ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 2, pp. 1-14
DOI:10.33168/JSMS.2021.0201

Classification of Attacks through the Type of Protocol Using Data Mining

Parlindungan Tampubolon, Abba Suganda Girsang*

Computer Science Department, Bina Nusantara University, Jakarta, Indonesia 11480

agirsang@binus.edu (Corresponding author)

Abstract. Consideration of the security risks in the implementation system must be designed properly. The existence of the intrusion detection system (IDS) will help us classify which one is included in normal access or attack traffic. Therefore, the IDS system itself must always be updated for its database, because the method of attacking a system continues to develop. This study wants to make a comparison of three learning models, Gradient Boosting, Logistic Regression, and Support Vector Machine to classify detection system. The test was carried out using IDS KDD CUP 1999 log, by detecting the access protocol. From the results obtained, SVM is high enough for accuracy, but in fewer class conditions. However when the number of classes increases, SVM processing takes longer than others. Therefore, based on the performance and duration computation, the gradient boosting is outperform than the others.

Keywords: Supervised Learning, Gradient Boosting, Logistic Regression, SVM, DDOS.

1. Introduction

Technology development currently grows so fast and each platform innovates so that it can be accepted by industry or personal use. When implementing the application on an existing server, we must also be able to understand the risks from the security side. When our application can be accessed by the public were for example using the HTTP protocol, then we must be able to prepare for any steps so that our application can be used properly. Content also has fragility. Many attack methods can be done if we don't pay attention to the standards when coding or importing the library on the website application. Among them is 1. Phishing: This method usually does not attack the actual server. However, hackers will create a web that has the same appearance. If we detail, we will see the difference in the link address that we access. But sometimes the user unconsciously doesn't pay attention to the link. 2. Distributed Denial of Service: The attack carried out is to put more load on the destination server. So that the server cannot be accessed by other users simultaneously. The protocol used is ICMP. In simple terms, we can do this attack using ping. In Windows, if we use command prompt tools, we can run the command "ping -l 65527", this is an example of the maximum value for the load on one session from the client to the access server.

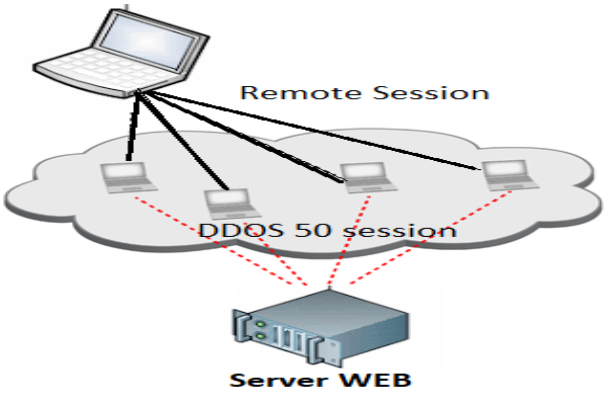

Fig. 1: DDOS attack

4. Defacing: This attack is also frequently carried out, which takes advantage of the vulnerability of content that has been published on the server. The SQL Injection method is an often applied loophole. Which takes advantage of loopholes, against SQL characters that can easily be added to page links. If the attacker can access the database, they will change the content on the hosting. These include index.php, home.html, or the page associated with the initial display. 5. Security Policy: System standardization must also be considered. Usually in companies using the hardening method, but in its implementation. This also has to be continuously updated, each system usually issues a list of Common Vulnerabilities and Exposures (CVE).

An intrusion Detection System (IDS) is a method that can be used to detect

suspicious activity in a system or network. IDS can conduct inspections of inbound and outbound traffic within a system or network, perform analysis and look for evidence of intrusion attempts.

2. Theoretical Backgrounds

Algorithm learning models have been implemented in different models, including Support Vector Machines, Naive Bayes, linear regression, logistic regression, linear discriminant analysis, decision trees, k-nearest, neighbor algorithm, gradient boosting machine. Each model has its advantages, but in this thesis, the author uses three algorithm models as a comparison to the predictions that will be used on the log data sheet.

2.1. Logistic Regression

Logistic regression to find the value true or false. It can be shown Eq. (1).

$$Y = \alpha + \beta X + \mathcal{E} \tag{1}$$

Y= Dependent Variable; α = Constanta; β = regression coefficient; X = error model

Linear regression is a method for modeling the relationship problem between an independent variable and the dependent variable. For the model, we can apply it to the case of knowing the character of Facebook users, who are pro-government or opposition people. Surely we can detect it by seeing which pages are followed, and how the Facebook users update each comment or column on their wall.

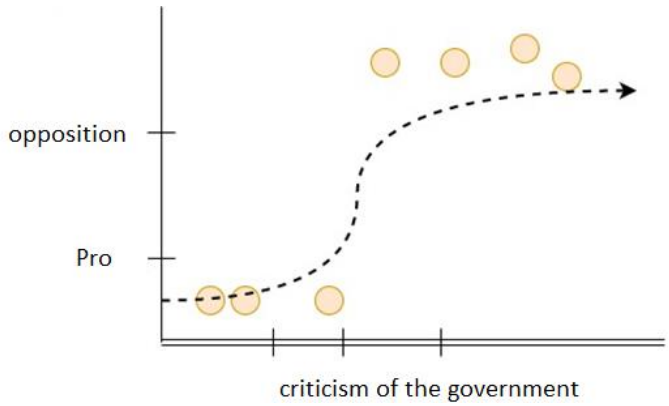

Fig. 2: Scatter Plot analysis of the pros and cons of government

In Fig. 2, the curve is made in the shape of the letter "S", a different shape from linear regression. In the table description, we can declare the option with a value of 0 and the pro-government is a value of 1. This is where the curve represents Facebook users who are in the pro or opposition category. If on this curve there are users who are very frontal in criticizing the government, then these users will be grouped into opposition. Fig. 3, is the following curve

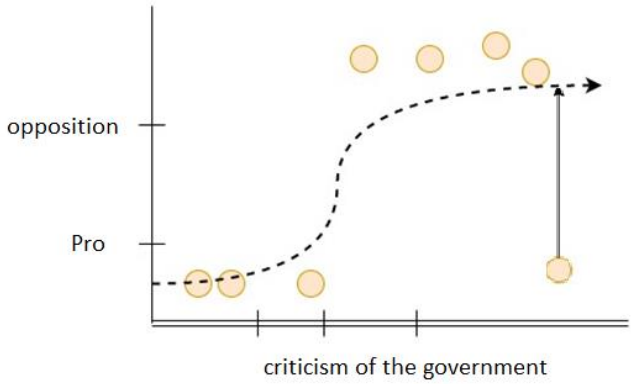

Fig. 3: Scatter Plot of social media users are very critical

2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) was first introduced by Vladimir Vapnik, Bernhard Boser, and Isabelle Guyon in 1992 at the Annual Workshop on Computational Learning Theory. SVM is the 10 best method category in data mining, this method makes SVM a promising new method for data classification, both linear and non-linear. The concept of SVM is a combination of computational concepts that have existed for decades, such as the hyperplane. The basic concept of the Support Vector Machine (SVM), namely maximizing the hyperplane limit (maximal margin hyperplane), here is an example of an image from a data sheet for Ubuntu and Redhat Linux users, which uses a non-standard hyperplane model 2.3 (a) and in Fig. 2.3 (b) uses hyperplane with a maximum margin which will produce better generalizations to the classification method.

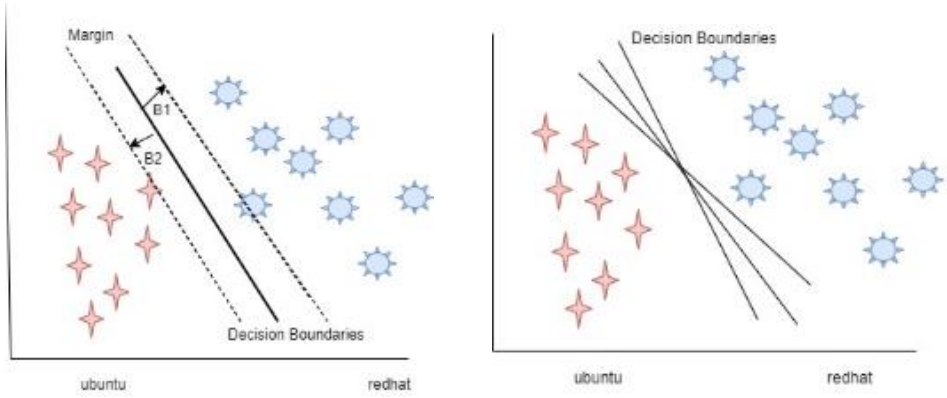

Fig. 4: Decision Boundary data sheet for Linux users

SVM is used to find the best hyperplane by maximizing the distance between classes. The hyperplane is a function that can be used to separate classes. To define

boundaries the simplest form of the linear model is as follows as shown Eq. (2).

$$y_{(x)} = wTx + w0 \quad (2)$$

Where *x* is the input vector, *w* is the weight vector and w0 is bias. Thus, the decision boundary is $y_{(x)} = 0$, which is a dimensionless hyperplane (D-1). An input vector x will be classified into class 1 (R_1) if $y_{(x)} \geq 0$, and class 2 (R_2) if $y_{(x)} < 0$. While the characteristics of the hyperplane: **1.** If x_A and x_B lie on the decision boundary (DS), then $y(x_A) = y(x_B) = 0$ or $w_T(x_A - x_B) = 0$, so it is perpendicular to all vectors in DS. In other words, w determines the orientation of the DS. **2.** The distance from the starting point to DS is $-w0 / ||w||$. In other words, w0 specifies the DS location. **3.** The distance of any vector x to DS and with the direction of w is $y_{(x)} / ||w||$

2.3. Gradient Boosting Machine

Gradient boosting is a machine learning algorithm that uses an ensemble of a decision tree to predict values. Gradient boosting can handle complex patterns and data when the linear model cannot handle it. Model development is done using the boosting method, namely by creating a new model to predict the error / residual from the previous model. New models are added until no more error fixes can be made. This algorithm is called gradient boosting because it uses gradient descent to minimize errors when creating a new model. Another study then made minor changes to regularized objectives on GBM. XGboost has many advantages including, it can perform parallel processing which can speed up computation, has high objective setting flexibility, built-in cross-validation, has regularization features, and overcomes splits when negative loss. This scalability is due to the optimization of the previous algorithm. This success was proven when XGBoost became one of the methods that were being widely applied to various cases in machine learning. Before we create it in equation form, we try to simulate a simple table and how gradient boosting works. In Table 1, we set 2 tables with X and Y columns.

Table 1: Example of table x and y

***X* (input)**	***Y* (Prediction)**
4	60
8	70
12	80
16	90
20	100
24	110
28	120
32	180
36	140
40	160

After that, we will look for the residual value, namely by adding the column average of the prediction (*y*) with the name F0. After finding this value, we will get

the residual value, obtained from the value of $y_{(x)}$ - $f0_{(x)}$.

Table 2: Example table for residual values

X	***Y***	***F0***	***y-f0***
4	60	111	-51
8	70	111	-41
12	80	111	-31
16	90	111	-21
20	100	111	-11
24	110	111	-1
28	120	111	9
32	180	111	69
36	140	111	29
40	160	111	49

Then we can use the remainder of $F_{0(x)}$ to find the value of $h_{1(x)}$. $h_{1(x)}$ becomes the regression value that will try and subtract the residual from the previously obtained step. However, the output value of $h1_{(x)}$ will not be a predictive function of y; which in turn, will aid in the prediction of the successive function $F_{1(x)}$ to decrease the residual value. In table 3, we divide, where values <20 and values > 20 become different classifications. The value obtained is -31. From table 3, there we have calculated F_1 and y-f_1. Where the remainder of the increased output (y-$F_{1(x)}$) is used to create the next tree $h_{2(x)}$ and the increased output is $F_{2(x)}$ which is obtained by adding $F_{1(x)}$ and $h_{2(x)}$.

Table 3: Example of Table tree regression on Gradient Boosting

X	***Y***	***F₀***	***y-f₀***	***h₁***	***F₁***	***y-f₁***
4	60	111	-51	-31	80	-20
8	70	111	-41	-31	80	-10
12	80	111	***-31***	-31	80	0
16	90	111	-21	-31	80	10
20	100	111	-11	-31	80	20
24	110	111	-1	31	142	-32
28	120	111	9	31	142	-22
32	180	111	69	31	142	38
36	140	111	29	31	142	-2
40	160	111	49	31	142	18

Will be repeated until it is found that the function of the error does not change, or the maximum number of predictions can be reached. It can be shown as Eq. (3), Eq (4), Eq. (5), and Eq. (6)

$$Fm+1_{(X)} = Fm_{(X)} + hm_{(X)} = y \quad (3)$$

OR

$$Hm_{(X)} = y - Fm_{(x)} \quad (4)$$

$$LMSE = ½ (Y - f_{(X)})2 \quad (5)$$

$$Hm(X) = -\frac{\partial L_{MSE}}{\partial F} = y - F_{(X)} \quad (6)$$

Hence, the residual for a given model is the negative gradient of the mean squared error (MSE) loss function and is a similar process to that of the gradient descent algorithm. Gradient increment does not change the distribution of the sample because a weak input x value trains a strong residual x-value error (pseudo residue).

3. Research Methods

The initial purpose of writing this thesis is to see the comparison of the prediction results from the three algorithm models. In the conceptual process until this thesis is completed, if the flow is made in the form of an image, it is as follows:

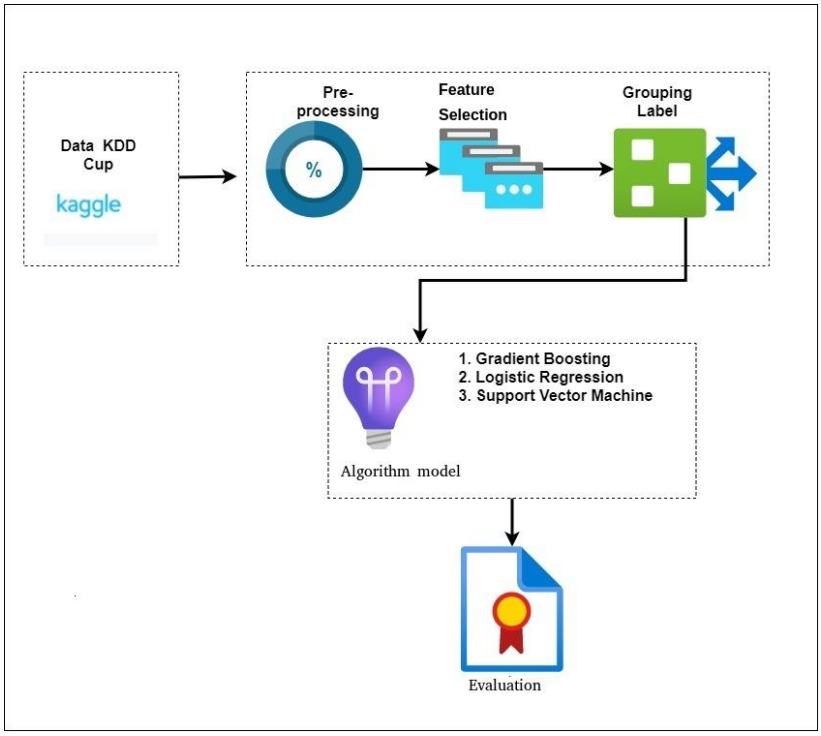

Fig. 5: Research Method

In Fig. 5, Research Method with Preprocessing is one of the important stages for data in the mining process. Data used in the mining process usually have values that must be uniform, for example, string values must be numeric.

3.1. Feature Selection

Due to the IDS data sheet taken from the Kaggle portal *[http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html]*, a selection of columns that are considered important in the machine learning prediction function is selected. This process aims to choose which features will greatly affect the prediction results later. Therefore, not all data is processed in the detection process using a learning model that will look for predictive values. Table 4 shows the feature selection used.

Table 4: Feature Selection

Index	**Feature name**	**Description**
1	Duration	Length of connection
2	Service	Destination service (ftp, telnet...)
3	Source bytes	Number of bytes from source to destination
4	wrong fragments	Number of wrong fragments
5	Count	Number of connections as the current connection to the same host at a given interval
6	Urgent	Number of urgent packets
7	compromised	Number of compromised states
8	service count	Number of connections as the current connection to the same service at a given interval

3.2. Evaluation

In the classification, each test will calculate the accuracy with a confusion matrix. Confusion matrix for calculating large can be shown Table 5.

Table 5: Confusion matrix

		Actual	
		True	**False**
Prediction	True	TP	FP
		(true positive)	(false positive)
	False	FN	TN
		(false negative)	(true negative)

Precision, recall and F1-score can be shown as Eq. (8), Eq. (9) and Eq. (10)

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

$$F_1\text{-score} = \frac{2*PrecisionxRecall}{(Precision+Recall)} = \frac{2TP}{2TP+FP+FN} \quad (10)$$

F_1-score shows the weighted harmonic mean comparison of the weighted average of precision and recall. Confusion matrices can be used to measure performance in binary classification problems as well as multiclass classification problems. Binary classification only produces two-class outputs (labels), such as "Yes" or "No", "0" or "1" for each given input data. The main class is usually denoted as positive data and the others as negative data. The metrics used in this study are overall accuracy, average precision, average recall, average F1-score with classification reports, and confusion matrix. In imbalance class, precision, recall, and F_1-score more describe the accuracy of prediction. The confusion matrix will provide detailed information from each class between true to prediction. Classification reports provide precision, recall, F_1-score information per class.

4. Results And Discussion

Raw data sheet from IDS doesn't have a definition for each column, it will be defined manually, with the following information.

colnames=["duration","protocol_type","service","flag","src_bytes","dst_bytes","land","wrong_*fragment","urgent","hot","num_failed_logins","logged_in","num_compromised","root_shell","su_attempted","num_root","num_file_creations","num_shells","num_access_files","num_outbound_cmds","is_host_login","is_guest_login","count","srv_count","serror_rate","srv_serror_rate","same_srv_rate","diff_srv_rate","srv_diff_host_rate","una1","una2","dst_host_count","dst_host_srv_count","dst_host_same_srv_rate","dst_host_diff_srv_rate","dst_host_same_src_port_rate","dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate","dst_host_rerror_rate","dst_host_srv_rerror_rate","result"]*

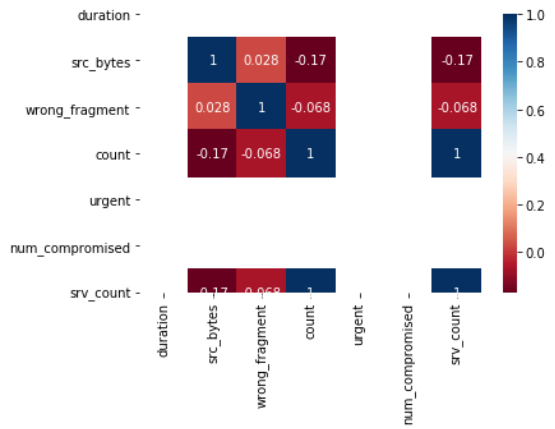

Fig. 6: Plot of correlation to Pearson feature extraction

Fig. 6 shows how each data effect in visual form using seaborn. Then the program will perform an iterative method where the model data is obtained from the variable models in the previous line script. First of all, do a fit model of the previous training data. Model fit states the level of fit (fit) of the research model with the ideal model for that research. This model is imaginary, cannot be described but exists. Then define y_pred as the prediction of the test process on the X_test variable. The next step is to find the accuracy of the tested y_test and y_pres values, where the variable a score. The value obtained will be inputted into the score. The results of the algorithm model carried out on 3 different protocols can be shown Table 6. Based on table 6, it shows that the gradient boosting model is the one with the highest prediction. Because of the 3 different conditions, the highest boosting gradient of the 3 learning models. It should be noted, in the TCP protocol, when running the SVM model, the processing time until the data is found reaches 26 minutes. While the Gradient boosting and logistic regression models only need 2 seconds.

Table 6: Result three model Algorithms

Protocol	**Model**	**Accuracy**	**Services**
ICMP	Gradient Boosting	99.99	4
	logistic Regression	99.95	
	SVM	100	
UDP	Gradient Boosting	71.48	5
	Logistic Regression	71.03	
	SVM	71.01	
TCP	Gradient Boosting	99.87	58
	logistic Regression	88.78	
	SVM	99.21	

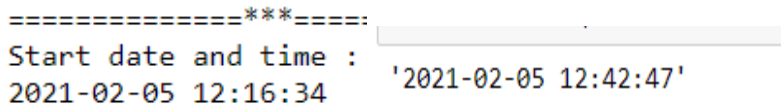

Fig. 7: Time stamp start and finish of the SVM algorithm model

Fig. 7 shows the result of the Support Vector Machine process since the start and end process. Therefore, even though the predicted value in the SVM model is also high. However, according to the research results especially on time process, it is not a recommendation for checking attack traffic on the system.

5. Conclusion

Based on the experiment, we can make the following conclusions: **1**. Each learning model has a different predictive value, in the case of detection the ICMP protocol was

found to use the highest SVM algorithm model, reaching 100% for its prediction. However, when trying to use a different protocol with more service types, the process to get predictions from the SVM algorithm model is very slow. **2**. The use of the IDS data sheet using the Gradient Boosting algorithm model, Regression Logistics, and SVM, for the 2 protocols has a predictive value above 90%. However, in terms of speed and prediction performance, the authors recommend using gradient boosting. **3**. The number of classes in the protocol affects the results of each algorithm model, as we know in the first test the use of the protocol is only 4 flags, and the result of SVM is 100%, but in testing with the UDP protocol already using 5 Flags, the use of a logistic learning model regression with higher prediction reached 71.01%.

Reference

Aishwarya, c., Venkateswaran, N., Sreekar, T. S., and Sreeja, V. (2020). Intrusion Detection System using KDD Cup 99 Dataset. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 99.9996.

Azar, A., Sebt, M., Ahmadi, P., and Rajaeian, A. (2013). A model for personnel selection with a data mining approach: A case study in a commercial bank. *SA Journal of Human Resource Management*, 11(1).

Chen, P. C. (2009). A Fuzzy Multiple Criteria Decision Making Model in Employee Recruitment. *IJCSNS International Journal of Computer Science and Network Security*, 9(7), 113-117.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD'16 Proceedings of the 22nd ACM SIGKDD International*, 785-794.

Chien, C. F., and Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34(1), 280-290.

Dartigue, C., Jang, H. I., and Zeng, W. (2009). A New data-mining based approach for network. *Proc. of Seventh Annual Communication Networks and Services Research Confrence*, 372-377.

Duda, O, R., and Hart, P. E. (1973). Pattern Classification and scene Analysis. *Stanford Research Institue*.

Fernandes, A. A., Filho, D. B., Rocha, E. C., and Nascimento, W. d. (2020). Read this paper if you want to learn. *Revista de Sociologia e Política*, 2.

Gao, X., Wen, J., and Zhang, C. (2019). An Improved Random Forest Algorithm forPredicting Employee Turnover. *Mathematical Problems in Engineering*, 1-12.

Guion, R. M., Highhouse, S., and Doverspike, D. (2016). *Essentials of Personel Assesment and Selection.* New York: Routledge.

Haskett, J. (2000). *Corporate Culture and Performance. New York: Free Press.*

Karatop, B., Kubat, C., and Uygun, O. (2014). Talent management in manufacturing system using fuzzy logic approach. *Computers and Industrial Engineering*.

Kelemenis, A., and Askounis, D. (2010). A New TOPSIS-based multi-criteria approach to personel selection. *Expert Systems with Applications*, 4999–5008.

Khorami, M., and Ehsani, R. (2015). Application of Multi Criteria Decision Making approaches for personnel selection problem: A survey. *International Journal of Engineering Research and Applications*, 5(5), 14-29.

Kim, H.-Y. (2018). Statistical notes for clinical researchers: covariance and correlation. *Restor Dent Endod*, 1.

Kumari, K., and Yadav, S. (2018). Linear Regression Analysis Study. *Journal of the Practice of Cardiovascular Sciences*, 33-34.

Luque, A., Carrasco, A., Martin, A., and de las Heras, A. (2019). The Impact of Class Imbalance in Classification Performance Metrics Based on The Binary Confusion Matrix. *Pattern Recognition*, 216-231.

Lytvyn , V., Vysotska, V., Pukach, P., Bobyk, I., and Pakholok, B. (2016). A Method for Constructing Recruitment Rules Based on The Analysis of a Specialist's Competences. *Eastern-European Journal of Enterprise Technologies*, 4-16.

Magdalena, L. (2015). Fuzzy Rule-Based System. In *Evolutionary Computation and Constraint Satisfaction* (pp. 203-218). doi:10.1007/978-3-662-43505-2_13

Mammadova, M. H., and Jabrayilova, Z. G. (2018). Decision-Making Support in Human Resource Management Based on Multi-Objective Optimization. *TWMS Journal of Applied and Engineering Mathematics*, 9(1), 55-72.

Mammadova, M., and Jabrayilova, Z. (2014). Application of Fuzzy Optimization Method in Decision-Making for Personnel Selection. *Intelligent Control and Automation*, 190-204.

Masum, A. K., Beh, L. S., Azad, A., and Hoque, K. (2018). Intelligent Human Resource Information System (iHRIS): A Holistic Decision Support Framework for HR Excellence . *The International Arab Journal of Information Technology*, 121-130.

Michel, V., Gramfort, A., Varoquaux, G., Evelyn, E., Keribin, C., and Thirion, B. (2011). A supervised clustering approach for fMRI-based inference of brain states. *ScienceDirect Journals*.

Morgan, D. B. (2019). *Management Strategies for Reducing Voluntary Employee Turnover in Small* . Walden University, Doctoral Dissertation . Retrieved from

Nasteski, V. (2015). An overview of the supervised machine learning methods. *ICTACT Journal on Soft Computing*, 4.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., and Clune, J. (2017). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*.

Pal, A. K., and Pal, S. (2013). Evaluation of Teacher's Performance: A Data Mining Approach. *International Journal of Computer Science and Mobile Computing*, 2(12), 359-369.

Patel, H. H., and Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering*, 74.

Pedregosa, F., Varoquaux , G. , Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2827.

Schneider, S., Greenberg, S., Taylor, G., and Kremer, S. (2020). Three Critical Factors Affecting Automated Image Species Recognition Performance for Camera Traps. *Ecology and Evolution*, 1-15.

Serhadlıoğlu, G., Güngöra, Z., and Kesen, S. E. (2009). A fuzzy AHP approach to personnel selection problem. *Applied Soft Computing,* 9(2), 641-646.

Shapiro, G. N., and Allman, E. (1999). Sendmail Evolution: 8.10 and Beyond. *Proceedings of the FREENIX Track.* 2-3.

Sokolova, M., and Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Informations Processing and Management*, 427-437.

Tabak, M., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., . . . Bought. (2019). Machine Learning to Classify Animal Species in Camera Trap Images: Applications in Ecology. *Methods in Ecology and Evolution*, 585-590.

Tai, W.-S., and Hsu, C.-C. (2006). A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method. *9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press. doi:10.2991/jcis.2006.46

Verzello, R. J., and Reuter III, J. (1982). International Student Edition. *Tokyo : The McGraw-Hill Companies*.

Wirth, R., and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). Citeseer.

Zadeh, L. A. (1985). Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. *IEEE Transaction on Systems, Man, and Cybernetics*, 15(6), 754-763.

Zin, T. T., Phyo, C. N., Tin, P., Hama, H., and Kobayashi, I. (2018). Image Technology based on Cow Identification System using Deep Learning. *International MultiConference of Engineers and Computer Scientists.* Hongkong.