Arms Dynamics Classification System for South East Asia Region by Using K-NN Algorithm

Zul Indra\(^1\), Azhari Setiawan\(^2\,\ast\), Andhik Beni Saputra\(^3\), Yessi Jusman\(^4\)

\(^1\)Department of Informatics Engineering, Faculty of Engineering, Universitas Abdurrab, Pekanbaru, 28291 Riau, Indonesia
\(^2\)Department of International Relations, Faculty of Social and Political Sciences, Universitas Abdurrab, Pekanbaru, 28291 Riau, Indonesia
\(^3\)Department of Government Studies, Faculty of Social and Political Sciences, Universitas Abdurrab, Pekanbaru, 28291 Riau, Indonesia
\(^4\)Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta, Bantul, Daerah Istimewa Yogyakarta 55183, Indonesia

*Corresponding author : azhari.setiawan.univrab.ac.id

Abstract. Machine learning is one of the areas of research that has received great attention in various fields of science. One of the most important applications of machine learning is its ability to perform data analysis. Machine learning plays an important role in finding patterns from data and providing predictions based on existing data. This research aims to apply the concept of machine learning to analyze the arms trade for the Southeast Asia region, which is currently rarely discussed. The research tries to make a prediction system for the arms dynamics based on the political and economic strength of a country. There are three main data sources in this study, namely wordbank.org for the economic strength variable, polity5 for political variable and sipri.org for the arms dynamics variable. The algorithm chosen is the K-NN algorithm which is proven to be one of the best and simplest algorithms in handling data classification. Based on the research that has been carried out, it can be concluded that this study succeeded in developing a data classification system that can be used to predict arms trade with a fairly high degree of accuracy.

1. Introduction
Nowadays, in the era of information technology, data is a very valuable asset that has an important role in our lives. Data is able to provide a wealth of information needed by modern humans to solve various problems. For example, data can be used to predict weather, stock prices, congestion points, and the spread of disease to predict sports events. One method that is closely related to data processing in solving human problems is machine learning. Machine Learning is the study of algorithms and statistical models of computer systems which was first mentioned by Arthur Samuel in 1959 [1], [2]. Machine learning has become one of the most important parts of scientific studies since it has been used for many other studies besides computer science and information technology.

Machine learning is part of artificial intelligence used to effectively perform certain tasks without using explicit instructions, relying only on patterns and inference [3]. It is closely related to computational statistics, which focuses on making predictions based on pattern recognition such as email filtering and computer vision [4]. However, pattern is essential for social sciences, especially
political science and international relations. Therefore, based on research conducted by Arthur Samuel, it was found that machine learning would be very useful for all kinds of strategic matters such as economics, politics, and socio-cultural analysis. It can be applied to aggregate information from large numbers of cases and to use the laws of probability generalize well beyond those cases [5]. The implementation of a machine learning algorithm not only can generate a description of the association of a variable or phenomenon but also the probability that the variable or phenomenon will occur. Thus the community can get better understanding about power relations, the behavior of international actors, and how the international system works regarding the study of international relations.

However, research on the implementation of machine learning algorithms are still less discussed in international relations studies. Over the past 10 years, most research has tried to seek inferential insights such as cause and effect, probability, and effects using various regression models [6]–[15]. Previous related studies have not applied prediction features for their analysis. Most studies try to find inferential insights such as cause and effect, probability, and effects using various regression models. So it can be concluded that prediction or forecasting methods are still less discussed. In addition, a large majority of studies focus on interaction, international organizations, conflict resolution, and international political economy, etc. The Southeast Asian region is also far from the epicenter of the discussion. Hence, the lack of implementation and discussion of machine learning for the Southeast Asia region can be identified as a research gap for this study.

This research is intended to conduct research related to the dynamics of the arms trade in Southeast Asia using machine learning algorithms in order to bridge the research gap. By using the $k$-NN algorithm, this research is expected to produce a forecasting system for the arms trade in Southeast Asia. This research tries to discover the patterns of arms trade dynamics among Southeast Asia countries which were associated with its political and economic posture. In terms of dataset, as one of the strategic regions in the world, there are several countries outside this region that also play a major role in the arms trade in this region, such as the United States, Russia, China, Japan and the Republic of Korea. Therefore, the discussion and dataset used in this study also included these countries in addition to countries in the Southeast Asia region.

2. Methodology

As previously mentioned, this study aims to develop a classification system for predicting arms trade in the Southeast Asia region. Generally, the classification algorithm mainly consists of two stages namely training and classification. The data classification process usually begins with a set of training data that is previously labeled with a set of classes such as high, medium and low. Finally, the classification algorithm will classify the test data into the class that is determined as the best class for it. The overall architecture for this study is shown in Figure 1.

![Figure 1. Research Flowchart](image-url)
### 2.1. Data Collection

The first stage in this research is data collection. This study uses 2 groups of input variables, namely political posture and economic posture in classifying the arms trade. This study then collected data for all these variables for each country in the Southeast Asia region plus several countries that had played major roles in ASEAN, such as the United States, Russia, China, Japan and South Korea. In terms of data source, this study uses three main database sources, namely Polity5, SIPRI and the World Bank. The overall data structure is presented in Table 1.

| Data Category      | Variables                                                                 | Source                      |
|-------------------|---------------------------------------------------------------------------|-----------------------------|
| Input             | Political Posture                                                         | Polity5,                     |
|                   | Political System, Regime Durability, State Fragility Index, Effectiveness   | INCSR, Center for Systemic  |
|                   | Score, Legitimacy Score, Security Effectiveness, Security Legitimacy,       | Peace                       |
|                   | Political Effectiveness, Political Legitimacy, Economic Effectiveness,      |                             |
|                   | Economic Legitimacy, Social Effectiveness and Social Legitimacy             |                             |
| Economic Posture  | GDP, GDP per Capita, GDP Growth and Inflation                             | Worldbank.org               |
| Output            | Total of Arms Export                                                      | SIPRI.org                   |
|                   | Total Number of Arms Export for Aircraft, Air defense system, Artillery,    |                             |
|                   | Engines, Missiles, Naval weapons, Sensors, Ships and Others                |                             |
|                   | Total of Arms Import                                                      | SIPRI.org                   |
|                   | Total Number of Arms Import for Aircraft, Air defense system, Artillery,    |                             |
|                   | Engines, Missiles, Naval weapons, Sensors, Ships and Others                |                             |

### 2.2. Preprocessing

The second stage in this study is data pre-processing. This stage is intended to convert variable data from numeric data to categorical data. The output variables will be converted into three categories, namely low, middle and high. The low category is intended for output variables whose value is smaller than the bottom edge. The middle category is intended for data that has between the bottom and top edges. While the high category is given to variables whose value is greater than the top edge. The entire flowchart for this pre-processing stage is illustrated in Figure 2.
2.3. Data Classification
The basic concept of this k-NN Algorithm is to select the k other nearest data ("neighbors") that surround it and then assign the test data to the most suitable neighbor [16]. The neighbor selection process is carried out by applying the Euclidian distance equation. After that, the k -NN algorithm will sort the resulting distance from the Euclidian distance equation and choose the k data with the closest distance. The class of k data with majority member data is defined as the category for the test data. Suppose there is more than one member of k sharing a class, then the accumulative distance of the data for each class. Finally, the k-NN algorithm will select the class with the smallest accumulated distance to be used as a category for the test data. The flowchart for k-NN algorithm is shown in figure 3:
3. Results and Discussions

3.1. Data Collection and Preprocessing
As explained in section 2, this research is started from the data collection and preprocessing stages. This data collection stage was carried out by observing data for 10 countries in the Southeast Asia region. In addition, data collection was also carried out on 5 other countries that are considered influential in the arms trade in this region. The dataset for this research can be seen in Table 2.

Table 2. Dataset of Arms Dynamics

| Country Name         | Year Observation | Total Data |
|----------------------|------------------|------------|
| Brunei Darussalam    | 1984-2018        | 35         |
| Cambodia             | 1960-2018        | 59         |
| China                | 1960-2018        | 59         |
| Indonesia            | 1960-2018        | 59         |
| Japan                | 1960-2018        | 59         |
| Laos                 | 1960-2018        | 59         |
| Malaysia             | 1960-2018        | 59         |
| Myanmar              | 1960-2018        | 59         |
| Philippines          | 1960-2018        | 59         |
| Republic of Korea    | 1960-2018        | 59         |
| Russian Federation   | 1960-2018        | 59         |
Based on the dataset presented in table 2, it can be seen that this study uses as many as 861 rows of data sourced from observations since 1960 to 2018. Each country has 59 data rows with the exception of Brunei Darussalam which only has 35 rows of data. This is because data observations for this country have only been carried out since the year of this country independence, precisely in 1980. Then this data is divided into two parts, namely 700 lines of training data. The remaining 161 lines of data will be used as testing data to evaluate the accuracy of the classification system. The next process is to preprocess this dataset so that the data that has been collected can be processed using the K-NN algorithm. The results of the preprocessing stage can be seen in Table 3.

| Country     | Period      | Total |
|-------------|-------------|-------|
| Singapore   | 1960-2018   | 59    |
| Thailand    | 1960-2018   | 59    |
| United States| 1960-2018 | 59    |
| Vietnam     | 1960-2018   | 59    |
| **Total**   |             | **861** |

3.2. Classification System

As mentioned in section 3, the classification process is carried out by applying the K-NN algorithm. Before the data is classified into the appropriate class, the distance between the training data and the testing data is calculated by applying the Euclidean distance equation. Each testing data will be calculated its distance to all existing training data. Hence, this research generates a total of 112700 of multiplication results which is obtained from the 700 training data and 161 testing data. An example of the results of calculating the Euclidean distance for testing data is shown in Table 4.

| ID Training Data | ID Testing Data | Euclidean Distance | Class |
|------------------|-----------------|---------------------|-------|
| 231              | 701             | 993014788854.12     | low   |
| 318              | 701             | 993014788854.05     | high  |
| 162              | 701             | 9908745851.26       | low   |
| 298              | 701             | 99014788854.028     | low   |
| 695              | 701             | 9898672708.006      | low   |
| 320              | 701             | 989261883985.13     | high  |
| 381              | 701             | 9857437672.185      | low   |
| 412              | 701             | 9819190670.05       | low   |
| 670              | 701             | 98014788854.027     | high  |
| 613              | 701             | 973014788854.72     | high  |

Based on the calculations shown in table 4, it can be seen that there are 6 data that are classified as low and 4 data that are classified as high. The process of determining the class for testing data is carried out by taking the number of k data which is considered the closest neighbor. As an example, testing data will be classified as data with low class if k is set to 3. This is because from the 3 closest
neighbors of testing data, the two closest neighbors are data with low class and only one of the closest neighbors is high class. The next process is to evaluate the accuracy of the classification system by varying the number of k. This study varied the k value from 3 to 17 to get the best accuracy. The performance of this classification system can be seen in Table 5 and Figure 4.

| Number of k | Accuracy of Arms Import (%) | Accuracy of Arms Export (%) |
|-------------|-----------------------------|-----------------------------|
| 3           | 75.18                       | 76.21                       |
| 4           | 76.24                       | 75.74                       |
| 5           | 77.53                       | 77.03                       |
| 6           | 78.98                       | 78.82                       |
| 7           | 80.17                       | 80.59                       |
| 8           | 81.5                        | 82.86                       |
| 9           | 82.36                       | 83.38                       |
| 10          | 84.87                       | 84.57                       |
| 11          | 85.31                       | 86.22                       |
| 12          | 88.56                       | 88.97                       |
| 13          | 89.72                       | 88.02                       |
| 14          | 89.21                       | 87.65                       |
| 15          | 88.56                       | 85.95                       |
| 16          | 85.31                       | 84.05                       |
| 17          | 84.87                       | 83.23                       |

Figure 4. Performance Accuracy of Classification System

Based on Table 5 and Figure 3, it can be seen that the accuracy of the classification system will increase with the addition of the k value. The increase in the level of accuracy will occur until the
classification system achieves the best performance in classifying the data. In this study, the classification system achieved the best accuracy when k was 13 for total arms exports, which was 89.72%. As for the total arms imports, the classification system achieves the best accuracy when k is set to 12 with an accuracy rate of 88.97%. Therefore it can be concluded that the research succeeded in developing a classification system for arms dynamics in southeast asian region with a fairly good level of accuracy.

4. Conclusion
This research is conducted to discover the patterns of arms trade dynamics among Southeast Asia countries which were associated with its political and economic posture by utilizing k-NN classification algorithm. This conducted research discovered that utilizing k-NN classification algorithm can classify arms dynamics into three classes such as low, middle and high. The classification system achieved the best accuracy when k is set to 13 for total arms exports, which was 89.72%. As for the total arms imports, the classification system achieves the best accuracy when k is set to 12 with an accuracy rate of 88.97%. This level of accuracy is good enough to be applied in the classification and prediction process of arms dynamics in Southeast Asia. Further research can be carried out using other classification algorithms such as the decision tree algorithm for comparison. In addition, it is necessary to carry out further research involving additional variables such as size of the country, state conditions and so on.

5. References
[1] J. McCarthy and E. Feigenbaum, “In Memoriam Arthur Samuel: Pioneer in Machine Learning,” AI Magazine, vol. 11, no. 3, p. 11, 1990.
[2] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” IBM J., vol. July, 1959.
[3] C. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[4] J. H. Friedman, “Data Mining and Statistics: What’s the Connection?,” Comput. Sci. Stat., vol. 29, no. 1, pp. 3–9, 1998.
[5] B. F. Braumoeller and A. E. Sartori, “Empirical-Quantitative Approaches to the Study of International Relations,” in Cases, Numbers, Models: International Relations Research Methods, D. Sprinz and Y. Wolinsky, Eds. Michigan: University of Michigan Press, 2004, p. 139.
[6] P. E. Shea and P. Poast, “War and Default,” J. Conflict Resolut., vol. 62, no. 9, pp. 1876–1904, 2018.
[7] M. Digiuseppe and P. Poast, “Arms versus Democratic Allies,” Br. J. Polit. Sci., vol. 48, no. 4, pp. 981–1003, 2018.
[8] C. Kaoutzanis, P. Poast, and J. Urpelainen, “Not letting ‘bad apples’ spoil the bunch: Democratization and strict international organization accession rules,” Rev. Int. Organ., vol. 11, no. 4, pp. 399–418, 2016.
[9] P. Poast, M. J. Bommarito, and D. M. Katz, “The Electronic World Treaty Index: Collecting the Population of International Agreements in the 20th Century,” Ssrn, pp. 1–29, 2015.
[10] D. Wiens, P. Poast, and W. R. Clark, “The Political Resource Curse: An Empirical Re evaluation,” Polit. Res. Q., vol. 67, no. 4, pp. 783–794, 2014.
[11] P. Poast, “Central Banks at War,” Int. Organ., vol. 69, no. 1, pp. 63–95, 2014.
[12] P. Poast and J. Urpelainen, “Fit and Feasible: Why Democratizing States Form, not Join, International Organizations,” Int. Stud. Q., vol. 57, no. 4, pp. 831–841, 2013.
[13] W. R. Clark, S. N. Golder, and P. Poast, “Monetary institutions and the political survival of democratic leaders,” Int. Stud. Q., vol. 57, no. 3, pp. 556–567, 2013.
[14] J. Erasey, “Causal Inference with Observational Data Introduction to Matching,” in Analytics, Policy, and Governance, New Haven: Yale University Press, 2015, pp. 1–37.
[15] W. D. Berry et al., “Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?,” *Am. J. Pol. Sci.*, vol. 54, no. 1, pp. 248–266, 2018.

[16] J. Jaafar, Z. Indra, and N. Zamin, “A category classification algorithm for Indonesian and Malay news documents,” *J. Teknol.*, vol. 78, no. 8–2, 2016.