CONIC SEMESP

23° Congresso Nacional de Iniciação Científica

TÍTULO: RESULTADOS DA COPA DO MUNDO DE FUTEBOL MASCULINO 2022: ANÁLISE PREDITIVA ORIENTADA A DADOS

CATEGORIA: CONCLUÍDO

ÁREA: CIÊNCIAS EXATAS E DA TERRA

SUBÁREA: Computação e Informática

INSTITUIÇÃO: INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS - IFMG

AUTOR(ES): MATHEUS HENRICK DIAS

ORIENTADOR(ES): CARLOS ALEXANDRE SILVA, DANILO BOECHAT SEUFITELLI

Realização: SEMESP

IES parceiras: UNIFRAN Universidade de Franca, Unit Centro Universitário Tiradentes, Facens

1. RESUMO

Neste trabalho, apresentamos uma abordagem para análise preditiva de resultados da Copa do Mundo de Futebol Masculino, utilizando quatro algoritmos de aprendizagem supervisionada aplicados aos dados históricos das seleções antes do início do torneio de 2022. O objetivo foi obter uma estimativa precisa acerca do time vencedor, mesmo antes do início da primeira partida, utilizando dados de jogos das edições anteriores. Os experimentos demonstram que as seleções tendem a seguir padrões de desempenho consistentes a médio e longo prazo, mesmo que ocorram mudanças no elenco e na comissão técnica. Além disso, os métodos de aprendizagem supervisionada mostraram-se eficazes para prever resultados nesse cenário com variáveis limitadas. Essa abordagem realizada proporciona uma visão promissora para a previsão de resultados esportivos, permitindo uma análise antecipada e otimista dos resultados da Copa do Mundo de Futebol, além de demonstrar possibilidade de utilização das ferramentas de inteligência artificial na solução desse tipo de problema.

2. INTRODUÇÃO

O futebol é considerado o esporte mais popular do mundo, com cerca de 250 milhões de jogadores, além de uma base sólida de atletas profissionais[1]. A Copa do Mundo de Futebol é o evento mais prestigiado no cenário do futebol profissional, atraindo bilhões de espectadores e despertando interesses sociais e econômicos. Com o avanço tecnológico e a digitalização dos registros esportivos, uma enorme quantidade de dados é coletada em cada partida, incluindo estatísticas de desempenho e informações contextuais. No entanto, uma compreensão profunda desses dados requer o uso de ferramentas computacionais modernas. Com isso, tais dados são o cerne da aplicação de técnicas de análise preditiva, que revelam padrões e tendências que podem influenciar o desempenho das equipes e as chances de vitória.

[1] FIFA: https://www.fifa.com/

A análise preditiva no futebol tem relevância tanto social quanto econômica. A Copa do Mundo de Futebol atrai uma audiência global massiva, com bilhões de pessoas assistindo à final. Esse alcance reflete o interesse comum das pessoas pelo futebol, independentemente de sua nacionalidade ou cultura. Além disso, as apostas esportivas são uma vertente importante deste cenário, impulsionadas pela globalização e democratização das práticas esportivas. O mercado de apostas esportivas tem apresentado um crescimento significativo, e a análise preditiva pode fornecer *insights* valiosos para os apostadores. No Brasil houve um crescimento no volume de apostas de 2 bilhões de reais em 2018 para 7 bilhões em 2022. Globalmente, este mercado foi avaliado em 59,6 bilhões de dólares em 2020, podendo alcançar cerca de 127,3 bilhões de dólares em 2027[2].

O futebol é capaz de comover multidões de fãs ao redor do mundo e instigar forte apelo financeiro. Para explorar as receitas, é comum a presença de bolsas de apostas na Europa e Estados Unidos, que oferecem uma plataforma de negociação para que apostadores atuem como “corretores” e apostem a favor ou contra eventos esportivos. No Brasil, as apostas vem crescendo devido a uma brecha legal explorada pelas casas de apostas, que não poderiam atuar legalmente no país . A área de predição de resultados esportivos possui alto potencial de aplicações. O avanço da inteligência artificial e a disponibilidade de uma quantidade crescente de dados no campo esportivo criam oportunidades para explorar e aprimorar abordagens de análise preditiva. Um exemplo disso é o estudo realizado por Csat’o (2020), que investigou a imparcialidade das partidas durante as eliminatórias da Copa do Mundo por meio de cálculos de probabilidade. Além disso, Iskandaryan et al. (2020) abordaram os impactos das condições climáticas nos resultados do futebol, utilizando técnicas de aprendizado de máquina.

Conforme abordado por Bunker e Susnjak (2022), a previsão de desfechos esportivos não constitui um tópico recente na literatura. Entretanto, o pioneiro estudo que incorporou técnicas de aprendizado de máquina é de natureza relativamente nova, datando de meados de 1996 . Recentemente, LIMA (2022) aplicou algoritmos de aprendizado em dados de cinco ligas nacionais de futebol (Brasil, Inglaterra,

[2] O Mercado de Apostas Esportivas: https://bit.ly/globo-apostas

Itália, Espanha e França). O objetivo foi encontrar a melhor combinação entre algoritmo e liga visando otimizar a precisão das previsões. De forma similar, Peconick (2028) utilizou Redes Neurais Artificiais (RNAs) para prever os resultados da Copa de 2018. Ela destaca que as RNAs são boas ferramentas de predição alcançando resultados similares ao mercado de apostas.

Neste contexto, este trabalho apresenta uma análise exploratória de dados históricos das seleções de futebol, utilizando algoritmos de aprendizagem supervisionada, como Decision Tree, Random Forest, K-Nearest Neighbors e Gaussian Naive Bayes, para prever resultados da Copa do Mundo de Futebol de 2022. Ao analisar variáveis históricas, buscamos identificar padrões de desempenho no médio e longo prazo. Além disso, discutimos os desafios e limitações enfrentados nesse processo, destacando a relevância da análise preditiva para diferentes públicos envolvidos no futebol.

O restante deste artigo é organizado como segue. A introdução e revisão bibliográfica são apresentados na Seção 2 trazendo mais contexto sobre o tema. Os objetivos são citados na Seção 3. A metodologia é detalhada na Seção 4 onde são explicadas as bases de dados e os algoritmos utilizados. O desenvolvimento é explicado na Seção 5, apresentando e discutindo os resultados obtidos. Os resultados da análise preditiva são apresentados na Seção 6, junto às propostas para trabalhos futuros e as principais limitações e ameaças à validade. Por fim, na Seção 7, são fornecidas as considerações finais.

3. OBJETIVOS

O objetivo deste trabalho é explorar o potencial preditivo de métodos de aprendizagem supervisionada aplicados a dados prévios ao início da copa do mundo de futebol Masculino de 2022. Busca-se avaliar o desempenho de ferramentas de Inteligência Artificial na predição de resultados e a modelagem dos dados históricos mais eficiente para este contexto.

4. METODOLOGIA

A metodologia deste trabalho é dividida em quatro etapas: coleta dos dados, limpeza e enriquecimento dos dados, seleção dos algoritmos de treinamento, e por fim, a análise preditiva. A seguir, cada uma das etapas do estudo são apresentadas e discutidas.

A base de dados foi construída unindo dois datasets disponíveis na plataforma Kaggle[3] no formato CSV. Tais bases estão em domínio público na plataforma, e são atualizadas à medida que as partidas acontecem. A Tabela 1 cita, explica e exemplifica todas as colunas de ambos conjuntos de dados.

Tabela 1. Descrição dos metadados do conjunto de dados Dataset 1.

Dataset 1		
Coluna	**Armazena**	**Exemplo**
date	Data de quando ocorreu a partida	1872-11-30
home_team	Nome do time da casa	Scotland
away_team	Nome do time visitante	England
home_score	Placar do time da casa	0.0
away_score	Placar do time visitante	4.0
tournament	Tipo de torneio	Friendly
city	Cidade onde a partida ocorreu	Glasgow
country	País onde a partida ocorreu	Scotland
neutral	Território neutro (se a partida foi realizado fora do país das duas seleções)	true / false
Dataset 2		
Coluna	**Armazena**	**Exemplo**
rank	Posição ocupada no ranking FIFA	1,2,3
country_full	Nome completo do país	Madagascar, Qatar
country_abrv	Abreviação do nome do país	MAD, QAT
total_points	Quantidade de pontos no momento da publicação	18.0, 28.0, ...
previous_points	pontos totais no último ranking	0.0, 4.0, ...
rank_change	Mudança no ranking desde a última publicação	1,0,-1
confederation	Confederações da Fifa as quais os países fazem parte	CAF, AFC, ...
rank_date	Data de publicação do ranking	1992-12-31

O primeiro conjunto de dados (Dataset 1), intitulado de *International football results from 1872 to 2022*, possui cerca de 44.060 registros, e inclui colunas como datas das partidas e número de gols marcados. É importante destacar que o resultado da partida é calculado pela diferença entre gols marcados por cada equipe, sendo essa a informação mais importante em nossa modelagem. Além disso, é necessário identificar quais variáveis explicativas podem ter influência sobre o resultado do jogo. Alguns exemplos comuns incluem o desempenho recente das equipes, estatísticas

[3] Kaggle: https://www.kaggle.com/

de posse de bola, número de chutes a gol, histórico de confrontos diretos, ranking das seleções, entre outros. Neste trabalho são utilizadas as variáveis de desempenho em jogos anteriores e o ranking de seleções da FIFA, que podem ser obtidos antes do início do campeonato. O segundo conjunto de dados (Dataset 2), *FIFA World Ranking 1992-2022* possui 64.127 registros. Este dataset inclui colunas como ranking da FIFA em determinada data, o nome do país e sua abreviação, a confederação ao qual o país faz parte, o total de pontos atual e anterior a edição do ranking utilizado para definir a colocação, além da variação da seleção no ranking. Com tais dados, é possível definir a posição de cada time na data de cada partida disputada na base de dados anterior.

Após a junção dos dados, foi necessário aplicar uma série de tarefas para prover a limpeza e padronização dos dados. Constavam valores divergentes nos dados, tais como países e suas respectivas seleções que mudaram de nome (por razões políticas), siglas ao invés de nomes completos, ou ainda erros de grafia. Além disso, foi construída uma série temporal entre 2002 a 2022 para traçar um perfil mais condizente com a realidade atual das equipes. Esse intervalo considera a variação na convocação de jogadores ao longo de 20 anos, que compreendem o período entre as últimas cinco Copas, que é o tempo médio de participação dos principais jogadores de cada equipe. Assim, a base de dados construída é composta por todas as partidas disputadas pelas seleções durante o período da série temporal, na qual foi possível correlacionar o ranking de classificação das equipes correspondente à data em que a partida foi disputada. Em seguida, realizou-se a preparação dos dados de treinamento, os quais consistem em pares de entrada-saída, ou seja, as variáveis explicativas e os resultados reais dos jogos passados. Cada jogo foi representado por um conjunto de valores para as variáveis explicativas e um valor correspondente para o resultado (0: derrota, 1: empate, 2: vitória).

Após os ajustes dos dados, a próxima etapa requereu escolher os melhores algoritmos de aprendizado de máquina. Para a predição dos resultados foram escolhidos quatro algoritmos de aprendizagem supervisionada: Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN) e Gaussian Naive Bayes (GNB). Tais algoritmos foram selecionados em virtude de suas generalizações e por

demandarem baixo tempo de processamento, mesmo com grande volume de dados. A nível de implementação, foi utilizado a biblioteca scikit-learn versão 1.2.2, que contempla implementações dos algoritmos escolhidos.

Por fim, foram implementados os modelos preditivos utilizando Python 3.10.11. Os dados foram divididos em um esquema de treino e teste aleatoriamente. Especificamente, 30% dos dados foram utilizados para teste, enquanto os outros 70% foram para o treinamento do modelo. A análise dos resultados da fase teste foi medida pela acurácia utilizando cross validation. Tal métrica possibilita a análise em cenários onde a base de dados foi separada aleatoriamente entre treino e teste. Desta forma, implementamos uma função que recebe o modelo e a métrica analisada, e aplicamos a validação cruzada por meio do algoritmo K-Fold com 10 folds. Para realizar a previsão, utilizamos o modelo treinado para estimar a probabilidade de vitória, empate ou derrota para cada equipe em um determinado confronto. Com base nessas probabilidades, pode-se tomar decisões informadas, como apostas ou análises táticas.

5. DESENVOLVIMENTO

Nesta seção apresentamos e discutimos os resultados obtidos. Este trabalho foi dividido em duas etapas: treino e teste, que serão detalhadas a seguir.

Inicialmente, criamos um modelo utilizando os quatro classificadores (DT, RF, KNN e GNB), e os treinamos com dados de partidas passadas compreendidas entre 2002 e 2018 e ranking das seleções. Em seguida, simulamos rodadas para definir se o time da casa venceria, perderia ou empataria a partida isoladamente. Em seguida, utilizamos o modelo treinado para predizer os resultados da base de teste. Executamos dez rodadas de predição, e os algoritmos DT e RF retornaram resultados diferentes a cada execução, enquanto os algoritmos KNN e GNB que retornaram o mesmo resultado. A acurácia média dessas simulações foi de 78,81% para GNB, 65,89% para RF, 60,44% para KNN e 59,38% para DT. Também analisamos o desempenho dos classificadores utilizando a curva ROC (Receiver Operator Characteristic) para verificar a taxa de falsos positivos e verdadeiros positivos. Os métodos GNB e RF demonstraram boa capacidade de generalização

dos dados de treino, sobressaindo aos demais algoritmos. A Figura 1 ilustra este resultado.

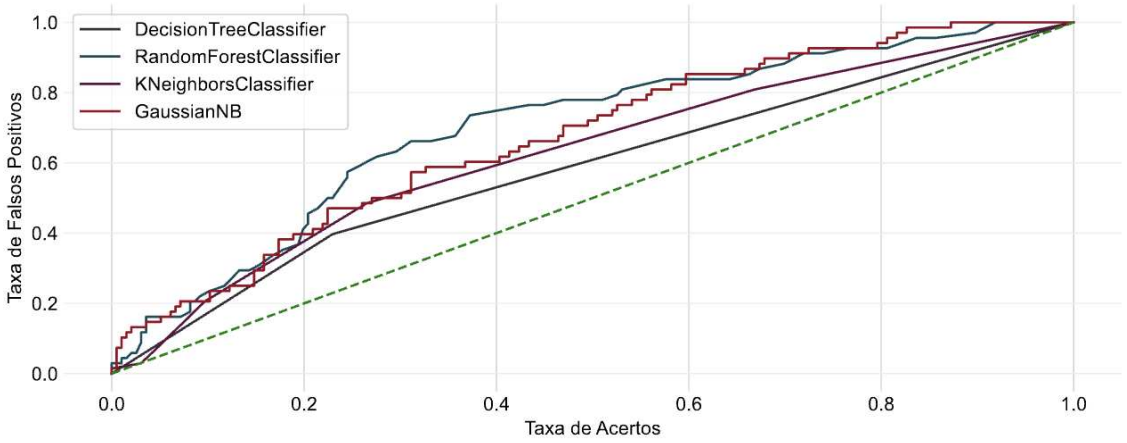

Figura 1. Curva ROC dos algoritmos DT, RF, KNN e GNB. A curva representa o desempenho dos classificadores na previsão dos resultados, sendo que uma taxa de acertos mais próxima de 1 indica um melhor desempenho. Note que a linha tracejada em verde representa um classificador aleatório.

Utilizamos o modelo treinado para predizer o resultado dos jogos da Copa de 2022. Neste caso, os dados de entrada para o modelo contemplam todos os jogos desta edição. Simulamos as condições disponíveis antes do início do campeonato, ou seja, foram utilizados os confrontos reais, incluindo as partidas eliminatórias. Os resultados divergentes dos reais (obtidos através do algoritmo de predição) para a fase de “mata-mata” não foram considerados. Incluímos apenas os jogos que realmente aconteceram no torneio. Dessa forma, obtivemos o vencedor ou empate de cada partida segundo cada classificador, e nessa circunstância o modelo KNN alcançou a melhor acurácia, seguido pelo modelo RF. Os algoritmos GNB e DT ficaram empatados. Apesar do algoritmo KNN obter melhor performance que o RF na edição atual da Copa, contrariamente ao treinamento, essa diferença foi de cinco acertos, sendo que estes dois modelos se mostraram mais eficientes que os demais nesse contexto de variáveis limitadas. Um fato curioso é que o algoritmo Gaussian Naive Bayes foi muito eficiente, generalizando os acertos do time visitante, mas não conseguiu predizer nenhum acerto do time da casa ou empate. A Figura 2 ilustra o desempenho dos algoritmos. Note que acertos_home_win significa predições

corretas de vitória do time da casa, acertos_draw predições corretas de empate, e acertos_away predições corretas de vitória do time visitante.

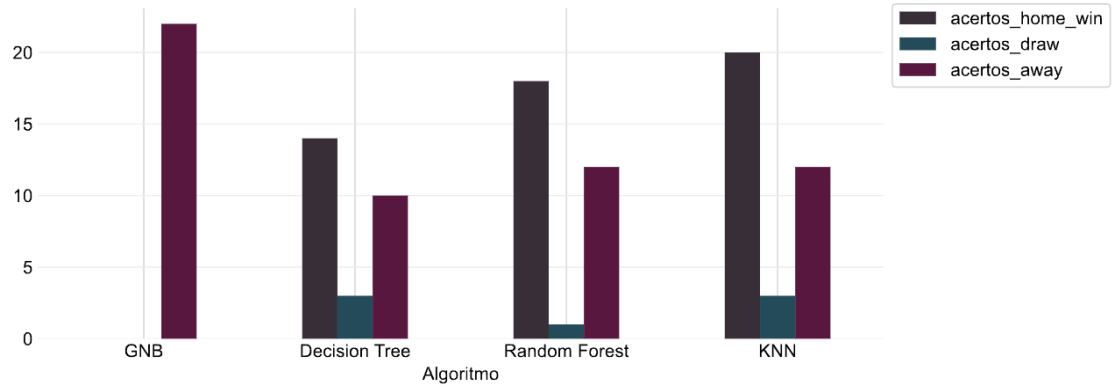

Figura 2. Desempenho de acertos para o time da casa (home win), empate (draw) e time visitante (away) por algoritmo.

6. RESULTADOS

Neste trabalho, apresentamos uma análise do desempenho de um modelo de aprendizado de máquina utilizando os algoritmos de classificação Decision Tree, Random Forest, K-Nearest Neighbors e Gaussian Naive Bayes. Investigamos a capacidade preditiva de tais algoritmos em resultados da Copa do Mundo de 2022 utilizando dados históricos de partidas e o ranking das cinco edições anteriores do torneio. Os algoritmos de aprendizado supervisionado mostraram-se uma ferramenta relevante para lidar com problemas de predição, mesmo em contextos complexos como o futebol, além de reafirmar a possibilidade das predições baseadas em inteligência artificial. Os resultados sugerem que as equipes tendem a apresentar um desempenho semelhante em recortes de tempo curtos, como a série temporal utilizada de 20 anos (5 edições da copa). Os classificadores KNN, RF e DT apresentaram resultados mais sólidos do que GNB. Este último, obteve um bom desempenho no treino, mas acertou somente as vitórias para times visitantes.

Por fim, conclui-se que os algoritmos de aprendizado de máquina têm um grande potencial para auxiliar e prever os resultados da Copa do Mundo. Além disso, melhorias na construção da base de dados podem corroborar com problemas complexos de predição no mundo do futebol. Afinal, a modelagem de dados é crucial

na solução do problema, e a qualidade do resultado está diretamente relacionada a esses dados. Como proposta para trabalhos futuros, espera-se acrescentar outros dados determinantes para o resultado das partidas ao modelo tais como odds das casa de aposta, jogadores convocados para a competição e suas estatísticas de habilidade, além de escalações e formações táticas das equipes de modo a aprimorar as predições. Além disso, pretendemos reavaliar o algoritmo GNB para melhor compreender o baixo desempenho na predição da Copa de 2022, contrariando o desempenho no treino.

7. CONSIDERAÇÕES FINAIS

Fatores imprevisíveis como lesões, decisões arbitrárias e momentos de inspiração individual podem influenciar o resultado final das previsões do futebol. Além disso, a intensidade emocional e a pressão de uma competição tal como a Copa do Mundo, adicionam elementos difíceis de quantificar aos modelos preditivos. Além das dificuldades inerentes ao esporte, o uso de dados anteriores ao início do campeonato também apresenta seus próprios desafios. Embora esses dados sejam acessíveis, organizá-los em um conjunto de treinamento requer um estudo detalhado de datas para combiná-los com dados históricos. A obtenção de um conjunto de dados coeso e abrangente é crucial para construir modelos de análise preditiva mais precisos.

8. FONTES CONSULTADAS

BUNKER, Rory; SUSNJAK, Teo. The application of machine learning techniques for predicting match results in team sport: A review. **Journal of Artificial Intelligence Research**, v. 73, p. 1285-1322, 2022.

CSATÓ, László. Quantifying the unfairness of the 2018 FIFA World Cup qualification. **International Journal of Sports Science & Coaching**, v. 18, n. 1, p. 183-196, 2023.

HASSAN, Amr et al. Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. **Sensors**, v. 20, n. 11, p. 3213, 2020.

ISKANDARYAN, Ditsuhi et al. The effect of weather in soccer results: an approach using machine learning techniques. **Applied Sciences**, v. 10, n. 19, p. 6750, 2020.

LIMA, João Henrique Martins. **Aplicação de machine learning para apostas esportivas: uso de regressão logística, SVM, árvore de decisão e Naive Bayes**. 2022. Trabalho de Conclusão de Curso.

LIRA, PEM et al. Os desafios para a regulamentação das apostas esportivas frente ao sistema jurídico brasileiro. 2018.

MIGLIORATI, Manlio; MANISERA, Marica; ZUCCOLOTTO, Paola. Integration of model-based recursive partitioning with bias reduction estimation: a case study assessing the impact of Oliver's four factors on the probability of winning a basketball game. **AStA Advances in Statistical Analysis**, v. 107, n. 1-2, p. 271-293, 2023.

NGUYEN, Nguyen Hoang et al. The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. **Journal of Information and Telecommunication**, v. 6, n. 2, p. 217-235, 2022.

PECONICK, Laura Defranco Ferreira. Inteligência artificial aplicada à previsão de jogos de futebol. 2018.

PURUCKER, Michael C. Neural network quarterbacking. **Ieee Potentials**, v. 15, n. 3, p. 9-15, 1996.

RICHTER, Chris; O'REILLY, Martin; DELAHUNT, Eamonn. Machine learning in sports science: challenges and opportunities. **Sports Biomechanics**, p. 1-7, 2021.

SADOCCO, Rafael Rodolfo Sartorelli; PINTO, Thais Bueno; DA SILVA, Gladistone Soares Lopes. A ENTRADA DOS SITES DE APOSTAS ESPORTIVAS NO MERCADO BRASILEIRO. **Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)**, v. 5, n. 1, 2021.

VAN DER MAAS, Mark; CHO, S. Ray; NOWER, Lia. Problem gambling message board activity and the legalization of sports betting in the US: A mixed methods approach. **Computers in Human Behavior**, v. 128, p. 107133, 2022.