BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2022 Volume 16 Issue 1 Page 057–064
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss1pp057-064

INTERPRETABLE PREDICTIVE MODEL OF NETWORK INTRUSION USING SEVERAL MACHINE LEARNING ALGORITHMS

Muhammad Ahsan[1*], Arif Khoirul Anam[2], Erdi Julian[3], Andi Indra Jaya[4]

[1,2,3,4] *Statistics Department, Faculty of Science and Analytic Data, Institut Teknologi Sepuluh Nopember Kampus ITS Sukolilo, Surabaya, 60111, Indonesia*

Corresponding author e-mail: [1] muh.ahsan@its.ac.id*

Abstract. *Network intrusion is any unauthorized activity on a computer network. Attacks on the network computer system can be devastating and affect networks and company establishments. Therefore, it is necessary to curb these attacks. Network Intrusion Detection System (NIDS) contributes to recognizing the attacks or intrusions. This paper explains the factors that influence network attacks. Some machine learning methods are used such as are logistic regression, random forest XGBoost, and CatBoost. The best model is chosen from these models based on its accuracy level. Classification modelling is divided into two types, namely using a dummy and not using dummy variables. The best method for predicting network intrusion is a random forest with a dummy variable that has an Area Under Curve (AUC) value of 92.31% and an accuracy of 90.38%.*

Keywords: *classification, intrusion, machine learning, network.*

Article info:

Submitted: 02nd September 2021 *Accepted: 28th January 2022*

How to cite this article:
Muhammad Ahsan, Arif Khoirul Anam, Erdi Julian and Andi Indra Jaya, "INTERPRETABLE PREDICTIVE MODEL OF NETWORK INTRUSION USING SEVERAL MACHINE LEARNING ALGORITHMS", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 1, pp. 057-064, Mar, 2022.

1. INTRODUCTION

A network intrusion is any unauthorized activity that takes place on a computer network. To detect an intrusion, defenders must have a clear understanding of how attacks work. With the growth of network technologies, network attacks have drastically increased in quantity. Intrusion detection is necessary for today's computing environment because it is impossible to track current and potential threats and vulnerabilities in our computing systems. The environment is constantly evolving and changing the field of science with new technologies and the Internet. [1].

An attack on a network computer system can be catastrophic and affect the network and the establishment of the company. We need to stop these attacks and the Intrusion Detection System will work to identify them. Without a Network Intrusion Detection System (NIDS) to recognize the activities in the network, this might result in irremediable harm to a system's network. An intrusion attack is an attack where an attacker enters your network [1]. Several methods have been employed in monitoring attacks. These methods are based on machine learning, data mining, and statistical methods. For machine learning and data mining methods, some investigations have been conducted such as in reference [2]–[4]. Furthermore, references [5]–[10] are research that uses statistical methods to detect an anomaly in a network.

The application of the logistic regression model has played a significant role in different fields. The logistic regression algorithm is used when we want to classify data items into categories [11], [12]. Usually, the target variable is binary which means it only contains data classified as 1 or 0. The main objective of the logistic regression algorithm is to find the best fit that is diagnostically reasonable to describe the relationship between the target variable and the predictor variables. Random forest classifier uses an ensemble learning method for classification which uses multiple decision trees during the training phase and outputs an average prediction of individual trees [13]. This classifier forms forests with a random number of trees. Normal decision tree algorithms are rule-based and rely only on a set of prediction rules on the dataset. In contrast to this, random forest classifiers instead of using Gini Index or information gain for calculation of the root node, find the root node, and splits the features randomly. Some advantages of random forest classifier are that it has good predictive performance on supervised learning algorithms, provides a reliable feature importance estimate, and offers an efficient estimate of the test error without incurring the iterative model training costs associated with cross-validation [14].

XGBoost is an ensemble algorithm that belongs to the category of boosting algorithm with three typical integration methods, namely bagging, boosting, and stacking. The main idea of this algorithm is to transform features by growing and adding trees constantly. Each time a tree is added, the new function learned will adjust to the residual from the last prediction. So when the training is completed, k trees are obtained, and then the sample scores can be predicted. According to the characteristics of the sample, each tree produces a leaf node, and each leaf node corresponding to the final score will be added to the corresponding score; so that, the predictive value of the sample can be obtained. Some of the advantages of the XGBoost algorithm are regularization which prevents the model from overfitting, parallel processing, handling missing values, cross-validation, and effective tree pruning [15].

CatBoost is a supervised machine-learning algorithm for classifying categorical data using gradient boosting on decision trees [16]. Initially, a series of under-fitted shallow decision tree models are built on sampled training datasets. the decision tree is formed with a top-down approach by dividing the training dataset into similar instances. Homogeneity between instances is measured by entropy. Decision trees act as weak learners during this ensemble learning method. After creating the decision trees, each tree model tries to reduce the residual error in the prediction using a log-loss function. The weighted cumulative sum of these predictions gives the final predicted value in the classifier with learning rate one. Two important algorithmic advances introduced in CatBoost are the implementation of ordered boosting, a permutation alternative based on the classic algorithm, and an innovative categorical feature processing algorithm.[17].

This paper will explain the factors that influence network attacks. Some machine learning methods used are logistic regression, random forest XGBoost, and CatBoost. The best model will be chosen from these models based on its accuracy. This paper is expected to be a reference in the field of data processing so that benefits can be taken.

2. RESEARCH METHODS

There are several studies on NIDS. Many statistical methods have been applied to developing intrusion prediction, such as discriminant analysis and logistic regression. Advanced machine learning methods including random forest, XGBoost, and CatBoost can also be applied.

2.1. K-Fold Cross-Validation

The K-Fold Cross Validation (KCV) method is a reliable method for predicting error in a concept. This method is widely used by researchers to reduce the bias that occurs because of taking data samples to be used. KCV repeatedly splits data into training data and testing data, where each data has the opportunity to become data testing [18]. The most commonly used *k* value is 10 because it is the most feasible value to get the best error estimate [19]. The data is divided into 10 parts, while 9 parts are used as training data and the other part becomes testing data, then repeated 10 times, so that each data has the opportunity to become training data as well as testing data.

2.2. Feature Selection

In machine learning and statistics, feature selection is the method of selecting a subset of relevant options to use in building the model. Feature selection techniques are used for shorter training times, simplification models to make them easier to interpret, to avoid the curse of dimensionality, and increased generalization by reducing overfitting [20].

2.3. Data Source

The data used in this paper is secondary data, namely UNSW NB15. The data was retrieved from the Kaggle website https://www.kaggle.com/mrwellsdavid/unsw-nb15.

2.4. Variable of Interest

The given training dataset has dimensions of 82,332 × 44 and the testing dataset has dimensions of 175,341 × 44. The variables are explained in Table 1.

Table 1. Variable of Interest

Var	Name	Type
X_1	proto	Non-Metric
X_2	state	Non-Metric
X_3	dur	Metric
X_4	sbytes	Metric
X_5	dbytes	Metric
X_6	sttl	Metric
X_7	dttl	Metric
X_8	sloss	Metric
X_9	dloss	Metric
X_{10}	service	Non-Metric
X_{11}	Sload	Metric
X_{12}	Dload	Metric
X_{13}	Spkts	Metric
X_{14}	Dpkts	Metric
X_{15}	swin	Metric
X_{16}	dwin	Metric
X_{17}	stcpb	Metric
X_{18}	dtcpb	Metric

Var	Name	Type
X_{19}	smeansz	Metric
X_{20}	dmeansz	Metric
X_{21}	trans_depth	Metric
X_{22}	res_bdy_len	Metric
X_{23}	Sjit	Metric
X_{24}	Djit	Metric
X_{25}	Stime	Timestamp
X_{26}	Ltime	Timestamp
X_{27}	Sintpkt	Metric
X_{28}	Dintpkt	Metric
X_{29}	tcprtt	Metric
X_{30}	synack	Metric
X_{31}	ackdat	Metric
X_{32}	is_sm_ips_ports	Non-Metric
X_{33}	ct_state_ttl	Metric
X_{34}	ct_flw_http_mthd	Metric
X_{25}	is_ftp_login	Non-Metric
X_{36}	ct_ftp_cmd	Metric
X_{37}	ct_srv_src	Metric
X_{38}	ct_srv_dst	Metric
X_{39}	ct_dst_ltm	Metric
X_{40}	ct_src_ ltm	Metric
X_{41}	ct_src_dport_ltm	Metric
X_{42}	ct_dst_sport_ltm	Metric
X_{43}	ct_dst_src_ltm	Metric
Y_1	Label	Non-Metric

2.5. Data Structure

The data structure used in this study is as follows:

Table 2. Data Structure

X_1	X_2	X_3	…	X_{43}	Y_1
$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	…	$X_{43,1}$	$Y_{1,1}$
$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	…	$X_{43,2}$	$Y_{1,2}$
$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	…	$X_{43,3}$	$Y_{1,3}$
⋮	⋮	⋮	⋮	⋮	⋮
$X_{1,n}$	$X_{2,n}$	$X_{3,n}$	…	$X_{43,n}$	$Y_{1,n}$

3. RESULTS AND DISCUSSION

3.1. Preprocessing

Before analyzing the data, it is necessary to pre-process the data first. Based on Figure 1, Data is known to have no missing value so further analysis can be done.

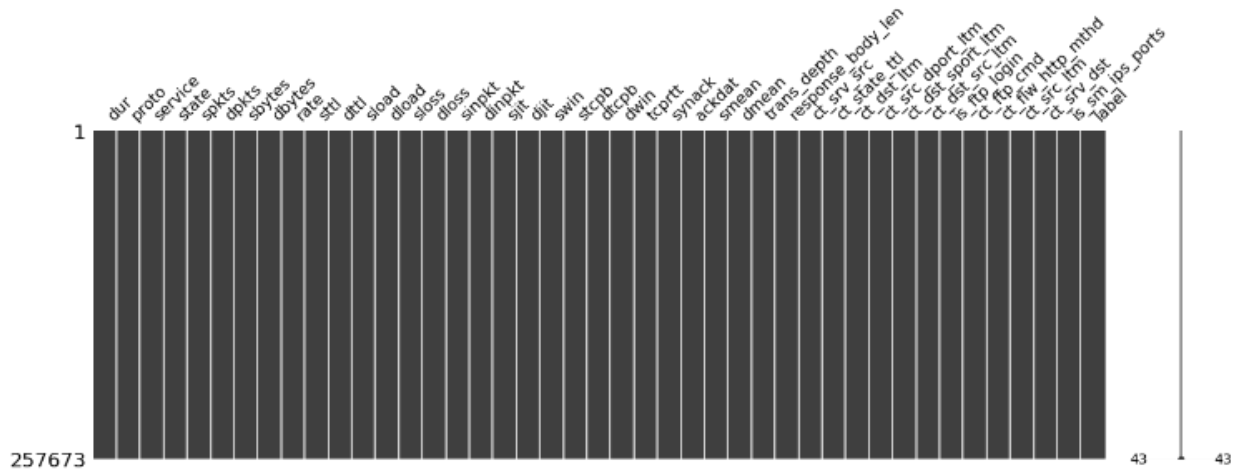

Figure 1. Missing Value Full Dataset

Predictor variables in this study consisted of numerical and categorical variables. Response variables used in this study consisted of 2 categories, namely normal and attack records. The characteristics of the labels in the training and testing data are as follows.

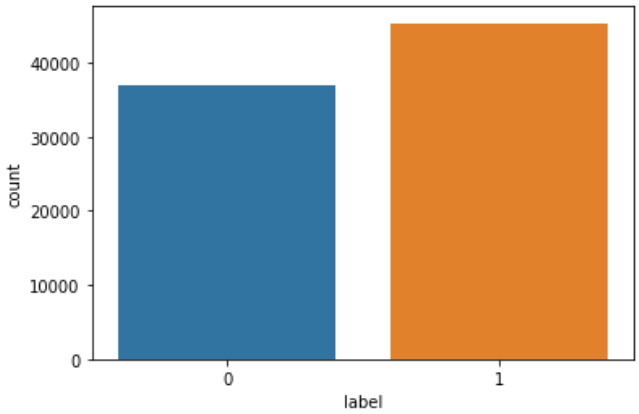

Figure 2. Training Labels

Based on Figure 2, The number of normal and attack categories is relatively balanced, although it appears that the attack categories are slightly higher. Because the data is balanced, there is no need to do data balancing.

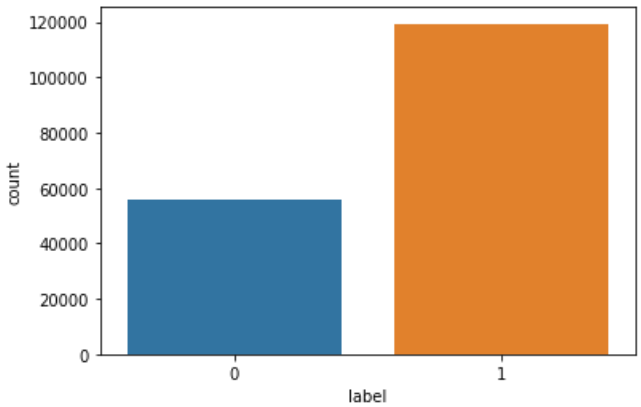

Figure 3. Testing Labels

Figure 3 shows that the label on the testing data has a significant difference between the normal and attack categories. The attack category has double the normal category.

3.2. Classification Modeling

In this study, classification modeling will be divided into two types, namely using a dummy and not using dummy variables. The dummy variable is used because several categorical variables are thought to improve the accuracy of the model. The methods used for classification are CatBoost, XGBoost, random forest, logistic regression, and linear discriminant analysis. The selection of the best method is based on the AUC value and the highest accuracy in the testing data.

Table 3. Classification Results

Model	Without Dummies		With Dummies	
	AUC	Accuracy	AUC	Accuracy
CatBoost	0.9217	0.9014	0.9222	0.9018
XGBoost	0.9208	0.9006	0.9218	0.9020
Random Forest	0.9208	0.9006	0.9231	0.9038
Linear Regression	0.8178	0.8151	0.7936	0.7435
LDA	0.8877	0.8792	0.8881	0.8678

Based on Table 3, it is known that the AUC value and the highest accuracy using the CatBoost method without dummy variables. Meanwhile, the random forest method has the highest AUC value and accuracy with a dummy variable. In conclusion, random forest is the best method for predicting network intrusion.

4. CONCLUSIONS

In this research, some machine learning methods are used such as are logistic regression, random forest XGBoost and CatBoost are used to classify the intrusion in the network. The UNSW NB15 dataset is known to have no missing values. The number of normal and attack categories in training is relatively balanced. The best method for predicting network intrusion is to use a random forest classifier with a dummy variable that has an AUC value of 92.31% and an accuracy of 90.38%. For the next studies, performing feature engineering on the data and tuning parameters can be used to get the best accuracy of the model.

ACKNOWLEDGMENT

Acknowledgments are addressed to the Institut Teknologi Sepuluh Nopember who has provided support so that this research can be realized properly.

REFERENCES

[1] M. Gandhi and S. K. Srivatsa, "Detecting and preventing attacks using network intrusion detection systems," *Int. J. Comput. Sci. Secur.*, vol. 2, no. 1, pp. 49–60, 2008.
[2] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," *Expert Syst. Appl.*, 2017, doi: 10.1016/j.eswa.2016.09.041.
[3] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No. 99CB36344)*, 1999, pp. 120–132.
[4] N. Devarakonda, S. Pamidi, V. V. Kumari, and A. Govardhan, "Intrusion Detection System using Bayesian Network and Hidden Markov Model," *Procedia Technol.*, vol. 4, pp. 506–514, 2012, doi: 10.1016/j.protcy.2012.05.081.
[5] M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo, and H. Khusna, "T2 Control Chart based on Successive Difference Covariance Matrix for Intrusion Detection System," in *Journal of Physics: Conference Series*, 2018, vol. 1028, no. 1, p. 12220.
[6] M. Ahsan, M. Mashuri, H. Kuswanto, and D. D. Prastyo, "Intrusion Detection System using Multivariate Control Chart Hotelling's T2 based on PCA," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 5, pp. 1905–1911, 2018.
[7] M. Ahsan, M. Mashuri, M. H. Lee, H. Kuswanto, and D. D. Prastyo, "Robust adaptive multivariate Hotelling's T2 control chart based on kernel density estimation for intrusion detection system," *Expert Syst. Appl.*, 2020, doi: 10.1016/j.eswa.2019.113105.
[8] M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo, and H. Khusna, "Multivariate T2 Control Chart Based on James-Stein and Successive Difference Covariance Matrix Estimators for Intrusion Detection," *MJS*, vol. 38, no. Sp2, pp. 23–35, 2019.
[9] M. Ahsan, M. Mashuri, and H. Khusna, "Intrusion Detection System Using Bootstrap Resampling Approach Of T2 Control Chart Based On Successive Difference Covariance Matrix," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 8, pp. 2128–2138, 2018.
[10] M. Mashuri, M. Ahsan, M. H. Lee, and D. D. Prastyo, "PCA-based Hotelling's T2 chart with Fast Minimum Covariance Determinant (FMCD) Estimator and Kernel Density Estimation (KDE) for Network Intrusion Detection," *Comput. Ind. Eng.*,

p. 107447, 2021.
[11] I.-C. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2473–2480, 2009.
[12] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics Med. Unlocked*, vol. 17, p. 100179, 2019.
[13] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier," *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 10–19, 2012.
[14] S. Hegelich, "Decision trees and random forests: Machine learning techniques to classify rare events," *Eur. Policy Anal.*, vol. 2, no. 1, pp. 98–120, 2016.
[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(pp. 785–794)," *New York, NY, USA ACM*, vol. 10, no. 2939672.2939785, 2016.
[16] A. Sau and I. Bhakta, "Screening of anxiety and depression among seafarers using machine learning technology," *Informatics Med. Unlocked*, vol. 16, p. 100228, 2019.
[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *arXiv Prepr. arXiv1706.09516*, 2017.
[18] E. Gokgoz and A. Subasi, "Comparison of decision tree algorithms for EMG signal classification using DWT," *Biomed. Signal Process. Control*, vol. 18, pp. 138–144, 2015.
[19] M. Berthold and D. J. Hand, *Intelligent data analysis*, vol. 2. Springer, 2003.
[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.