E-ISSN: 2476-9606
Abstract Proceedings International Scholars Conference
Volume 7 Issue 1, October 2019, pp. 1774-1785

https://doi.org/10.35974/isc.v7i1.1628

Exploratory Data Analysis Towards Terrorist Activity in Indonesia Using Machine Learning Techniques

Green Arther Sandag
Computer Science Department, Universitas Klabat
greensandag@unklab.ac.id

ABSTRACT

Terrorism Activity is the subject of the talks in various countries, especially in Indonesia. The activities of terrorism are carried out in various ways using suicide bombs, violent action that aimed to demoralize by creating fear to the society and national security. In Indonesia, according to Kompas news website recorded there were 10 suicide bombings occurred in the past 6 years and took many casualties in every event. With this, it certainly gives a threat to the people in Indonesia in terms of physical, moral and even in terms of national security. To overcome this problem, it is necessary to increase the national security so that terrorism can be prevented, and it will not happen again. This study is aimed to conduct an exploratory data analysis and predict terrorist activity in Indonesia using K-Nearest Neighbor (KNN), and ¬k-fold cross-validation. In this research, data selection, data cleaning, data reduction was carried out and feature selection process which aimed to find out the most influential data attributes. Based on the result of the analysis to predict the terrorist activity, the result of the accuracy was obtained with a value of k = 8 at 88.86%.

Keywords: Prediction, K-Nearest Neighbor, K-Fold Cross-Validation

INTRODUCTION

Indonesia is located in the Asian Continent and is astronomically located between 980 East Longitude (BT) 1410 East Longitude (BT). Indonesia has located a latitude between 60 North Latitude (LU) to 110 South Latitude (LS). The total area of Indonesia is 1,906,240 km2 with 33 provinces. The Province is divided into 288 districts, 88 cities, 617 sub-districts, 88 cities, 617 sub-districts, and 69,007 villages. The population in Indonesia based on the result of a census in 2000 was 202.9 million people, was made Indonesia the fourth country with the largest population in the world. According to The Book of World Ranking, the 1984 edition noted that Indonesia is a country that has the 11th most powerful security defense.

Security is a basic need used as a safeguard and protection of the national benefit in a country, where national security must be based on Pancasila. The National benefit is a dominant factor in the national security of a country. The safety of mankind has become something that should be fighting for every country including Indonesia. Therefore, security is not only deal with traditional threats but also aimed to protect the safety of all mankind.

One of the threats occur in society is the act of terrorism. The action of terrorists aimed to create fear in the community and usually carried with coordinated attacks. besides, terrorist strategy to determine the right location for their act. The location tends to be a celebrity to have a large psychological impact on the community.

Terrorist acts in Indonesia had taken a place in different patterns and strategies. Terrorist act in the traditional way has changed to a modern pattern, where terrorist used a concept of phantom call network which can connect to other terrorist groups. The pattern of the terrorist attacks in Indonesia continues to occur in different ways, including bombs and committing acts of violence in public so they can demoralize.

To help the researchers find out the activity of the terrorist in Indonesia, the researcher will make a prediction using data mining. Data mining is used to find out pattern or trend which aim to obtain something useful by doing data mining, which has larger in number, through data mining several things such estimation, description, prediction, clarification, and association [4]. In this study, researchers will use the K-Nearest Neighbor method which is an algorithm in data mining aimed to classify new objects based on attributes and training data.

Research conducted by Elovici used a data mining technique to detect terrorist by using web traffic content in to relate the information of terrorism, the information used by the system to detect the real-time of suspect users involved in the terrorist activity. The Receiver-Operator-Characteristics (ROC) analysis shows methodology can perform an intrusion detection system. Based on the related research, the prediction has been done using the K-Nearest Neighbor method conducted by Hutami et al, which aims to examine and implement the K-Nearest neighbor methods for the sales forecasting and get the result of sales predictions in a CV. Octo Agung Jepara with a minimum error rate. The result of this study has error rates or 6% of MSE and 94% accuracy.

Another study conducted by Yustanti aims to predict the selling price of the land-based on the measurable factors where the approach used to predict the land selling prices is the K-Nearest Neighbor algorithm. This study has an 80% accuracy rate [7]. The purpose of this study the researcher will predict the terrorist activity, especially in Indonesia, for Indonesian citizens can have accurate information about the act of terrorism. With accurate information, Indonesians will be more vigilance and the police with national security will enhance vigilance and safety in the country.

METHODS

Table 1. **Global Terrorism Dataset**

Parameter	Description	Value
success	1 = Attack; 0 = No Attack	Integer
crit2	1 = "Yes" the incident complies with the criteria 2; 0 = "No" the incident not complies with the criteria 2 or no indication.	Integer
suicide	1 = "Yes" the incident was suicide bombings; 0 = "No" no indication that incident was suicide bombings	Integer
INT_LOG	1 = "Yes" it was International logistics attack; 0 = "No" domestic logistics attack; the nationality of the doer group is the same as the attacked location.	Integer
Ransom	1 = "Yes" the incident involve with a ransom; 0 = "No" the incident did not involve a ransom; u = "Unknown" the incident is not known whether the incident involved with a ransom. [[NULL] not applicable.	Integer
Vicinity	1 = "Yes" the incident occurred around the relevant city; 0 = "No" the incident is in the city.	Integer
crit1	1 = "Yes" the incident complies with the criteria 1; 0 = "No" the incident does not comply with the criteria 1 or no indication about the incident comply with criteria 1.	Integer
crit3	1 = "Yes" the incident complies with the criteria 3; 0 = "No" the incident does not comply with the criteria 3.	Integer
ishostkid	1 = "Yes" the victims were taken hostage or kidnapped; 0 = "No" the victims were not taken hostage or kidnapped; u = "Unknown" not known if the victims were taken hostage or kidnapped.	Integer
specificity	This field identifies geospatial resolution in latitude and longitude fields. The most specific resolution that available throughout the dataset is the city, village, or city where the attack occurred.	Integer
Guncertain1	1 = "Yes" the perpetrator attributes to this incident is a suspect; 0 = "No" the perpetrator attributes to this incident is not a suspect.	Integer
doubter	1 = "Yes" what are the doubts in the incidents is an act of terrorism; 0 = "No" basically there is no doubt whether the incident was an act of terrorism.	Integer
weaptype1_txt	Until four types of a weapon were recorded in every incident.	Integer
nwound	This field record the total number of the victims of non-fatal injuries confirmed by both perpetrators and victims.	Integer
targtype1_txt	Target field/type of victims catch the common types of target/victims.	Integer

Nkill	This field stores a total of mortality confirmed in the incident. Including all victims and attackers who died	Integer
gname	This field contains the group name for an attack	Integer
attacktype1_txt	General Attack Method	Integer
weapsubtype1_t xt	This field records value more specifics in most types of weapons identified above.	Integer
provstate	This variable records the name (event time) from the first order in the subnational administrative area where the event occurred.	Integer
targsubtype1_txt	The subtype variable target capture more specific target categories and provides the next designation level for each target.	Integer
weapdetail	This field records all information related to the type of weapon used in the incident.	Integer
Corp1	This is the name of a corporate entity or government target institution. If the target element is not specified, "Unknown" is registered.	Integer
Terget1	Certain people, buildings, installations etc., who target and/or become victims and part of the entities mentioned above.	Integer

Research Design is a process that will be carried out in this study. The first process is to retrieve data from Kaggle in Global Terrorism Dataset. The next stage is done by pre-processing data by doing data selection, data cleaning, and data reduction then the dataset is divided into 80%, consist of 160 data as training data and 20% consisting of 152 data is testing data, then it will proceed to the next process is by using algorithm K-Nearest Neighbor, after that which of the performance evaluation will be known based on the data that has been tested. The following is a research design picture:

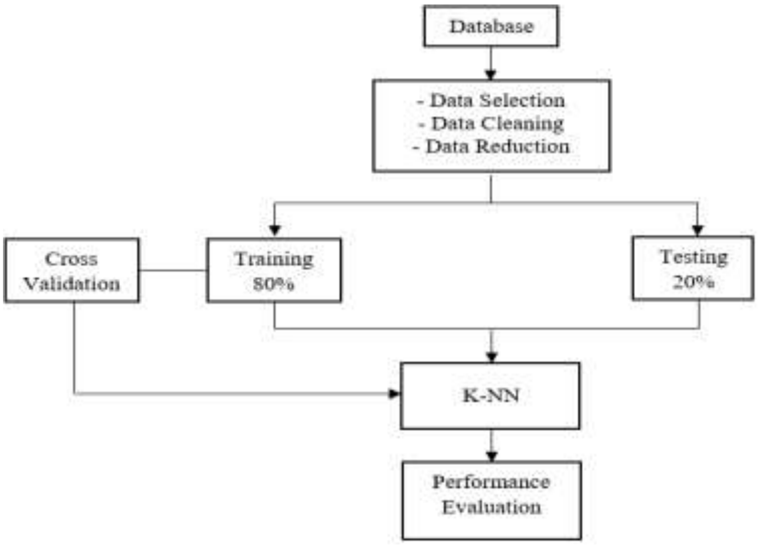

Figure 1. Research Design

Data Preprocessing is divided into 3 parts, which are Data Selection, Data Cleaning, and Data Reduction. Data Selection is data taken from the database where the data used by the researchers is only data that is appropriate with the research [8]. Data Cleaning is a process of cleaning data so that the data to be used is suitable for the needs [9]. Data Reduction is the process of deleting data that is incomplete in attribute so that it can be reduced but can produce accurate data [10]. In this case, the researcher conducts a data selection process by choosing terrorist activities that only occur in Indonesia, while data cleaning process, researchers replace data that does not have a value on an attribute using Rapidminer. Data reduction is done by selecting attributes that are inappropriate with the research where previously there were 135 attributes, but after data reduction, there are 24 attributes taken by the researcher because these 24 attributes are considered appropriate for predictions about terrorist activities in Indonesia.

Cross-Validation is one of the techniques used to evaluate a model by dividing the original sample into a training set that aims to train the model and other samples to set a test by evaluating the model. In cross-validation, the original sample will divide randomly in k equal size subsample will be used as data testing and the remain will be used for training data. The process had been done in cross-validation by repeating as many times as in this case using multiplication, with each sub-sample used once as validation data [11]. K-fold validation will run the study by dividing the data into k partitions and will be tested many times to estimate the accuracy of the estimation, the reason researchers use k-fold cross-validation is because cross-validation techniques can process data accurately [12].

K-Nearest Neighbor (KNN) is a clarification method used in research data on the closest distance to the object. The K-NN algorithm is a method which in this case is used to classify objects based on the closest training example. K-NN is known to be a Lazy learning algorithm because these algorithm functions close locally with its calculations delayed until the classification process [13]. The following is the formula for K-Nearest Neighbor:

$$d\ (x, y) = \sqrt{\sum_{i=1}^{n}(xi - yi)}$$

Where:

d = distance	i = total data	n = amount of data
x = first point.	y = end point	

In the performance evaluation stage, the performance test of the classifier will be performed. Calculating the recall value is the success rate of the system in recovering of the information, precision is the presentation of the accuracy of the information provided by the computer according to the user's request and accuracy is the presentation of the correct number of datasets based on the method used in performance evaluation. The following is the formula for recall, precision and accuracy is the presentation of the correct number of datasets base on the method used in performance evaluation [14]:

***Accuracy*:**

$$Accuracy = \frac{TN+TP}{TN+FB+FN+TP}$$

***Recall*:**

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

Precision:

$$Precision = \frac{TP}{TP+FN}$$

Where:

TP	: *True Positive* Value	TN	: *True Negative* Value
P	: *Positive* Value	FP	: *False Positive* Value
N	: *Negative* Value	FN	: *False Negative* Value

***RMSE*:**

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(pi-ai)}{n}}$$

Where:

pi = prediction value output for i

ai = actual value output for i

n = total example data

Feature Important is a technique for selecting important attributes in research by removing irrelevant attributes in the data that will be used in research [15]. In this study, the researcher used an *information gain ratio* which used to determine the rank level of existing attributes to help researchers predict the activities of the terrorist in Indonesia.

$$Info\ (D) = -\sum_{I=1}^{C} p_I log_2\ (pi)$$

Where:
c = total value of attribute target
pi = total sample for class i

RESULTS

In this section the researcher will conduct an analysis of the Global Terrorism Dataset, in order to analyze what the researcher will do is by conducting preprocessing data where the existing data will be selected, cleaned it reduced the attributes that are not needed in this study so that feature selection can be performed to find out which attributes have the greatest influence and at last phase will be analyze using K-Nearest Neighbor algorithm.

Visualizing Terrorist Attack in Indonesia.

Figure 2. shows the activity of terrorists in Indonesia, Terrorist Activity occurs in an area that is prone to conflicts such as Aceh, Poso, Maluku, Papua and mostly in Java Island. The least of the terrorist activity is in Kalimantan Island.

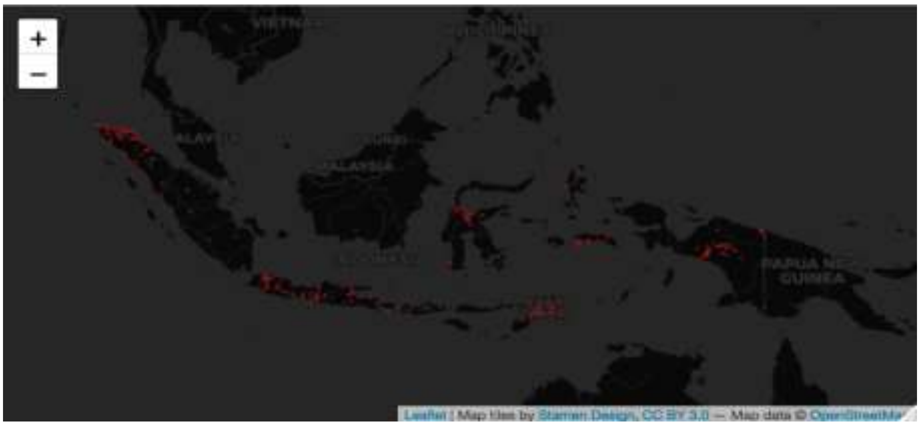
+
−
PAPUA NEW GUINEA
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Figure 2. Activity of terrorists in Indonesia

Based on Figure 3, it provided information that the target1 attribute with a weight of 0.485 has the highest influence to predict terrorist activity in Indonesia. The target1 attribute has the greatest influence because target1 is an attribute that explains the target of the terrorist attack so it determines what types of attack will be used and determine the number of individuals who will be the victims in an attack. Whereas, the crit2 attribute is the attribute that has the least influence with weight which is only 0.0001. Based on the result of observations from researchers, the reason why suicide attributes have the least influence because of the intentions and the individual not intent to perform attacks to escape but not affect the total of the individuals who are the victims of the terrorist attacks.

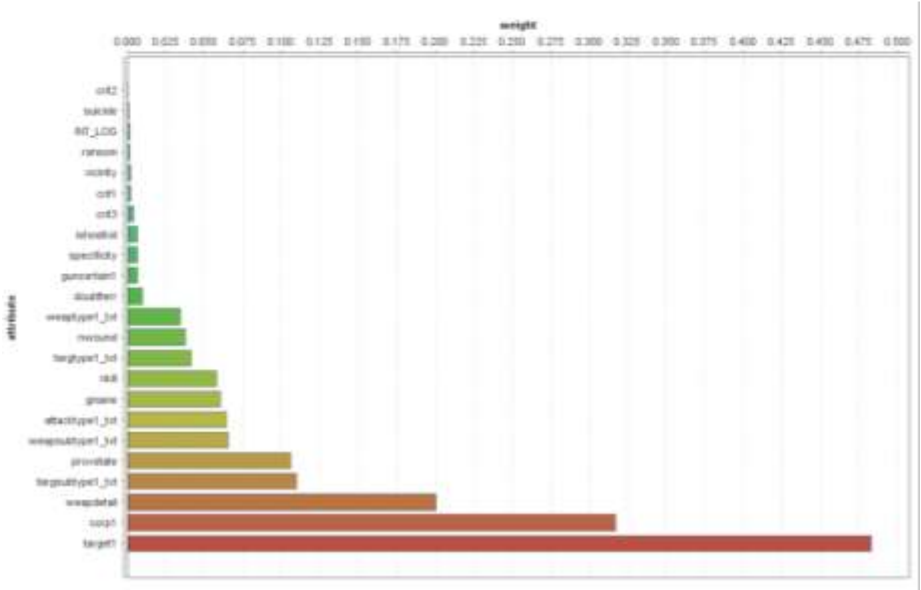

Figure 3. Feature Importance

Model

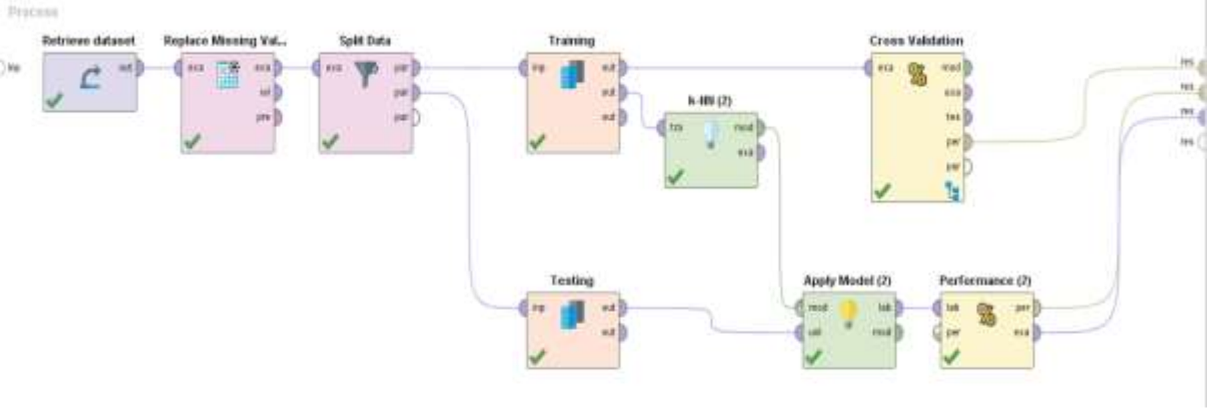

Figure 4. Final Model

Figure 4 shows the modeling run by Rapidminer software to obtain prediction results. In this model, researcher input a dataset to be used based on the researcher's needs which in this case is global terrorism dataset had been done a preprocessing in the previous stage and split the data used to divide the 80% of training data or in this case, there are 160 row data used in training data. While the remaining 20% or 152 rows of other data is used for data training. Whereas the data is divided, the next step is the data will be tested independently and tested using k-fold cross-validation. For independent testing the training data and data testing have been divided will be directly tested using K-NN algorithm and measure the performance to determine the results. Other tests, instead of using the same algorithm such as K-NN the performance measurements are performed to determine the percentage value of the accuracy, recall, precision, and RMSE (Root Mean Square Error).

Performance Evaluation and confusion matrix using Cross-Validation Global Terrorism Dataset

Table 2. **Performance Evaluation using Cross-Validation Results**

Algorithm K-Nearest Neighbor	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**RMSE**
K = 7	88.18%	72.86%	72.99%	0.342
K = 8	88.86%	73.69%	74.44%	0.333
K = 9	88.02%	73.82%	72.66%	0.344
K = 10	87.34%	70.89%	73.15%	0.349

In the K-NN algorithm, there is a value of k which is important because it can affect the performance of the K-NN algorithm. If the determination of the value of k is too small it will affect the result of the classification that influences by noise, but if the determination of the value of k is too high it will affect the noise in the data. To get a good value in determining the value of k, which is the amount of the data neighbor. It can be determined using optimization parameters that can be seen based on the level of accuracy obtained from each value k, which can use k-fold cross-validation, to determine a value of k to be used [16]. Based on the k-fold cross-validation test for the global terrorism dataset was conducted and found that k = 8 obtained the highest accuracy so that k= 8 will be used in the K-NN algorithm to predict terrorist activity in Indonesia.

Table 2 shows the result of performance evaluation for using cross-validation. A result of the performance evaluation of the tests conducted by the researcher showed a value of k = 8 accuracy level was 88.86%, recall was 73.69%. precision was 74.44% and RMSE was 0.333. For the second higher k value is the value k= 7 with an accuracy of 88.18%, 72.86%, precision 72.99% and RMSE 0.342. The third highest value is k = 9 with accuracy value 88.02%, recall 73.82%, precision 72.66% and RMSE 0.344 and the last k = 10 produces an accuracy value of 87,34%, recall 70.89%, precision 73.15% and RMSE 0.349.

Table 3. **Confusion Matrix**

	true 1	true 0	class precision
pred. 1	520	67	88.59%
pred. 0	13	9	40.91%
class recall	97.56%	11.84%	

A confusion matrix is a data mining method used to calculate the accuracy of the data in this case in Table 3 shows the result of confusion matrix with k = 8 that the correct predictions made on value 1 are amount 520 and predictions that are incorrect as much as 67. While the prediction result that is the true value of 0 carried out 13 and 0 for prediction value for wrong is as much as 9. From the result of this confusion matrix, the percentage of the result of independent accuracy is 88.68%.

Table 4. **Performance Evaluation for Independent Test Result**

Algorithm K-Nearest Neighbor	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**RMSE**
K = 7	86.84%	65.41%	68.95%	0.304
K = 8	88.82%	64.29%	72.42%	0.308
K = 9	87.50%	65.79%	70.65%	0.300
K = 10	87.50%	63.53%	70.36%	0.300

Table 4 shows the results of the performance evaluation for an independent test. The result obtained from tests conducted by the researcher shows that the value of k = 8 the accuracy level is 88.82%, recall is 64.29%. precision is 72.42% and RMSE is 0.308. The second highest k value is value of k= 9 with accuracy of 87.50%, recall 65.79%, precision 70.65% and RMSE 0.300. Third highest value is k= 10 with accuracy value 87.50%, recall 63.53%, precision 70.36% and RMSE 0.300 and last k = 7 produces accuracy value of 86.84%, recall 65.42%, precision 73.15% and RMSE of 0.349.

Table 5. **Confusion Matrix for Independent Test**

	true 1	true 0	class precision
pred. 1	129	13	90.85%
pred. 0	4	6	60.00%
class recall	96.99%	31.58%	

Table 5 shows result of confusion matrix with a value of k = 8, the result of prediction of value 1 which has a true value of 129 and the wrong is 13. While the prediction of value 0 which is true is 4 and the wrong predictions are 6. According to the results, it is found the accuracy value is 88.82%, recall 64.29%, precision 72.42% and RMSE 0.308.

DISCUSSION

The results obtained in this study expected to be a reference for other researchers who will conduct further research related to terrorist activities in Indonesia either performing analytical activities or making an application to predict terrorist activities and additional information from the research that had performed will provide advice for security forces to enhance national security.

Conclusion

According to the analysis, the researcher proved the result using the K-NN algorithm independently is different from the result of K-NN algorithm testing which added the use of k-fold cross-validation in predicting terrorist activity in Indonesia. The evidenced of the result obtained by doing a comparison between the best value of k, found that value of k = 8 values is the best in this study by generating the value of accuracy using k-fold cross-validation of 88.86%, recall 73.69%, precision 74.44% and RMSE 0.333. While independent testing with k = 8 produces an accuracy value of 88.82%, recall 64.29%, precision 72.42% and RMSE value (root mean square error) of 0.308.

REFERENCES

Rachmawati, I. (2019). *Profil Negara Indonesia Lengkap*, Retrieved from: https://portal-ilmu.com/negara-indonesia/#

Darmono, B. (2016). Konsep Dan Sistem Keamanan Nasional Indonesia, *Jurnal Ketahanan Nasional*, 15(1), 1–42.

Sanur, D. (2018). Terorisme: Pola Aksi dan Antisipasinya. *Kajian Singkat Terhadap Isu Aktual dan Strategis*, 10(10), 25-30.

Nasari, F., & Sianturi, C. J. M., (2016). Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat. *CogITo Smart Journal,* 2(2), 108–119.

Hutami, R., & Astuti, E. Z. (2016*). Implementasi Metode K-Nearest Neighbor Untuk Prediksi Penjualan Furniture Pada CV. Octo Agung Jepara*. (tugas akhir). Fakultas Ilmu Komputer. Universitas Dian Nusantara Semarang, Semarang, Indonesia.

Elovici, Y., Kandel, A., Last, M., Shapira, B., & Zaafrany, O. *Using Data Mining Techniques for Detecting Terror-Related Activities on the Web*, p. 13.

Yustanti, W., (2012). Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika, Statistika, & Komputasi*, 9 (1), 57–68.

Asriningtias, Y. & Marhadiyah, R. (2014). Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa. *Jurnal Informatika*, 8(1), 837–848.

Pratiwi, R. W. & Nugroho, Y. S. (2016). Prediksi Rating Film Menggunakan Metode Naïve Bayes. *Jurnal Teknik Elektro*, 8(2), 60-63. doi: https://doi.org/10.15294/jte.v8i2.7764

Fithri, D. L. and Darmanto, E. (2014). Sistem Pendukung Keputusan untuk Memprediksi Kelulusan Mahasiswa Menggunakan Metode Nave Bayes. in *Prosiding SNATIF Ke-1 Tahun 2014*, 319–324.

Sandag, G. A., Leopold, J. & Ong V. F. (2018). Klasifikasi Malicious Websites Menggunakan Algoritma K-NN Berdasarkan Application Layers dan Network Characteristics. *Cogito Smart Journal*. 4(1), 37-45. doi: 10.31154/cogito.v4i1.100.37-45

Murtopo, A. A. (2016). Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes. *CSRID (Computer Science Research and Its Development Journal)*, 7(3), 145–154.

Banjarsari, M. A., Budiman, I. & Farmadi, A. (2015). Penerapan K-Optimal Pada Algoritma KNN Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer FMIPA UNLAM Berdasarkan IP Sampai Dengan Semester 4. *KLIK-Kumpulan Jurnal Ilmu Komputer*, 2(2), 159–173.