para prever lesões potenciais em jogadores, utilizando métricas de esforço físico, frequência de partidas e recuperação. Além disso, investigar padrões táticos nos dados das partidas, como movimentação em quadra e comportamento em situações críticas (por exemplo, *break points*), pode oferecer insights valiosos.

Estudos comparativos entre esportes, como tênis, badminton e squash, podem generalizar as descobertas e identificar padrões comuns ou específicos de cada modalidade. Por fim, a automação e escalabilidade do *pipeline* de análise, integrando-o com plataformas de computação em nuvem como GCP ou AWS, garantirá a capacidade de processar grandes volumes de dados em tempo real.

Este trabalho demonstrou como a combinação de dados históricos e técnicas avançadas de aprendizado de máquina pode oferecer insights poderosos e aplicáveis na prática esportiva. Os avanços futuros descritos têm o potencial de ampliar ainda mais o impacto dessas técnicas no esporte, ajudando a moldar o futuro do tênis e de outras modalidades.

Referências

Association of Tennis Professionals. *ATP Tour*. 2025. Accessed: 2024-12-27. Disponível em: <https://www.atptour.com>. Citado na página 9.

BERGSTROM, C. T.; WEST, J. D. *Why scatter plots suggest causality, and what we can do about it*. 2018. Disponível em: <https://arxiv.org/abs/1809.09328>. Citado na página 15.

BODEMER, O. Enhancing individual sports training through artificial intelligence: A comprehensive review. Institute of Electrical and Electronics Engineers (IEEE), ago. 2023. Disponível em: <http://dx.doi.org/10.36227/techrxiv.24005916.v1>. Citado na página 4.

BRANCO, P.; TORGO, L.; RIBEIRO, R. P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 49, n. 2, ago. 2016. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/2907070>. Citado na página 5.

BRITO, A. V. et al. The influence of kinematics on tennis serve speed: An in-depth analysis using xsens mvn biomech link technology. *Bioengineering*, v. 11, n. 10, 2024. ISSN 2306-5354. Disponível em: <https://www.mdpi.com/2306-5354/11/10/971>. Citado na página 7.

CHMAIT, N.; WESTERBEEK, H. Artificial intelligence and machine learning in sport research: An introduction for non-data scientists. *Frontiers in Sports and Active Living*, v. 3, 2021. ISSN 2624-9367. Disponível em: <https://www.frontiersin.org/journals/sports-and-active-living/articles/10.3389/fspor.2021.682287>. Citado na página 7.

COHEN, M.; SLOANE, M. Predictive modeling and statistical analysis in sports. May 2015. Pomona College CS190 Thesis. Disponível em: <https://cs.pomona.edu/classes/cs190/thesis_examples/Cohen-Sloane.15.pdf>. Citado 2 vezes nas páginas 4 e 7.

DAVILA, F.; PAZ, F.; MOQUILLAZA, A. Usage and application of heatmap visualizations on usability user testing: A systematic literature review. In: MARCUS, A.; ROSENZWEIG, E.; SOARES, M. M. (Ed.). *Design, User Experience, and Usability*. Cham: Springer Nature Switzerland, 2023. p. 3–17. ISBN 978-3-031-35702-2. Citado na página 16.

HALDER, R. K. et al. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. 1 2024. Disponível em: <https://dro.deakin.edu.au/articles/journal_contribution/Enhancing_K-nearest_neighbor_algorithm_a_comprehensive_review_and_performance_analysis_of_modifications/26785429>. Citado na página 12.

K., A.; K., S. Machine learning and artificial intelligence in sports performance: A comprehensive review. In: . [S.l.: s.n.], 2023. Citado na página 7.

LOUPPE, G. *Understanding Random Forests: From Theory to Practice*. 2015. Disponível em: <https://arxiv.org/abs/1407.7502>. Citado na página 11.

SACKMANN, J. *tennis_atp: ATP tennis databases, files, and algorithms*. 2025. GitHub repository. Disponível em: <https://github.com/JeffSackmann/tennis_atp>. Citado na página 9.

SACKMANN, J. *tennis_wta: WTA tennis databases, files, and algorithms*. 2025. GitHub repository. Disponível em: <https://github.com/JeffSackmann/tennis_wta>. Citado na página 9.

SAMPAIO, T. et al. Applications of machine learning to optimize tennis performance: A systematic review. *Applied Sciences*, v. 14, p. 5517, 2024. Citado na página 7.

TAN, X. Enhanced sports predictions: A comprehensive analysis of the role and performance of predictive analytics in the sports sector. *Wireless Personal Communications*, v. 132, p. 1613–1636, 2023. Citado 2 vezes nas páginas 4 e 7.

VERDONCK, T. et al. Special issue on feature engineering editorial. *Machine Learning*, v. 113, p. 1–12, 08 2021. Citado na página 10.

Women's Tennis Association. *WTA Tour*. 2025. Accessed: 2024-12-27. Disponível em: <https://www.wtatennis.com>. Citado na página 9.