Análise de Predição de Campeões de Futebol Utilizando Machine Learning

Marcelo soares

25/09/2024

Introdução

A motivação deste estudo foi entender como estatísticas no futebol podem ser utilizadas para extrair informações preditivas. O objetivo inicial foi levantar quais dados são importantes para predizer o próximo campeão das principais ligas de futebol Europeias, utilizando o database European Soccer Database, do Kaggle. Futuramente, este modelo pode ser aprimorado com métricas mais detalhadas, como posse de bola, estilo de jogo,média de público em jogos em casa, e número de jogadores lesionados. Este trabalho usa dois algoritmos de aprendizado supervisionado: Random Forest e K-NN, para predizer se um time será campeão.

Fundamentos Teóricos e Metodológicos

Random Forest é um método de aprendizado baseado em árvores de decisão, onde várias árvores são criadas e os resultados são combinados para melhorar a acurácia. O modelo é robusto contra *overfitting*, especialmente quando há muitas features disponíveis.

K-NN (K-Nearest Neighbors) classifica novos pontos de dados com base na proximidade a outros pontos, utilizando a distância euclidiana. É um método simples, mas pode ser menos eficiente quando há muitas variáveis.

Aplicação

1. Pré-Processamento (Limpeza e Preparação dos dados)

A preparação dos dados foi realizada em várias etapas para garantir que apenas informações relevantes fossem mantidas e que novas features úteis fossem adicionadas.

Inicialmente, apenas as ligas de interesse foram selecionadas: Inglaterra, Itália e Alemanha, abrangendo as temporadas de 2008/2009 a 2015/2016.

Em seguida, as partidas foram filtradas para incluir apenas aquelas das ligas selecionadas. As features mais importantes foram escolhidas, resultando em um dataframe contendo as seguintes colunas: season, date, league_id, home_team_id, away_team_id, home_team_goal, away_team_goal. Essa filtragem garantiu que o conjunto de dados fosse mais manejável e focado.

As informações dos times foram unificadas com os dados das partidas, formando um dataframe que indicava quais times participaram de cada jogo e quantos gols cada um marcou.

A partir do dataframe unificado, foram calculados os gols marcados e sofridos por cada time, tanto em casa quanto fora. Novas colunas foram criadas para representar esses totais:

- Gols Marcados em casa
- Gols sofridos em casa
- Gols marcados fora de casa
- Gols sofridos fora de casa

Um cálculo foi realizado para determinar quantos pontos cada time acumulou ao longo da temporada. Isso foi feito comparando o resultado de cada partida e atribuindo pontos com base no desempenho:

3 pontos para uma vitória,
1 ponto para um empate,
0 pontos para uma derrota.

Com base nessa pontuação, uma nova coluna chamada is_champion foi adicionada ao dataframe. Essa coluna indica se um time foi campeão (1) ou não (0) ao final da temporada, proporcionando uma visão clara de quais times se destacaram.

Essas etapas de limpeza e preparação dos dados foram fundamentais para garantir que a análise subsequente fosse precisa e informativa, facilitando a exploração do desempenho dos times nas ligas de interesse.

2. Aplicação do Modelo

Configuração

Para a análise, foram construídos dois modelos de Machine Learning: Random Forest e KNN (k-nearest neighbors).

Os dados foram divididos em 80% para treinamento e 20% para teste, utilizando o parâmetro random_state=123. O random_state garante a reprodutibilidade dos resultados ao controlar a aleatoriedade na divisão dos dados. Isso significa que, ao usar o mesmo valor para random_state, a divisão dos dados será a mesma a cada execução, permitindo comparações consistentes entre diferentes execuções do modelo.

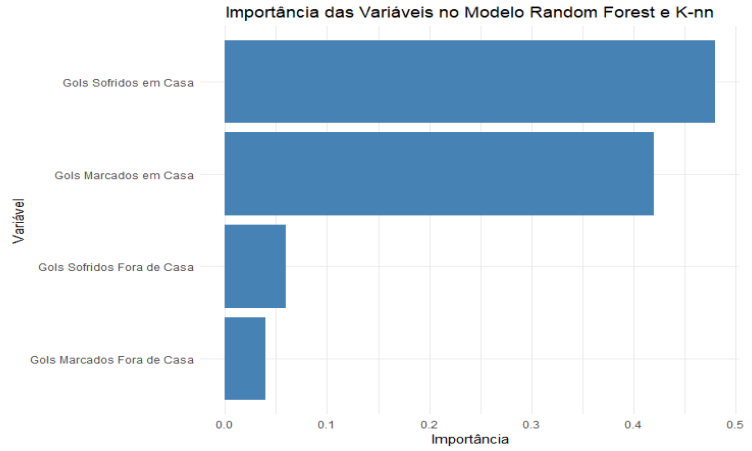
A variável is_champion foi escolhida como alvo de estudo, enquanto os gols marcados e sofridos (tanto em casa quanto fora) foram utilizados como critérios de predição.

Os modelos escolhidos foram Random Fores e K-nn. A escolha do Random Forest foi fundamentada em sua robustez e capacidade de lidar com overfitting, enquanto o K-nn foi selecionado pela sua simplicidade e eficácia em classificações, apesar de ser suscetível a underfitting.

3. Avaliação do Modelo

As variáveis mais importantes para ambos os modelos, listadas da mais para a menos significativa, foram Gols sofridos em casa, Gols marcados em casa, Gols sofridos fora de casa e Gols marcados fora de casa:

O modelo Random Forest foi ajustado (tuned), mas não houve diferença significativa nos resultados, já que a acurácia e a curva de aprendizado permaneceram semelhantes.



- Na matriz de confusão, os resultados foram:

Random Forest: i) 1640 verdadeiros não-campeões
ii) 3 não-campeões classificados como campeões
iii) 2 campeões classificados como não campeões
iv) 22 verdadeiros campeões

Knn: i) 1632 verdadeiros não-campeões
ii) 11 não campeões classificados como campeões
iii) 12 campeões classificados como não campeões
iv) 12 verdadeiros campeões

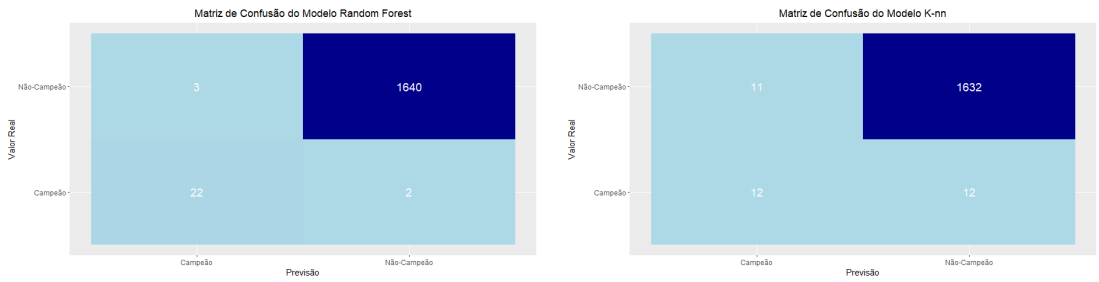


Gráfico 2 - Matriz Confusão para Random Forest (Esq) e K-nn (Dir)

Box-Plot: Comparação de Times Campeões e Não-Campeões

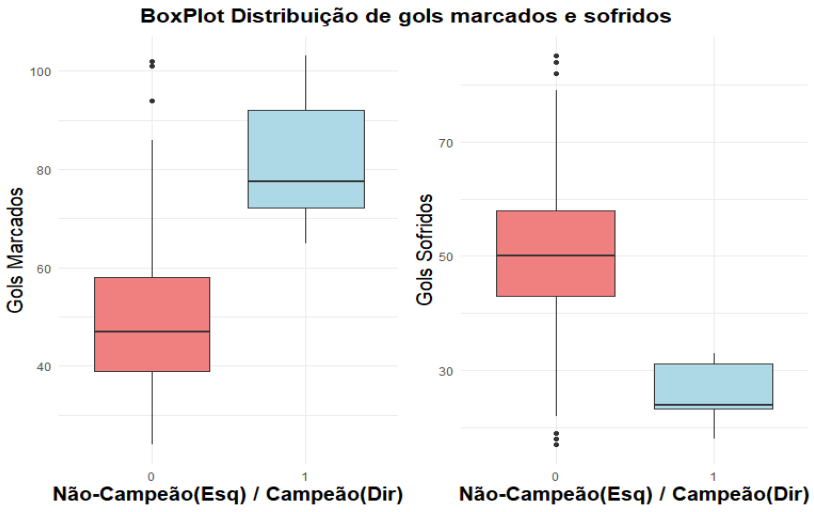
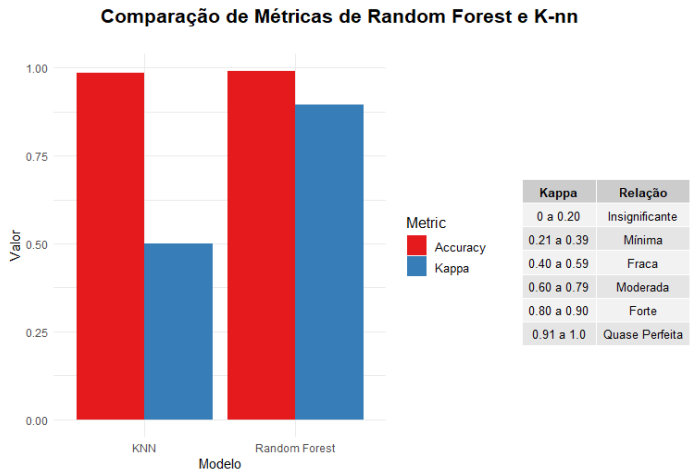


Gráfico 3 - BoxPlot dos Gols Marcados (Esq) e Sofridos (Dir)

Comparação de Acurácia, Kappa e Aprendizado entre Modelos

Random Forest: - Acurácia: 0.99
- Kappa: 0.896
- No gráfico de boundary, que visualiza a separação entre as classes (neste caso, times campeões e não campeões), o modelo conseguiu classificar eficientemente os times. No entanto, alguns times classificados como campeões ficaram na área de não campeões.

KNN:- Acurácia: 0.986
- Kappa: 0.50
- As curvas de aprendizado divergiram no final do processo, onde a training score subiu, enquanto a cross-validation diminuiu. Esse comportamento indica que o modelo está se ajustando demais aos dados de treinamento (overfitting), levando a uma performance pior em dados não vistos.



Kappa	Relação
0 a 0.20	Insignificante
0.21 a 0.39	Mínima
0.40 a 0.59	Fraca
0.60 a 0.79	Moderada
0.80 a 0.90	Forte
0.91 a 1.0	Quase Perfeita

Gráfico 6 - Acurácia e Kappa dos Modelos de Aprendizado Random Forest e K-nn

Conclusão

O modelo de Random Forest apresentou uma acurácia de 0.997 e um coeficiente Kappa de 0.89, indicando um alto nível de concordância nas classificações feitas. Já o modelo K-NN teve uma acurácia de 0.98, mas com um Kappa mais baixo (0.5), sugerindo que as classificações corretas foram mais aleatórias.

A alta acurácia pode ser atribuída ao uso das variáveis Gols Sofridos e Gols marcados, que têm um impacto direto no resultado do jogo e são bons preditores de sucesso. No entanto, há indícios de overfitting no modelo de Random Forest, pois a acurácia é extremamente alta. O modelo pode não generalizar bem para novos dados.

Influência da Distribuição das Classes: A diferença de distribuição entre classes (com muito mais times não campeões do que campeões) pode ter influenciado o desempenho de ambos os modelos. Modelos tendem a ser mais propensos a prever a classe majoritária (não campeões, neste caso) em situações de desbalanceamento, o que pode resultar em uma menor acurácia na identificação da classe minoritária (campeões). Essa disparidade pode levar a uma matriz de confusão com muitos falsos negativos, como observado no KNN.

Essas análises demonstram a importância da escolha adequada de modelos, ajuste fino e consideração da distribuição de classes ao aplicar Machine Learning em problemas de classificação.

Para aprimorar os resultados, seria interessante incluir estatísticas mais detalhadas (como posse de bola, condições do time, etc., além de equilibrar a distribuição entre as classes campeão e não campeão) e verificar a robustez do modelo em relação à mudança dessas variáveis.

Contribuições da Equipe

Marcelo José Soares - 100% (Escolha do database, pré-processamento e preparação dos dados, desenvolvimento do modelo, avaliação das métricas e resultados e interpretação das informações)

Referências

- sklearn Random Forest Documentation
- sklearn K-nn Documentation
- About Kappa
- Pre-processing and Predicting with Random Forest
- Random Forest Algorithmn
- European Soccer Database
- IBM Knn