Random Forest e seus frutos

O que podemos aprender com uma *Random Forest*

Marcelo Reis, Bernardo Lara, Raphael Caetano, Bruno Zandona
marceloestevesreis@gmail.com,be1212lara@gmail.com,caetano.missiaggia@gmail.com,bruzan57@gmail.com
PUC Minas, Belo Horizonte, Minas Gerais, Brasil

ABSTRACT

Neste artigo, intitulado "Random Forest e seus frutos: O que podemos aprender com uma Random Forest," descrevemos um estudo que aplicou o algoritmo Random Forest a uma nova base de dados após a fase anterior, O estudo incluiu etapas de pré-processamento para garantir a qualidade dos dados e do modelo. O objetivo principal foi prever o valor "overall" de jogadores no contexto do FIFA 20, destacando a importância dos atributos de habilidade, como Dribbling e Shooting, na classificação dos jogadores. O modelo obteve uma alta acurácia global de aproximadamente 92.13%, demonstrando sua robustez e confiabilidade. O código do projeto está disponível no Google Colab para facilitar a replicação. Este estudo ilustra como o uso de algoritmos de aprendizado de máquina pode fornecer informações valiosas para orientar decisões no universo do futebol.

KEYWORDS

Random Forest, algoritmos, base de dados, pré-processamento, balanceamento.

ACM Reference Format:
Marcelo Reis, Bernardo Lara, Raphael Caetano, Bruno Zandona. 2023. Random Forest e seus frutos: O que podemos aprender com uma *Random Forest*. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

Na disciplina de Inteligência Artificial, foi proposto um trabalho no qual deveríamos implementar algum algoritmo visto em sala na base de dados escolhida. Na primeira etapa implementamos o algoritmo CART na base '*Drugs A, B, C, X, Y for Decision Trees*', para a segunda etapa escolhemos implementar o Random Forest, numa nova base. Para isso, utilizamos a plataforma Kaggle e encontramos a base '*FIFA 20 complete player dataset*'. Definimos como objetivo a análise dos atributos do jogador e a previsão de seu overall, para conseguimos apontar a importância de compreender quais atributos individuais contribuem mais para o sucesso de um jogador pode orientar decisões de recrutamento, treinamento e estratégias táticas.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DESCRIÇÃO DA BASE DE DADOS

A base de dados "FIFA 20 complete player dataset" fornece informações detalhadas sobre jogadores de futebol presentes no jogo FIFA 20, incluindo vários atributos que descrevem suas habilidades e características. Abaixo está uma descrição dos atributos presentes nesta base de dados e suas distribuições:

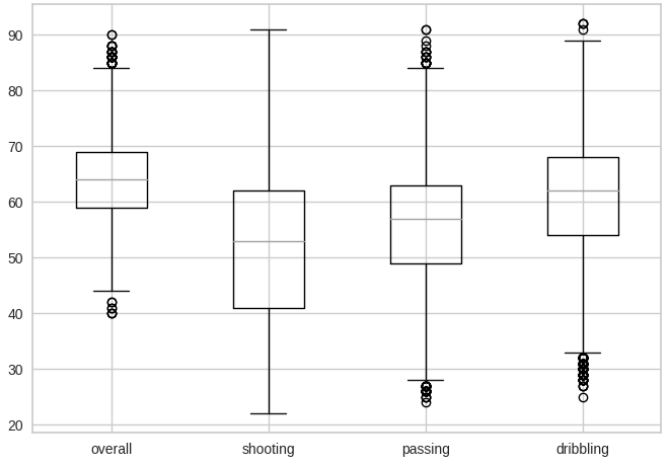

Figure 1: Distribuição de dados na base.

- `age`: A idade do jogador.
- `height_cm`: A altura do jogador em centímetros.
- `weight_kg`: O peso do jogador em quilogramas.
- `overall`: O "overall" é uma métrica que representa a qualidade geral do jogador. É uma avaliação composta de várias habilidades e características do jogador.
- `skill_moves`: O número de habilidades que o jogador possui. É uma indicação de quão habilidoso ele é em dribles e movimentos especiais.
- `pace`: A velocidade do jogador, que inclui sua velocidade de aceleração e velocidade máxima.
- `shooting`: Habilidade de finalização e precisão nos chutes.
- `passing`: Habilidade de passe e precisão nas jogadas de passes.
- `dribbling`: Habilidade de drible e controle de bola.
- `defending`: Habilidade de defesa, incluindo marcação e desarme.
- `physic`: A força física do jogador, que afeta sua resistência, força no choque e capacidade de resistir a desarmes.

3 ETAPAS DE PRÉ-PROCESSAMENTO

Um componente essencial de qualquer análise de dados é o pré-processamento, que envolve a preparação dos dados para análises posteriores. No contexto do nosso estudo com o conjunto de dados "FIFA 20 complete player dataset," foram realizadas várias etapas de pré-processamento para garantir a qualidade e a integridade dos dados.

3.1 Remoção de Registros com Atributos em Branco

Uma das primeiras etapas de pré-processamento envolveu a identificação e remoção de registros que continham valores em branco ou nulos em qualquer um dos atributos. A presença de valores ausentes pode afetar a qualidade das análises e modelagem de dados, tornando importante realizar essa limpeza. A eliminação desses registros garante que os dados restantes sejam mais confiáveis e consistentes.

3.2 Eliminação de "Overalls" com Apenas um Registro

Observamos que, durante a análise dos dados, alguns valores do atributo "overall" continham apenas um registro. Esses casos não forneciam informações úteis para a análise estatística, e sua inclusão poderia distorcer as métricas. Portanto, optamos por eliminar esses "overalls" com apenas um registro.

Essas etapas de pré-processamento garantiram que o conjunto de dados utilizado em nossa análise estivesse mais limpo e coeso. Com dados mais completos e confiáveis, estávamos prontos para avançar para as etapas subsequentes, como a modelagem de aprendizado de máquina e interpretação dos resultados.

4 DESCRIÇÃO DOS MÉTODOS UTILIZADOS

Neste projeto, utilizamos o algoritmo Random Forest para a tarefa de previsão do valor "overall" de jogadores no contexto do FIFA. O Random Forest é um algoritmo de aprendizado de máquina que faz parte da família de métodos de ensemble, combinando múltiplas árvores de decisão para melhorar o desempenho de previsões.

Este projeto foi desenvolvido na plataforma Google Colab, que oferece um ambiente de desenvolvimento integrado para Python, facilitando a execução e colaboração em projetos de aprendizado de máquina.

Para a configuração do modelo Random Forest, definimos o hiperparâmetro ‘maxfeatures‘ como ’auto’. Esse valor significa que o modelo seleciona automaticamente o número de características a serem consideradas em cada divisão de árvore, que é igual à raiz quadrada do número total de características disponíveis. Essa configuração ajuda a diversificar as árvores no ensemble, tornando o modelo mais robusto e eficaz na previsão.

5 RESULTADOS

Os resultados da análise do modelo de classificação são apresentados a seguir, com base nas métricas de precisão, revocação, F1-Score e matriz de confusão para cada classe, bem como na importância dos recursos e na acurácia geral.

5.1 Métricas de Classificação

As métricas de classificação são essenciais para avaliar o desempenho do modelo em cada classe individualmente. Os valores de precisão, revocação e F1-Score são calculados para cada classe, fornecendo informações detalhadas sobre a qualidade da classificação:

Classe	Precisão	Revocação	F1-Score	Suporte
...	...	...	...	...
60	1.00	1.00	1.00	221
61	0.99	0.99	0.99	196
62	1.00	1.00	1.00	221
63	1.00	1.00	1.00	242
64	1.00	1.00	1.00	266
65	0.98	0.99	0.99	242
66	0.95	0.99	0.97	220
67	0.97	0.95	0.96	219
68	0.97	0.97	0.97	208
69	0.97	0.94	0.95	155
70	0.90	0.97	0.93	143
...	...	...	...	...

5.2 Importância dos Recursos

A importância dos recursos revela a contribuição de cada atributo no processo de classificação. Valores mais altos indicam que o atributo é mais importante para a tarefa de classificação. Os valores de importância de recursos são apresentados abaixo:

[0.03306797, 0.03404989, 0.03505775, 0.58557399, 0.00763203, 0.04274659, 0.05246457, 0.04747328, 0.04942533, 0.06700021, 0.0455084]

Logo, é possível notar que os atributos mais importantes para o modelo são:
Dribbling e Shooting. Esses dois atributos desempenham um papel significativo na classificação de jogadores no conjunto de dados "FIFA 20 complete player dataset."

1. Dribbling (0.0670): A habilidade de driblar e controlar a bola é um dos principais determinantes na classificação dos jogadores. Quanto melhor o jogador for no drible, mais provável é que ele seja classificado em uma categoria específica de jogadores. Isso sugere que a destreza e agilidade no controle de bola têm um impacto substancial na avaliação dos jogadores.

2. Shooting (0.0525): A precisão e a capacidade de finalização dos chutes também têm uma influência considerável na classificação. Jogadores com excelentes habilidades de finalização e precisão têm maior probabilidade de serem classificados em categorias específicas, indicando a importância do desempenho nas jogadas de ataque.

A análise da importância dos recursos destaca a relevância desses dois atributos-chave no processo de classificação. No entanto, é importante lembrar que essas conclusões são específicas para o modelo e o conjunto de dados em questão. Dependendo do contexto e das necessidades da análise, outras características podem ser mais ou menos importantes.

5.3 Acurácia Geral

A acurácia global do modelo é de aproximadamente 92.13%, o que indica que o modelo classifica corretamente 92.13% dos exemplos no conjunto de dados.

É digno de destaque o desempenho sólido e genuíno demonstrado pela base de dados, mesmo quando enfrentamos o desafio de uma categoria 'overall' com um número restrito de instâncias. Um exemplo notável desse desempenho é evidenciado pela classe 'overall' 49, a qual, mesmo contando com apenas 33 instâncias, conquistou um impressionante F1-Score de 0,85.

Esses resultados ressaltam a robustez do modelo e a confiabilidade das previsões. Eles indicam que, mesmo diante de um conjunto de dados com limitações de amostragem, as classificações obtidas são substanciais e confiáveis, fornecendo insights valiosos e precisos sobre as categorias 'overall' dos jogadores.

É importante notar que a interpretação dos resultados pode variar dependendo do contexto do problema e dos objetivos da análise. Portanto, a avaliação detalhada das métricas de classificação é fundamental para compreender o desempenho do modelo em cada classe individualmente.

6 CÓDIGO DESENVOLVIDO

Código no Google Colab: https://colab.research.google.com/drive/19kPFjYCr2Lc2U5oXVVQoKpF2fa7VOJvW?usp=sharing

7 REFERÊNCIAS

Kaggle, 2021. FIFA 20 Complete Player Dataset. Disponível em: https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?select=players_15.csv

Kaggle, 2021. Drugs A, B, C, X, Y for Decision Trees. https://www.kaggle.com/datasets/pablomgomez21/drugs-a-b-c-x-y-for-decision-trees