Projeto 3 - Treinamento de modelo de classificação de asteroides potencialmente perigosos (PHA)

Este projeto consiste em utilizar dados de asteroides coletados da NASA via Small-Body DataBase API para treinar um modelo de classificação capaz de identificar asteroides potencialmente perigosos.

Para realizar esta tarefa, utilizamos o modelo de classificação SVC (Support Vector Classification), o qual foi selecionado por possuir uma boa adequação a espaços de alta dimensão (features numerosas).

Obtenção de dados

Nesta etapa utilizamos duas funções principais para realizar requisições através da API SBDB disponibilizada pela NASA.

request_asteroids_des: responsável por obter uma lista com as designações atribuídas a asteroides detectados nos anos especificados.

request_asteroids_data: responsável por receber a lista com as designações dos asteroides e obter os dados referentes a eles.

Dados alvo de nossas requisições:

- des: designação primária
- pha: objeto potencialmente perigoso
- neo: objeto próximo da Terra
- e: excentricidade
- a: semi-eixo maior (au)
- q: distância do periélio (au)
- i: inclinação (deg)
- om: longitude do nó ascendente (deg)
- w: argumento do periélio (deg)
- ma: anomalia média (deg)
- tp: tempo de passagem do periélio (TDB) formatado como dia juliano
- per: período orbital (d)
- n: movimento médio (deg/d)
- ad: distância do afélio (au)
- moid: distância mínima de intersecção orbital em relação à Terra (au)
- moid_jp: distância mínima de intersecção orbital em relação a Júpiter (au)
- t_jup: Júpiter Tisserand invariante
- rms: valor quadrático médio do ajuste da órbita
- H: magnitude absoluta

Estes dados foram obtidos acessando as chaves “object”, “orbit” e “phys_par” do json retornado pela API.

Exploração e tratamento de dados

Essa etapa foi a responsável por verificar a existência de valores “NA” e “null” em nossos dados, bem como converter a tipagem das variáveis para tipos mais convenientes para o posterior treinamento do modelo.

Através da plotagem da matriz de correlação, foram selecionadas as variáveis com menor dependência com as outras, reduzindo a dimensionalidade de nossos dados e aumentando o desempenho computacional sem perdermos informação relevante para o modelo. Após este processo, houve uma nova seleção de variáveis levando em consideração seus significados físicos para este problema, resultando na eliminação de 3.

Também foram exploradas e analisadas as características (dimensões, distribuição e etc) dos nossos dados. Isto nos permitiu detectar um desbalanceamento no número de asteroides classificados como pha e não pha, levando a um posterior tratamento com uma divisão adequada para alimentar o modelo com dados de treino balanceado, evitando quaisquer vieses.

Modelo

Modelo: SVC (Support Vector Classification)
Biblioteca utilizada: scikit-learn
Métricas: acurácia, precisão e recall

O treinamento do modelo consistiu nas seguintes etapas:

1ª - Normalização dos dados

2ª - Balanceamento e separação em dados de treino e de teste
- dados de treino: 1900
- dados de teste: 400

3ª - Treinamento

4ª - Teste e avaliação das métricas

Resultados:

Acurácia: 93.75%
Precisão: 89.24%
Recall: 99.5%