Universidad Técnica Federico Santa María

Departamento de Física / Optativo Avanzado

Clasificación de Asteroides Cercanos a la Tierra con Random Forest y XGBoost

Cristián Núñez
Isidora Morales
Miyaray Arenas

Curso de Machine Learning a cargo de

Profesor(a): Pia Amigo

27 de noviembre de 2025

1. Introducción

El *Machine Learning* es clave en Astronomía y Astrofísica para clasificar datos de catálogos y surveys. Este proyecto clasifica *Near-Earth Objects* (NEOs) y *Potentially Hazardous Asteroid* (PHA), asteroides que pueden impactar la Tierra. Usamos datos del *JPL Small-Body Database* (NASA) con parámetros físicos y orbitales de asteroides. El problema es una clasificación binaria: predecir si un objeto es NEO y PHA. Comparamos dos algoritmos de árboles: *Random Forest* y *XGBoost*.

2. Datos y metodología

Dataset original con 10^6 objetos y ~ 45 columnas, con variable objetivo `PHA`. Se seleccionaron características orbitales y físicas, imputando valores faltantes y usando partición entrenamiento–prueba estratificada para mitigar el desbalance.

3. Modelos utilizados

Random Forest

El *Random Forest* consiste en un ensamble de árboles de decisión entrenados sobre subconjuntos bootstrap del dataset y subconjuntos aleatorios de variables. Sus ventajas incluyen: modelar relaciones no lineales, robustez frente a outliers moderados y la posibilidad de obtener importancias de variables de forma directa.

Se ajustaron hiperparámetros como el número de árboles, profundidad máxima y tamaño mínimo de hojas, y se utilizó `class_weight` para penalizar la clase minoritaria (NEO). El entrenamiento se realizó usando validación cruzada estratificada sobre un subconjunto del conjunto de entrenamiento para reducir el costo computacional.

XGBoost

XGBoost es un algoritmo de *gradient boosting* sobre árboles que construye los modelos de manera secuencial, de modo que cada nuevo árbol corrige los errores de los anteriores. Permite controlar de forma precisa la regularización y el desbalance mediante hiperparámetros como `learning_rate`, `max_depth`, `subsample`, `colsample_bytree` y `scale_pos_weight`.

Para el ajuste de XGBoost se utilizó una búsqueda aleatoria de hiperparámetros sobre un subconjunto estratificado del dataset (del orden de 10^5 objetos), equilibrando rendimiento y costo de cómputo.

4. Resultados

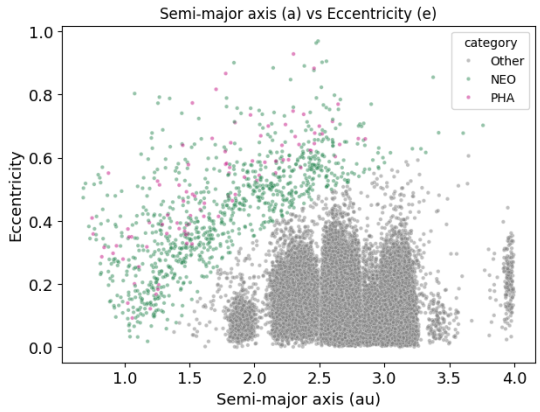
Ambos modelos superan ampliamente a un clasificador trivial que siempre predice "no NEO". El Random Forest, con pesos balanceados, logra un ROC–AUC elevado y un buen compromiso entre *recall* y *precision* de la clase NEO. XGBoost, tras el ajuste de hiperparámetros, presenta en general un ROC–AUC algo mayor y una métrica F1 superior para la clase positiva, sin embargo este último modelo no clasificó correctamente a 2 objetos PHA. Así, es preferible optar por un modelo de Random Forest en comparación a XGBoost debido a que la alta complejidad de est último sobreajustó los datos y mal clasificó a dos asteroides.

El análisis de importancia de variables en ambos modelos resalta de forma consistente a `NEO` y q como las características más relevantes, seguidas del semieje mayor a, la excentricidad e y la inclinación i. Las variables puramente fotométricas (H) y físicas (diámetro, albedo) aportan información secundaria. Estos resultados son coherentes con la definición dinámica de NEO: la

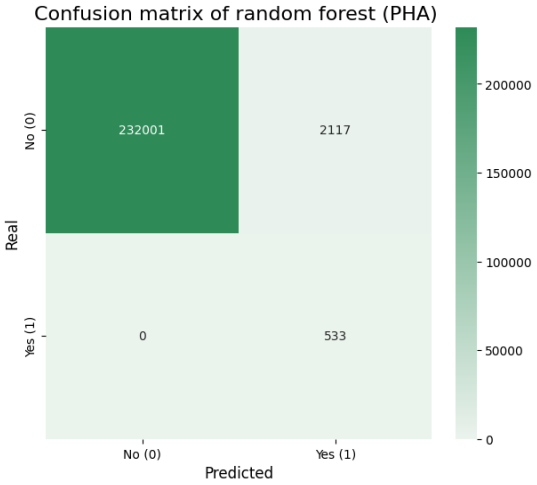
geometría orbital domina la clasificación. Finalmente, se obtiene el supuesto esperado: los PHA corresponden a un subconjunto de los NEO estando relativamente cerca y con excentricidades no circulares.

5. Figura de análisis exploratorio

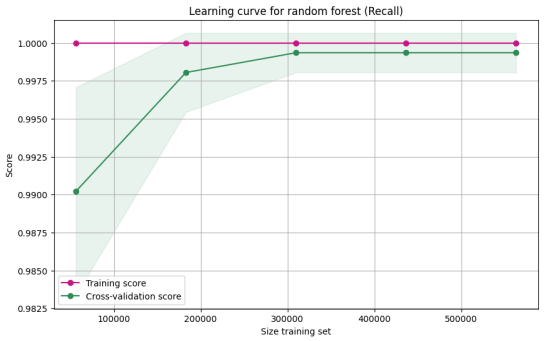
En la Figura 1 se presenta una síntesis de los resultados tras aplicas los modelos de Machine Learning en este proyecto.



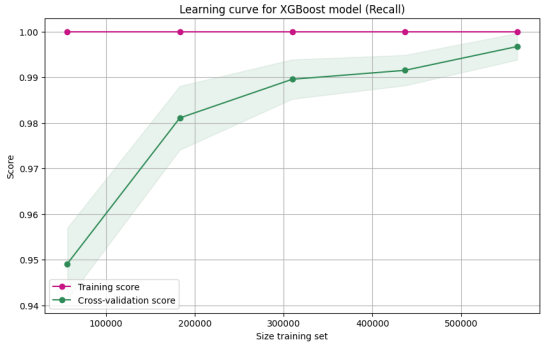
(a) Visualización de NEO y PHA según la relación entre semi eje mayor y excentricidad



(b) Histograma de a.



(c) Curva de aprendizaje para Random Forest



(d) Curva de aprendizaje para XGBoost

Figura 1: Curvas de Aprendizaje de los modelos empleados

6. Conclusiones

Modelos de Machine Learning basados en árboles, como Random Forest y XGBoost, clasifican eficazmente asteroides cercanos a la Tierra usando parámetros orbitales y físicos del JPL Small-Body Database. XGBoost supera a Random Forest en desempeño, pero presenta sobreajuste al clasificar erróneamente a 2 asteroides. Por lo tanto, se recomienda Random Forest por su robustez y simplicidad.

El trabajo futuro incluye mejorar el manejo del desbalance de clases, incorporar validación temporal y explorar algoritmos como LightGBM o CatBoost. Integrar estos modelos en pipelines de surveys de próxima generación priorizará objetos para seguimiento y caracterización, complementando métodos dinámicos tradicionales.

7. Apéndice

Repositorios de GitHub

C. Núñez Haz click aquí ! o bien copia la siguiente URL en tu navegador de preferencia:
`https://github.com/ccanunez/Machine-Learning-2025-2.git`

I. Morales Haz click aquí ! o bien copia la siguiente URL en tu navegador de preferencia:
`https://github.com/Isi-mrls/Introduction-to-the-machine-learning/tree/9190ea68f88c3dfa4e19Final_project`

M. Arenas Haz click aquí ! o bien copia la siguiente URL en tu navegador de preferencia:
`https://github.com/miyarayarenas-debug/AST332/tree/main/Proyecto%20Final`