Este projeto é parte do meu portfólio de exercícios de aplicações práticas de ML e IA visando praticar e internalizar conceitos, ferramentas e técnicas vistas em cursos e conteúdo acadêmico

Visando experimentar os principais algoritmos de Machine Learning optei como conjunto de dados a base de jogadores do FIFA, escolha motivada por minha familiaridade com o jogo e paixão por futebol que facilitam a compreensão dos dados em questão, assimilação de conceitos e resolução de problemas

Fiz o carregamento do Dataset de jogadores do FC 24 por meio de um Web Scraper escrito em Java utilizando as bibliotecas HtmlUnit e Commons CSV da Apache para extração dos dados do site SoFIFA e escrita do CSV respectivamente

O projeto responsável por isso pode ser visitado em https://github.com/RenanFR/sofifa-web-scraping

Como primeiro desafio escolhi a classificação da posição do jogador com base em seus atributos e parti para a Engenharia de Features seguindo as etapas abaixo:

* Redução de dimensionalidade agregando as avaliações específicas de cada habilidade em colunas com a média para cada aspecto do jogo (Ataque, habilidade técnica, movimentação, força, mentalidade, defesa e goleiro). Isso, pois, existem 34 atributos distintos para cada valência (Como chute, cruzamento, drible, aceleração, visão, salto, etc.) e a tese inicial é de que as médias por posição podem ser o suficiente para classificar a posição de um jogador
* Remoção de outros atributos sem correlação com a posição como ID
* Como a idade do jogador em atividade pode ser um previsor útil para descobrir a posição (Visto que atletas mais longevos geralmente estarão nas posições de goleiro e zagueiro), transformei a data de nascimento na idade numérica criando uma coluna que subtrai o ano presente nela do atual
* Conversão de atributos financeiros do jogador como salário, valor de mercado e clausula de rescisão de String (Como foram originalmente capturados) para numérico
* Foi necessário realizar a codificação (Encoding) da variável de pé dominante (Destro ou canhoto) usando a estratégia de One-hot

Em seguida executei o treinamento do modelo de árvore de decisão utilizando a biblioteca sklearn

A acurácia obtida foi de apenas 0.51 e a aderência aos dados de treinamento é de 0.92, mostrando a necessidade de refinamento e otimização tanto do pré-processamento quanto dos parâmetros do algoritmo, o que farei em exercícios posteriores

Problemas detectados

* O SoFIFA armazena várias versões de sua base de dados de jogadores com base nas atualizações periódicas da EA, na data da extração dos dados foi utilizada como base a versão de 26/09/2023, ao fazer porém, uma comparação das pontuações de ataque do jogador Bukayo Saka usando CSV e site retroagindo a versão para a que foi utilizada há algumas discrepâncias como 82 para cruzamento no CSV, enquanto no site está como 83. Outros atributos e registros foram impactados como por exemplo, a clausula de liberação do mesmo jogador consta como €158M no CSV e €195.5M no site
  + Não se sabe se se trata de um bug na navegação por histórico de versão no site ou se foi algo do lado do código do Web Scraper