UTFSM - AST332

Cristián Núñez

Isidora Morales

Miyaray Arenas

Machine Learning Final Project

Potential risk of asteroids trajectories

2025

Final Semester

a brief summary

Cite this paper

Hossain, M.S., Zabed, M.A. (2023). Machine Learning Approaches for Classification and Diameter Prediction of Asteroids. In: Ahmad, M., Uddin, M.S., Jang, Y.M. (eds) Proceedings of International Conference on Information and Communication Technology for Development. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore.
DOI

MIR SAKHAWAT HOSSAIN · UPDATED 5 DAYS AGO

147 | Code | Download

Asteroid Dataset

NASA JPL Asteroid Dataset

Data Card | Code (51) | Discussion (2) | Suggestions (0)

About Dataset

Story Behind This Dataset

I am an Astronomy and Astrophysics Researcher. As a Mathematics background I am a data science, machine learning, and deep learning enthusiast. Nowadays Machine Learning is solving so many problems in Astronomy and Astrophysics fields. Asteroid is nice topic for Machine Learning projects like classification and regression problems.

Usability
10.00

License
Database: Open Database, Cont...

Update frequency
Weekly

Tags

Index

1

Context

Why?

- Identify Near-Earth Objects (NEO) with high impact risk
- Great for future studies in asteroids orbits
- Improvement in simulations

2

Goal

So...?

- Create a simple ML method
- Identify potentially risk asteroids

3

Methodology

How?

- Dataset + Features + Outliers handling
- **Model:** Random Forest and XGBoost
- **Limitations:** High imbalance over features (few NEOs)

4

Results

Got it

- **Random Forest**: Robust model and great handling non related features
- **XGBoost:** Gradient boosting modeling complex interactions

5

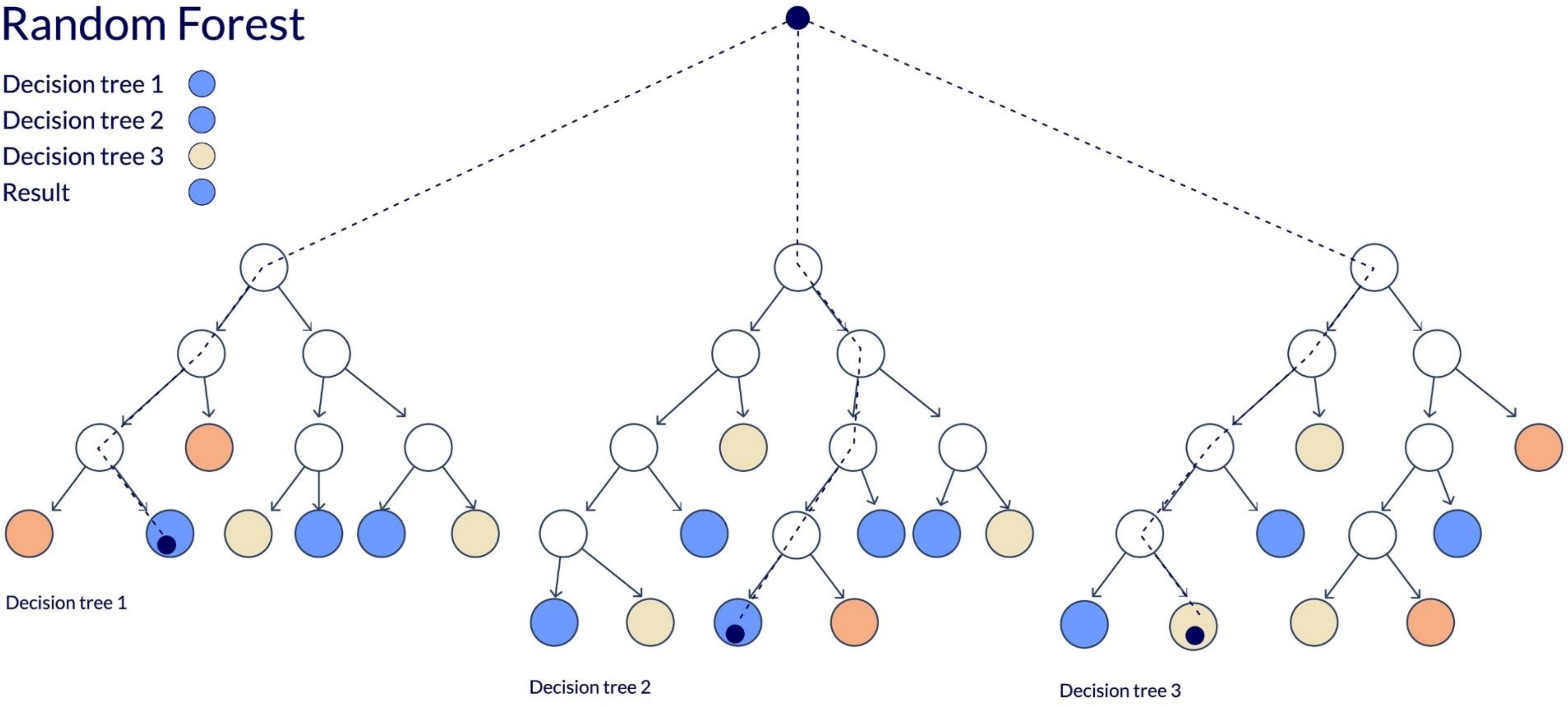
Conclusion

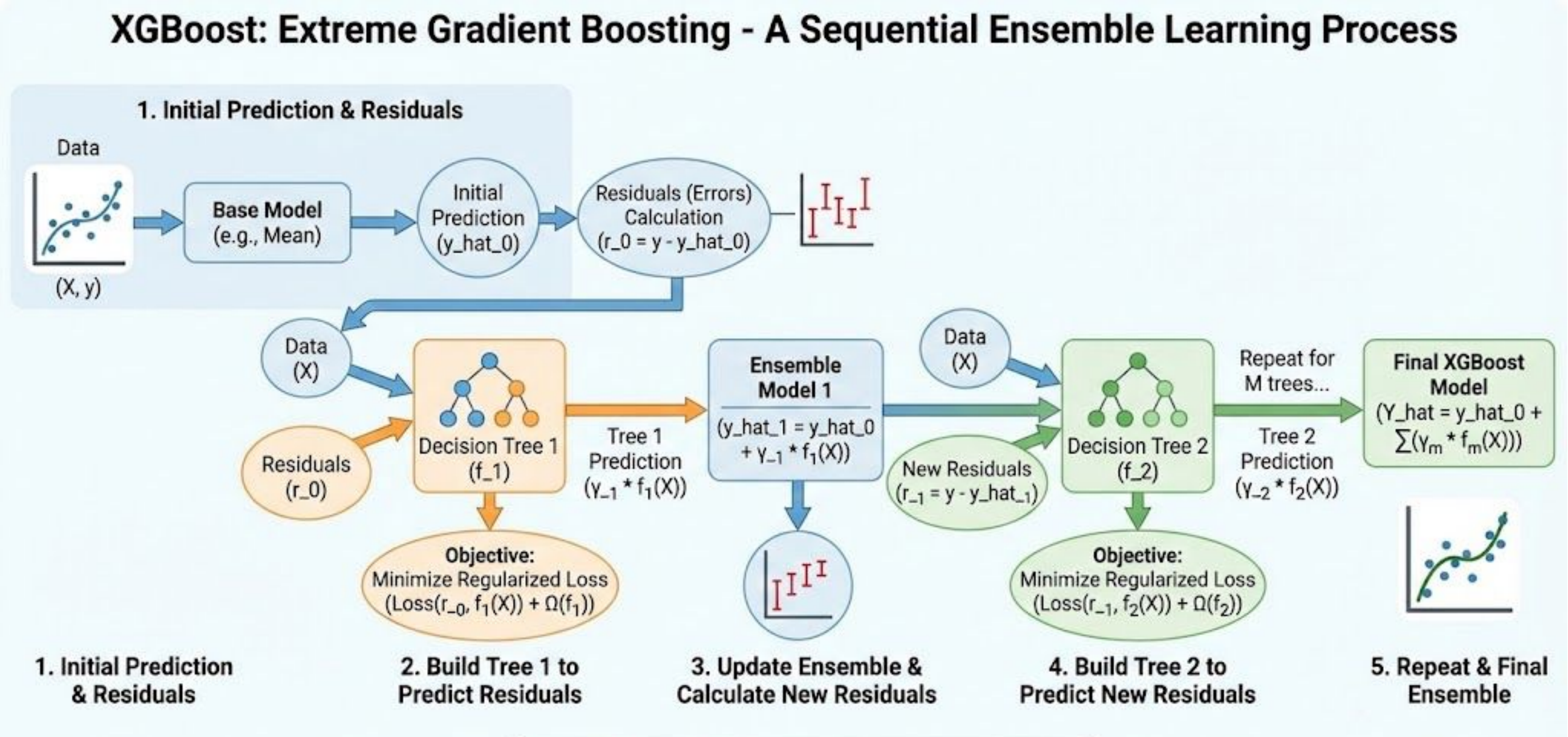
Let's cross over !

- Trade-off between the two models
- Random Forest achieved perfect recall
- XGBoost tends to overtfit

Random Forest

Decision tree 1
Decision tree 2
Decision tree 3
Result





Loss Function (L): Measures error
Regularization (Ω): Controls complexity (prevents overfitting)
Learning Rate (γ): Scales each tree's contribution

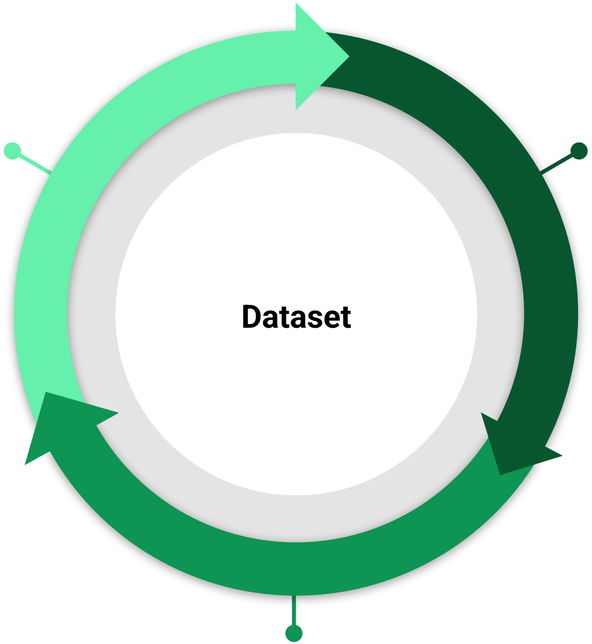
from the **JPL Small-Body Database** (NASA)

- Over 1 millions objects
- **Density filter** was used to study the correlations between features
- **Variable objects** written as NEO and Not NEO
- **NEO** are Near-Earth Objects and **PHA** corresponds to Potentially Hazardous Asteroid

DATA PREPARATION

ORBITAL PARAMETERS

eccentricity, semi-major axis, perihelio, inclinations, etc.



FEATURES AND FLAGS

NEO and PHA

PHYSICAL PROPERTIES

absolute magnitude, diameters, albedo, etc.

nevertheless there is a huge imbalance in the data which we have to resolve to make good estimations.

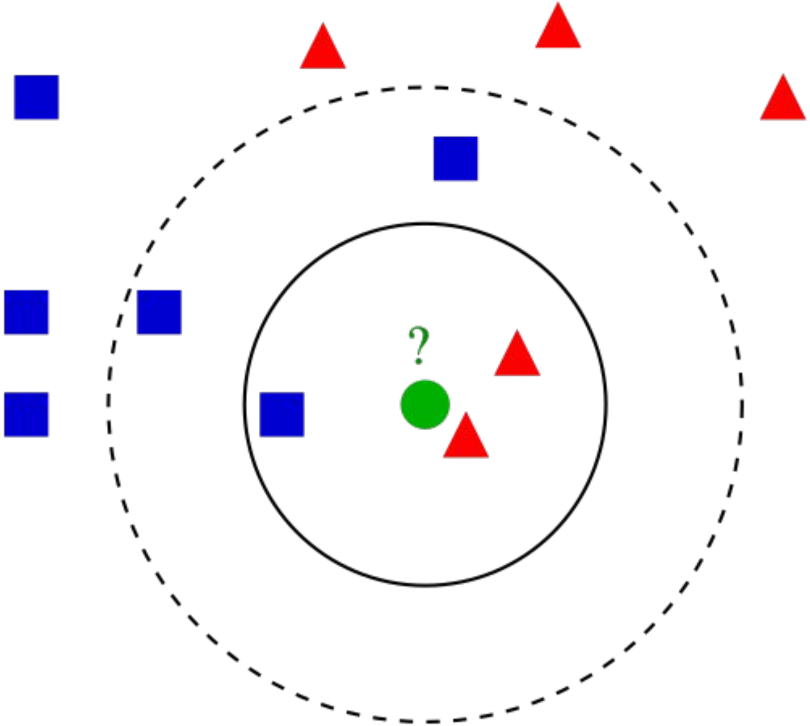
Relationship between NEO and PHA

1 **NEO**

may be also PHA but these type are usually non important asteroids

2 **PHA**

could also be NEOs which leads to certain high risk impact objects



Visual scheme

parametric physical statistics

5

Magnitude (bright) absolute magnitude parameter

Semi-major axis of the orbit

Eccentricity of the orbit

these two are great parameters

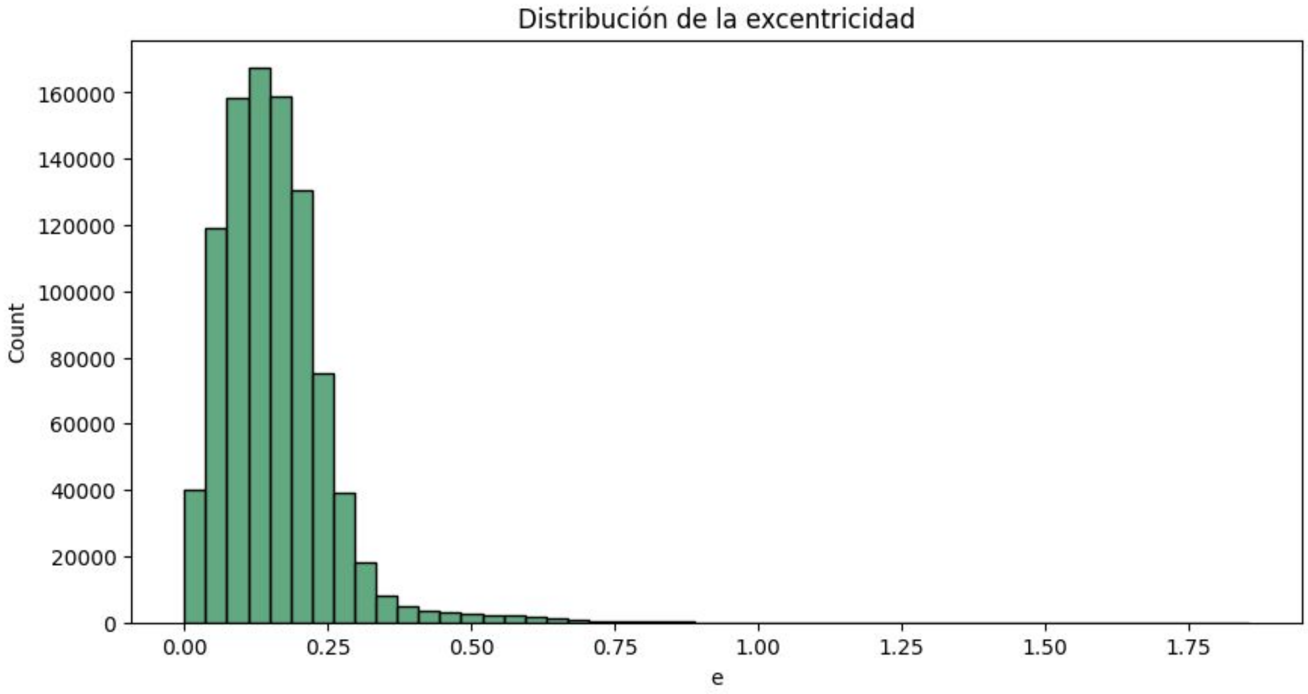
Inclination angle between the x-y ecliptic plane

Moid Earth minimum orbit intersection distance

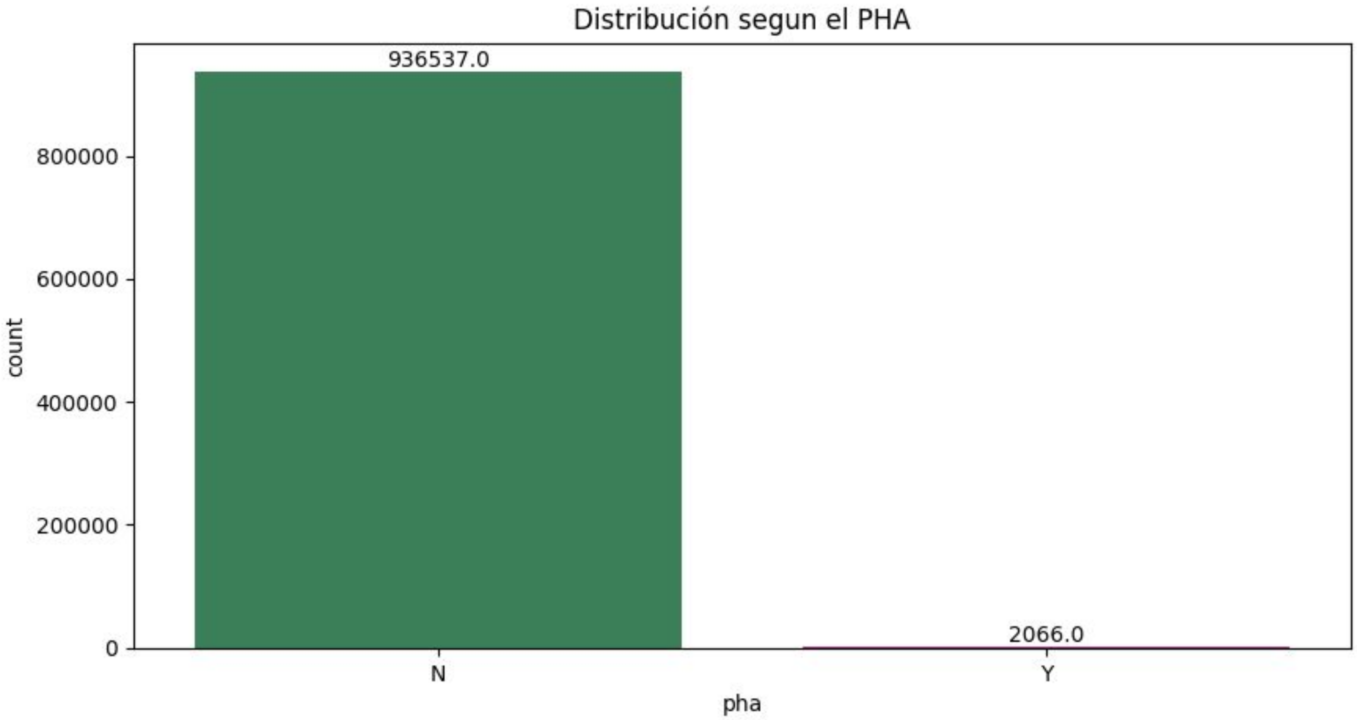
Others like multimode, intensity, deviation, shape asymmetry and ellipticity

Exploratory Data Analysis

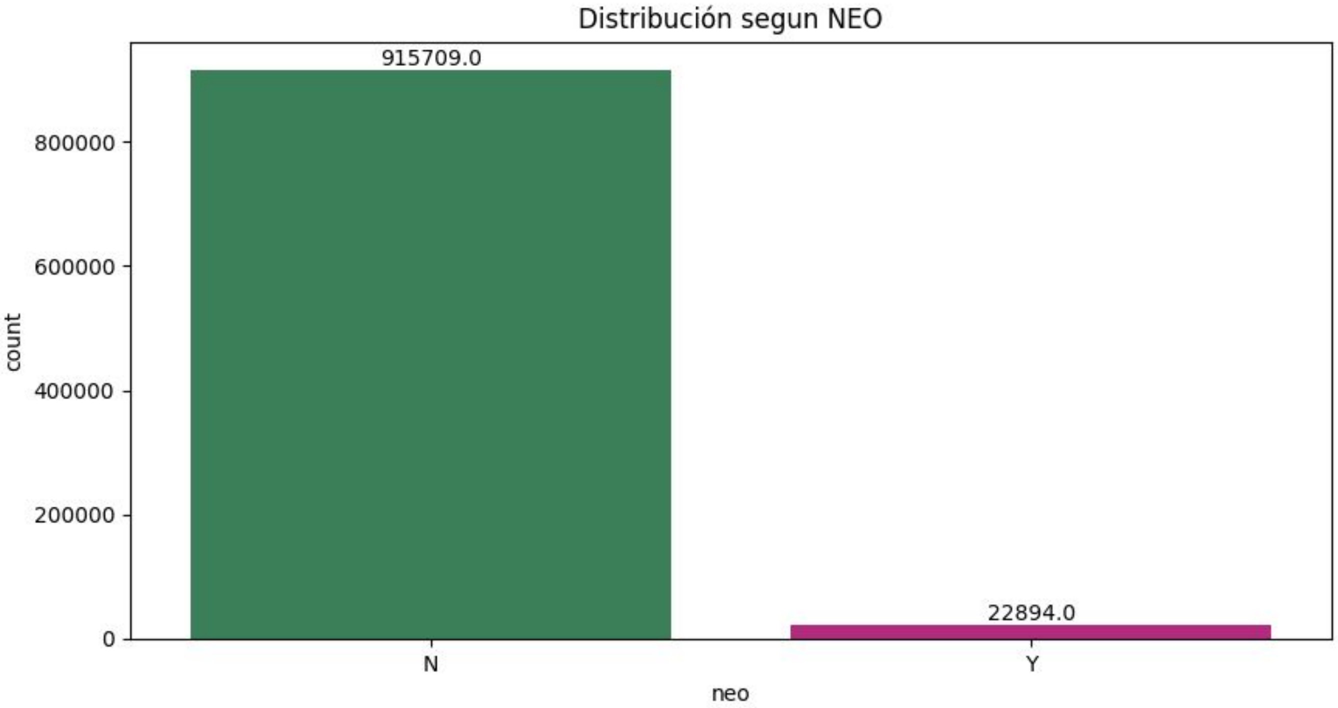
DATASET



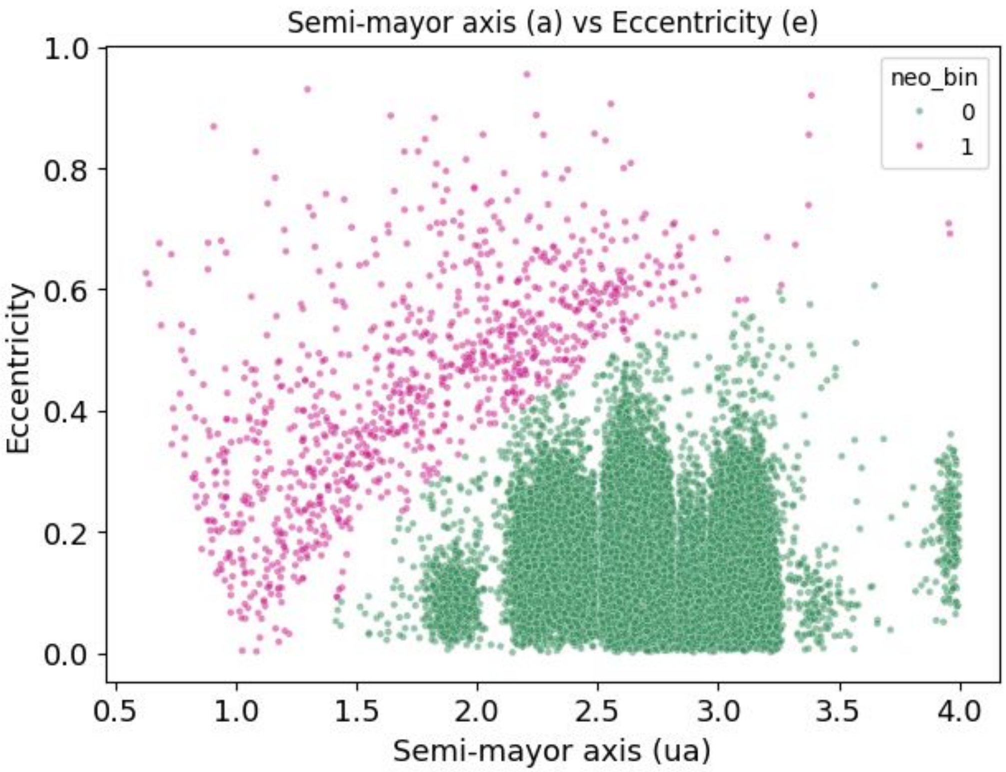
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



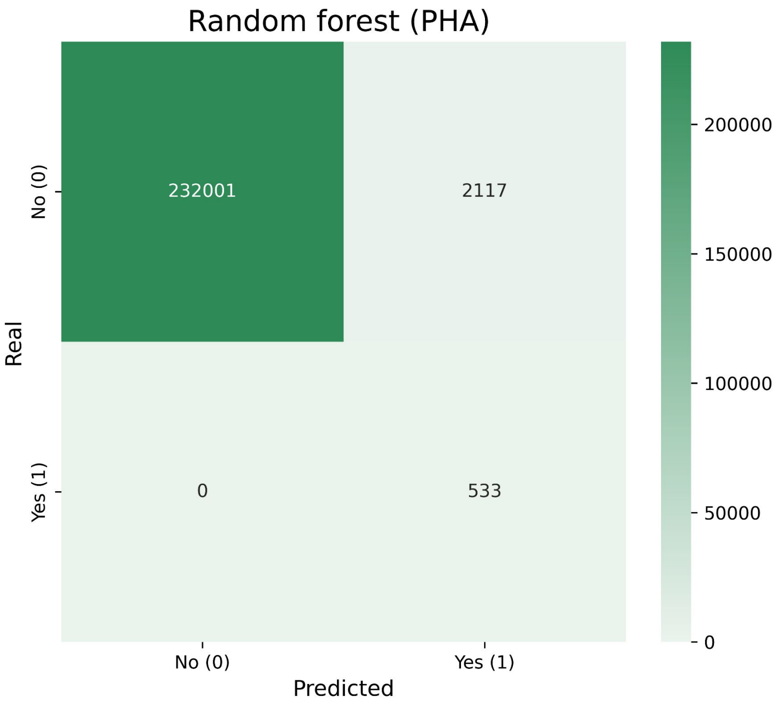
Results

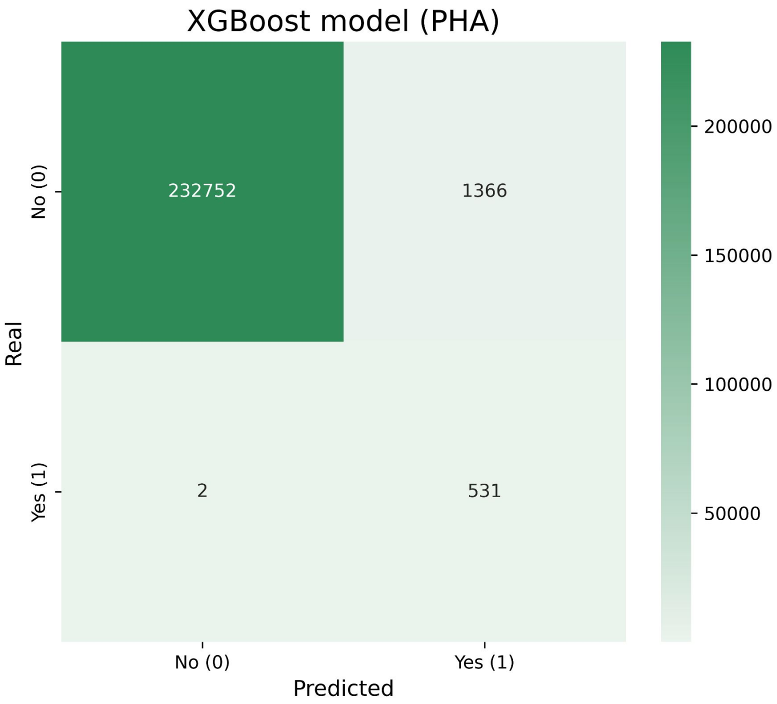
Confusion Matrix

ROC Curve

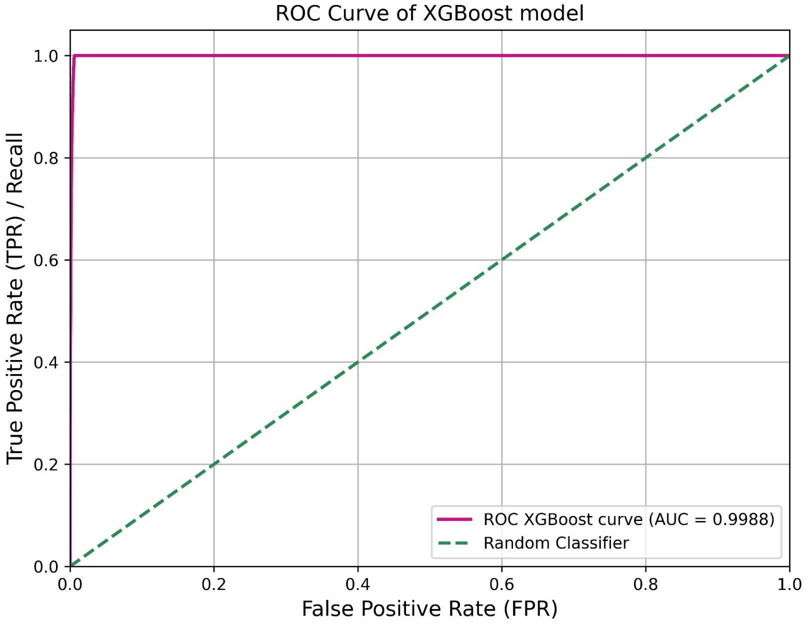
Learning Curves

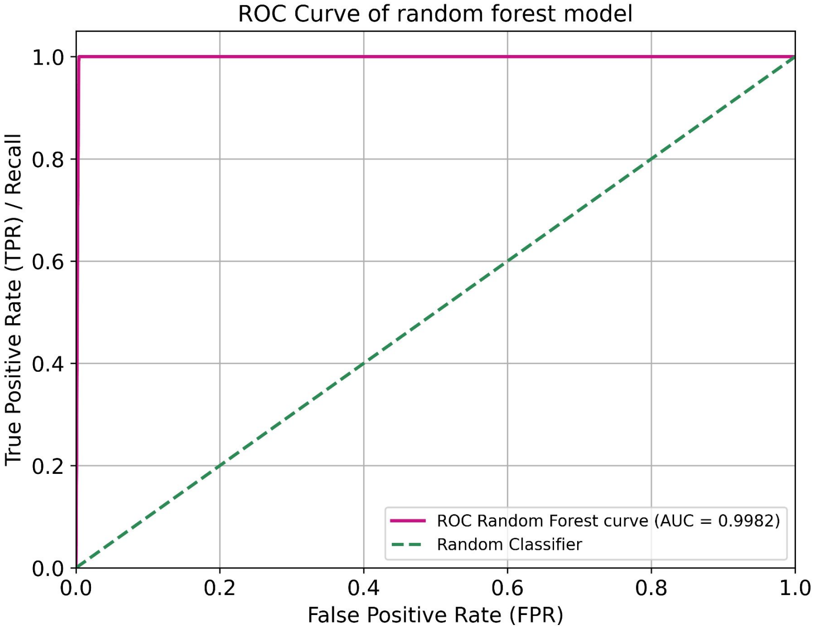
Confusion Matrix





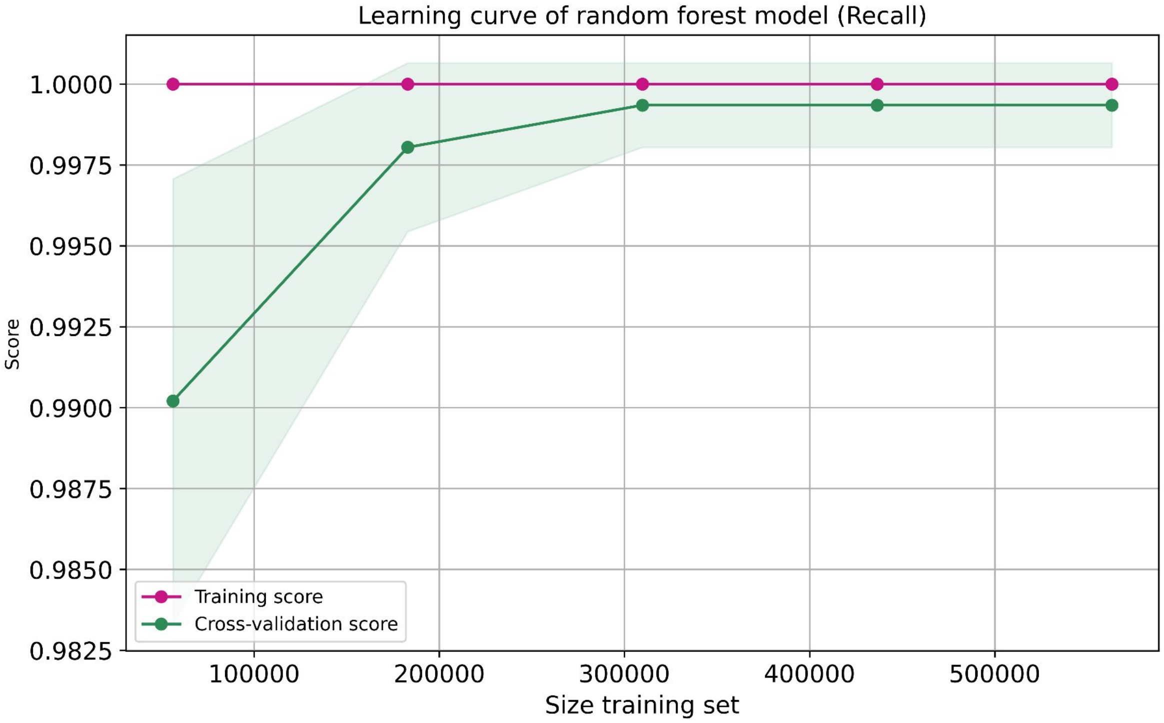
ROC Curves

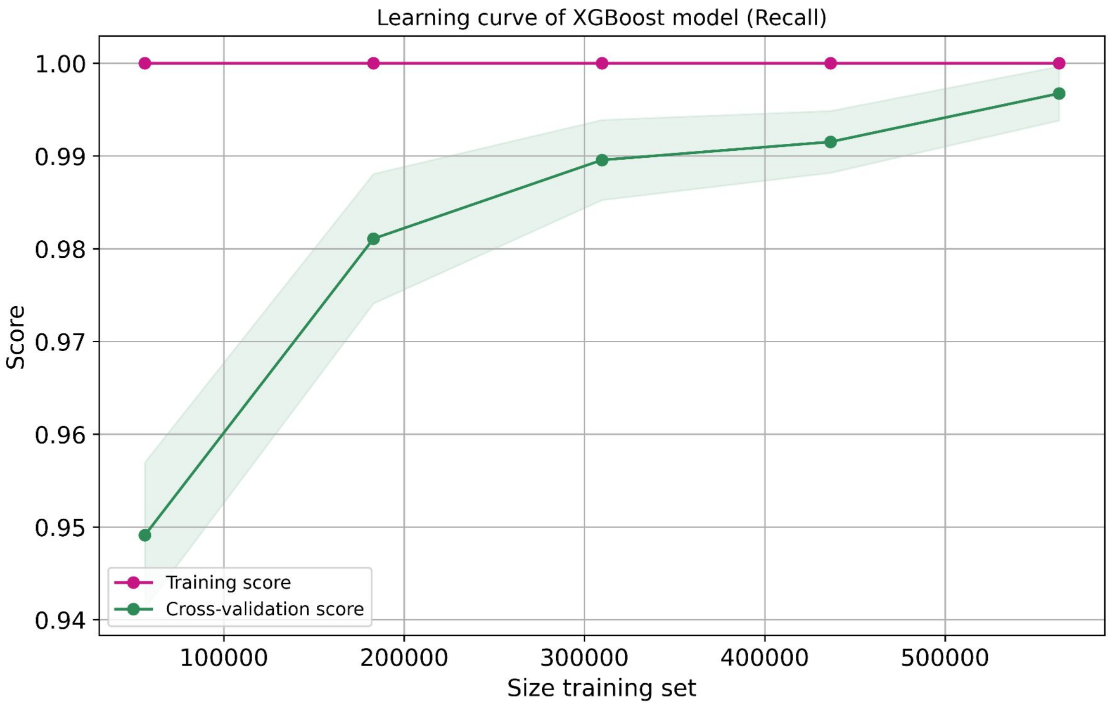

ROC Curve of XGBoost model
True Positive Rate (TPR) / Recall
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
ROC XGBoost curve (AUC = 0.9988)
Random Classifier


ROC Curve of random forest model
True Positive Rate (TPR) / Recall
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
ROC Random Forest curve (AUC = 0.9982)
Random Classifier

Learning Curve
Random Forest

- ★ High capacity model
- ★ To a sufficiently large training set, low variance, and low bias.
- ★ No **overfitting**





Learning Curve
XGBoost

- ★ Perfect fit to training data
- ★ Strong boosting model
- ★ Quickly stabilization
- ★ Greater tendency to **overfitting**

How well does each model behaves?

Random Forest

PERFECT RECALL FOR PHA

Identified all dangerous asteroids

ROBUST & STABLE

Minimal train-validation gap

MORE FALSE POSITIVE

2.117 false alarms

XGBoost

HIGHER AUC (0.9988)

Superior discrimination

LOW FALSE POSITIVE

Only 1.366 false alarms

2 FALSE NEGATIVES

Missed 2 dangerous asteroids

Conclusion

6 Interpretability

I like it

- Simple models
- Both OK but with a few exceptions

7 Scalability

Now what?

- 99% of objects is not bad
- Would be better with more data

8 Robustness

It works b-

- Random Forest needs only one separation
- Both methods work across multiple training/test

9 Novelty

I trust you

- ROC-AUC were prioritized instead of accuracy
- Random Forest reached *perfect recall*

10 Transferability

Works in every classification sample

- RF stabilizes with low variance
- XGBoost fits to perfectly which leads to overfitting
- Great example of transfer learning in astronomy

References

★ https://www.ibm.com/think/topics/random-forest

★ https://www.ibm.com/es-es/think/topics/xgboost?mhsrc=ibmsearch_a&mhq=xgboost

★ https://www.kaggle.com/datasets/ankit1743/asteroids-dataset

UTFSM - AST332

Cristian Nuñez
Isidora Morales
Miyaray Arenas

Machine Learning Final Project

Potential risk of asteroids trajectories

2025

Final Semester