Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 6, No. 2, pp. 62-69, 2025

# Prediction of Asteroid Hazard Distance Through Earth’s Orbit Using K-Nairest Neighbor Method

Syahrul Firdaus[1*], Wina Witanti[2], Melina[3], Asep Id Hadiana[4]

*[1,2,3,4]Department of Informatics, Faculty of Science and Informatics, Universitas Jenderal Achmad Yani, Cimahi, Indonesia*

**Corresponding author email: Syahrul.firdaus@student.unjani.ac.id*

**Abstract**

The National Aeronautics and Space Administration (NASA) is the U.S. government agency that is responsible for space program. NASA observes objects in space, including asteroids. Asteroids are small, rocky objects that orbit the sun with irregular shapes and are also called planetoids. The Government agencies observe space objects including asteroids. In terms of the infinite number of objects in space that will cross Earth's orbit, prediction is needed to determine the danger and its level when they are crossing Earth's orbit. Prediction is a process to know what will happen in the future which is aimed to find out the approximate asteroids that will cross the earth in the future. In this study, data mining classification techniques and the K-Nearest Neighbor algorithm are used to create a prediction system for the threat of asteroids while crossing the earth. Classification is a grouping by classifying items into designated class labels, building a classification model from the data set, building a model that is used to predict future data. To determine the distance of the asteroid's threat throughout the earth, data mining classification techniques and the K-Nearest Neighbor algorithm are used. The results are 57.71% accuracy, 54.89% precision, 81.42% recall, and 47.45% missclassification rate.

*Keywords:* Data Mining, Classification, K-Nearest Neighbor, Nasa, Asteroids, Prediction.

## 1. Introduction

Hundreds of tonnes of asteroid debris fall on Earth each day, most in the form of micrometeorites and sand-sized particles, some of which reach the ground. Asteroids are small rocky planets. Asteroids are equivalent to planetesimals, which are the earth's components of the planet. Asteroids represent the stage of a rocky road to planet formation. Most survived from the original protoplanetary disk destined to become a planet but the small, undocumented bodies of the terrestrial planet-forming asteroid regions are destined to wander into the sun of the system at the whim of the larger dynamic bodies (Asphaug, 2009). Prediction is a process for estimating a variable in the future. Prediction is divided into 3 parts, namely long term, medium term and short term. Short-term predictions are predictions made by paying attention to data patterns, and require a short period of time for changes based on the factors that make up the data pattern, while medium and long-term predictions are used for strategic planning. Medium-term forecasts help prepare for expansion and forecast demand. Long-term prediction serves to ensure long-term availability. Prediction is usually used to find information from a large amount of data, so data mining is needed. One method that can make predictions is the K-Neirest Neighbor method, which classifies objects based on the learning data closest to the object (Budianto, 2024). The formulation of the problem in this study is how to accurately predict the distance of the threat of asteroid hazards that will cross Earth's orbit in the future using the K-Neirest Neighbor method.

### 1.1 Objectives

The research aims to apply the K-Nearest Neighbor method using classification techniques and the K-Nearest neighbor method to predict the threat of asteroids crossing the earth's orbit for the future periode accurately.

## 2. Literature Review

Some authors have conducted the research about the K-Nearest Neighbor method is a study by Li Yu Hu and other. The focus of their work is used K-Neirest Neighbor classification based on the measurement of the distance between the test data and the training data. It is assumed that the selected distance function can affect the classification accuracy. The results of the study show that the K-Nearest Neighbor classifier is the best (Hu et al., 2016). A study by Tamil Nadu who applied data mining to predict Hazardous Asteroid Classification through Various Machine Learning Techniques. This paper proposes classifying asteroid threats using machine learning techniques and using an algorithm with the RBF kernel approach. From the results of his research obtained an accuracy of more than 90% (Si, 2020). A study by Kurt Varmuza. This paper about KNN Classification-Evaluated By Repeated Double Cross Validation: Recognition Of Minerals Relevant for Comet Dust. In this study, Repeated Multiple Cross was applied by optimizing and evaluating the empirical multivariate calibration model with an evaluation strategy adapted to the K-Nearest Neighbor classification. Repeated Multiple Cross-Validation principles were applied to classify 17 mineral groups associated with cometary dust particle composition characterized by peak heights at 20 selected masses in the time-of-flight secondary ion mass spectrum (TOF-SIMS). The results showed the predictability of 15 mineral levels was 95% N, 2 types 75 and 85% (Varmuza et al., 2014). A study by Jayasri on Big data analytics in the health sector with data mining and classification techniques. The conclusion of this study is the purpose of this task is to more accurately predict the development of diabetes datasets in order to find the optimal solution for the patient. To perform this analysis, MapReduce platform is used in addition to the proposed hierarchy algorithms such as the hierarchical decision attention network, AA and outlier-based multiclass classification. With this diabetes patients are classified and there insulin levels are determined. From the comparison chart it is clear that the proposed hierarchical algorithm shows improved performance. The confusion matrix performance indicators used are precision, recall and F-score, which is 0.99 executed on data set. In the future, this algorithm will be allowed to cloudcomputing structure for better access and real performance time (Jayasri and Aruna, 2022). A study by hawkins' on Spacecraft Guidance Algorithms for Asteroid Intercept and Rendezvous Missions. Conclusions from the research the feedback guidance algorithms described in this paper are applicable to a wide variety of asteroid missions. This paper lays a solid foundation for further research on asteroid interception missions. This includes topics such as waypoint guidance and strategies for dealing with the irregular gravitational fields of larger asteroids (Hawkins, Guo, and Wie, 2012). Furthermore, research that has a relationship with this research is from Malti Bansal's research on A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning Research Conclusions. In this white paper, all relevant questions that may arise during the research of these algorithms are addressed in terms of their origins, definitions, methods of execution, real-time applications related to sufficient new evidence, and subsequent benefits. This paper highlights and trusts the Future ML Algorithms and Future Scope of Artificial Intelligence, and Their Role in Automation and Overall Development in Human Aspects (Bansal, Goyal, and Choudhary, 2022).

## 3. Methodology

In this paper, there are the eight steps of the methodology process.
Step 1: Data Collection, collect data to carry out the data analysis process.
Step 2: Selection Data, at this stage perform data selection. The selected data is data on the prediction of the distance of the threat of asteroid hazards across the earth's orbit. Selected data used in the data mining process are stored separately in a file, separate from operational data.
Step 3: Data Preprocessing, From the data that has been selected, it is necessary to clean data such as double data, check for inconsistent data, and correct the data.
Step 4: The Transformation stage is the stage of changing the data selected so that the data is suitable for the data mining process.
Step 5: Data Mining, Data Mining used classification technique. Classification is one form of data analysis The resulting model Describes important data classes. Classification predicts the category of the class label. Classification is Find the model or feature Explain or differentiate the term Data class with the aim of being able to estimate the object class if the label is not known (Sari, 2020). Predictions have similarities such as classification and forecasting, but predictions find new value in the future. future by observing past data (Depan 2021).
Step 6: K-Nearest Neighbor, method are algorithms aimed at classifying new data objects. The process of classifying new data objects is trained using tributes and training sample data. New object classification modeling is based only on memory. This method works by searching for the $k$ training data objects that are closest to the specified test data and selecting the class with the most votes.
Step 7: K-Fold Cross Validation, this technique is primarily used to make model predictions and estimate the accuracy of the predictive model will be when run in practice. One method of cross-validation is the k-fold cross-validation. It divides the data into k parts of a dataset of the same size (Xiong and Yao 2021).
Step 8: Confusion Matrix, Confusion matrix is a commonly I used the visualization tool for supervised learning. Each column in the matrix is an example of a predictive class, but each row represents an event in the actual class.

## 4. Data Collection

The data obtained before the knowledge discovery process in the database was 90,837 but after the data was processed through the knowledge discovery process in the database it became 3,500 data. The processed data will later be transformed so that the system can process it properly. The following Table 1 contains examples of data that have passed the knowledge discovery process in the database.

**Table 1**: Pure Data Acquisition

| No | Name | Diameter Min | Diameter Max | Relative Velocity | Miss Distance | Orbiting Body | Sentry Object | Absolute Magnitude | Hazardous |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 162635 (2000 SS164) | 1.198271 | 2.679415 | 13569.25 | 54839744 | Earth | FALSE | 16.73 | FALSE |
| 2 | 277475 (2005 WK4) | 0.2658 | 0.594347 | 73588.73 | 61438127 | Earth | FALSE | 20 | TRUE |
| 3 | 512244 (2015 YE18) | 0.72203 | 1.614507 | 114258.7 | 49798725 | Earth | FALSE | 17.83 | FALSE |
| 4 | (2012 BV13) | 0.096506 | 0.215794 | 24764.3 | 25434973 | Earth | FALSE | 22.2 | FALSE |
| 5 | (2014 GE35) | 0.255009 | 0.570217 | 42737.73 | 46275567 | Earth | FALSE | 20.09 | TRUE |
|  | …. | …. | …… | ……. | …… | ……. | …… | …… | …… |
| 3.500 | (2021 TW7) | 0.039862 | 0.089133 | 27024.46 | 59772131 | Earth | FALSE | 24.12 | FALSE |

## 5. Results and Discussion

### 5.1. Result

From the results of data acquisition as much as 90,837 data, a knowledge discovery process in the database has been carried out, the data that have been selected and will be used amounted to 3,500 data, after that the data will be executed a data transformation process, which changes the Prediction Data for the Distance of Asteroid Hazards Passing Earth Orbit to make it more easy to use using the K-Nearest Neighbor method. The K-nearest neighbor method is an algorithm aimed at classifying new data objects. The process of classifying new data objects is trained using attribute and training example data. The new object classification modeling is, which is based solely on memory. This method works by finding k training data objects that are closest to the specified test data and selecting the class with the most votes and then evaluating it using K-Fold Cross Validation. K-Fold Cross Validation Extended K-Fold Cross Validation is an improvement over the traditional *k*-validation cross-validation to evaluate the exploratory predictions of its model strength. The first sample is sorted according to material property values and then divided evenly into k subsets. This technique is primarily used to make model predictions and estimate the accuracy of the predictive model will be when run in practice. One method of cross-validation is the k-fold cross-validation the data into k parts of data sets of the same size.

### 5.2 Discussion

#### 1) Data Mining

Data mining is a series of processes for acquiring knowledge or patterns based on a combination of data. Data mining solves the problem by analyzing the data that is already contained in the database. Data mining, also known as Knowledge Discovery. In a database, is an activity that involves collecting and found in historical data regular and interaction patterns in large datasets. Database is programming language to access, edit and manipulate data, can use the Standard Query Language (SQL) programming language (Melina et al., 2020). Output results based on data mining can be used to improve future decision making. The advent of data mining is based on the amount of data stored in the database. The data is getting bigger and bigger. Data mining seeks to overcome this phenomenon by discovering new and useful information. Expression data mining is sometimes referred to as knowledge discovery (Nathan and Scobell, 2012).

### 2) K-Nearest Neighbor

K-nearest neighbors are the number of nearest neighbors used as points to classify new data or objects. It is recommended to use odd numbers when determining the number of $k$ values (Cheng, 2021).

$$dist\left(x_{1,}x_2\right)=\sqrt{\sum_{i=1}^{n}(x_{1i}-x_{2i})^2} \tag{1}$$

Where:
$dist\ (x_1, x_2)$ : distance between object $x_{1i}$ and $x_{1i}$
$x_{1i}$ : test data
$x_{2i}$ : training data
$n$ : data dimension

### 3) Confusion Matrix

The confusion matrix contains real and expected information about the classification system. The confusion matrix is described by specifying the number of correctly classified test data and the number of classified test data. Based on the above confusion matrix (Solichin, 2019):

a. True Positives (TP) is the number of positive data records classified as positive values.
b. False Positives (FP) is the number of negative data records that are classified as negative values.
c. False Negatives (FN) is the number of positive data records classified as positive values.
d. True Negative (TN) is the number of negative data records classified as negative.

The value generated through the confusion matrix method is in the form of an evaluation as follows:

a. Accuracy, percentage of the number of data records correctly classified by the algorithm.

$$\frac{(TP+TN)}{amount\ of\ data}=Accuracy$$

b. Misclassification rate, percentage of the number of data records classified incorrectly predicted by the algorithm.

$$\frac{(FP+FN)}{amount\ of\ data}=Missclassification\ Rate$$

c. Recall, Percentages represent the success rate of the model in finding information.

$$\frac{(TP)}{TP+FN}=Recall$$

d. Precision, Percentages represent the level of accuracy between some of the requested data and the predictions provided by the model.

$$\frac{(TP)}{TP+FP}=Precision$$

### 5.1 Numerical Results

The following is an example of calculating the K-Nearest Neighbor algorithm using NASA data samples regarding the threat of asteroids crossing Earth's orbit in Table 2.

**Table 2:** Example Of K-Nearest Neighbor Training Data

| No | Name | Diameter Min | Diameter_ Max | Relative Velocity | Miss Distance | Orbiting Body | Sentry Object | Hazardous |
|---|---|---|---|---|---|---|---|---|
| 1 | (2016 AP164) | 1272 | 2845 | 11402 | 54551867 | Earth | FALSE | 0 |
| 2 | 418849 (2008 WM64) | 2141 | 4787 | 70111 | 29409957 | Earth | FALSE | 1 |
| 3 | (2005 ET70) | 1926 | 4306 | 87520 | 22879679 | Earth | FALSE | 0 |
| 4 | (2013 WR67) | 1058 | 2366 | 74163 | 10124669 | Earth | FALSE | 1 |

Table 3 shown a test data table, the table will be compared one by one with the training data, from these results will issue the Euclidean distance value.

**Table 3:** Example Of K-Nearest Neighbor Test Data

| No | Name | Diameter min | Diameter max | Relative velocity | Miss distance | Orbiting body | Sentry object |
|---|---|---|---|---|---|---|---|
| 1 | kepler | 3224 | 4363 | 5674 | 3467 | Earth | 0 |

In determining the classification results using the K-Neirest Neighbor method, it is necessary to go through several steps that must be passed, the following are the steps that must be followed to be able to determine the classification results along with examples in Table 4.

**Table 4:** Euclidean Distance Result

| No | Name | Diameter min | Diameter max | Relative velocity | Miss distance | Orbiting _body | Ranking | Euclidean | Label |
|---|---|---|---|---|---|---|---|---|---|
| 1 | (2016 AP164) | 1272 | 2845 | 11402 | 54551867 | Earth | 4 | 54548400.35679 | 0 |
| 2 | 418849 (2008 WM64) | 2141 | 4787 | 70111 | 29409957 | Earth | 3 | 29406560.621734 | 1 |
| 3 | (2005 ET70) | 1926 | 4306 | 87520 | 22879679 | Earth | 2 | 22876358.449821 | 0 |
| 4 | (2013 WR67) | 1058 | 2366 | 74163 | 10124669 | Earth | 1 | 10121434.154678 | 1 |

Table 5 shown, sort the data on the Euclidean distance from the smallest to the largest, in order to find out the value of K in the next step.

**Table 5:** Ascending Result Euclidean Distance

| No | Name | Diameter min | Diameter max | Relative velocity | Missdistance | Orbiting body | Ranking | Euclidean | Label |
|---|---|---|---|---|---|---|---|---|---|
| 1 | (2013 WR67) | 1058 | 2366 | 74163 | 10124669 | Earth | 1 | 10121434.154678 | 1 |
| 2 | (2005 ET70) | 1926 | 4306 | 87520 | 22879679 | Earth | 2 | 22876358.449821 | 0 |
| 3 | 418849 (2008 WM64) | 2141 | 4787 | 70111 | 29409957 | Earth | 3 | 29406560.621734 | 1 |
| 4 | (2016 AP164) | 1272 | 2845 | 11402 | 54551867 | Earth | 4 | 54548400.35679 | 0 |

The last step in Table 6 is to determine the predetermined K value. at a predetermined K value of 3. then the system will display 3 data from 4 data owned according to the data that has been sorted from the smallest to the largest.

**Table 6:** Determination of K. Value

| No | Name | Diameter min | Diameter max | Relative velocity | Missdistance | Orbiting body | Ranking | Euclidean | Label |
|---|---|---|---|---|---|---|---|---|---|
| 1 | (2013 WR67) | 1058 | 2366 | 74163 | 10124669 | Earth | 1 | 10121434.154678 | 1 |
| 2 | (2005 ET70) | 1926 | 4306 | 87520 | 22879679 | Earth | 2 | 22876358.449821 | 0 |
| 3 | 418849 (2008 WM64) | 2141 | 4787 | 70111 | 29409957 | Earth | 3 | 29406560.621734 | 1 |

## 5.2 Graphical Results

The following is a graphic that illustrates the results for each K value used in K-Fold Cross Validation.

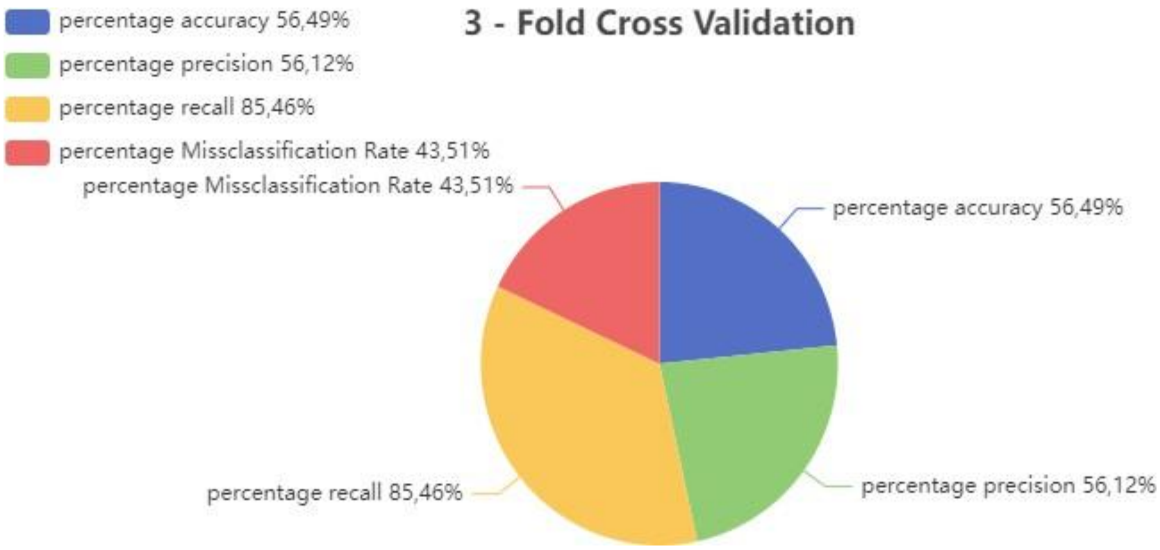

**Figure 1:** 3-Fold Cross Validation Graph Results

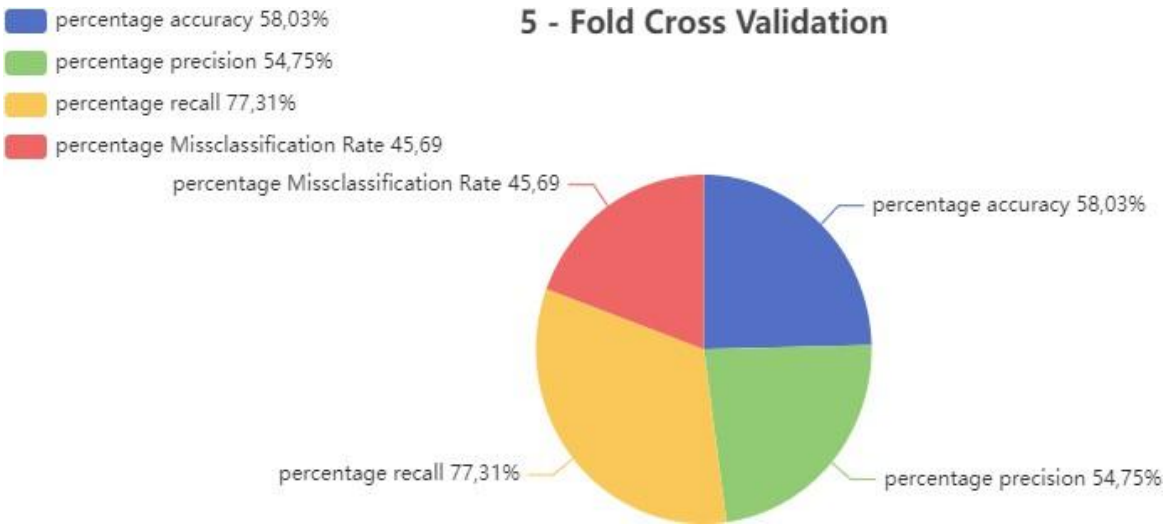

**Figure 2:** 5-Fold Cross Validation Graph Results

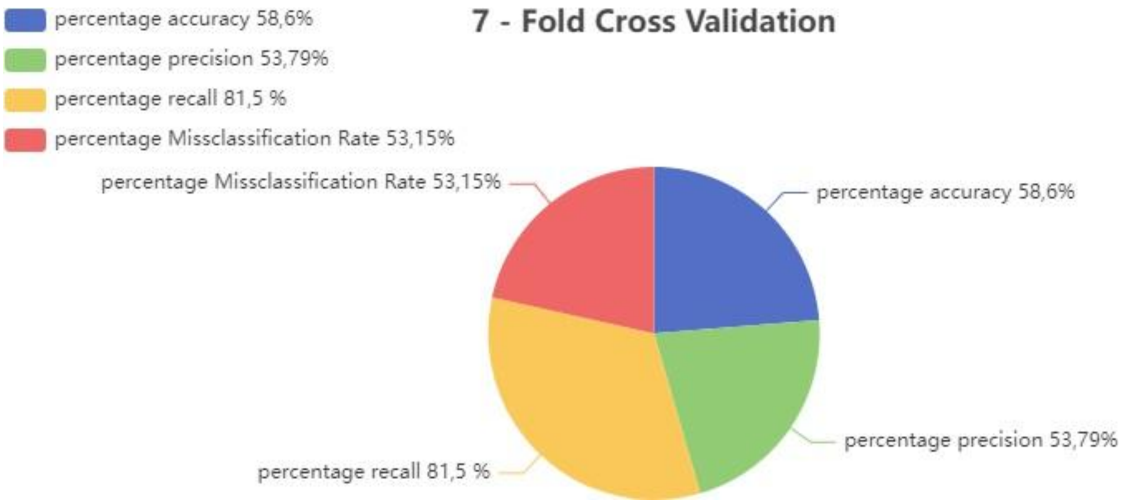

**Figure 3:** 7-Fold Cross Validation Graph Results

The average results of the K-Fold Cross Validation values are 57.71% accuracy, 54.89 precision, 81.42% recall and 47.45% error rate.

### 5.3 Validation

Data validation by K division cross-validation. Predicted data on the distance of a threat from an asteroid across the Earth's orbit is 3500 data, which is divided into training and test data and processed in multiple experiments: 3-fold cross-validation, 5-fold cross-validation. And 7-fold cross-validation. The K-Fold Cross Validation Score results perform an average percentage calculation of accuracy, search, accuracy, and degree of classification error. The results of the K-Fold Cross verification are as follows (Mahesh, et al. 2022).

Below in table 7 are the results of the 3, 5, 7 K-Fold Cross Validation values. In these results, an average will be sought to obtain an accurate predictive value.

**Table 7**. K-Fold Cross Validation Calculation

| Fold | Accuracy | Precision | Recall | Average Missclassification Rate |
|---|---|---|---|---|
| 3 | 56.49% | 56.12% | 85.46% | 43.51% |
| 5 | 58.03% | 54.75% | 77.31% | 45.69% |
| 7 | 58.6% | 53.79% | 81.5% | 53.15% |

In the table 8 below is the final result to find an accurate prediction.

**Table 8:** K-Fold Cross Validation Final Calculation

| No | Average accuracy | Average precision | Average recall | Average Missclassification Rate |
|---|---|---|---|---|
| 1 | 57.71% | 54.89% | 81.42% | 47.45% |

## 6. Conclusion

In outer space there are an infinite number of objects, one of which is an asteroid object. Asteroids are small rocky planetary bodies. Asteroids are equated with planetesimals, which are the terrestrial building blocks of planets. Asteroids are small rocky planetary bodies. Asteroids are equated with planetesimals, which are the terrestrial building blocks of planets. Prediction is usually used to find information from a large amount of data, so data mining is needed. One method that can make predictions is the K-Neirest Neighbor method, which classifies objects based on the learning data closest to the object. The purpose of this study is to be able to create a system that predicts the distance of the threat of asteroid hazards crossing the earth's orbit using the K-Nearest Neighbor method using classification techniques and the K-Nearest neighbor method so that it can predict the distance of the threat of asteroid hazards that will cross Earth's orbit for the future period accurately. The known conclusion is that it can be ascertained that by applying data mining method using K-nearest neighbor algorithm on 3500 data, accuracy, precision, recall, and classification errors are obtained with a fairly good percentage obtained, namely 57.71% accuracy, 54 precision, 89, recall 81.42%, and misclassification rate 47.45%. From these results, it can be said that K-Nearest Neighbor is accurate enough to analyze predictive data on the threat of steroids orbiting the earth.

## References

Asphaug, Erik. (2009). Growth and Evolution of Asteroids. *Annual Review of Earth and Planetary Sciences 37*, 413–48.

Bansal, Malti, Apoorva Goyal, and Apoorva Choudhary. (2022). A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning. *Decision Analytics Journal 3* : 100071. https://doi.org/10.1016/j.dajour.2022.100071.

Budianto, H., Darmawan, E., Krisdiawan, R. A., & Maulana, Y. (2024). Prediction Analysis of Buled Bread Production Using Holt's Exponential Smoothing. *International Journal Administration, Business & Organization*, *5*(4), 52-61.

Cheng, G., & Yan, Y. (2021). Sociodemographic, health-related, and social predictors of subjective well-being among Chinese oldest-old: a national community-based cohort study. *BMC geriatrics*, *21*, 1-13.

Depan, Halaman Sampul.(2021). "IMPLEMENTASI METODE K-NEAREST NEIGHBOR."

Hawkins, Matt, Yanning Guo, and Bong Wie. (2012). Spacecraft Guidance Algorithms for Asteroid Intercept and Rendezvous Missions. *International Journal of Aeronautical and Space Sciences 13(2)*, 154–69.

Hu, Li Yu, Min Wei Huang, Shih Wen Ke, and Chih Fong Tsai. (2016). The Distance Function Effect on K-Nearest Neighbor Classification for Medical Datasets. *SpringerPlus 5(1).*

Jayasri, N. P., and R. Aruna. (2022). Big Data Analytics in Health Care by Data Mining and Classification Techniques. *ICT Express 8(2),* 250–57. https://doi.org/10.1016/j.icte.2021.07.001.

Mahesh, T. R., Dhilip Kumar, V., Vinoth Kumar, V., Asghar, J., Geman, O., Arulkumaran, G., & Arun, N. (2022). AdaBoost ensemble methods using K‐fold cross validation for survivability with the early detection of heart disease. *Computational Intelligence and Neuroscience*, *2022*(1), 9005278.

Melina, Melina et al. (2020). Design and Implementation of Multi Knowledge Base Expert System Using the SQL Inference Mechanism for Herbal Medicine. *Journal of Physics Conference Series 1477*, 22007.

Minerals Relevant for Comet Dust. *Chemometrics and Intelligent Laboratory Systems 138*, 64–71. http://dx.doi.org/10.1016/j.chemolab.2014.07.011.

Nathan, Andrew J., and Andrew Scobell. (2012). Model Algoritma K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa. *Foreign Affairs 91(5)*, 1–9.

Sari, M., & Al Maki, W. F. (2023, December). Improving K-Nearest Neighbor Performance in Footwear Classification Using Leave One Out Cross Validation. In *2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)* (pp. 55-60). IEEE.

Si, Anish. (2020). Hazardous Asteroid Classification through Various Machine Learning Techniques. 5388–90.

Solichin, A. (2019, September). Comparison of decision tree, Naïve Bayes and K-nearest neighbors for predicting thesis graduation. In *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 217-222). IEEE.

Varmuza, Kurt et al. (2014). KNN Classification - Evaluated by Repeated Double Cross Validation: Recognition of

Xiong, Lei, and Ye Yao. (2021). Study on an Adaptive Thermal Comfort Model with K-Nearest-Neighbors (KNN) Algorithm. *Building and Environment 202*, 108026. https://doi.org/10.1016/j.buildenv.2021.108026.