Supervised Learning: Near Earth Objects Classification

Artificial Intelligence

Turma 10 - Grupo A2 91

Ntsay Zacarias up202008863

1 de Maio de 2024

Contents

1 Introdução

No âmbito dos estudos espaciais, a classificação de objetos celestes, como os asteróides, desempenha um papel crucial na compreensão das suas características e potenciais riscos. Este projeto tem como objetivo utilizar técnicas de aprendizagem automática *(Machine Learning)* para prever se os asteróides são perigosos com base nas suas características físicas e orbitais. O objetivo é desenvolver um modelo preditivo que melhore a precisão da classificação de asteroides utilizando vários algoritmos, como *Decision Trees,Neural Networks, K-NN e SVM.* Nesta iteração para simplicidade utilizei somente o Random Forest, uma Decision Tree, conhecido pela sua eficácia no tratamento de conjuntos de dados complexos e de alta dimensão.

2 Implementação/Processo

A implementação envolveu várias fases-chave:

2.1 Pré-processamento de Dados

A fase de pré-processamento dos dados é crucial para garantir a qualidade e a precisão das análises subsequentes. Este processo envolveu várias etapas estratégicas para refinar o conjunto de dados, detalhadas a seguir:

- **Limpeza de Atributos Redundantes ou Desnecessários:** Para simplificar o modelo e melhorar o desempenho computacional, atributos duplicados ou irrelevantes foram removidos. Isso incluiu a eliminação de medidas em diferentes unidades para o mesmo atributo (como diâmetro em metros, milhas e pés) mantendo apenas em quilômetros *(Km)* e unidades astronómicas *(AU)* que são padrão para análises astronômicas.

- **Unificação de Nomes e Eliminação de Variáveis Constantes:** O atributo '*Name*', que era redundante com '*Neo Reference ID*', foi removido, assim como '*Orbiting Body*' e '*Equinox*', que apresentavam valores constantes e, portanto, não contribuíam para a análise preditiva.

- **Tratamento de Valores Ausentes:** Após a limpeza inicial, a análise dos dados revelou a presença de valores ausentes em algumas colunas. Esses foram imputados utilizando métodos estatísticos apropriados para preservar a integridade do conjunto de dados sem introduzir viés significativo.

2.2 Seleção de Características e Treino do Modelo:

Um classificador Random Forest foi treinado, e as características foram selecionadas com base na sua importância.

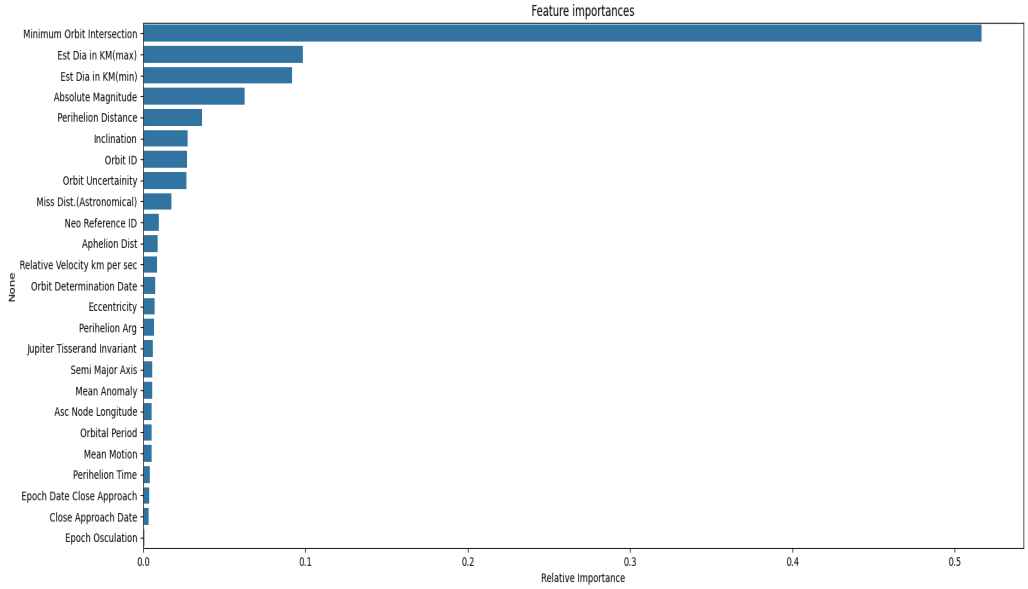

Figure 1: *Feature importances* em um modelo RF simples

2.3 Avaliação do Modelo

O modelo foi meticulosamente avaliado utilizando a técnica de ***cross-validation***, essencial para assegurar que o modelo é generalizável e robusto contra diferentes subconjuntos de dados. Esta abordagem divide o conjunto de dados em várias partes menores, permitindo que o modelo seja treinado várias vezes em diferentes segmentos dos dados e validado em outros, proporcionando uma visão abrangente do seu desempenho esperado em condições variadas.

Para otimizar ainda mais a performance do modelo, foi realizada uma ***hyperparameter tuning*** utilizando o método ***GridSearch***. Esta estratégia sistemática testa combinações de parâmetros pré-definidas, buscando aquelas que maximizam a eficácia do modelo. A *GridSearch* também ajuda a evitar o *overfitting*, assegurando que o modelo final seja tanto preciso quanto capaz de generalizar bem para novos dados.

O processo também incluiu uma extensa visualização de dados para analisar a importância das características e as métricas de desempenho do modelo, tais como precisão, exatidão, recuperação e a pontuação F1.

3 Resultados

A fase final de avaliação do modelo focou em validar a eficácia e a generalização do classificador Random Forest através de métricas detalhadas obtidas por validação cruzada e teste no conjunto de dados de teste. A ***cross-validation*** foi consistentemente alta, com scores variando de 99,39% a 99,85%, resultando em uma média impressionante de aproximadamente 99,60%. Este alto nível de desempenho também se refletiu nos resultados do conjunto de teste:

- **Precisão (Accuracy):** 99,79%

- **Precisão (Precision):** 99,12%
- **Recuperação (Recall):** 99,56%
- **Pontuação F1 (F1 Score):** 99,34%

3.1 Visualizações Detalhadas

As visualizações de dados desempenharam um papel crucial na interpretação dos resultados e na validação da robustez do modelo:

- **Matriz de Confusão** A matriz de confusão obtida revela um excelente desempenho do modelo. Dos casos testados, o modelo identificou corretamente 1178 verdadeiros negativos e 226 verdadeiros positivos. O número de falsos positivos foi extremamente baixo, com apenas 2 casos, e apenas 1 caso foi identificado incorretamente como falso negativo. Estes resultados mostram uma alta precisão e capacidade de recuperação do modelo.

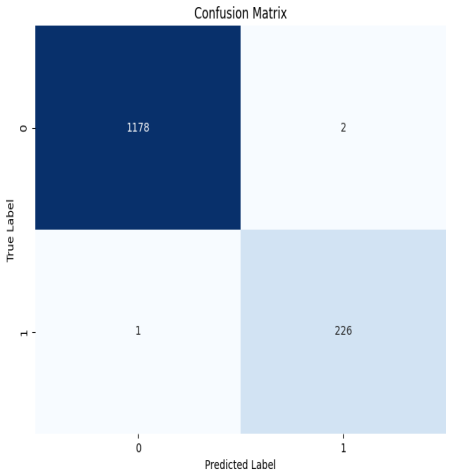

Figure 2: Matriz de Confusão

- **Curvas de Aprendizado:** Observando as curvas, notamos que a **Training Score** manteve-se elevada e estável, o que pode indicar que o modelo aprende eficientemente desde o início ou que pode existir algum nível de *overfitting*. Por outro lado, a **Cross-validation score** melhorou progressivamente, sugerindo uma boa capacidade de generalização à medida que mais dados foram incluídos.

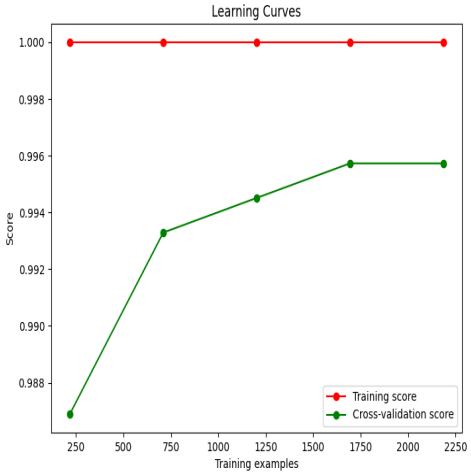

Figure 3: Curvas de Aprendizado

- **Curva ROC:** A Curva de Característica de Operação do Receptor (ROC) apresentada demonstra um desempenho exemplar do modelo, com uma Área Sob a Curva (AUC) de 1,00. Este resultado indica que o modelo tem capacidade perfeita de distinguir entre as classes positivas e negativas sem incorrer em erros

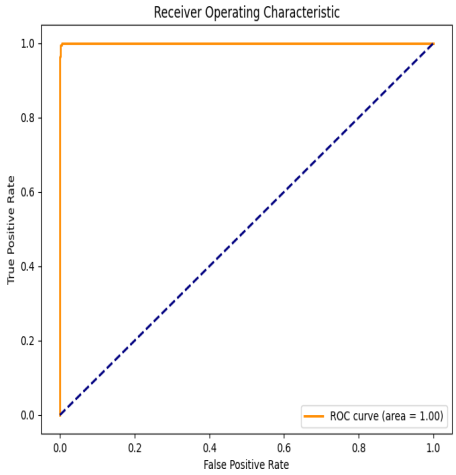

Figure 4: Receiver Operating Characteristic

4 Dataset

O dados foram fornecidos pela NASA a partir um serviço web RESTful *NeoWs (Near Earth Object Web Service)*.

Segue uma breve descrição de cada atributo dos dados fornecidos:

- **Neo Reference ID**: Identificador único para cada asteroide na base de dados, usado como referência principal.

- **Name**: Nome oficial ou designação do asteroide.
- **Absolute Magnitude**: Medida do brilho do asteroide, que é uma indicação do seu tamanho e refletividade.
- **Est Dia in KM(min)** e **Est Dia in KM(max)**: Diâmetro estimado do asteroide, apresentado como um intervalo mínimo e máximo em quilômetros.
- **Close Approach Date**: Data em que o asteroide passa mais próximo da Terra.
- **Epoch Date Close Approach**: Representação em formato de época (UNIX timestamp) da data em que o asteroide esteve mais próximo da Terra.
- **Relative Velocity km per sec**: Velocidade do asteroide em relação à Terra, medida em quilômetros por segundo.
- **Miss Dist.(Astronomical)**: Distância por que o asteroide passa da Terra, medida em unidades astronômicas (AU).
- **Orbit ID**: Identificador da órbita do asteroide, usado para rastrear diferentes observações da mesma trajetória.
- **Orbit Determination Date**: Data em que a órbita do asteroide foi determinada com precisão.
- **Orbit Uncertainity**: Medida da incerteza da órbita do asteroide, indicando quão bem sua trajetória é conhecida.
- **Minimum Orbit Intersection**: Distância mínima entre a órbita do asteroide e a órbita da Terra, um indicador crítico do potencial de colisão.
- **Jupiter Tisserand Invariant**: Um valor numérico que ajuda a classificar a órbita do asteroide em relação à influência de Júpiter.
- **Epoch Osculation**: Data de referência para os elementos orbitais dados.
- **Eccentricity**: Medida da excentricidade da órbita do asteroide, onde 0 significa uma órbita perfeitamente circular.
- **Semi Major Axis**: O maior eixo da órbita elíptica do asteroide.
- **Inclination**: Ângulo de inclinação da órbita do asteroide em relação ao plano eclíptico da Terra.
- **Asc Node Longitude**: Longitude do nó ascendente da órbita do asteroide.
- **Orbital Period**: Tempo que o asteroide leva para completar uma órbita ao redor do Sol.
- **Perihelion Distance**: Distância mais próxima do Sol no ponto da órbita do asteroide.
- **Perihelion Arg**: Argumento do periélio, que é o ângulo da órbita do asteroide no ponto de maior aproximação do Sol.

- **Aphelion Dist**: Distância mais afastada do Sol no ponto da órbita do asteroide.
- **Perihelion Time**: Momento em que o asteroide está no periélio, o ponto mais próximo do Sol.
- **Mean Anomaly**: Medida do ponto em que o asteroide se encontra na sua órbita em relação ao periélio.
- **Mean Motion**: Velocidade média do asteroide ao longo da sua órbita.
- **Equinox**: Sistema de coordenadas usado para representar os elementos orbitais do asteroide.
- **Hazardous**: Indicador de se o asteroide é considerado potencialmente perigoso para a Terra.

References

[1] Scikit-learn developers. *Model evaluation: quantifying the quality of predictions*. Disponível em: `https://scikit-learn.org/stable/modules/model_evaluation.html`.

[2] Medium - Analytics Vidhya. *Evaluating a Random Forest Model*. Disponível em: `https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56`.

[3] Analytics Vidhya. *Metrics to Evaluate your Classification Model to take the Right Decisions*. Disponível em: `https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/`. Acesso em: [01-05-2024].

[4] Kaggle - Will Koehrsen. *Intro to Model Tuning: Grid and Random Search*. Disponível em: `https://www.kaggle.com/code/willkoehrsen/intro-to-model-tuning-grid-and-random-search`.

[5] Bansal, L. *NASA Asteroids Classification*. Recuperado de Kaggle: `https://www.kaggle.com/datasets/lovishbansal123/nasa-asteroids-classification`.