UDC 004.852
IRSTI 20.53

https://doi.org/10.55452/1998-6688-2023-20-3-7-16

Duisek B.E*, Sarsembin D.D, Abdurazak K.A.

Kazakh-British Technical University, 050000, Almaty, Kazakhstan
*E-mail: be_duisek@kbtu.kz

COMPARISON AND ANALYSIS OF DIFFERENT MACHINE LEARNING METHODS ON ASTEROID DIAMETER PREDICTIONS BASED ON THE NASA SMALL CELESTIAL BODIES DATABASE

Abstract. The database of small celestial bodies NASA is provided by the Jet Propulsion Laboratory and represents the collected information about asteroids and comets, describing their parameters available for observation and determination, including physical ones, as well as their classification and data on the number and duration of observation. Many of these celestial techs have an incomplete description of their properties, which makes it difficult to predict their behavior and potential interaction with other objects in space, including man-made ones. This study proposes a solution to a certain part of the problems of asteroid exploration by finding a prediction of the diameter of asteroids based on information from the NASA database and the results of machine learning methods on processed data from the source. For this research, some of the most commonly used algorithms for implementing such prediction models have been selected, such as KNN, linear regression, random forest, decision trees, and gradient boosting. Applied machine learning algorithms were evaluated based on the results of diameter prediction accuracy, speed of training and prediction process, and square mean error rates. The study will help to choose the most optimal approach for predicting this feature of asteroids, describe the process of data pre-processing, while achieving the best performance of the model, and analyze the correlations between the properties of these celestial bodies.

Key words: machine learning, asteroid, prediction model, KNN, linear regression, random forest, decision tree, gradient boosting.

Introduction

The Solar System and space beyond are inhabited by plenty of small body objects that float in different directions and collide with other objects, which may result in the creation of potentially hazardous situations for our planet[7]. Hence numerous researchers track and collect data about asteroids to identify those objects that are the most threatening to the Earth. In 2009, the University of Glasgow proposed a paper “Multicriteria Comparison Among Several Mitigation Strategies for Dangerous Near-Earth Objects” with properties of objects that may help to evaluate and assess effective methods of identifying such. This paper contains about 90 variables that are taken into account to predict and define mathematics models for identifying dangerous near-Earth objects[16]. None of the proposed strategies used machine learning algorithms. In another article “Parameter estimation for optimal asteroid transfer trajectories using supervised machine learning” the authors used supervised machine learning techniques such as differential evolution algorithm, gaussian process regression to evaluate the trajectories of asteroids[17].

Our research is based on the current database of small celestial bodies presented by the Jet Propulsion Laboratory of California Institute of Technology consists of hundreds of thousands of asteroids and comets, and while some of them are well studied, some objects miss many valuable parameters, which can describe their future interactions with other bodies, while also assisting in prediction possible behavioral patterns[11]. Correlation with several attributes, such as categorical values of Potentially Hazardous Asteroids (PHA)[3] or semi-major axis, may help researchers to predict possible threats of previously unknown or under-researched asteroids or generally identify characteristics of their orbits[20]. However, it is worth considering that an asteroid's diameter also has a direct correlation with its mass. Mass is not one of the features we are taking into account in our study, but the distribution of masses of asteroids is a more complex topic due to the nature of mass measurement techniques, but the various mass distribution prediction methods have been applied for closely located asteroids for many decades now[9].

Table 1 – The embeddings for each column

name	Name of asteroid
a	Semi-major axis, in AU
e	Eccentricity
i	Inclination, in degrees
om	Longitude of the ascending node, in degrees
w	Argument of perihelion, in degrees
q	Perihelion distance, in AU
ad	Aphelion distance, in AU
per_y	Orbital period, in years
data_arc	Number of days spanned by the data arc, in days
condition_code	Orbit condition code
n_obs_used	Number of observations used
H	Absolute magnitude parameter
neo	Near-Earth Object flag, yes or no
pha	Potentially Hazardous Asteroid flag, yes or no
diameter	Object diameter, in kilometers
extent	Object tri-axial ellipsoid dimensions, in kilometers
albedo	Albedo
rot_per	Rotation period, in hours
GM	Product of the mass (M) multiplied by the gravitational constant (G)
BV	Color index B-V magnitude difference
UB	Color index U-B magnitude difference
IR	Color index I-R magnitude difference
spec_B	Spectral taxonomic type (SMASSII)
spec_T	Spectral taxonomic type (Tholen)
G	Magnitude slope parameter
moid	Earth minimum orbit intersection distance, in AU
class	Orbit class
n	Mean motion, in degrees/days
per	Orbital period, in days
ma	Mean anomaly, in degrees

Literature review

Basu(2019) used the Multilayer Perceptron algorithm to predict the diameter of asteroids[21]. It analyzed its performance, utilizing other machine learning methods on the same dataset, as in this paper. It appears that the methods that were used for comparison differ from the methods that will be used in this paper.

Recently Hossain & Zabed(2023) produced a comparison of machine learning algorithms for the classification and diameter prediction of asteroids[22]. For the task of diameter, predictions used the same machine-learning algorithms. However, only parameters of absolute magnitude H and albedo were used as inputs and it seems that no proper process of finding data correlation between parameters was conducted, for the task of diameter prediction.

It is clear that a thorough analysis of the dataset features correlation is needed for a more accurate forecast of asteroid diameter and comparison of the performance of machine learning algorithms. Moreover, studies involving this particular NASA dataset and forecasting models for diameter prediction are not that frequent and mostly set their objectives in other areas.

Main provisions

The main goal of the analysis of this particular dataset consists of data preprocessing [1] and feature identification through the profound examination of the correlation between the diameter and each of the columns.

Figure 1 – Relationship between data_arc and diameter

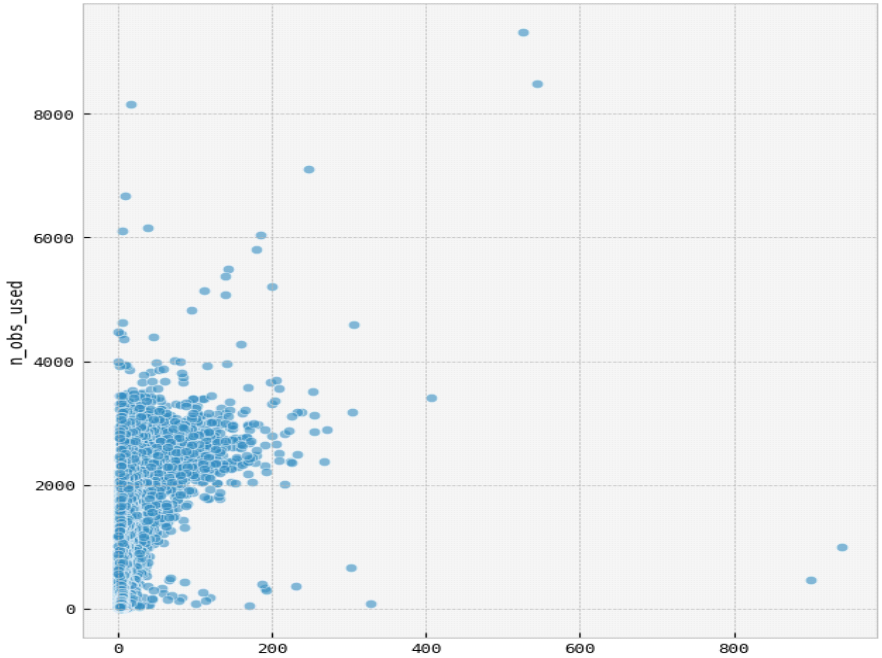

Figure 2 – Relationship between n_obs_used and diameter

Data

The dataset has 839736 entries and 27 columns. Out of all columns and their descriptions, which are depicted in Table 1., initially we dropped only 3 features: *name*, *data_arc*, and *n_obs_used* because all these fields will either result in overly biased results, in the case of *data_arc* or *n_obs_used* or just be useless in prediction since they are manually assigned names, in case of *name*. *n_obs_used* column represents the total number of observations of the distinct asteroid. At the same time, the *data_arc* feature refers to the total amount of days between the first and the last observation of an asteroid. Even though both fields have a high correlation with diameter, shown in Figure 1. for *data_arc*, and in Figure 2. for *n_obs_used*, they represent historical human activity. They will not contribute to predictions for newly discovered celestial objects.

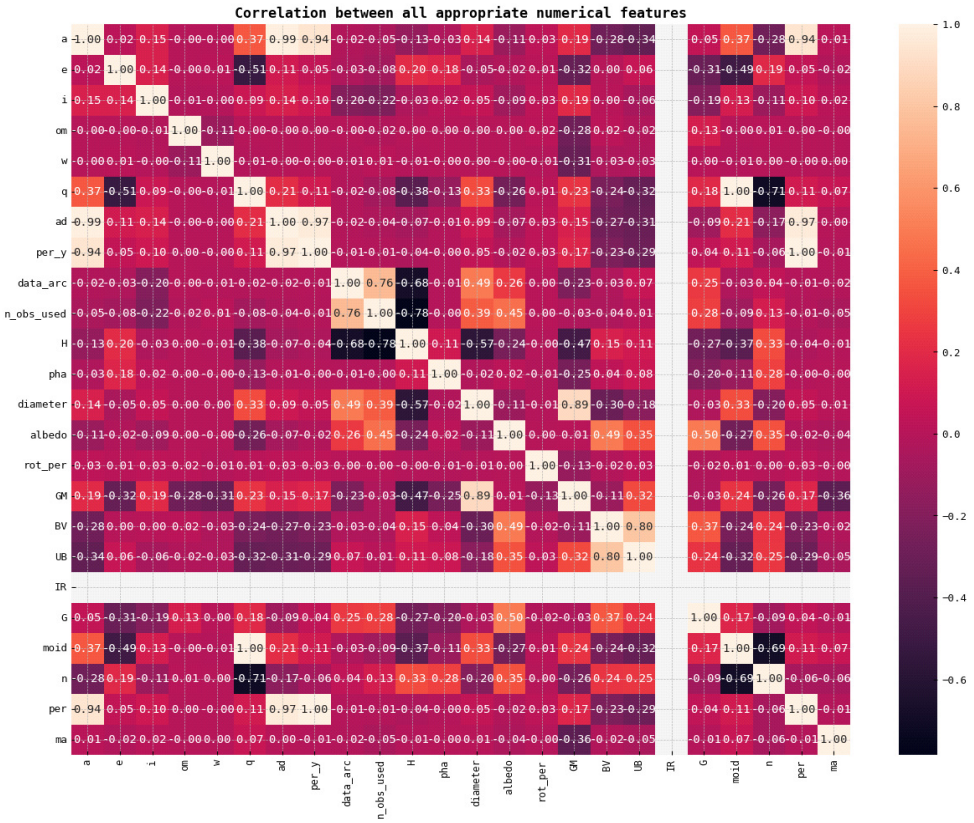

Figure 3 – Correlation heatmap for all numerical and categorical features

After this step, we began analyzing other numerical and categorical attributes by plotting the correlation heatmap for all features in order to find the most appropriate features for diameter prediction, depicted in Figure 3. To plot the correlation heatmap we used Pandas built-in .corr method which can use Pearson(standard) correlation coefficient [12], Kendall Tau correlation coefficient [15], and Spearman rank correlation [8]. According to this data, 11 out of 24 columns have a pairwise correlation between -0.1 and 0.1, which demonstrates their insignificance for the forecasting process, thus they were dropped alongside. Those fields are *e (Eccentricity)*, *ad (Aphelion distance)*, *i (Inclination)*, *om (Longitude of the ascending node)*, *w (Argument of perihelion)*, *per_y (Orbital period)*, *rot_per (Rotation period)*, *G (Magnitude slope parameter)*, *per (Orbital period)*, *pha (Potentially Hazardous Asteroid)* and *ma (Mean anomaly)*. Low correlation values for the aforementioned attributes may be related to the nature of asteroid formation or events that occurred before measurements were taken. Nonetheless, those features were excluded from further testing, improving overall prediction accuracy. Furthermore, we dropped all items with unidentified diameters, since in this research we are trying to train models with predefined desired prediction parameters for testing, resulting in dropping 702100 rows from the initial dataset, leaving 137636 items in the final iteration of a dataset.

Table 2 – Number of asteroids with missing values for given attributes

Column	***Rows with NaN value***
a	0
q	0
H	747
UB	136671
BV	136631
GM	137622
moid	0

diameter	0
albedo	1230
n	0
IR	137635

During the next phase of preprocessing, we counted all the asteroids with missing values for all remaining features. As demonstrated in Table 2., *UB (Color index U-B magnitude difference), BV (Color index B-V magnitude difference), GM (Product of the mass (M) multiplied by a gravitational constant (G)),* and *IR (Color index I-R magnitude difference)* fields have only 965, 1005, 14, and 1 non-missing values presented, correspondingly. While we may fill empty *UB*, *GM*, and *BV* cells with mean values of these columns since they have at least some amount of rows filled with data, it will negatively impact the mean square error during prediction, which is shown in Table 3. Thus, we are removing these columns from further processing and testing, alongside rows, which contain NaN values in remaining attributes, which means out of 11 columns, the resulting dataset only includes 7. After the deletion of all items with missing values, the total number of removed rows reached 1230, which is a maximum between *H* and *albedo*.

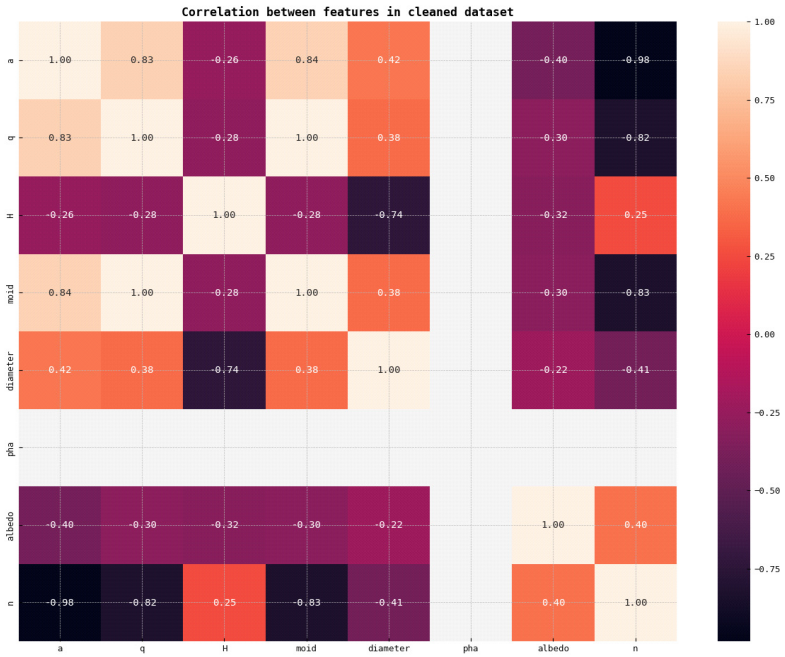

Figure 4 – Correlation heatmap for all numerical and categorical features after optimization

Outliers deletion is the final step in data standardization to achieve the most optimal accuracy rates for all the models this research uses as prediction models for the diameter of asteroids. Outliers were identified by calculating *Z-score* for every value inside the features and using list-wise deletion [10] with an absolute *Z-score* exceeding the value of 3, which equals 5333 deleted rows, and 130800 left after this step. Number 3 was taken as an arbitrary value, often used by models to find unusual entries in datasets. Z-score can be described as a statistical measurement, which depicts the connection between a value and a set of values mean [2]. Standard deviations from the mean are used to measure Z-score. Score formula:

$$Z = (x - \mu) / \sigma, \qquad (1)$$

where is the standard score, is the observed value, is the mean of the sample, is the standard deviation of the sample. In this case, the mean of the sample represents the average data on the column, while the standard deviation of the sample is the root-mean square of the difference between the given observation and the sample mean [14].

The new correlation heatmap depicted in Figure 4. shows far greater pairwise correlation values for diameter, implying our preprocessing had a significant effect on the prediction capabilities of our model.

Methods

During the training and prediction phases of this study, we were able to test several popular machine learning techniques as regressors, such as K-nearest neighbors (KNN) [4], linear regression [13], decision tree [5], random forests [6], and gradient boosting [18].

KNN: the KNN algorithm is a supervised learning classifier that utilizes proximity by producing classifications or predictions about how a particular data point will be grouped. It is non-parametric.

Linear regression: as for linear regression, and its application as the classifier, it can be characterized as a method, in which a variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to predict the value of the other variable.

Decision Tree: by constructing a decision tree, the decision tree classifier [19] develops the classification model. A test on an attribute is specified by each node in the tree, and each branch descending from that node represents one of the possible values for that property.

Random forests: as an ensemble learning technique for classification and regression, random forests build a large number of decision trees during the training phase. The class that the majority of the trees choose is the output of the random forest for classification problems. The mean or average forecast of each tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests.

Gradient boosting: this estimator allows for the optimization of any differentiable loss function and constructs an additive model in a forward stage-wise manner. A regression tree is fitted on the negative gradient of the provided loss function at each level.

Table 3 – R2 score, Root MSE (Mean Square Error), and execution time for each method (UB, GM, and BV)

Method (Regressor)	R2 score	R2 score, outliers removed	Root MSE	Root MSE, outliers removed	Execution time, in ms	Execution time, outliers removed, in ms
Linear regression	0.55	0.78	6.90	1.39	6.30	4.06
Decision tree	0.92	0.93	2.91	0.77	127.65	105.53
KNN	0.77	0.95	4.98	0.63	1357.18	72.94
Random forest	0.94	0.96	2.53	0.57	9528.88	7997.67
Gradient boosting	0.93	0.96	2.58	0.57	3057.80	2934.02

Table 4 – R2 score, Root MSE (Mean Square Error), and execution time for each method

Method (Regressor)	R2 score	Root Mean Square Error	Execution time, in ms
Linear regression	0.78	1.39	4.06
Decision tree	0.93	0.77	105.53
KNN	0.95	0.63	72.94
Random forest	0.96	0.57	7997.67
Gradient boosting	0.96	0.57	2934.02

Results and Discussion

In order to achieve a better understanding of per-model performance we conducted 3 separate sets of testing, in which predictions were made based on datasets with both *UB*, *GM*, and *BV* features removed and remained, then there was a removal of any outliers with Z-index score higher than 3. In this study, the authors use the R2 score as an indicator of accuracy. Authors can observe the difference between forecasting results of data before and after outlier removal with *UB*, *GM*, and *BV* attributes in Tables 3 and 4. Outliers had a significant impact on R2 score, which is calculated as the R2 score, which is a coefficient of determination, used as a regression score function, for some methods, such as linear regression and KNN, while other techniques only had improvement in root mean square error. Also, we have a major improvement in execution times. The most noticeable execution time inequality is represented by the difference in KNN execution times before and after outlier removal, from 1357.18 ms to 72.94 ms, which can be explained by a significant reduction in the total

number of rows. Overall, as shown in Table 3, with given initial input data, the decision tree and KNN have the best R2 score per execution time ratio, while gradient boosting and random forest both demonstrate very high R2 score and execution time, but lower root square mean error in comparison with other algorithms.

On the other hand, as Table 4. depicts, removing *UB*, *GM*, and *BV* columns, which almost fully consist of mean sample data values of the initial few items, resulted in a comparable performance, but a significantly better root mean square error indicator. Comparing all the methods in our final testing, all the methods except linear regression had a great R2 score in forecasting asteroid diameters. Linear regression, while being the least accurate one, still has the acceptable root mean square error value, and substantially lower execution time. KNN achieved the best overall performance, reaching a value of 0.95 for R2 score, which is 0.01 lower in comparison with random forest and gradient boosting, and had a reasonable execution time of 72.94, while the aforementioned random forest and gradient boosting exceed 2500 ms each.

Conclusion

This paper presented a profound description of building a model for forecasting asteroid diameters based on NASA's small body database. The main idea of the research was to identify pairwise correlations between dataset features and diameter and analyze several approaches to diameter prediction with the help of various machine learning algorithms.

With the given results, we may potentially forecast the diameters of many currently understudied asteroids and newly discovered ones. Applications to such predicted data can improve the identification of potentially hazardous asteroids, and generally enhance our understanding of the behavior of many small bodies we can not study due to technological limitations.

References

1 Alexandropoulos S.A., Kotsiantis S. and Vrahatis M. (2019) The Knowledge Engineering Review, 34, pp.1–33. https://doi.org/10.1017/S026988891800036X.

2 Altman E. (1968) The Journal of Finance, pp. 589–609.

3 Badescu. Asteroids: Prospective Energy and Material Resources. Springer Berlin, Heidelberg, 689 p.

4 Carruba V., Aljbaae S., Domingos R.C., Huaman M. and Barletta W. (2022) Celestial Mechanics and Dynamical Astronomy, 134, p. 36. https://doi.org/10.1007/s10569-022-10088-2.

5 Carruba V., Aljbaae S., Domingos R.C., Lucchini A. and Furlaneto P. (2020) Monthly Notices of the Royal Astronomical Society, 496(1), pp. 540–54. https://doi.org/10.1093/mnras/staa1463.

6 Chao H., Yue-hua M., Hai-bin Z. and Xiao-ping L. (2017) Chinese Astronomy and Astrophysics, 41(4), pp. 549–557. https://doi.org/10.1016/j.chinastron.2017.11.006.

7 Chapman C. and Morrison D. (1994) Nature, 367, pp. 33–40. https://doi.org/10.1038/367033a0.

8 Dodge, The Concise Encyclopedia of Statistics, Springer, New York, 2008, 616 p.

9 Donnison J.R. and Sugden R.A. (1984) Monthly Notices of the Royal Astronomical Society, 210(3), pp. 673–682. https://doi.org/10.1093/mnras/210.3.673.

10 Emmanuel T., Maupong T. and Mpoeleng. (2021) Journal of Big Data, 8, 140 p. https://doi.org/10.1186/s40537-021-00516-9.

11 Jet Propulsion Laboratory of California Institute of Technology, Small-Body Database Query. Retrieved May 3, 2023, from https://ssd.jpl.nasa.gov/tools/sbdb_query.html.

12 Kirch. Encyclopedia of Public Health, Springer, Dordrecht, 2008, 1600 p.

13 Kobayashi N., Oyamada Y., Mochizuki Y. and Ishikawa H., 14th IAPR International Conference on Machine Vision Applications (MVA) (Tokyo, 18-22 May 2015), p. 551–554.

14 Kotz S. and Johnson N. L. (1992) Breakthroughs in Statistics: Methodology and Distribution, Springer New York, NY, 600 p.

15 Lovric, International Encyclopedia of Statistical Science (Springer Berlin, Heidelberg), 79 p.

16 Sanchez P., Colombo C., Vasile M. and G. Radice. (2009) Journal of Guidance, Control and Dynamics, 32, pp. 121–142. https://doi.org/10.2514/1.36774.

17 Shang H., Wu X., Qiao D. and Huang X. (2018) Aerospace Science and Technology, 79, pp. 570–579. https://doi.org/10.1016/j.ast.2018.06.002.

18 Smirnov E.A. and Markov A.B. (2017) Monthly Notices of the Royal Astronomical Society, 469(2), pp. 2024–2031.https://doi.org/10.1093/mnras/stx999.

19 Steinbach M., Kumar V. and Tan P.-N. (2006) Introduction to Data Mining, Addison Wesley, Pearson, 165 p.

20 Wang, Y. (2023). Highlights in Science, Engineering and Technology, 39, pp. 201–208. https://doi.org/10.54097/hset.v39i.6527.

Information about authors

Duisek Bermagambet Erikuly (corresponding author)
Master student, School of Information Technology and Engineering, Kazakh-British Technical University, 59, Tole bi street, Almaty, 050000, Kazakhstan.
ORCID ID: 0009-0007-3508-8718
E-mail: be_duisek@kbtu.kz

Sarsembin Dauren Diyasovich
Master student, School of Information Technology and Engineering, Kazakh-British Technical University, 59, Tole bi street, Almaty, 050000, Kazakhstan.
ORCID ID: 0009-0008-7229-2985
E-mail: da_sarsembin@kbtu.kz

Abdurazak Kuanyshbek Abdurazakovich
Master student, School of Information Technology and Engineering, Kazakh-British Technical University, 59, Tole bi street, Almaty, 050000, Kazakhstan.
ORCID ID: 0000-0001-5743-5572
E-mail: ku_abdurazak@kbtu.kz

Авторлар туралы мәліметтер

Дуйсек Бермагамбет Ерикулы (корреспонденция авторы)
Магистрант, Ақпараттық технологиялар және инженерия мектебі, Қазақстан-Британ техникалық университеті, Төле би көш., 59, 050000, Алматы қ., Қазақстан.
ORCID ID: 0009-0007-3508-8718
E-mail: be_duisek@kbtu.kz

Сарсембин Даурен Диясович
Магистрант, Ақпараттық технологиялар және инженерия мектебі, Қазақстан-Британ техникалық университеті, Төле би көш., 59, 050000, Алматы қ., Қазақстан.
ORCID ID: 0009-0008-7229-2985
E-mail: da_sarsembin@kbtu.kz

Абдуразак Куанышбек Абдуразакович
Магистрант, Ақпараттық технологиялар және инженерия мектебі, Қазақстан-Британ техникалық университеті, Төле би көш., 59, 050000, Алматы қ., Қазақстан.
ORCID ID: 0000-0001-5743-5572
E-mail: ku_abdurazak@kbtu.kz

Информация об авторах

Дуйсек Бермагамбет Ерикулы (автор для корреспонденции)
Магистрант, Школа информационных технологий и инженерии, Казахстанско-Британский технический университет, ул. Толе би, 59, 050000, г. Алматы, Казахстан
ORCID ID: 0009-0007-3508-8718
E-mail: be_duisek@kbtu.kz

Сарсембин Даурен Диясович
Магистрант, Школа информационных технологий и инженерии, Казахстанско-Британский технический университет, ул. Толе би, 59, 050000, г. Алматы, Казахстан.

ORCID ID: 0009-0008-7229-2985
E-mail: da_sarsembin@kbtu.kz

Абдуразак Куанышбек Абдуразакович
Магистрант, Школа информационных технологий и инженерии, Казахстанско-Британский технический университет, ул. Толе би, 59, 050000, г. Алматы, Казахстан
ORCID ID: 0000-0001-5743-5572
E-mail: ku_abdurazak@kbtu.kz

Дүйсек Б.Е.*, Сарсембин Д.Д., Абдуразак К.А.

Қазақстан-Британ техникалық университеті, 050000, Алматы қ., Қазақстан
*E-mail: be_duisek@kbtu.kz

КІШІ АСПАН ДЕНЕЛЕРІ ТУРАЛЫ NASA ДЕРЕККӨРЫ НЕГІЗІНДЕ АСТЕРОИДТАРДЫҢ ДИАМЕТРІН БОЛЖАУ ҮШІН ӘРТҮРЛІ МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН САЛЫСТЫРУ ЖӘНЕ ТАЛДАУ

Аңдатпа. NASA кіші аспан денелерінің дерекқорын Jet Propulsion Laboratory ұсынады және ол астероидтар мен кометалар туралы жиналған ақпаратты, оларды бақылау және анықтау үшін қол жетімді параметрлерді, соның ішінде физикалық параметрлерді, сондай-ақ олардың жіктелуі, бақылау саны мен ұзақтығы туралы деректерді қамтиды. Бұл аспан денелерінің басым көпшілігінің қасиеттері толық сипатталмаған, бұл олардың мінез-құлқын және ғарыштағы басқа объектілермен, соның ішінде қолдан жасалған заттармен өзара әрекеттесуін болжауды қиындатады. Бұл зерттеу астероидтарды зерттеу мәселелерінің белгілі бір бөлігін NASA дерекқорынан алынған ақпарат пен бастапқы көзден өңделген деректерді пайдалана отырып, машиналық оқыту әдістерінің нәтижелері негізінде астероидтардың диаметрінің болжамын табу арқылы шешуді ұсынады. Бұл жұмыста осындай болжау модельдерін жүзеге асыру үшін ең жиі қолданылатын KNN, linear regression, random forest, decision tree және gradient boosting сияқты алгоритмдер таңдалды. Пайдаланылған машиналық оқыту алгоритмдері диаметрді болжау дәлдігінің, жұмыс жылдамдығының және орташа квадраттық қателік көрсеткіштерінің нәтижелері бойынша бағаланды. Зерттеу астероидтардың берілген көрсеткішін болжаудың ең оңтайлы тәсілін таңдауға көмектеседі, модельдің ең жақсы көрсеткіштеріне қол жеткізу үшін деректерді алдын ала өңдеу процесін сипаттайды және осы аспан денелерінің қасиеттері арасындағы корреляцияны талдайды.

Тірек сөздер: машиналық оқыту, астероид, болжау моделі, KNN, linear regression, random forest, decision tree, gradient boosting.

Дүйсек Б.Е.*, Сарсембин Д.Д., Абдуразак К.А.

Казахстанско-Британский технический университет, 050000, г. Алматы, Казахстан
*E-mail: be_duisek@kbtu.kz

СРАВНЕНИЕ И АНАЛИЗ РАЗЛИЧНЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ПРЕДСКАЗАНИЯХ ДИАМЕТРОВ АСТЕРОИДОВ НА ОСНОВЕ БАЗЫ ДАННЫХ МАЛЫХ НЕБЕСНЫХ ТЕЛ NASA

Аннотация. База данных малых небесных тел NASA предоставляется Jet Propulsion Laboratory и представляет собой собранную информацию об астероидах и кометах, описывая их доступные для наблюдения и определения параметры, в том числе физические, также их классификацию и данные по количеству и длительности наблюдений. Множество этих небесных тех имеют неполное описание их свойств, что делает затруднительным предсказание их поведения и потенциальное взаимодействие с другими объектами в космосе, в том числе и рукотворными. Данное исследование предлагает решение определенной части проблем по исследованию астероидов путем нахождения предсказания диаметра астероидов, основываясь на информации из базы данных NASA и результатах работы методов машинного обучения по обработанным данным из изначального источника. Для этой работы

были выбраны некоторые из наиболее часто используемых алгоритмов для реализации подобных моделей предсказания, такие как: KNN, linear regression, random forest, decision tree и gradient boosting. Использованные алгоритмы машинного обучения были оценены по результатам точности предсказании диаметра, скорости работы и показателям среднеквадратичных ошибок. Исследование поможет выбрать наиболее оптимальный подход для предсказания данного показателя астероидов, опишет процесс предварительной обработки данных для достижения лучших показателей модели и проанализирует корреляции между свойствами этих небесных тел.

Ключевые слова: машинное обучение, астероид, модель предсказания, линейная регрессия, случайный лес, дерево решений, повышение градиента.