UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LOURIVAL GONÇALVES PRATA NETTO

METODOLOGIA PARA ANÁLISE PREDITIVA DE RESULTADOS DE JOGOS ESPORTIVOS: UM ESTUDO DE CASO POR JOGOS DA NBA

CAMPINA GRANDE - PB

2023

LOURIVAL GONÇALVES PRATA NETTO

METODOLOGIA PARA ANÁLISE PREDITIVA DE RESULTADOS DE JOGOS ESPORTIVOS: UM ESTUDO DE CASO POR JOGOS DA NBA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Herman Martins Gomes

CAMPINA GRANDE - PB

2023

LOURIVAL GONÇALVES PRATA NETTO

METODOLOGIA PARA ANÁLISE PREDITIVA DE RESULTADOS DE JOGOS ESPORTIVOS: UM ESTUDO DE CASO POR JOGOS DA NBA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Herman Martins Gomes

Orientador – UASC/CEEI/UFCG

Marcus Salerno de Aquino

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 30 de Junho de 2023.

CAMPINA GRANDE - PB

RESUMO

Este artigo apresenta uma metodologia para análise preditiva de resultados de jogos esportivos, utilizando jogos da National Basketball Association (NBA) como estudo de caso. Especificamente, a NBA tem notoriedade no uso de análises de dados e estatísticas avançadas para melhorar o desempenho da equipe e do jogador. Os objetivos deste trabalho são analisar a disponibilidade e qualidade dos dados da NBA, selecionar variáveis para análise preditiva, aplicar técnicas de aprendizado de máquina para prever resultados de jogos da NBA e avaliar a eficácia da metodologia proposta. A metodologia inclui coleta de dados obtidos através da técnica de raspagem de dados utilizados no site oficial da NBA, pré-processamento dos dados, seleção de variáveis mais relevantes através da técnica PCA, modelagem e avaliação. Os resultados obtidos demonstram uma metodologia eficaz para análise preditiva de resultados de jogos esportivos, no tocante à identificação das variáveis mais importantes para prever resultados de jogos da NBA e estatísticas de acerto de predição. O presente trabalho, portanto, contribui para o avanço da análise preditiva em esportes e outros campos de aplicação.

METHODOLOGY FOR PREDICTIVE ANALYSIS OF RESULTS IN SPORTS GAMES: A CASE STUDY BY NBA GAMES

ABSTRACT

This article presents a methodology for predictive analysis of sports game results, using National Basketball Association (NBA) games as a case study. Specifically, the NBA is renowned for using data analytics and advanced statistics to improve team and player performance. The objectives of this work are to analyze the availability and quality of NBA data, select variables for predictive analysis, apply machine learning techniques to predict NBA game results and evaluate the effectiveness of the proposed methodology. The methodology includes data collection obtained through the data scraping technique used on the official NBA website, data pre-processing, selection of the most relevant variables through the PCA technique, modeling and evaluation. The results obtained demonstrate an effective methodology for predictive analysis of results of sports games, regarding the identification of the most important variables to predict results of NBA games and statistics of prediction success. The present work, therefore, contributes to the advancement of predictive analytics in sports and other fields of application.

Metodologia para análise preditiva de resultados de jogos esportivos: um estudo de caso por jogos da NBA

Lourival Netto
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
lourival.netto@ccc.ufcg.edu.br

Herman Martins
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
hmg@computacao.ufcg.edu.br

RESUMO

Este artigo apresenta uma metodologia para análise preditiva de resultados de jogos esportivos, utilizando jogos da *National Basketball Association* (NBA) como estudo de caso. Especificamente, a NBA tem notoriedade no uso de análises de dados e estatísticas avançadas para melhorar o desempenho da equipe e do jogador. Os objetivos deste trabalho são analisar a disponibilidade e qualidade dos dados da NBA, selecionar variáveis para análise preditiva, aplicar técnicas de aprendizado de máquina para prever resultados de jogos da NBA e avaliar a eficácia da metodologia proposta. A metodologia inclui coleta de dados obtidos através da técnica de raspagem de dados utilizados no site oficial da NBA, pré-processamento dos dados, seleção de variáveis mais relevantes através da técnica PCA, modelagem e avaliação. Os resultados obtidos demonstram uma metodologia eficaz para análise preditiva de resultados de jogos esportivos, no tocante à identificação das variáveis mais importantes para prever resultados de jogos da NBA e estatísticas de acerto de predição. O presente trabalho, portanto, contribui para o avanço da análise preditiva em esportes e outros campos de aplicação.

PALAVRAS-CHAVE

Análise preditiva, Machine learning, NBA, Jogos esportivos, Previsão de resultados.

DADOS E CÓDIGO

O seguinte repositório contém os dados coletados e notebooks Python para coleta de dados, treinamento e validação dos modelos investigados:

https://n9.cl/r2g0hV

1. INTRODUÇÃO

A análise de dados tem se tornado uma ferramenta cada vez mais importante em diversos campos, incluindo esportes. A análise de dados esportivos é usada para prever resultados, identificar padrões e melhorar o desempenho de jogadores e equipes. Nesse sentido, a *National Basketball Association* (NBA) tem sido uma das organizações esportivas líderes na adoção de análises avançadas e estatísticas para melhorar o desempenho dos times e jogadores.

A análise preditiva, na atualidade, tornou-se uma ferramenta poderosa para prever resultados em vários campos, incluindo esportes.[1] A análise preditiva pode ser usada para prever resultados de jogos esportivos de maneira mais precisa e informada, levando a melhores decisões estratégicas por parte dos times e jogadores.[2]

Este trabalho apresenta uma metodologia para análise preditiva de resultados de jogos esportivos, utilizando jogos da NBA como estudo de caso. A metodologia proposta inclui a coleta de dados, pré-processamento, seleção de variáveis, modelagem e avaliação. O objetivo principal deste trabalho é apresentar uma metodologia eficaz para a análise preditiva de resultados de jogos esportivos e avaliar a sua eficácia no tocante a taxa de acerto em relação às predições.

Na próxima seção discutem-se os trabalhos relacionados, visando poder contextualizar a relevância desse trabalho em relação ao conhecimento na área. Em seguida, descreve-se a solução proposta e a metodologia do trabalho. Posteriormente, são reportados os experimentos e resultados obtidos. Por fim, são tecidas considerações finais, destacando as experiências, lições aprendidas e agradecimentos.

2. TRABALHOS RELACIONADOS

A análise de dados tem sido cada vez mais utilizada no contexto esportivo, tanto para melhorar o desempenho dos atletas como para prever resultados de jogos.[3] Nesse sentido, a NBA tem sido um dos pioneiros no uso de análises avançadas e estatísticas para melhorar o desempenho das equipes e jogadores.[4]

No artigo "NBA Game Result Prediction Using Feature Analysis and Machine Learning", os autores propõem um método para prever resultados de jogos da NBA, fundamentando-se na análise de características das equipes e dos jogadores, e fazendo uso de técnicas de aprendizado de máquina. Os resultados indicam que ao comparar o desempenho e os modelos derivados de diferentes conjuntos de características relacionadas aos jogos de basquete, foi descoberto que características como DRB, TPP, FT e TRB são fatores significativos que influenciam os resultados dos jogos.

Esses fatores foram incluídos no modelo, o que resultou em um aumento de 2-4% na taxa de precisão de previsão.[5]

Em outro estudo, intitulado "Predictive Analysis on eSports Games: A Case Study on League of Legends (LoL) eSports Tournaments", os autores propõem uma metodologia semelhante para prever resultados de torneios de eSports. O estudo usa técnicas de análise preditiva para identificar os fatores que mais influenciam os resultados dos jogos, e apresenta resultados promissores na previsão de resultados de torneios de LoL ao utilizar modelos de regressão logística e Árvores de decisão, ambos apresentando ser eficiente para tal. Além disso, ao considerar informações adicionais que existem internamente no jogo, chegou a conclusão de que é necessário implementar um modelo estatístico mais complexo para melhorar a precisão das previsões.[6]

Em "A data-driven prediction approach for sports team performance and its application to National Basketball Association", os autores desenvolvem uma abordagem baseada em análise de envoltória de dados (DEA) e técnicas orientadas por dados para prever o desempenho de equipes esportivas. O estudo utiliza análise de regressão logística multivariada e análise de eficiência de portfólio de jogadores baseada em DEA. A abordagem é aplicada a um dos times da National Basketball Association, com resultados promissores, com a precisão chegando a ser de quase 82,15%, o R2 de McFadden de 0,3791 e a estatística de LR de 150,6680, também confirmando um bom ajuste para a previsão do desempenho das equipes.[7]

Outro estudo, intitulado "A gamma process based in-play prediction model for National Basketball Association games", propõe uma metodologia para prever resultados de jogos da NBA em tempo real, com base na análise de dados dos jogos em andamento. O estudo usa técnicas de modelagem baseada em processos estocásticos, os quais são modelos matemáticos que descrevem a evolução aleatória de um sistema ao longo do tempo, para prever a pontuação final dos jogos. O artigo apresenta resultados promissores na previsão de resultados de jogos da NBA em tempo real, obtendo como principal métrica o RMSE abaixo de 19.62.[8]

Por fim, em "The use of data mining for basketball matches outcomes prediction", os autores propõem uma metodologia para prever resultados de jogos de basquete, usando técnicas de mineração de dados para identificar padrões nos resultados históricos de jogos. Os resultados mostram que a metodologia proposta tem um bom desempenho na previsão de resultados de jogos de basquete, chegando a predizer corretamente 67% dos vencedores entre as 778 partidas disputadas durante a temporada regular de 2009/2010.[9]

Esses estudos mostram que a análise preditiva pode ser uma ferramenta poderosa para prever resultados em esportes, incluindo a NBA. Com base nesses trabalhos, propomos uma metodologia para a análise preditiva de resultados de jogos esportivos, utilizando jogos da NBA como estudo de caso. A metodologia proposta inclui a coleta de dados, pré-processamento, seleção de variáveis, modelagem e avaliação, com o objetivo de fornecer uma metodologia eficaz para a previsão de resultados de jogos esportivos.

3. SOLUÇÃO

3.1 Descrição

A solução desenvolvida consiste nas seguintes etapas: aquisição de dados, pré-processamento, filtragem, engenharia de características, redução de dimensionalidade, treinamento de modelos e análise de resultados. (Figura 1)

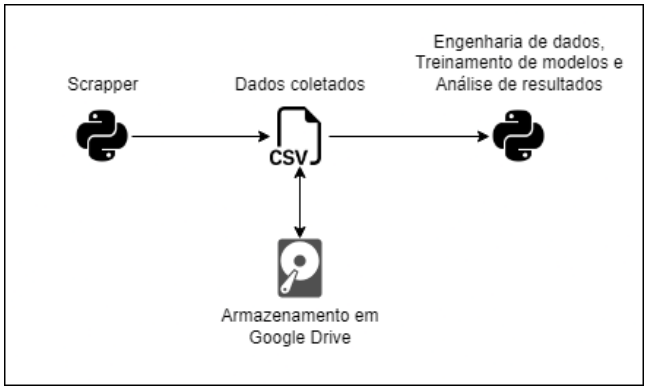

Figura 1. Arquitetura da solução.

A solução foi desenvolvida utilizando dois notebooks Jupyter. O primeiro notebook, responsável pela coleta de dados, utiliza a técnica de raspagem de dados para coletar informações sobre os jogos da NBA. Ele acessa o site oficial da NBA, coleta os dados de todas as partidas da temporada regular realizadas desde a temporada 2018/19 até a 2022/23, e exporta em um arquivo *Comma-Separated Values* (CSV). Esse notebook também faz a limpeza e a preparação dos dados para serem utilizados posteriormente. Já o segundo arquivo, responsável pela apresentação dos dados e treinamento dos modelos, também realiza a análise exploratória dos dados coletados e os apresenta em gráficos e tabelas. Em seguida, é realizado o treinamento de diversos modelos de aprendizado de máquina, alguns mais simples e outros mais complexos e robustos.

O segundo notebook também avalia a performance dos modelos treinados e apresenta os resultados das previsões em forma de gráficos e métricas. Dessa forma, é possível avaliar quais modelos apresentam melhor desempenho na previsão dos resultados de jogos da NBA.

Essa solução pode ser útil para técnicos, jogadores e torcedores que desejam fazer previsões mais precisas e informadas sobre os resultados dos jogos da NBA.

3.2 Ferramentas utilizadas

3.3 Notebook Python

No desenvolvimento dos notebooks python foram utilizados o Google Colab[1] e o Jupyter Notebook[2] que são ambientes de programação interativos, baseados em navegador, que permitem a criação e execução de códigos em linguagens de programação como Python, R, Julia, entre outras.

O Jupyter Notebook é um aplicativo de código aberto que permite a criação e compartilhamento de documentos que contenham códigos, visualizações, anotações e textos explicativos. Ele permite a execução de códigos em blocos, o que facilita a interatividade e a experimentação, além de ser uma ferramenta popular para análise de dados e aprendizado de máquina.

O Google Colab, por sua vez, é uma plataforma de computação em nuvem gratuita fornecida pelo Google, que permite a criação e execução de notebooks Jupyter. Ele permite o acesso a GPUs e TPUs de alta potência, o que torna a plataforma ideal para a execução de códigos que demandam maior poder de processamento. Além disso, o Colab também permite o compartilhamento dos notebooks e o trabalho colaborativo entre os usuários.

Esses ambientes também fornecem bibliotecas auxiliares para a análise de dados e desenvolvimento dos modelos que foram amplamente utilizados nesse trabalho como: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Keras e entre outras.

4. METODOLOGIA

Inicialmente foi realizada a coleta de dados por meio da criação de um script em Python para raspar dados das estatísticas semanais dos times da NBA, utilizando o *Power Ranking* oficial da associação americana de basquete, que constitui das estatísticas de eficiência ofensiva e defensiva dos times, que são respectivamente as médias de pontos que um time produz e sofre a cada 100 posses de bola. Isso também aparece em estatísticas individuais. Neste caso, os números mostram a eficiência ofensiva e a defensiva da equipe com o jogador em quadra, o *Net Rating*, que é o saldo de pontos de um time a cada 100 posses de bola e o *Pace* no qual se refere ao número de posses de bola que uma equipe tem em média por partida ao longo da temporada. Também foi realizado a coleta dos resultados das partidas para poder associar ambos os dados, nos quais foram armazenados em um arquivo CSV. Na Tabela 1, tem-se a quantidade de jogos coletados por cada temporada.

Temporada	Amostras
18/19	1203
19/20	1017
20/21	1046
21/22	1230
22/23	1229

Tabela 1. Total de amostras de dados por temporada.

A próxima etapa consistiu na preparação dos dados para treinamento dos modelos de aprendizado de máquina. Nessa etapa, foram realizadas diversas operações de limpeza de dados, como remoção de dados duplicados, preenchimento de valores faltantes, normalização dos dados e transformação de variáveis categóricas em variáveis numéricas quando necessário. Também foram aplicadas técnicas manuais de engenharia de características, visando criar novas características a partir das originais, ao realizar transformações matemáticas , como também foi realizada uma Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) que é uma técnica estatística de redução de dimensionalidade que tem como objetivo encontrar padrões nos dados, reduzindo a complexidade dos mesmos, mantendo ao mesmo tempo a maior parte das informações relevantes. A técnica consiste em transformar um conjunto de variáveis correlacionadas em um conjunto menor de variáveis não correlacionadas, chamadas de componentes principais. Esses componentes principais são ordenados de forma a representar a maior quantidade possível de variância nos dados originais. A primeira componente principal representa a maior variação nos dados, a segunda representa a segunda maior variação, e assim por diante.[10] (Figura 2)

[1] https://colab.research.google.com/
[2] https://jupyter.org/

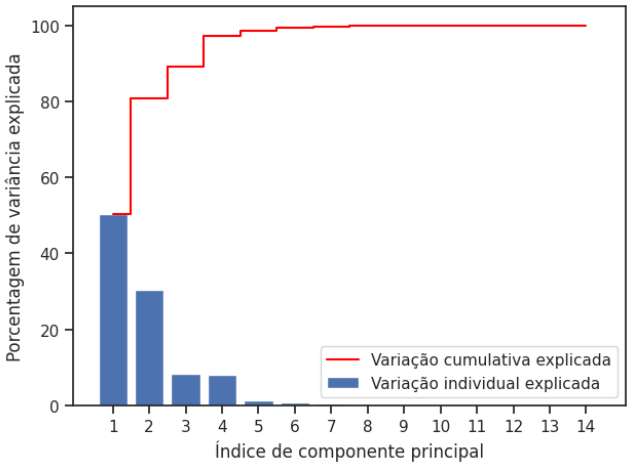

Figura 2. Gráfico demonstrando as características com a maior quantidade de variância dos dados originais, demonstrando a viabilidade de reduzir para 4 características, mantendo o maior número de variância.

Após isso foi feita a análise exploratória destes dados, na qual foram utilizadas as bibliotecas Pandas, Matplotlib e Seaborn para visualização e manipulação dos dados coletados. Nessa etapa, foram realizadas diversas análises estatísticas e visualizações para identificar padrões, tendências e possíveis correlações nos dados entre as estatísticas de Net e Defesa dos times, onde ao ter um aumento da estatística de defesa dos times, o Net tende a diminuir, como também uma relação mais direta das estatísticas ofensivas do time com a vitória do que as estatísticas defensivas, como podemos observar nas figuras 3 e 4.

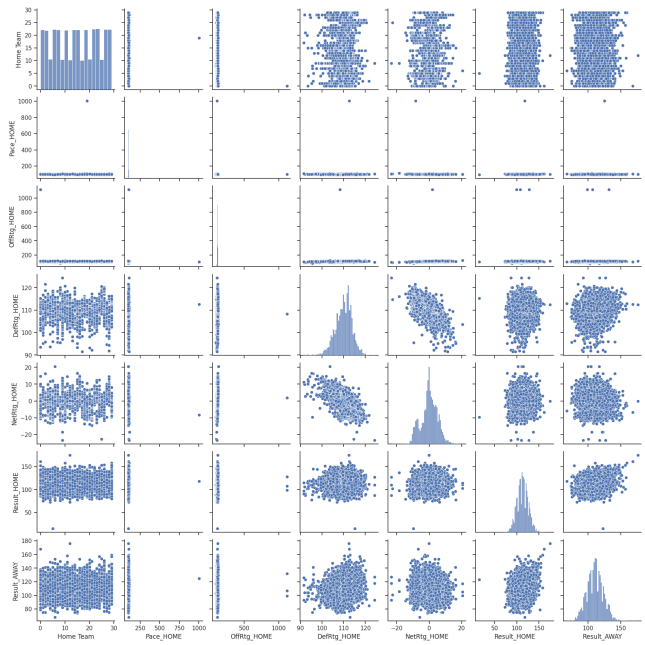

Figura 3. Exemplo matriz de gráficos para analisar a relação entre múltiplas variáveis numéricas.

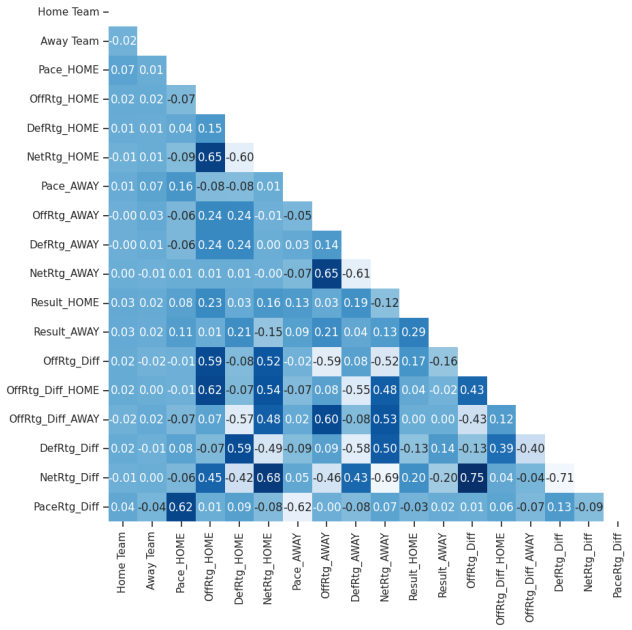

Figura 4. Gráfico de correlação entre as colunas do *DataFrame*, utilizando o coeficiente de correlação de Spearman.

Por fim, foi realizada a seleção e treinamento dos modelos de aprendizado de máquina, que visa obter a previsão do resultado final de uma partida, medindo a pontuação do time da casa e a pontuação do time visitante. Nessa etapa, foram selecionados quatro algoritmos diferentes: *Ridge*, *Random Forest*, e duas redes neurais, uma simples e uma contendo camada LSTM.

A próxima etapa consistiu na avaliação dos modelos e na apresentação dos resultados.

5. EXPERIMENTOS E RESULTADOS

Nessa etapa, foram analisadas as métricas de desempenho dos modelos treinados, como o *Mean Squared Error* (MSE), conhecido em português como Erro Quadrático Médio, *Mean Absolute Error* (MAE), conhecido em português como Erro Médio Absoluto e *R-squared* (R2), conhecido em português como Coeficiente de Determinação, além de visualizações das as curvas de aprendizado dos modelos em gráficos.

Foram experimentados 4 algoritmos de aprendizagem de máquina, sendo eles: (i) *Ridge*, que é uma variação da regressão linear que adiciona uma penalidade à função de perda, a fim de reduzir a complexidade do modelo e evitar o sobreajuste (*overfitting*); (ii) *Random Forest*, um algoritmo de aprendizado de máquina que é utilizado tanto para problemas de classificação quanto para problemas de regressão, sendo uma técnica de aprendizado

supervisionado que combina várias árvores de decisão individuais em um único modelo preditivo robusto e preciso; (iii) uma rede neural simples e (iv) uma rede neural contendo camada de "*Long Short-Term Memory*" (LSTM, Memória Longa de Curto Prazo, em português), que é uma variação da arquitetura de Rede Neural Recorrente (RNN) projetada para superar o problema do desvanecimento ou explosão do gradiente, que pode ocorrer em RNNs tradicionais ao lidar com sequências de dados de longo prazo. Para os algoritmos *Ridge* e *Random Forest*, foram realizados busca em grade, variando parâmetros como *alpha*, profundidade máxima, número de estimadores e entre outros, visando encontrar os modelos mais próximos do ideal a ser utilizado nos experimentos, em ambos os casos a função de perda utilizada foi a de erro médio absoluto (*Mean Average Error*, MAE). Já no algoritmo de rede neural simples, foi utilizado um modelo sequencial com 3 camadas, sendo a primeira de entrada, a segunda camada com 30 neurônios e a função de ativação de unidade linear retificada, e a terceira camada sendo a de saída com 2 neurônios, para as duas pontuações, como otimizador do modelo foi utilizado o RMSprop, e o cálculo da função de perda foi com base no erro médio quadrático. Para o último algoritmo foi utilizado apenas a camada de LSTM com 128 neurônios, a camada de saída com 2 neurônios, o otimizador escolhido foi o *Adam* e a função de cálculo de perda também foi a de erro médio quadrático.

Após separar 80% dos dados para treinamento e 20% para teste e validação, aleatoriamente, e realizado o treinamento, os seguintes resultados foram obtidos para os algoritmos *Ridge*, *Random Forest*, Rede Neural Recorrente simples e Rede Neural LSTM, com o conjunto de dados de teste com as características selecionadas:

- *Ridge* obteve o **MSE** de 121.18, **MAE** de 8.55 e **R2** de 0.08. Na Figura 5, apresenta-se a curva de aprendizagem do modelo *Ridge* obtida a partir de múltiplos treinamentos considerando um tamanho crescente do conjunto de treinamento.

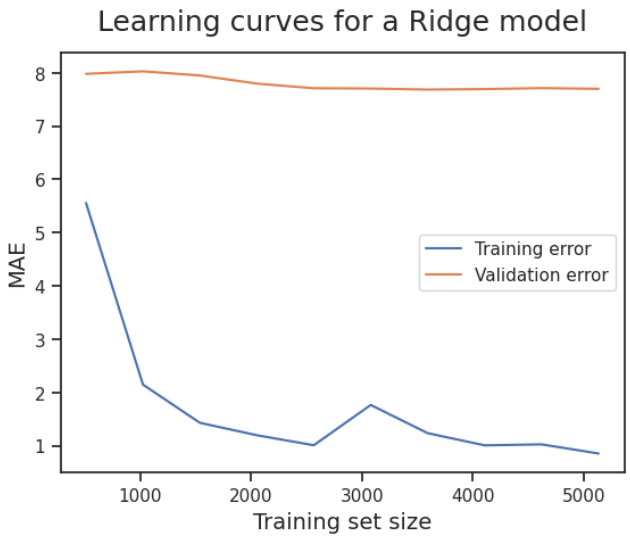

Figura 5. Gráfico de curva de aprendizagem do modelo Ridge.

- *Random Forest* obteve o **MSE** de 138.61, **MAE** de 9.42 e **R2** de -0.03. Na Figura 6, apresenta-se a curva de aprendizagem do modelo de regressão *Random Forest* obtida a partir de múltiplos treinamentos considerando um tamanho crescente do conjunto de treinamento.

Figura 6. Gráfico de curva de aprendizagem do modelo Random Forest.

- *RNN* simples obteve o **MSE** de 118.30, **MAE** de 8.37 e **R2** de 0.10. A Figura 7 contém a evolução da curva de perda (*loss*) do treinamento do modelo ao longo do número de épocas de treinamento, indicando que não houve *overfitting*.

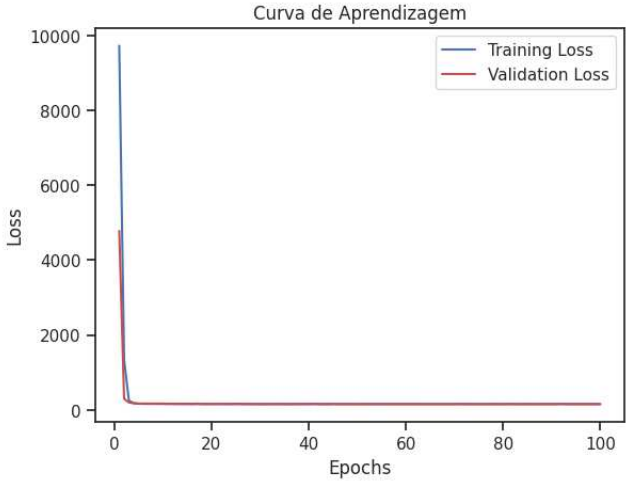

Figura 7. Gráfico de curva de aprendizagem do modelo RNN simples.

- *RNN* com camada LSTM, obteve o **MSE** de 146.49, **MAE** de 9.60 e **R2** de 0.04. A figura número 8 mostra como a curva de perda (loss) do modelo evolui ao longo

das épocas de treinamento, demonstrando que não houve *overfitting* do modelo.

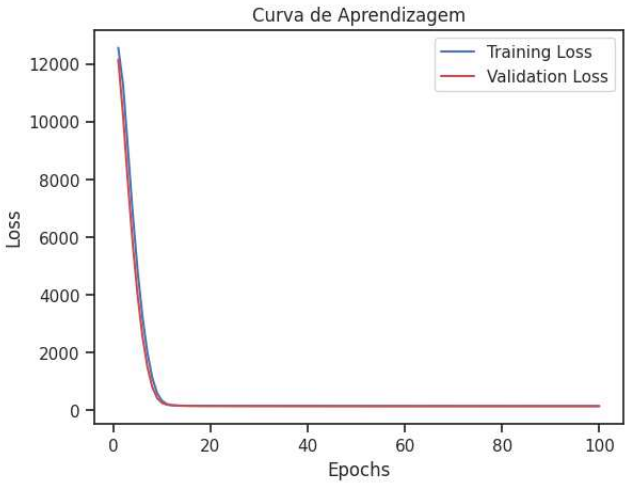

Figura 8. Gráfico de curva de aprendizagem do modelo RNN com camada de memória curta.

Em conclusão, com base nas métricas obtidas, o modelo RNN Simples apresentou o melhor desempenho geral, seguido pelo modelo Ridge. O modelo Random Forest teve um desempenho intermediário, enquanto o modelo RNN com Camada LSTM obteve o pior desempenho. É importante ressaltar que essas conclusões são baseadas nas métricas específicas utilizadas (MSE, MAE e R2). As métricas do modelo RNN simples apresentaram um desempenho semelhante ao modelo Ridge, porém um pouco melhor. O MSE de 118.30 indica um desvio médio quadrático um pouco menor em comparação com o modelo Ridge. O MAE de 8.37 indica um desvio absoluto médio ligeiramente menor, onde demonstra a diferença entre o valor previsto e o valor real. O R2 de 0.10 indica que o modelo consegue explicar cerca de 10% da variância dos dados, mostrando um ajuste um pouco melhor em relação ao modelo Ridge.

6. CONSIDERAÇÕES FINAIS

Nesta seção são apresentadas as principais conclusões obtidas ao final do trabalho. Além disso, descreve-se a experiência propiciada pelo processo de desenvolvimento dos Notebooks em Python para coleta e análise dos dados, e modelagem dos algoritmos. Por fim, são comentados os desafios e propostas para trabalhos futuros.

Em conclusão, este trabalho explorou a aplicação de diferentes modelos de aprendizado de máquina na análise preditiva de um determinado conjunto de dados. Os modelos avaliados incluíram Ridge, Random Forest, RNN Simples e RNN com Camada LSTM. Os resultados indicaram que o modelo RNN Simples teve o melhor desempenho, seguido pelo modelo Ridge. Ambos apresentaram capacidade razoável de predição, embora o poder de explicação ainda seja limitado. É importante ressaltar que as conclusões aqui apresentadas são específicas para o conjunto de dados e as métricas de avaliação utilizadas neste estudo. Outros fatores, como o tamanho do conjunto de dados, a qualidade dos recursos disponíveis e a complexidade do problema, podem influenciar os resultados e, consequentemente, as escolhas de modelo. Em suma, este trabalho contribuiu para a compreensão das capacidades e limitações dos modelos de aprendizado de máquina aplicados à análise preditiva. Os resultados obtidos oferecem uma base sólida para a seleção de modelos em tarefas semelhantes e destacam a importância da avaliação criteriosa e da escolha adequada de métricas de desempenho para um processo de modelagem eficaz.

6.1 Processo de desenvolvimento

Durante o desenvolvimento da metodologia, seguimos um processo iterativo que envolveu as seguintes etapas: revisão da literatura existente sobre análise preditiva de resultados esportivos, coleta e preparação dos dados históricos da NBA, implementação e treinamento de modelos de análise preditiva utilizando técnicas de aprendizado de máquina e avaliação dos modelos desenvolvidos por meio de métricas apropriadas, testes de validação cruzada e análise de sensibilidade.

6.2 Principais Desafios

Durante o trabalho, enfrentamos desafios significativos que impactaram o desenvolvimento da metodologia e os resultados obtidos. A dificuldade de obter dados através da técnica de raspagem no site oficial da NBA foi um dos principais desafios encontrados. A complexidade do domínio da análise preditiva de jogos esportivos, visto que o esporte é dinâmico e influenciado por múltiplos fatores, o que tornou difícil identificar as variáveis mais relevantes e capturar adequadamente a interação entre elas, juntamente com a modelagem e otimização dos modelos de aprendizado de máquina, também apresentaram desafios importantes. Esses desafios ressaltam a importância de lidar com dados confiáveis e de alta qualidade, compreender a complexidade do domínio esportivo e adotar uma abordagem cuidadosa na modelagem e otimização dos modelos.

6.3 Limitações e Trabalhos Futuros

Apesar dos esforços realizados, nosso trabalho apresenta algumas limitações importantes que podem ser abordadas em trabalhos futuros. Um dos principais pontos é o tamanho limitado do dataset utilizado. Embora tenhamos usado um conjunto considerável de dados históricos, é necessário ressaltar que a quantidade de dados ainda pode ser limitada em relação à complexidade do problema. Para superar essa limitação, um trabalho futuro poderia explorar a obtenção de um dataset mais robusto, abrangendo um período maior de tempo e incluindo informações adicionais relevantes.

Além disso, há oportunidades de melhorias no modelo de análise preditiva. Diferentes algoritmos de aprendizado de máquina, como SVM (Support Vector Machines) e outros algoritmos

baseados em *ensemble*, podem ser explorados. Técnicas avançadas de processamento de linguagem natural também podem ser aplicadas para analisar notícias e comentários de especialistas, a fim de incorporar informações externas relevantes.

Embora nosso estudo de caso tenha se concentrado nos jogos da NBA, a metodologia desenvolvida pode ser generalizada para outros esportes e ligas. Trabalhos futuros podem explorar a aplicação da abordagem em jogos de futebol, beisebol, futebol americano e outros esportes, ampliando o escopo e a aplicabilidade da metodologia desenvolvida.

7. AGRADECIMENTOS

Sinceramente, gostaria de expressar minha gratidão a todas as pessoas que me apoiaram ao longo deste trabalho de conclusão de curso. Agradeço a todos que me prestaram suporte, seja de forma ativa, ajudando com implementações e ideias para o desenvolvimento dos algoritmos, ou estando ao meu lado, sempre me incentivando a continuar.

Gostaria de deixar um agradecimento especial ao meu orientador, Herman Martins Gomes, pelo apoio e direções essenciais para o desenvolvimento deste TCC. Agradeço também à minha mãe, cujo amor, suporte e esperança foram fundamentais para tornar tudo isso possível. Minha família como um todo merece meu agradecimento pela ajuda e apoio prestados.

Aos meus amigos que me acompanharam nessa longa jornada, agradeço do fundo do coração. Embora sejam muitos para mencionar aqui, gostaria de destacar alguns amigos queridos, como Alex Alves, Igor Guimarães, Edson Weslley, Arthur Macena, Heriberto Junior, Flavio Quirino, Igor Franca, entre outros. Agradeço também a todos os colegas de graduação que, a todo momento, estiveram dispostos a me ajudar durante o curso, em especial Tulio, Guilherme e Carol, obrigado por me acolherem em seus grupos de estudos.

Todos esses momentos que compartilhamos e vivemos juntos durante esses anos serão eternamente lembrados em minha memória.

“Gratidão, integridade, honestidade, papo reto e só visão, eu sei que toda glória vai ser dada a Deus, mas não posso esquecer daquele que me deu a mão, daqueles, porque foi mais de um sem eles, lugar nenhum” - dos Santos, Lennon

“A amizade é uma predisposição recíproca que torna dois seres igualmente ciosos da felicidade um do outro “ - Platão

8. REFERENCES

[1] BAI, Z.; BAI, X. Sports Big Data: Management, Analysis, Applications, and Challenges. Complexity, v. 2021, p. 1–11, 30 jan. 2021.

[2] KESHTKAR LANGAROUDI, MILAD; YAMAGHANI, M. Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey. Journal of Advances in Computer Engineering and Technology, v. 5, n. 1, p. 27–36, fev. 2019.

[3] SARLIS, V.; TJORTJIS, C. Sports analytics — Evaluation of basketball players and team performance. Information Systems, v. 93, p. 101562, nov. 2020.

[4] KILCOYNE, S. The Decline of the Mid-Range Jump Shot in Basketball: A Study of the Impact of Data Analytics on Shooting Habits in the NBA. Honors Projects in Mathematics, 1 nov. 2020.

[5] Thabtah, Fadi, et al. “NBA Game Result Prediction Using Feature Analysis and Machine Learning.” Annals of Data Science, vol. 6, no. 1, 3 Jan. 2019, pp. 103–116, https://doi.org/10.1007/s40745-018-00189-x.

[6] Wang, Tian. Predictive Analysis On Esports Games: A Case Study On League of Legends (lol) Esports Tournaments. 2018. https://doi.org/10.17615/ez9n-t517

[7] Li, Yongjun, et al. “A Data-Driven Prediction Approach for Sports Team Performance and Its Application to National Basketball Association.” Omega, 25 Sept. 2019, p. 102123, www.sciencedirect.com/science/article/pii/S0305048319302002, https://doi.org/10.1016/j.omega.2019.102123.

[8] Song, Kai, and Jian Shi. “A Gamma Process Based In-Play Prediction Model for National Basketball Association Games.” European Journal of Operational Research, vol. 283, no. 2, June 2020, pp. 706–713, https://doi.org/10.1016/j.ejor.2019.11.012. Accessed 14 July 2021.

[9] Miljković, D., et al. “The Use of Data Mining for Basketball Matches Outcomes Prediction.” IEEE Xplore, 1 Sept. 2010, ieeexplore.ieee.org/document/5647440. Accessed 3 Mar. 2021.

[10] BOSCHETTI, A.; MASSARON, L. Python Data Science Essentials : a Practitioner’s Guide Covering Essential Data Science Principles, Tools, and Techniques, 3rd Edition. Birmingham: Packt Publishing Ltd, 2018.

[11] NBA. Site da *National Basketball Association*, 2023. The official site of the National Basketball Association. Disponível em: https://www.nba.com/. Acesso em: 06 jan. 2023.

[12] Beautiful Soup. Beautiful Soup Documentation. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 08 jan. 2023.