Machine Learning Approaches for Classification and Diameter Prediction of Asteroids

Chapter · January 2023

DOI: 10.1007/978-981-19-7528-8_4

Mir Sakhawat Hossain
Kabi Nazrul Government College, Dhaka, Bangladesh

Machine Learning Approaches for Classification and Diameter Prediction of Asteroids

Mir Sakhawat Hossain[1] and Md. Akib Zabed[2]

[1] International Astrostatistics Association, Milano, Italy,
s.hossain18@gmail.com
[2] Florida International University, Miami, USA
mkhan149@fiu.edu

Abstract. In Astronomy, the size of data is increasing day by day and is becoming more complex than in previous years. Even it is also found in the study of Asteroids. There are millions of asteroids to study its classification and calculating its diameter to determine their characteristics. These will help us to know which are potentially hazardous asteroids for Earth. We have applied Machine Learning methods to classify the asteroids and predict their diameter. For classification task, we have implemented kNN classifier, Logistic Regression classifier, SGD classifier, and XGBoost classifier algorithms. For the prediction of diameter, we have used Linear Regression, Decision tree, Random Forest, Logistic regression, XGBoost regression, kNN, and Neural network models. We have depicted a comparative analysis of our results. Applying these approaches, we have gained significant 99.99 percent accuracy for asteroid classification task.

Keywords: asteroid classification, diameter prediction, machine learning modelling.

1 Introduction

The discovery rate of asteroids especially Near Earth Objects is on an average of 1000 per year[5]. The number of known asteroids has increased from 10000 to 750000 with their elements and this number is continuously increasing day by day[4]. This rate is increasing by some important factors i.e., highly developed ground-based surveys like Catalina Sky Survey, Palomar Transient Factory, Pan-STARRS-2, etc., and space satellite-based surveys like NEOSM, NEOCam, GAIA, etc. Asteroids families come from a common origin. The resultants of origin can be a product of a collision or rotational fission of a parent body or the satellites[12]. Asteroid families originated by collisions are identified in the domain of proper elements[6] like constant s of motion over time scales of Myr[7]. For classifying asteroids hierarchical clustering method (HCM) is widely used in proper element domain like proper semi-major axis a, proper eccentricity e, and proper inclination i. In this method, using a distance metric two asteroids in domains of proper elements or frequencies are calculated. If the distance of the second object from the first is less than the primary characteristics which are called cut off, the object is assumed as the member of the first object asteroid family. This process is repeated until a second new asteroid is found[14]. In the last years, there are many machine learning algorithms have been applied to solve different types of classification problems in Astronomy. The traditional HCM method is highly computationally intensive. So, we have applied different Machine Learning classification and regression algorithms to classify the asteroids and predict their diameter of them. Our main contributions are given below.

1. Feature Analysis on asteroid data and gained some important features which are mostly important to understand the characteristics of asteroids.
2. Effectively and accurately classified asteroids using different Machine Learning algorithms which has outperformed the traditional statistical approaches.
3. Predicted the diameter of asteroid efficiently using Machine Learning regressors.

This paper is organized as follows: In section II, we have given a review of related works for illustrating the previous works already happened in this astronomy field. Section III describes the methodology to classify asteroids and predict the diameter of asteroids. The result and analysis of our study is presented in section IV. And finally, we have concluded our work in section V, including discussion points for future works.

2 Literature Review

In this section, we are presenting some of the previous works related to asteroid group detection and identification, and the uses of Machine Learning algorithms to detect and classify different asteroid related data. A Machine Learning approach for identification of asteroid family groups was presented in[6]. V. Carruba et al. applied hierarchical clustering algorithms to identify and classify different collisional asteroid families where they outperformed the accuracy of traditional methods[2]. They identified 6 new asteroid families and 13 new clumps providing an accuracy of 89.5% and showed performance evaluation metrics for identifying different asteroid groups. Different regression algorithms and Lambert baselines predictor had been applied in[8] to give a solution to global optimization problems. A. Mereta et al. implemented application of machine learning regressors to evaluate final spacecraft mass after optimal low-thrust transfers between two Near Earth objects. They showed comparative analysis by applying different regression models: Decision Tree, Random Forest, Extra Trees, Gradient Boosting, AdaBoost, and Bagging and also Neural Networks to estimate the spacecraft mass. Their accuracy outperformed the commonly implemented Lambert's estimation, but they couldn't completely remove the maximum initial mass (m*).
There are lots of studies already happened to understand different characteristics of asteroids. Most of the authors preferred Machine Learning algorithms and regression models to detect or predict or classify different objects or asteroids on the asteroid data. V. Pasko proposed a model to predict combination of orbital

parameters for the undiscovered PHAs (Potentially Hazardous Asteroids)[10]. He used Support Vector Machine algorithms with a non-linear kernel function named Radial Basis Function (RBF) to find boundaries of different PHA subgroups For clustering purpose, the author used Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. In 1982, J.K. Davies et al. presented a method for asteroid classification in[3]. They studied previous works related to asteroid classification and provided a solution based on Numerical Taxonomy. Numerical Taxonomy programs was traditionally used in Microbiology and the authors represented a dendrogram for 82 asteroids. They didn't mention about any accuracy of their classification task.
A statistical model had been introduced[1] for asteroids classification task in 1987. MA Barucci et al. used G-mode analysis which is a multivariate statistical approach to classify 438 asteroids. They separated seven taxonomic units of asteroids with 99.7% confidence level by using their method. They classified asteroids into nine classes based on eight-color photometry data and geometric albedo of asteroids. A taxonomic classification model for asteroids was presented in M Popescu et al.[11]. M Popescu et al. used k-nearest neighbors algorithm (KNN) and probabilistic methods to classify 18265 asteroids on MOVIS catalog which contains near-infrared colors. They found 70% matching for identifying silicaceous and carbonaceous types while doing comparison with existing model on the basis of SDSS survey. From the above studies, we can say that there are lots of works existing related to classification of asteroids. But no diameter prediction model is available for asteroid data. Here, we have proposed a model to classify 13 asteroid groups using Machine Learning regression algorithms and predict the diameter of asteroids efficiently.

3 METHODOLOGY

Our methodology is divided into four parts. They are: (i) Description of dataset, (ii) Preprocessing of dataset, and (iii) Different models for asteroids classification (iv) Diameter prediction models.

3.1 Description of Dataset

The dataset consists of 958524 rows and 45 columns. The columns consist of significant orbital parameters of the asteroids like semi-major axis a, eccentricity e, inclination i, perihelion distance q, aphelion distance ad, period as days per, period as years per-y, etc. The columns also consist of some other important characteristics like absolute magnitude H, geometric albedo as albedo, potentially hazardous asteroid PHA or not, near-earth objects NEO or not. The dataset is available in the Jet Propulsion Laboratory of NASA[9]. Orbital elements of Asteroids describe the state of cartesian position and velocity in a conic orbit in space at a specific epoch. These Keplerian elements express the osculating orbit (Tangent orbit approximating the actual orbit) of an object. We have assumed these elements as machine learning features. Here we have described them.

Absolute Magnitude: The visual magnitude denoted by H of an asteroid is measured from where observation was placed at unit heliocentric and geocentric distances at zero phase angle.

Diameter: Diameter of an Asteroid at unit Kilometer.

Albedo: Albedo is the ratio of the light received by a physical body to the light reflected by the body at zero phase angle with the same position and apparent size. Its value ranges from 0 to 1.

Eccentricity: Eccentricity denoted by e is an orbital parameter that describes the structure of orbit. It is a ratio between foci to semi-major-axis.

Semi-major-axis: Semi-major-axis of an elliptical orbit is half of the major axis. It is denoted by a.

Perihelion Distance: Perihelion distance denoted by q is the closest distance between the orbiting body (Here eg. Asteroids) and the sun.

Inclination: Inclination denoted by i is the angle between a vector normal to the orbiting body's orbit plane and the ecliptic plane as the reference plane.

Longitude of the ascending node: The longitude of the ascending node denoted by om is the angle between the inertial frame vernal equinox and the point of passing up through the reference frame.

Argument of Perihelion: Argument of Perihelion denoted by w is the angle between the ascending node line and the perihelion point in the direction of orbit.

Mean Anomaly: Mean Anomaly denoted by ma is the product of mean motion of orbiting body and past perihelion passage.

Aphelion Distance: Aphelion Distance denoted by the ad is the farthest distance between the orbiting body (Here eg. Asteroids) and the sun.

Mean Motion: Mean Motion denoted by n is the angular speed at unit degree per day to complete one orbit around an ideal ellipse.

Period: Sidereal Period at unit day.

3.2 Preprocessing of Dataset

The dataset that was collected was very well defined and well-arranged but it was needed to perform some preprocessing tasks for a better model outcome. In the dataset, some data were missing that created hindrance for machine learning modeling. We substituted the missing values with median values. We did not remove any row. Then we have analyzed the features of all attributes we have got in our dataset.

3.3 Feature Analysis

For our modelling at first we have dropped the unnecessary features which are not related to the classifications based on Pearson correlation among the features and classes. We have used two methods ExtraTreesClassifier Table 2 and Chi Square Table 1a to select important features. Then we have computed Variance Inflation Factor (VIF) Table 1b of the rest of the features. In this step, we have also dropped 6 features which may cause over fitting. We then checked the VIF factor above of 8 of the rest features. Based on feature importance, VIF Factors and Correlations we have chosen the most important 8 features. These features are absolute magnitude H, object diameter as diameter, geometric albedo as albedo, eccentricity e, semi-major axis a, perihelion distance q, inclination i, and mean motion n. And we have shuffled the dataset 5 times and then split it into train and test set with an 8 : 2 ratio.

(a) Chi Square Method

Feature Name	Score
moid ld	36815.078985
q	21762.640286
albedo	17466.459663
e	9153.018613
n	6853.726787
i	3213.704662
H	2197.257868
diameter	2084.505512
ma	287.077008
per	284.962249

(b) Variance Inflation Factor (VIF)

Feature Name	VIF Factor
H	2.16
diameter	1.71
albedo	1.29
e	1.20
a	1.01
q	1.64
i	1.12
n	1.86

Table 1: Feature Extaction

3.4 Different Models for Asteroids Classification

In our machine learning modeling for asteroids classification task, we have applied supervised learning algorithms. We have applied K-Nearest Neighbors

Table 2: Extra Tree Classifier Method

Feature Name	Score of Feature Importance
H	0.053427
diameter	0.079550
albedo	0.145152
e	0.076104
a	0.100125
q	0.119450
i	0.024435
om	0.000915
w	0.000996
ma	0.001626
ad	0.050871
n	0.151369
per	0.112984
moid ld	0.082996

(KNN), Logistic Regression, Stochastic Gradient Descent (SGD), and Extreme Gradient Boosting (XGBoost).

K-Nearest Neighbors: K-Nearest Neighbors is a simple non-parametric classification method. It is used for both classification and regression problems. For the classification method, the output of this algorithm indicates a class. Here a target is classified according to the plural vote of its closest neighbors, with this target being assigned to the class which is the most common among its K Nearest Neighbors where K is a positive integer. In regression problems, the target is a numerical value. The calculation process is the same but the output will be the predicted numerical value. In our modeling, we have applied the KNN algorithm with default hyperparameters. Then we have applied k-fold cross-validation in this model.

Logistic Regression: Logistic Regression is a probabilistic algorithm where the probability of a class or numerical target is predicted. The mathematically logistic regression model has a dependent variable which can be one of the two possible values are labeled 0 or 1. In this model, the logarithmic values of odds for the labeled 1 are a linear combination of one or more independent variables. Each of the independent variables can be a binary variable like two classes. The corresponding probability of the binary variable varies. But our target class is multi-label classes. So we used one vs rest classifier which is a heuristic method for the multi-label classification tasks. Then we have implemented k-fold cross-validation in the Logistic Regression model.

Stochastic Gradient Descent: Stochastic Gradient Descent (SGD) is a linear classifier where this is optimized by the SGD method. Here it is used to minimize the cost function. This value of logistic regression cannot be computed directly but it is possible with the SGD method. In this procedure, we descend along with the cost function towards local minima to the global minimum. Since it is a logistic regression, here we have also applied one vs. rest classifier and the k-fold cross-validation method in this model.

Extreme Gradient Boosting: Extreme Gradient Boosting (XGBoost) is an ensemble learning model based on a decision tree that uses a gradient boosting framework. Trees are added to the ensemble each time and fit to minimize prediction errors by the prior model using any arbitrary differentiable loss function and gradient descent optimization. To handle multi-label target class, we have used one vs. rest classifier. Then we have implemented k-fold cross-validation in this model. Our overall procedure of applying machine learning models is illustrated in Fig. 1.

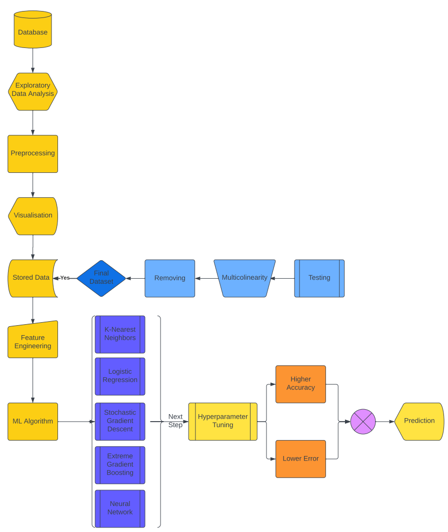


Fig. 1: Overview of methodology

3.5 Diameter Prediction Models

For the diameter prediction task, we have taken only absolute magnitude H and geometric albedo as inputs. We have analyzed features of different columns of our dataset and finally, we have taken H and albedo columns as inputs and predicted diameter of asteroids as output using 7 different Machine Learning models. For predicting the numeric value, we have implemented regression models: Linear regression, Decision tree, Random Forest, Logistic Regression, k-Nearest Neighbors, XGBoost, and the Neural Network model. Linear Regression, Decision Tree, and Random Forest models are very popular regressors that are mostly used for predicting numeric values. We have used Logistic Regression, KNN, and XGBoost for both classification and prediction tasks. The working principle of these models is the same as the classification task. But in the prediction task, we have got only one predicted numeric output which is the value of diameter where we have found different classes as the output before. A Neural network model takes inputs and creates a network of neurons. These inputs are passed to hidden layers and fully connected layers. We have used a neural network model and passed H and albedo as inputs to predict the diameter as output in the output layer. We have used default hyperparameters for the learning process in all the models. While training and testing, we have calculated mean squared error, mean absolute error, and root mean square error for all the models.

4 RESULT AND ANALYSIS

To evaluate the implemented models for the classification of asteroids, we have calculated the accuracy and the performance evaluation metrics which include precision, recall, and F1-score of each of the models. The following formulas (Eq. 1 to Eq. 4) are being used to calculate these metrics.

$$Accuracy : A = ((TP + TN)/(TP + TN + FN + FP)) * 100\% \tag{1}$$

$$Precision : P = TP/(TP + FP) \tag{2}$$

$$RecallRate : R = TP/(TP + FN) \tag{3}$$

$$F1 - Score : F1 = (2 * Precision * Recall)/(Precision + Recall) \tag{4}$$

Firstly, we have classified the asteroids into 13 classes using KNN, Logistic Regression, SGD, and XGBoost models. Table 3. shows the accuracy and class-wise evaluation metrics of the K-Nearest Neighbors model. We have achieved 99.42% accuracy using KNN model. And we have got 98.20% accuracy using k-fold cross-validation in this model. After cross-validation, the average precision, average recall, and average F1-score are 0.987, 0.982, 0.984 respectively. The accuracy and class-wise evaluation metrics of the Logistic Regression model are presented in Table 4. We have got 94.73% accuracy using the Logistic Regression model using the One Vs Rest classifier. After doing k-fold cross-validation in this model, we have achieved 95.06% accuracy, and the average precision, average recall, and average F1-score are 0.965, 0.960, 0.960 respectively.

Then, we have classified the asteroids using the Stochastic Gradient Descent model. Table 5. depicts the accuracy and class-wise evaluation metrics of the SGD model. 90.22% accuracy has been achieved using the SGD model. After using k-fold cross-validation in SGD model, we have gained 92.14% accuracy, average precision: 0.936, average recall: 0.937, and average F1-score: 0.932.

And finally, the accuracy and class-wise evaluation metrics of the XGBoost model are illustrated in Table 6. Here, we have achieved a significant accuracy which is 99.99% using the XGBoost model. We have got the same accuracy using k-fold cross-validation in XGBoost model. And the average precision, average recall, and average F1-score are 0.999, 0.999, 0.999 respectively.

Table 3: Accuracy, Precision , Recall and F1 Score for KNN

Class Name	Accuracy	Precision	Recall	F1 Score
AMO		0.99	0.98	0.98
APO		0.99	0.99	0.99
AST		0.93	0.93	0.93
ATE		0.99	0.99	0.99
CEN		0.99	0.95	0.97
HYA		1.00	1.00	1.00
IEO	99.42%	1.00	1.00	1.00
IMB		1.00	1.00	1.00
MBA		1.00	1.00	1.00
MCA		0.99	0.99	0.99
OMB		0.96	0.86	0.91
TJN		1.00	1.00	1.00
TNO		1.00	1.00	1.00

Table 4: Accuracy, Precision , Recall and F1 Score for Logistic Regression

Class Name	Accuracy	Precision	Recall	F1 Score
AMO		0.05	0.02	0.03
APO		0.98	0.98	0.98
AST		0.00	0.00	0.00
ATE		0.99	0.99	0.99
CEN		0.47	0.23	0.31
HYA		1.00	1.00	1.00
IEO	94.73%	0.86	0.75	0.80
IMB		0.99	0.99	0.99
MBA		0.98	1.00	0.99
MCA		0.38	0.41	0.41
OMB		0.96	0.51	0.67
TJN		0.99	0.99	0.99
TNO		1.00	1.00	1.00

By analyzing the presented results (from Table I to Table IV), we can say that the overall accuracy and performance of the XGBoost model is the best compared to other Machine Learning regression models to classify asteroid groups into 13 categories. Furthermore, we can understand the accuracy of a model by generating ROC-AUC curve. So, we have generated ROC curves for all the applied models. Here, we are presenting the ROC curves of the XGBoost model which is the best model for classifying asteroids. The macro-average ROC curve is illustrated in the following Fig. 2a. For macro-average, ROC AUC score of the model is 0.99996. And Fig. 2b presents the micro-average ROC curve of XGBoost model. For weighted by prevalence, ROC AUC score is 1.00000 for the XGBoost model.

For evaluating the applied models for diameter prediction, we have calculated mean squared error (mse), mean absolute error (mae) and root mean square error

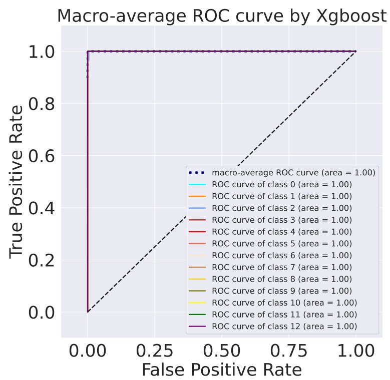
Table 5: Accuracy, Precision , Recall and F1 Score for SGD

Class Name	Accuracy	Precision	Recall	F1 Score
AMO	90.22%	0.05	0.02	0.03
APO		0.98	0.98	0.98
AST		0.00	0.00	0.00
ATE		0.99	0.99	0.99
CEN		0.47	0.23	0.31
HYA		1.00	1.00	1.00
IEO		0.86	0.75	0.80
IMB		0.99	0.99	0.99
MBA		0.98	1.00	0.99
MCA		0.38	0.14	0.21
OMB		0.96	0.51	0.67
TJN		0.99	0.99	0.99
TNO		1.00	1.00	1.00

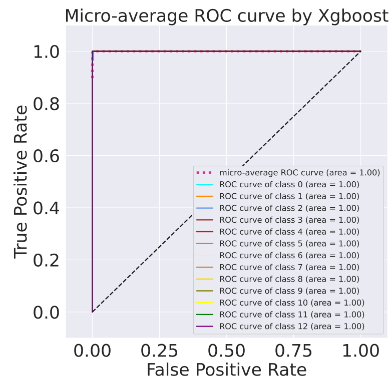
Table 6: Accuracy, Precision , Recall and F1 Score for XGBoost

Class Name	Accuracy	Precision	Recall	F1 Score
AMO	99.99%	1.00	1.00	1.00
APO		1.00	1.00	1.00
AST		0.96	0.90	0.93
ATE		1.00	1.00	1.00
CEN		1.00	0.98	0.99
HYA		1.00	1.00	1.00
IEO		1.00	1.00	1.00
IMB		1.00	1.00	1.00
MBA		1.00	1.00	1.00
MCA		1.00	1.00	1.00
OMB		1.00	1.00	1.00
TJN		1.00	1.00	1.00
TNO		1.00	1.00	1.00

(rmse) for each of the models. For predicting numeric data, the lower mse, mae, and rmse determine the higher accuracy of the prediction model. A comparative analysis is depicted in Table 7 to show the mse, mae, and rmse we have got implementing different Machine Learning regression models: (i) Linear Regression, (ii) Decision Tree, (iii) Random Forest, (iv) Logistic Regression, (v) XGBoost, (vi) K-Nearest Neighbors, and (vii) Neural Network model for predicting the diameter of the asteroids. We have achieved the lowest mse and rmse using XGBoost model which are 1.84 and 1.36 respectively, and lowest mae: 0.49 using the K-Nearest Neighbors model. Overall, we can say that the XGBoost model is the best model for diameter prediction as it has the lowest rmse among all the models.



(a) Macro Average ROC curve of XGBoost Model



(b) Micro Average ROC curve of XGBoost Model

Fig. 2: ROC curve of XGBoost Model (X-Axis:False Positive Rate, Y-Axis:True Positive Rate)

Table 7: mse, mae and rmse for different Machine Learning models

Model Name	**mse**	**mae**	**rmse**
Linear Regression	58.91	2.76	7.55
Decision Tree	2.84	0.55	1.68
Random Forest	2.05	0.53	1.43
Logistic Regression	60.69	1.45	7.79
XGBoost	1.84	0.55	1.36
KNN	2.21	0.49	1.49
Neural Network	4.74	0.49	2.18

5 Conclusion

We have applied eight Machine Learning models. Among these models, XGBoost plays the best performance for both classification and prediction tasks. The main purpose of this work is to find out the best way of asteroid classification and its diameter prediction. Here we have also got that Machine Learning algorithms can achieve the most accurate class and diameter. So classical methods that are computationally intensive and time-consuming can be replaced with this model. In the future, we will find out the best model to detect potentially hazardous asteroids by applying different Machine Learning algorithms and Deep Learning models.

Acknowledgment

We are grateful to the Jet Propulsion Laboratory of NASA for its amazing Asteroid Database and Sciserver[13] revolutionary data science platform for Astronomy and other scientific fields run by John Hopkins University.

Bibliography

[1] M.Antonietta Barucci, M.Teresa Capria, Angioletta Coradini, and Marcello Fulchignoni. Classification of asteroids using g-mode analysis. *Icarus*, 72(2):304–324, nov 1987.

[2] V Carruba, S Aljbaae, and A Lucchini. Machine-learning identification of asteroid groups. *Monthly Notices of the Royal Astronomical Society*, 488(1):1377–1386, 07 2019.

[3] J.K. Davies, N. Eaton, S.F. Green, R.S. McCheyne, and A.J. Meadows. The classification of asteroids. *Vistas in Astronomy*, 26:243–251, jan 1982.

[4] F. E. DeMeo, C. M. O'D. Alexander, K. J. Walsh, C. R. Chapman, and R. P. Binzel. *The Compositional Structure of the Asteroid Belt*. University of Arizona Press, 2015.

[5] Martin Elvis, C. Beeson, J. Galache, F. DeMeo, I. Evans, J. Evans, N. Konidaris, J. Najita, L. Allen, E. Christensen, and T. Spahr. Linnaeus: Boosting near earth asteroid charaterization rates. 45:208.24, October 2013.

[6] K. Hirayama. Families of asteroids. *Japanese Journal of Astronomy and Geophysics*, 1:55, January 1922.

[7] Knezevi´c, Z. and Milani, A. Proper element catalogs and asteroid families. *A&A*, 403(3):1165–1173, may 2003.

[8] Alessio Mereta, Dario Izzo, and Alexander Wittig. Machine learning of optimal low-thrust transfers between near-earth objects. In *Lecture Notes in Computer Science*, pages 543–553. Springer International Publishing, 2017.

[9] NASA. Jpl small-body database search engine, April 2020.

[10] Vadym Pasko. *Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning*. Springer International Publishing, Cham, February 2018.

[11] Popescu, M., Licandro, J., Carvano, J. M., Stoicescu, R., de León, J., Morate, D., Boaca, I. L., and Cristescu, C. P. Taxonomic classification of asteroids based on movis near-infrared colors. *A&A*, 617:A12, sep 2018.

[12] P. Pravec, D. Vokrouhlický, D. Polishook, D. J. Scheeres, A. W. Harris, A. Galád, O. Vaduvescu, F. Pozo, A. Barr, P. Longa, F. Vachier, F. Colas, D. P. Pray, J. Pollock, D. Reichart, K. Ivarsen, J. Haislip, A. LaCluyze, P. Kušnirák, T. Henych, F. Marchis, B. Macomber, S. A. Jacobson, Yu. N. Krugly, A. V. Sergeev, and A. Leroy. Formation of asteroid pairs by rotational fission. *Nature*, 466(7310):1085–1088, aug 2010.

[13] M. Taghizadeh-Popp, J.W. Kim, G. Lemson, D. Medvedev, M.J. Raddick, A.S. Szalay, A.R. Thakar, J. Booker, C. Chhetri, L. Dobos, and M. Rippin. SciServer: A science platform for astronomy and beyond. *Astronomy and Computing*, 33:100412, oct 2020.

[14] Vincenzo Zappala, Alberto Cellino, Paolo Farinella, and Zoran Knezevic. Asteroid families. i. identification by hierarchical clustering and reliability assessment. *American Astronomical Society*, 100:2030, December 1990.