JAES JOURNAL OF APPLIED ENGINEERING SCIENCE www.engineeringscience.rs

ISTRAŽIVANJA I PROJEKTOVANJA ZA PRIVREDU

Indexed by

IMPLEMENTATION OF MACHINE LEARNING FOR HUMANA ASPECT IN INFORMATION SECURITY AWARENESS

Valentina Siwi Saridewi

University of Indonesia, Faculty of Engineering, Department of Electrical Engineering, Depok, Indonesia

Riri Fitri Sari

University of Indonesia, Faculty of Engineering, Department of Electrical Engineering, Depok, Indonesia

Key words: *classification, clustering, python, information security awarenessy*

doi:10.5937/jaes0-28530

Cite article:

Siwi Saridewi V., Sari Fitri R. (2021) IMPLEMENTATION OF MACHINE LEARNING FOR HUMAN ASPECT IN INFORMATION SECURITY AWARENESS, *Journal of Applied Engineering Science*, 19(4), 1126 - 1142, DOI:10.5937/ jaes0-28530

Online aceess *of full paper is available at: www.engineeringscience.rs/browse-issues*

doi:10.5937/jaes0-28530

Paper number: 19(2021)4, 895, 1126-1142

IMPLEMENTATION OF MACHINE LEARNING FOR HUMAN ASPECT IN INFORMATION SECURITY AWARENESS

*Valentina Siwi Saridewi, Riri Fitri Sari**

University of Indonesia, Faculty of Engineering, Department of Electrical Engineering, Depok, Indonesia

This research discussed our experience in implementing machine learning algorithms on the human aspect of information security awareness. The implementation of the classification and clustering approach have been conducted by creating a questionnaire, creating dataset, importing data, handling incompleted and imbalanced data, compiling datasets, feature scaling, building models, and subsequently evaluating machine learning models. Datasets are generated based on the collection of questionnaire result of the distributed questionnaire related to the Human Aspects of Information Security Questionnaire (HAIS-Q) to the stakeholder of an Indonesian institution. Models as results of algorithms implementation through the classification approach has been evaluated by several methods, such as: k-fold Cross Validation analysis, Confusion Matrix, Receiver Operating Characteristics, and score calculation for each model. A model of the Support Vector implementation in the classification has an accuracy of 99.7% and an error rate of 0.3%. Models of clustering implementation are used to determine the number of clusters that can optimally divide the dataset. The model of the DBSCAN algorithm on the clustering approach has an adjusted rand index value of always close to 0.

Key words: classification, clustering, python, information security awareness

INTRODUCTION

The growth of Information and Communication Technology (ICT) in Indonesia is overwhelming, especially during the Covid-19 pandemic. Many offline activities become online, thereby increasing the internet usage. Threats and attacks on the cyber world are multiplaying with the increase of the ICT usage in society. This is related to the security aspects of the cybersecurity model [2]. A threat is a vulnerability to control physical infrastructure, information sources, workers, and market values. Information security intends to maintain the availability of information resources when stored, transferred, and processed data against attacks attempt to destroy data and access by unauthorized parties [3]. McCumber Cube (MC) shows a three dimensional cybersecurity model that consists of status information, critical information characteristics, and security measurements [4]. The dimension of security measurement consists of technology, policy and application, and human aspects. Human aspects of these dimensions involve education, training, and awareness.

One of the causes of vulnerability to information security is related to the human aspect. It occurs due to technical errors or social engineering practice [5]. The responsibility for information security rests with the individual or group that arranges and uses that information. The information held can be used to make crucial decisions that need to prioritize confidentiality, integrity, and availability. It shows that the human aspect has a significant role. Awareness of information security emerges from understanding the threats and vulnerabilities in the usage of ICT. The understanding will encourage effective security control measures.

Machine learning was already been extensively implemented in various applications, one of which is cybersecurity. It is applied to observe data related to phenomena in the cyber world. Data is used to be studied to gain an understanding of events to build a model. The resulting model was then used to predict new data and detect anomalies to observed the phenomena [6]. Generally, machine learning is implemented for intrusion detection, malware analysis, and spam detection [7].

The purpose of our research is to implement machine learning algorithms on the human aspects of Information Security Awareness (ISA). We built models based on the point of view of classification and clustering. We determined the appropriate model based on the results of a questionnaire referred to the Human Aspects of Information Security Questionnaire (HAIS-Q) for each approach.

The challenges in this research are as follows: multidimensional data; achieve balances of data in each class of model for the classification; performance of the algorithm; define parameters, and determine the number of clusters for clustering. Many instruments used on HAIS-Q caused the dataset to have multidimensional characteristics. Multidimensional data affects the performance of the algorithm in machine learning. It is necessary to select valuable instruments and assure the minimum risk of losing information. Define parameters used in the algorithm for each approach. Finally, models are analyzed within the field of information security awareness.

The main contribution of our research is related to the information security awareness, especially on the human aspect. These contributions include implementing the dataset that has multidimensional characteristics from the point of view of classification and clustering. Thus,

*riri@eng.ui.ac.id

we produce models, determine the optimal epsilon in the DBSCAN algorithm, and generate analysis of the model of information security awareness.

INFORMATION SECURITY ON HUMAN ASPECTS AND MACHINE LEARNING

Information security awareness research

In this work, we presented some literature reviews on information security awareness. We performed a thorough literature review on the theory validation, statistical approaches, and classification methods. The literature review identified some factors that have the main effects of the respondent behavior on information security. Alohali et al. have proposed several strategies to improve information security awareness [8]. They found that individuals need to understand the risks and change their behavior in improving information security awareness.

Literature review on the theory of validation method utilized partial least squares structural equation modeling to validate measurements and hypotheses through analysis of the employee information security awareness in the workplace [9]. Bauer et al. work shows that information security behavior within employees increases through internal and external ways. It has an impact on improving their attitudes towards information security and norms.

Research in building information security models has been done with the Analytic Hierarchy Process [10] to map employee information awareness levels in an institution. Another research related to information awareness utilized machine learning. It used classification approaches using Principle Component Analysis (PCA) [11] and the J48 algorithm [12]. The research that uses PCA has identified areas of student attention toward information security. It is related to personal, social, institutional, and technological conditions. Carella's work related to security awareness used the J48 and Random Committee algorithms to map user profiles to predict phishing victims [12].

The Human Aspects of Information Security Questionnaire (HAIS-Q) is a popular tool used in the information security awareness field. Most research relates to HAIS-Q use a statistical approach [13] [14] [15] [16] to map parameters on questionnaire results. Mapping applied to employees of an institution or area has been conducted.

HAIS-Q questionnaire as a reference in building questionnaire to form our dataset. HAIS-Q consists of 63 statements which created a multidimensional dataset. The multidimensional dataset has an impact on the time and complexity of the algorithm on machine learning performance. The Hughes phenomenon shows the performance of algorithms related to data dimensions. It stated that the predictive ability and effectiveness of the algorithm decrease with the increase of data dimensions. The more features used will lower the performance and accuracy of the algorithm. Feature selection has been used to reduce irrelevant data, unnecessary feature extraction, and determine a subset of relevant features based on specific criteria [17] [18].

Therefore, the dataset is necessary to adjust the number of features to fit data processing requirements. The reduction is considering the risk of the lost information to a minimum. This is performs through feature selection. HAIS-Q parameters were reduced by selection features using the Principle Component Analysis and Multi Cluster Feature Selection (MCFS) approaches [19]. The feature selection resulted in 21 questionnaire statements which are to be taken based on the MCFS score of knowledge, attitude, and behavior. Each score is related to parameters in each sub area such as knowledge, attutide and behavior.

This research uses the MCFS results to collect data by new questionnaire and build a dataset and models. Models is built using several algorithms with classification and clustering approaches. Classification is part of supervised learning and clustering is part of unsupervised learning in machine learning. In the classification approach, a model is selected based on the most accurate result. Models on the clustering approach have different handling mechanism to analyze the clusters from the dataset.

Information security

Implementation of security is a process of protecting and securing assets. Information can be an asset that has value and evolve. Security in information consists of several components, i.e., confidentiality, integrity, and availability. The National Institute of Standards and Technology (NIST) defines information security. It is an effort to protect information and information systems by ensuring the confidentiality, integrity, and availability of business access, usage, exposure, disruption, modification, or destruction caused by unauthorized parties [20]. John McCumber developed the MC model that consists of 3 dimensions to show relationships between computer security and communication to describe elements involved [4]. The model can adapt to the information environment regardless of the technology involved. Each dimension consists of 3 layers, i.e., information status, critical information characteristics, and security measurement. Information status dimension related to information protection is according to the conditions traversed. Information condition related to transmission, storage, and processing of data. Dimensions of the characteristics of critical information is associated with the cybersecurity principles which include confidentiality, integrity, and availability. The security measurement dimension is related to the assurance that it maintained the crucial information.

The dimensions of security measurement consist of technology, policy and application, and human aspects. Technology can be either hardware or software used to ensure the characteristics of critical information in protecting information under many conditions. Information security policy arranges procedures and guidelines. It intends to protect valuable information.

The human factor comprises of education, training, and awareness of security actions. Technology and policy are very dependent on education, training, and awareness related to security. Someone who understands threats and vulnerabilities in information systems should take forceful measures to control.

Policies must be able to regulate and limit the distribution of information resources selectively and securely. Guaranteed information sources distribution has to do with the awareness of individuals or society that use information. Individuals or society of information usage has a responsibility to protect the information in making crucial decisions.

Information security measurement

Parsons et al. developed HAIS-Q as a measurement tool for information security awareness that focuses on the human aspect [21].

Human being is one aspect of vulnerability in information security. Information security awareness consists of two factors. First, an understanding of safety information security behavior. Second, individual commitment and behavior to implement information security properly. It is consistent with the Knowledge Attitude Behavior (KAB) model [21].

The research conducted by Parsons holistically developing HAIS-Q uses the Knowledge Attitude Behavior (KAB) model approach. HAIS-Q consists of 7 focus areas: keyword management, email usage, internet usage, social media usage, device devices, information handling, and incident reporting. Each focus area consists of 3 subareas. Knowledge influences attitudes are reflected in individual behavior. Therefore, HAIS-Q statements are arranged in the order of knowledge, attitudes, and behavior towards each subarea. Each subarea has three statements that represent the interconnected KAB model. So, it makes HAIS-Q produced 63 statement instruments.

The HAIS-Q research shows a correlation between knowledge level and the increasing attitude of information security awareness. Information security knowledge level on an individual has an impact on the behavior in implement information security policies and procedures. HAIS-Q is designed in a modular manner so that the focus area can be adjusted according to the usage requirements.

The scale of information security awareness is viewed from a management perspective [14] [22] which consists of 3 levels: good, average, and low. The information security awareness level is good, indicating that individuals have an understanding of information security awareness. Individuals need the reinforcement of information security awareness on the average level. The lower level is individuals who need guidance on information security awareness.

Classification and clustering

Machine learning is a computer programming system built to optimize algorithms performance of using data or experience as the input to produce a model that functions to predict specific criteria [23] [18]. The model as an output represents how the algorithm has learned about data. Models can be analyzed deeper to provide predictions on new data, decisions, and insights.

The dataset used to construct a model consists of samples. Samples contain features and labels. Feature as variables (x). The label is determinants or things that need to predict (y) [24]. In machine learning, it has two categories, i.e., with and without the label. The data usage in statistical analysis is to read trends when machine learning use to find patterns so that it does not require explicit statistical evidence [18].

Machine learning can be categorized based on the input type in the process and output as needed. It includes supervised learning and unsupervised learning [25] [26]. One of the methods for supervised learning classification is to study the mapping of datasets into predictive models. It uses discrete values that correspond to the determinants associated with labels. The number of classes in classification can vary with a minimum number of two (binary) up to more than two (multiclass). In multiclass, each sample can only be a member of a class. Some algorithms in this category include logistic regression, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF).

Logistic regression is a linear probability modeling process on a discrete classification. It has some results ranged from 0 to 1.

Algorithm k-Nearest Neighbors (k-NN) is a simple algorithm based on k number nearest neighbors using distances such as Euclidean, Manhattan, Minkowski, or Mahalanobis. The value of *k* increases then the boundary between classes becomes seamless. The chance for misclassification increases and underfitting rises [27]. Rules related to k value, i.e., can not exceed the number of data train, an odd number to avoid the problem of two classes, does not represent a multiple of the classes number.

Support Vector Machine (SVM) algorithm comes under classification algorithm for analyzing data and recognizing patterns. It uses hyperplane lines to separate the two classes optimally, then a maximum margin is obtained [26].

Hyperparameter setting is a process to specify the best parameter combination to use in a model based on a dataset. The purpose is to get a model which yields optimal results that the performance scores have a high accuracy degree.

Various approaches to evaluate classification model performance, such as Cross Validation (CV) of data separation, Confusion Matrix (CM), Receiver Operating Characteristics (ROC), and scores include accuracy, precision, recall, and others. The evaluation model goals are as follows: knowing the predictive performance of the model against data that has never been used, improving the predictive performance of a model, identifying machine learning algorithms that are appropriate for the problem, and choosing the model with the best performance [28]. k-fold cross validation is a procedure for estimating the ability of a model for new data by dividing the data train into k number folds.

Confusion Matrix describes classification model performance which consists of actual and predictive data based on the algorithm used [28]. CM in a multiclass datasets is shown in a two-dimensional, matrix for each class. True negative (TN) indicates that actual and predictive values are false. In this case, its value shows the achievement of the model in recognizing samples and not classifying them into unsuitable classes. True positive (TP) is the success of a model in predicting classification samples into proper class so the actual and predicted values are true. Prediction errors consist of 2 types, such as false positive (FP) as the error type 1 and false negative (FN) as the error type 2. Error type 1 is an error that the prediction is true while actual test data is false. Error type 1 fails to detect samples in certain classes. Error type 2 provides prediction errors by giving an incorrect value to the actual value as true. FN is an error in recognizing samples in certain classes. The model is considered good with high TN and TP values and the prediction error values on FN and FP are low.

Receiver Operating Characteristic (ROC) is a tool to describe the performance of classification algorithm models. It is shown in the form of two-dimensional graphics. ROC presents a true positive rate and a false positive rate. Area Under Curve (AUC) shows the probability that a positive result chose randomly will be rated higher by the classifier than a negative one chose randomly. AUC presents the area calculation under the ROC curve with a range of values from 0 to 1.

Accuracy is one of the assessments of classification models. Accuracy shows the percentage of success of the model correctly predicted (Equation 1). Its calculations are based on the ratio of precise classifications to the total number of samples in the dataset. True positive (TP) is the sample number in which predictions and actual are true. True Negative (TN) is the number of samples in which predictions and actual values are false. False Negative (FN) is when the number of samples is true in the prediction but false on the actual.

$$accuracy = \frac{(TP+TN)}{(TN+TP+FN+FP)} \tag{1}$$

Sensitivity/true positive rate/recall in Equation 2, describes the success of the model in predicting samples positively [26]. Specificity/true negative rate is the opposite of recall which is defined as a successful model to predict that the sample does not include a specific class. Models that do not produce FN will have a recall of 1.

$$sensitivity = \frac{TP}{(FN+TP)} \tag{2}$$

Prediction values describe the predictive model performance. Prediction of a positive value/precision (Equation 3) illustrates the accuracy of the model in predicting it in a specific class.

$$precision = \frac{TP}{(TP+FP)} \tag{3}$$

F1 score shows the combination of precision and recall values. It calculates the average weights of precision and recall that serve to evaluate the quality of model output.

$$F1\ score = 2\left(\frac{(precision*sensitivity)}{(precision+sensitivity)}\right) \tag{4}$$

Balanced accuracy describes the accuracy of a data balance, particularly imbalanced data in multiclass through the average recall value in each class. Balanced accuracy is good if it reaches 1.

$$balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{5}$$

Clustering is one of the methods in unsupervised learning. It only has input data without determinant or divides the dataset into several groups that have similarities. The clustering algorithm has many categories [29] that include partitional clustering, hierarchical clustering, and density-based clustering [30]. Partitional clustering uses an algorithm with a concept to group data with data points center (centroids) of corresponding groups. The advantages of this algorithm are low time complexity and computational efficiency. The weakness of clustering is necessary to determine the number of clusters. The clustering results are sensitive to the number of clusters. One of the clustering algorithms included in this type is K-Means.

K-Means algorithm divides the dataset into several groups with a similar variant (inertia) that only uses Euclidean distance. It has two purposes, i.e., to determine the center point of each cluster and each sample included in a cluster. The model uses this algorithm to determine the number of clusters. Some approaches to determine the number of clusters are silhouette analysis, davies bouldin score, calinski harabasz index, and the elbow method [31] [26] [32]. Silhouette analysis provides data about the distance between the group and the resulting group. The silhouette plot shows the distance between each point in a cluster with the points on the neighbor cluster. The distance measurement is between -1 to 1. The coefficient value that is getting closer to 1 show that a sample position is farther from the neighbor cluster. A value of 0 indicates that a sample position is in the nearest cluster where a negative value indicates that the sample is in the wrong cluster.

The Davies Bouldin Index (IDB) is used to guide the cluster search algorithm. It shows the similarity in the distance between clusters and the internal clusters. IDB values 0 means the lowest on an index that indicates the closest value as a good partition. The Calinski Harabasz Index (CHI) approach uses the ratio of dispersions number between clusters also the intercluster dispersions for all clusters.

Hierarchical clustering is an algorithm that constructs a hierarchy between data as cluster trees for grouping. Some of the advantages are data can be grouped based on some forms and attributes. Hierarchical clustering can detect hierarchical relationships between clusters, and has a high scalability. The disadvantages are the number of clusters must be determined, and the time complexity required is generally high. One algorithm in this type is agglomerative clustering.

Agglomerative clustering starts the clustering samples in a dataset, then combines the closest cluster or have similarities into one [29]. The process stops until it forms a cluster. Visualization in determining the number of clusters in agglomerative clustering uses a dendrogram that describes the cluster structure with a bottom up approach.

Density based clustering groups samples are based on the density of area considered is a cluster. The advantages are that it is very efficient in clustering and suitable for constantly changing data. The disadvantage of this density-based clustering is that the quality is low for varied region's densities and those with a high data dimension. This algorithm requires a high memory for large scale data. Besides, cluster results are sensitive to the parameters used. One algorithm in this type is Density Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN is an algorithm that identifies clusters based on the measurement of the distribution of sample densities in an area [33]. It uses the concept of core samples in high-density areas as the main component.

Some performance metrics have been used as an approach to determine the validity of the model generated by the algorithm in clustering [34]. Samples in a cluster have more similarities compared with other samples. Adjusted Rand Index (ARI) is a function for measuring the similarity of two labels by ignoring permutations and normalization opportunities. ARI value is perfect if it has a value of 1. The ARI value is low for a negative value.

Metric evaluation is a way to determine the validity of the model generated by the algorithm in clustering through homogeneity, completeness, and of v-measure [34]. The homogeneity indicator shows that the sample is only part of a group. The completeness indicator shows that each sample in the dataset has become part of a cluster. Harmonic homogeneity and completeness are the result of the v-measure. The range of evaluation values of the metric is between 0 to 1. The perfect value on the evaluation of the metric is 1, and 0 is the opposite. Cardinality clusters provide a picture of the number of samples in a group [24].

Framework and open source tools

Build models in machine learning can be created base on the framework development guidance. One of the most popular frameworks is the CRoss Industrial Standard Process for Data Mining (CRISP-DM) [26]. CRISP-DM process consists of defining business, defining data, preparing data, modeling, evaluating, and implementing.

Python is a powerful open source programming language intended to be used for various requirements to be applied for many application domains. Commonly, Python is used in machine learning research. It is supported by package libraries in accordance with requirements such as scikit-learn, NumPy, SciPy, Pandas, seaborn, etc. Anaconda manages various versions of Python in research environment and provides data science libraries. Jupyter Notebook has been used as an editor on Anaconda to support Python programming language.

METHODOLOGY

The research uses a framework for both classification and clustering which is shown in Fig. 1. It consists of a business definition, data definition, datasets compilation, build a model, evaluation, and implementation. Influential phases in the framework are the business definition and the data definition. The business definition phase is related to understanding the overall research objectives, research problem definition, the requirements usage data, and the evaluation method. The data definition phase gathers data based on specific requirements. The data collected is relevant to the research objectives and business definition.

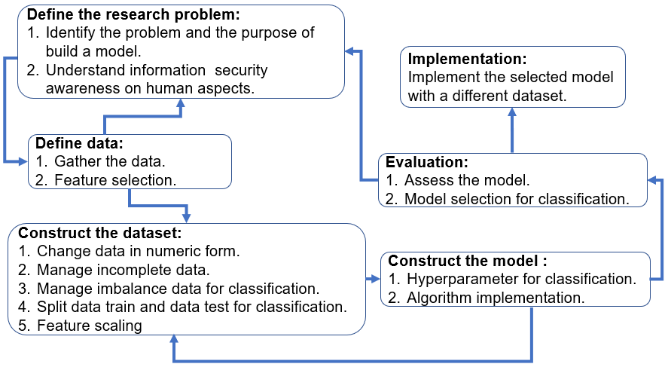

Define the research problem:
1. Identify the problem and the purpose of build a model.
2. Understand information security awareness on human aspects.
Implementation:
Implement the selected model with a different dataset.
Define data:
1. Gather the data.
2. Feature selection.
Evaluation:
1. Assess the model.
2. Model selection for classification.
Construct the dataset:
1. Change data in numeric form.
2. Manage incomplete data.
3. Manage imbalance data for classification.
4. Split data train and data test for classification.
5. Feature scaling
Construct the model :
1. Hyperparameter for classification.
2. Algorithm implementation.

Figure 1: Research framework

Data collection

Data were collected through an online questionnaire. The questionnaire is intended for Indonesian residents. This questionnaire has been compiled based on the results of the MCFS feature selection. The purpose of data collection is to provide data to form a dataset. Data collection was carried out from April 19 to May 6, 2020. The respondents involved in the questionnaire were 488 people.

The questionnaire has 21 features as presented in Table 1 and Table 2. It refers to HAIS-Q and three additional features, i.e., education, occupation, and gender. The respondents which consists of 49% male and 51% female. The respondent education composition consists of 1% elementary school, 12% junior high school, 30% high school, 6% diploma, 28% undergraduate, 18% graduate, and 5% postgraduate. The respondent occupation composition consists of 37% students, 4% manage household, 27% private employees, 16% government employee/military personnel/police, 7% professionals, and 8% entrepreneur.

Table 1: Feature of HAIS-Q in this research

Knowledge	Attitude	Behavior
1. Similar password 2. Secure password 3. Privacy settings 4. Posts 5. Device security 6. Disregard rule	1. Unknown email link 2. Unsafe access 3. Consequences 4. Shoulder surfing	1. Sharing password 2. Known email link 3. Email attachment 4. Download file 5. Give information online 6. Free wifi 7. Dispose document 8. Removable storage 9. Leave document 10. Unusual behavior 11. Report incident

Table 2: The questionnaire

No	Questionnaire
1.	I avoid using the same password for multiple application accounts. Scale: Agreement
2.	I never give password information of an account, even to the closest person. Scale: Frequency
3.	In my opinion, a password that consists of lowercase letters is sufficiently safe. Scale: Agreement
4.	I have accessed a link in an e-mail/SMS from a well-known sender which lead to a website that has the potential for dangerous risks. Scale: Frequency
5.	Nothing will happen if I access the link in the e-mail/SMS from the unknown person. Scale: Agreement
6.	I have opened attachments from the unknown person. Scale: Frequency
7.	I have downloaded a file/application on a device from an accountable/official website. Scale: Frequency
8.	I feel safe accessing adult sites, gambling sites, torrent-based download sites, or other similar sites. Scale: Agreement
9.	I ensure the security of a website (URL, HTTPS, VPN, digital certificate, and others) before providing information online. Scale: Frequency
10.	The default setting on social media is safe to protect personal information. Scale: Agreement
11.	In my opinion, uploads/posts on social media will not have any impact. Scale: Agreement
12.	Avoid uploading/posting related personal information (children's photos, ID cards, vehicle registration plates, and others) or data related to work social media. Scale: Agreement
13.	It is prohibited to leave unattended devices in public space. Scale: Agreement
14.	I have used free Wi-Fi in public spaces to access crucial data (financial transactions, work e-mails, and others). Scale: Frequency
15.	I feel uncomfortable when a stranger notice my device when I open my device. Scale: Agreement
16.	In my opinion, documents can be thrown directly into the trash. Scale: Agreement
17.	I will ignore a USB drive from an unknown personel. Scale: Agreement
18.	I do not have problem to leave documents or screen open even though it is easy to be access/ read by others. Scale: Agreement
19.	I will not report any suspicious actions regarding an information security by anyone. Scale: Frequency
20.	Any suspicious actions related to information security even though carried out by the known should be reported to the authorities. Scale: Agreement
21.	I have made a security incident report to the authorities such as on social media, complaint to Indonesia Telecommunication Regulatory body (BRTI) services, Police, The Financial Services Authority (OJK), lapor.go.id, and others against violation/fraud. Scale: Frequency

In building a model for classification and clustering processes in Fig. 2. we presented a flowchart to be used as a guidance for compiling a dataset in a classification. The compilation consists of: verify data completeness, define label, define feature (X) and class (y), manage unbalanced data, and divide the dataset into train data and test data. Feature scale has as part of transformation sample using standardization techniques. Classification algorithm implementation for the model begins by using Logistic Regression, k-NN, SVM, Naïve Bayes, decision tree, and random forest algorithm. The Hyperparameter process aims to obtain a combination of parameters with the best accuracy results using the grid search method and dataset. Score accuracy calculation in the research use, such as k-fold CV, CM, and ROC have been applied to each model result. The evaluation phase of the flowchart begins by examining and comparing the scores between the resulting models. A model with the best score on several criteria selected as a resulting model through a classification process.

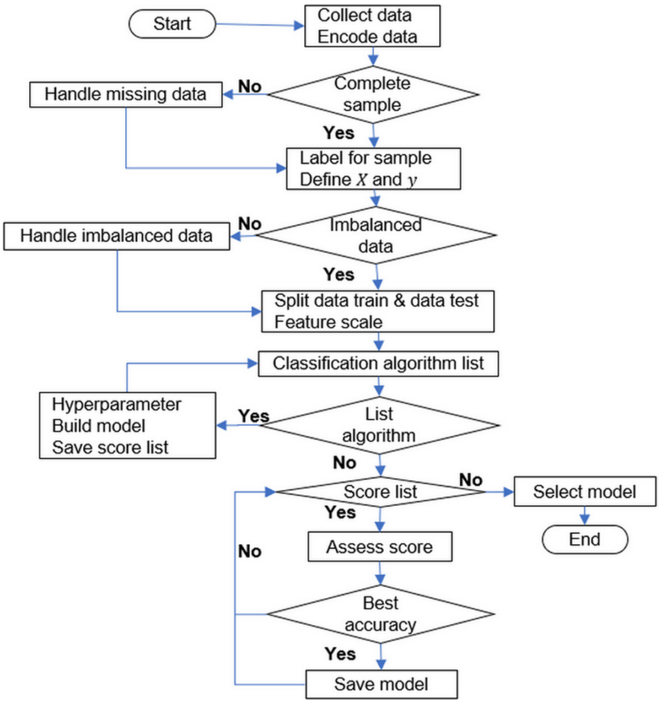


Figure 2: Classification flowchart

Fig. 3 shows the dataset compile phase that consists of arranging and transforming the dataset. The dataset composed process use the clustering algorithm by checking the completeness of the sample and determining X. It is different from the classification because in clustering there is no labeling process for samples. In addition, this process does not determine the variable y. The purpose of the process of clustering is to determine the number of clusters in the model. A feature scale in the transformation process used the standard technique. The modeling phase starts with registering several clustering algorithms that to be used, namely: K-Means, agglomerative, and DBSCAN. Algorithms have been chosen based on popularity and library support. Each algorithm goes through a process to determine the cluster parameter (cluster) and eps. Each score in the model has been used for evaluation and analysis without choosing a model.

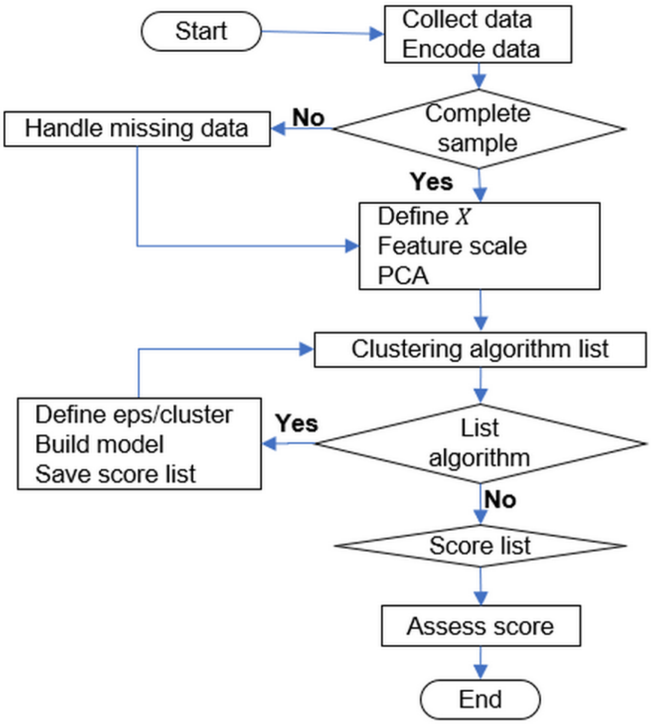


Figure 3: Clustering flowchart

IMPLEMENTATION AND ANALYSIS

Classification Implementation

The online questionnaire form has been distributed to be filled in by respondents. We ensure that respondents fill in the data completely to reduce the risk of having incomplete data. Incomplete data can have an impact on the dataset arrangement and consume time to handle. The process continues with a defined class for each sample that uses weighting.

Label in this research uses an information security awareness level coded in numeric form. Code 1 represents the good condition, 2 means average, and 3 means low. The use of these three types of labels makes the dataset have multiclass classification characteristics that a sample only has one label of available labels. Defined X consists of 24 features and the label of sample as y.

Examination balance of sample number in the class is defined y result as shown in Fig. 4. The composition consists of 121 samples in class 1 (25%), 346 samples in class 2 (71%), and 21 samples in class 3 (4%) from a total of 488 samples. It shows that the sample number is imbalanced between classes. The minority class is class 3 which has a much different sample quantity compared to class 2 that is the majority class with a ratio of 1:16. This condition at the next stage will affect the accuracy of the model built, especially for the minority class, so that data imbalance handling is required.

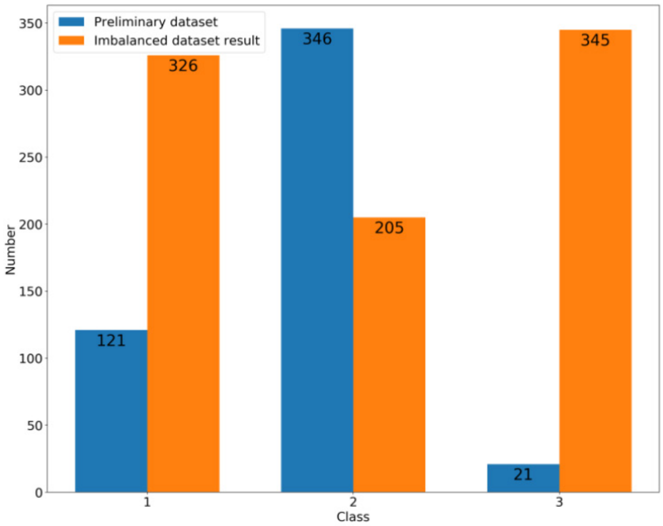


Figure 4: Imbalanced data result

Data imbalances handling uses imblearn.combine.SMOTEENN as the library on python programming. It has effects on managing the sample quantity in each class. The managing imbalanced data result produced new samples for class 1 and class 3 and reduced samples in class 2. The current composition sample quantity in each class is as follows: 326 samples in class 1 (37.2%), 205 samples in class 2 (23.4%), and 345 samples in class 3 (39.4%). Therefore, there are a total of 876 samples in the dataset.

Dataset is divided into train data and test data using train_test_split with a ratio of 7:3. The train data consists of 613 samples, and the test data consists of 263 samples (30%) from a total of 876 samples on datasets. The feature scaling process used by the StandardScaler. It means removing the mean on the features and adjusting the scale to the variant units so that data normally distributed.

Our experiments use algorithms for classification that is stored in lists. The list consists of logistic regression, k-NN, SVM, Naïve Bayes, decision tree, and random forest. The hyperparameter process uses the GridSearchCV before building a model to obtain a combination of parameters with high accuracy. The parameter combination result is selected based on the best_score_ to choose the best accuracy with a minimal loss and best_params_ that provides the best parameter combination. The parameter combination is implemented in the algorithm.

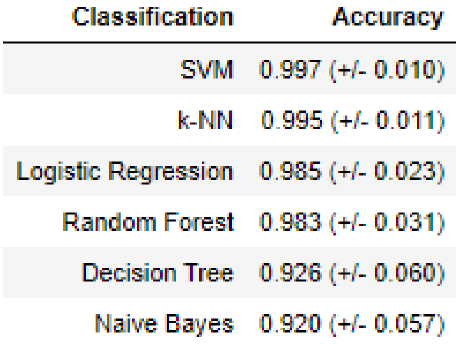
Classification model analysis

The dataset compiled through an imbalanced data process is then used to build the model. Models is the implementation result of several classification algorithms. We have selected some models as part of the evaluation phase within this research framework. The selected model is based on the results of the k-fold analysis of CV, CM, ROC, and the calculation of the score for each model.

k-Fold cross validation

k-Fold CV in this research is used to measure the ability of a model by dividing the dataset. We used StratifiedKFold from the scikit-learn library which applied the k-Fold CV. So, it divided the dataset with the same percentage samples in each class as a complete set. Each model uses parameters result from the hyperparameter process in the application of CV. k-fold CV score in Table 3 that used the StratifiedKFold shows that some models have an average accuracy value of more than 0.90. The imbalanced data handling and hyperparameter processes have an impact on result scores. The highest accuracy value has been gained from the SVM model with an accuracy value of 0.997.

Table 3: Model score is based on the k-fold CV



Classification	Accuracy
SVM	0.997 (+/- 0.010)
k-NN	0.995 (+/- 0.011)
Logistic Regression	0.985 (+/- 0.023)
Random Forest	0.983 (+/- 0.031)
Decision Tree	0.926 (+/- 0.060)
Naive Bayes	0.920 (+/- 0.057)

Confusion matrix (CM)

Confusion Matrix (CM) is used to measure the accuracy of the model. against the dataset. CM learns the data train that already has a label then applies the pattern to the test data. The prediction results on the test data are then compared with the labels existed in the test data. This research used multilabel_confusion_matrix from scikit learn library that is applied to the multiclass. This is suitable for the dataset. CM uses 3 classes according to the label of the training data. It calculated sample number that was successfully predicted in each class and have a minimum error of predictions. The criteria is that the models successfully predict that the sample that has the same predicted value as the actual label in the test data.

The number of class samples in TN that are almost ideal in class 1 is the logistic regression model and SVM. Both models have a sample prediction error that is considered part of class 1. TN prediction in class 2 reaches an ideal value of 201 samples in the k-NN model and SVM. Therefore, both models successfully predict samples that did not have class 2 characteristics. We predicted TN in class 3 for 164 samples by random forest and SVM algorithms. The test data is successfully predicted TP in class 1 by logistic regression and SVM for 102 samples. Class 2 in SVM and random forest made prediction errors only by one sample. Class 3 is predicted successfully by SVM and random forest.

Prediction error type 1 (FP) in class 1 is due to an error predicting samples that should not be part of class

1. The type 1 error frequently happened in the decision tree model by 2.66%. The Naïve Bayes model predicted errors in class 2 and class 3 by 9 samples each. Type 2 error is due to an incorrect prediction so that the sample belongs to a class. FN errors in CM results include 7 samples in class 1 for the Naïve Bayes model, 14 samples in class 2 for the decision tree model, and 2 samples in class 3 for the Naïve Bayes model.

The models selected with TN and TP values were higher than the FN and FP values based on the CM results. Table 4 shows that a low average value of FN and FP in the SVM model. SVM has an average prediction error in all classes of 0.25% when the highest prediction error is in the decision tree model which is 5.58%.

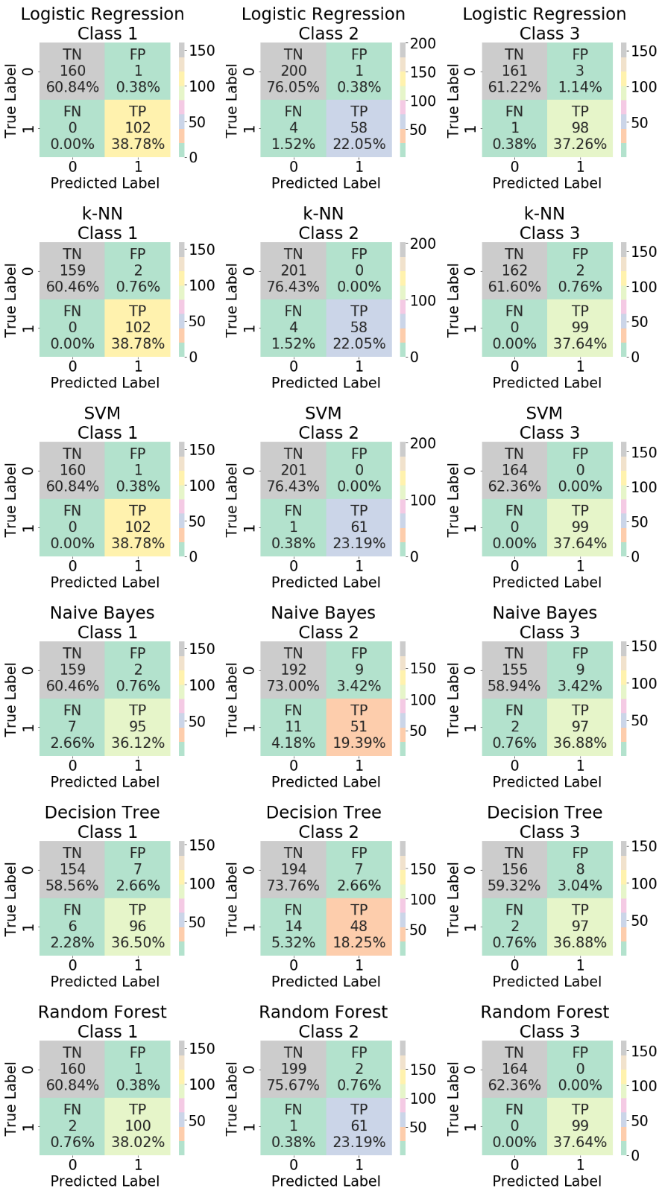


Figure 5: Confusion Matrix results

Table 4: TNTP and FPFN average

Classification	Avg TNTP	Avg FPFN	%Avg FPFN
SVM	262,33	0,67	0,25
Random Forest	173,33	2,00	1,14
k-NN	260,33	2,67	1,01
Logistic Regression	259,67	3,33	1,27
Naïve Bayes	249,67	13,33	5,07
Decision Tree	248,33	14,67	5,58

Receiver operating characteristic

The performance of models is presented by a two-dimensional graph ROC. It shows the true positive rate and the false positive rate. Model selection is based on ROC by comparing the values each model and select the highest value among them. The SVM and random forest (Fig. 6) models have ROC that shows the true rate is positive value of 1 as the value so that the Area Under Curve (AUC) is 100%. It shows that the model can accurately predict positive samples.

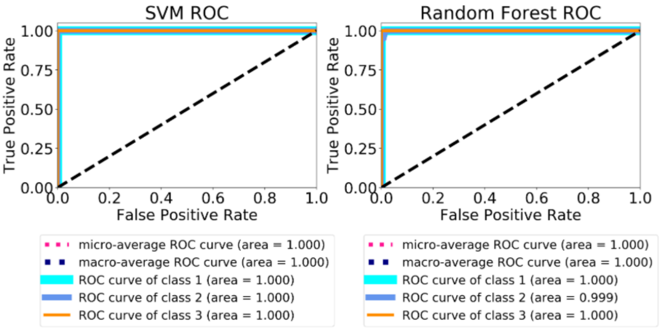


Figure 6: ROC SVM and Random Forest

Score classification model

The highest accuracy score of 99.7% shown in Table 5 was achieved by the SVM model. It has only 1 sample in the test data failed to predict. The error rate in making predictions during the classification on the SVM is 0.3%. The highest misclassification is 5.6% in making predictions on the decision tree.

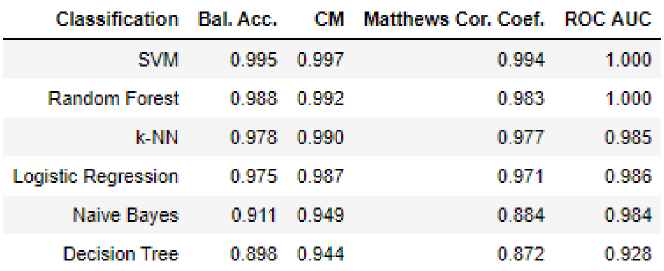
Sensitivity is calculated using sklearn.metrics.recall_score as the scikit learn library uses manage of the macro parameter. It is used because this is related to the multiclass dataset and the metrics calculation for each class. The highest sensitivity on the SVM model is 99.5%, while the decision tree is only 89.8%. The specificity of all models was above the 99% threshold. The top three specificity percentages are the SVM model which is 99.8%, i.e. 99.5% for the random forest model, and 99.2% for the k-NN model. The combination of sensitivity and specificity can be seen from the value of the G-Mean score with the lowest value on the decision tree of 0.928.

False positive rate/fall out represents the percentage of type 1 error prediction. It describes the actual value as false but is predicted to have the true value. Fall out is frequently happen in the decision tree of 0.042 and Naïve Bayes of 0.037. False negative rate/miss rate describes a type 2 prediction error that indicates an actual value is true but predicted as a false value. The decision tree has a miss rate of 10.2% while the SVM miss rate is 0.5% and a fall out of 0.2%.

The positive value prediction shows the performance of a model in predicting test data. All models in this research have a positive predictive value above 90% as a threshold, i.e: 99.7% of the SVM model, 98,7% of the the k-NN model, 98,6% of the random forest model, and 98,1% of the logistic regression model. F1 score is 90,3% for the decision tree model.

In Table 6, the lowest value of balanced accuracy is 0.898 for the decision tree, whereas the highest value if from the SVM model of 0.995. Matthews correlation coefficient is a measure that shows the multiclass classification quality. The quality shows the correlation between train data and test data on the model. The correlation value between train data and test data on the SVM model is 0.994, while the lowest value of 0.884 from the Naïve Bayes model. ROC AUC using roc_auc_score shows the area under the ROC curve with the parameter value of One-vs-the-rest that is ovr as multiclass and macro as average. roc_auc_score value indicates that SVM and Random Forest both have the best value of 1.

Table 6: Score for classification models



Classification	Bal. Acc.	CM	Matthews Cor. Coef.	ROC AUC
SVM	0.995	0.997	0.994	1.000
Random Forest	0.988	0.992	0.983	1.000
k-NN	0.978	0.990	0.977	0.985
Logistic Regression	0.975	0.987	0.971	0.986
Naive Bayes	0.911	0.949	0.884	0.984
Decision Tree	0.898	0.944	0.872	0.928

Classification models selection

The model in classification as results of algorithms implemented on the dataset. The dataset conducts through handling imbalanced data and hyperparameters processes. Both processes affect the model accuracy and error rate. Based on CV results, a model built using the SVM algorithm has a higher value and the lowest deviation. Model selection through CM results based on TN and TP values is high while FP and FN values must be minimum. Based on the CM results, the SVM model has the lowest error value compared to others. The AUC results show that SVM and random forest have 1 as the value of roc_auc_score using the average parameter with the macro and weighted value. Overall, the resulting model of SVM's algorithm implementation is superior to the other models using our research datasets.

Classification model analysis on information security awareness

The implementation phase in this research framework used the SVM algorithm model. It conducted selection based on the evaluation of the classification model. The model shows a prediction that divides the sample into three (3) classes using the test data. Class 2 is very dominant to represent the level of information security awareness. Samples on class 2 as the average level show that people require strengthening they information security awareness. The lowest number of samples is in class 3. It represents people that have a low level of information security awareness.

The SVM model has crucial features (Table 1). It was selected based on eigenvector (Fig. 7) that contains: leave documents, email attachments, consequences, and unsafe access to a website. The test data shows that the behavior of samples from society has a concern to save the documents. It shows the practice of not leaving documents or screen open even though it is easy to reread/access. The email attachment feature provides information that the society is aware of avoid to opening email attachments from unknown senders. The consequence feature is related to social media usage. Post attitude towards the statement that posting on social media does not have an impact is rational. All participants have a great awareness of the consequences of uploading information on social media. The participant's attitude toward the awareness of suspicious websites is high. The test data shows that the participant is distrustful the link in an

Table 5: Performance metrics for classification models

Metric	SVM	RF	k-NN	LR	NB	DT
Recall	0,995	0,988	0,978	0,975	0,911	0,898
Specificity	0,998	0,995	0,992	0,990	0,963	0,958
G-Mean	0,996	0,991	0,985	0,983	0,937	0,928
Fall Out	0,002	0,005	0,008	0,010	0,037	0,042
Miss Rate	0,005	0,012	0,022	0,025	0,089	0,102
Accuracy	0,997	0,992	0,990	0,987	0,949	0,944
Misclass.Rate	0,003	0,008	0,010	0,013	0,051	0,056
Precision	0,997	0,986	0,987	0,981	0,915	0,910
F1 Score	0,996	0,987	0,982	0,978	0,912	0,903

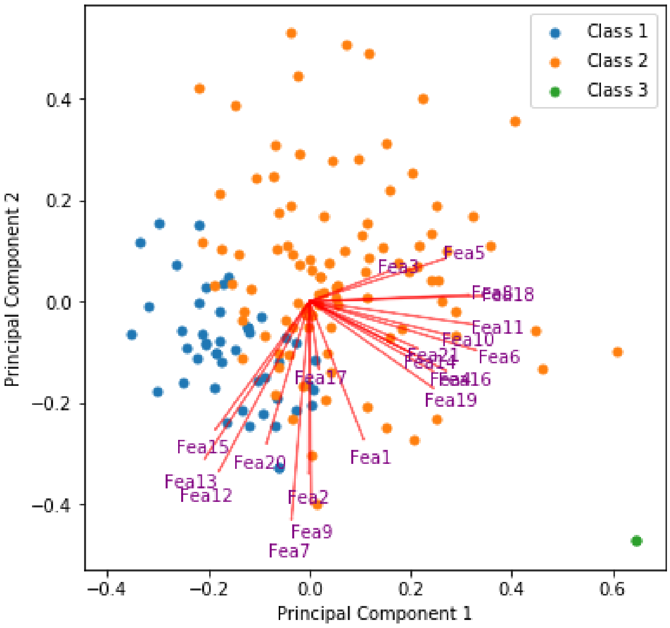


Figure 7: Biplot SVM test data model

email or Short Message Service (SMS) even it is from a well-known sender.

Features used in this research are relevant to one of the KAB components. The eigenvector score shows the relationship between features with each area of KAB. The eigenvector score obtained is based on the knowledge as shown in Table 7. The highest value is gained through the privacy setting feature. It shows that the … already know the privacy settings on social media that default settings are not safe to protect private information. The smallest eigenvector value on a feature disregard the rule. It shows that in terms of vulnerability in public, we ignore security when there is a suspicious action related to information security even though it is done by a well-known person.

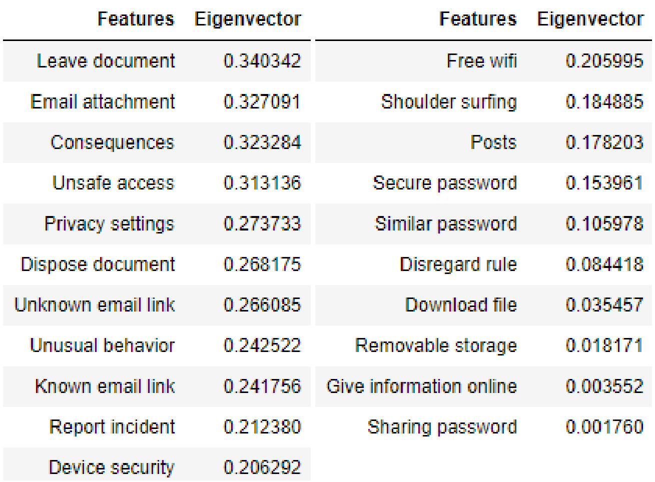
In term of the attitude evaluation, eigenvector scores are high on consequences features and low on shoulder surfing features. The consequences feature is related to the attitude in society to upload/post something on social media that will induce positive or negative impacts. The shoulder surfing feature shows that the public uncomfortable if others notice the opening of a device or important document in a public place. It exposes the vulnerability of data fraud through shoulder surfing that is not recognized by the public.

The behavior area has high eigenvector scores on the feature of leaving document and low scores on sharing password features. The participants understands that leaving a document or the device screen open is a hazardous activity. Thus leads to data leakage. The public has a vulnerability to password information that can be leaked to other people.

Visualization of model prediction results use the Principal Component Analysis (PCA). The substantial feature in PCA 1 was to leave documents, and PCA 2 was sharing the password. Table 7 shows features that such as the sharing password, to give information online, and removable storage have a low eigenvector. It illustrates that the participant still has vulnerabilities in sharing password information behavior, providing information on the site behavior, and the use of portable media behavior.

Information security awareness in sharing password behavior is lacking in some cases. It is because this action is still being done even with the closest person. The participant needs to be educated on website security to provide information online. The behavior related to removable storage shows that people feels vulnerable in using it, especially toward those with the unidentified owner. This behavior provides an opportunity for removable storage usage as bait to commit violations in cybers space.

Table 7: Eigenvector score for SVM model used test data



Features	Eigenvector	Features	Eigenvector
Leave document	0.340342	Free wifi	0.205995
Email attachment	0.327091	Shoulder surfing	0.184885
Consequences	0.323284	Posts	0.178203
Unsafe access	0.313136	Secure password	0.153961
Privacy settings	0.273733	Similar password	0.105978
Dispose document	0.268175	Disregard rule	0.084418
Unknown email link	0.266085	Download file	0.035457
Unusual behavior	0.242522	Removable storage	0.018171
Known email link	0.241756	Give information online	0.003552
Report incident	0.212380	Sharing password	0.001760
Device security	0.206292		

Clustering implementation

The clustering algorithms used to build the model include K-Means, agglomerative, and DBSCAN. Each algorithm has a method for determining the number of clusters. The objective of the clustering method is different from the classification that impacts the dataset preparation process. The dataset preparation process to form the clustering model consists of data import, incomplete data handling, feature scaling, and feature selection. The feature selection approach used in this method uses PCA with the value of n_components of 2 (n_components=2). Thus is intended to summarize data.

K-Means

The first thing that is necessary to do in using the k-Means algorithm is to determine a suitable number of clusters. There are some approaches to determine the number of clusters, such as the silhouette, calinski harabasz, elbow method, and bouldin davies is shown in Fig. 8. In our research, we try to find clusters that are suitable for the k-means algorithm. Find the number of clusters with various approaches next getting the average value of the elbow location.

The elbow method is used to determine K-Means algorithm parameters related to clusters number based on the inertia_value, which is the total sample distance squared to each center of the nearest clusters. As the number of clusters increases, the distortion decreases, the samples will be reduced in each cluster, and will be closer to the center point.

Several approaches scores mentioned above were combined using StandardScaler for comparison purposes. The comparison results show that silhouette and calinski harabasz scores tend to increase with the increase of cluster number. On the other hand, the elbow method and the bouldin davies scores tend to decrease with the increase of cluster number.

In this research, the number of cluster determination is based on the average location of the elbow method. The point of the elbow is an intersection that is considered the optimal usable value. Locating the elbow point on the curve through the maximum point curve involves axis values and curve direction on each graph. The elbow method is implemented using the KneeLocator. knee from Kneed library [35] which is applied in some of approaches. The number of cluster increases and the elbow location changes accordingly. The number of clusters in the graph is between 2 to 11, with the average value of 4 for the elbow location. Thus, we use 4 clusters that are considered optimal in the model with the K-Means algorithm.

PCA visualization for the K-Means model shows the distribution of samples in 4 clusters. This can be seen in Fig. 9. A feature that is very influential on Principal Component 1 is the leaving a document feature, whereas the Principal Component 2 has the largest eigenvector value for downloading file as the most influential feature. Each cluster has been analyzed so that each group has several factors to become a characteristic associated with KAB (Fig. 10).

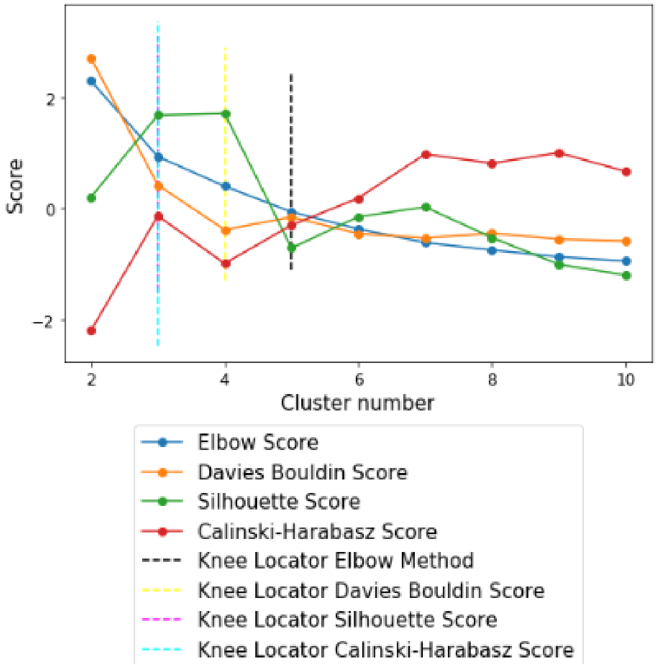


Figure 8: The number of clusters scores

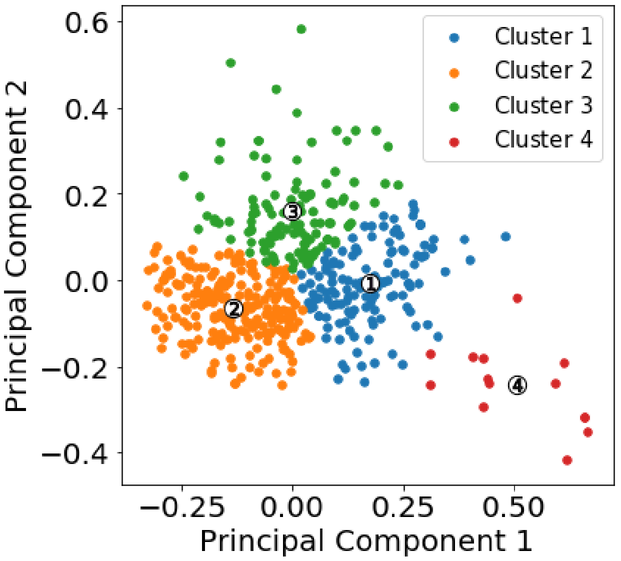


Figure 9: K-Means model visualization

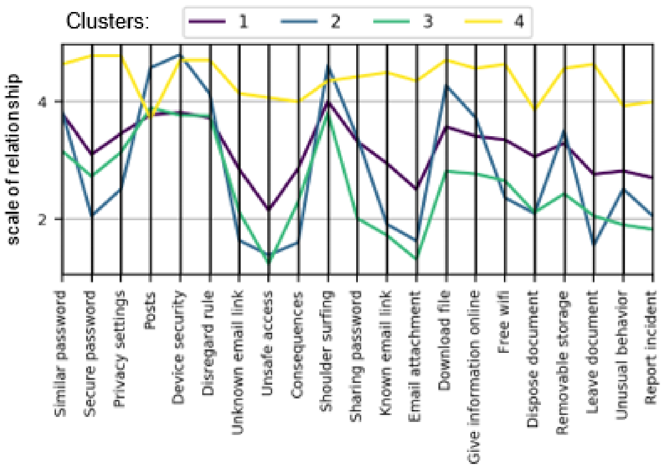


Figure 10: Parallel coordinates of K-Means model

The knowledge area in Cluster 1 on the secure password feature has a vulnerability whereas, other features related to this area are in a reasonably good condition. Cluster 2 has vulnerabilities features in the secure password and privacy settings. However samples in this area have some concerns about issues of posts information and device security. Cluster 3 has a vulnerability in password security and has some concerns about the posts feature. Cluster 4 is the group in which the samples have a good knowledge of information security awareness. Unfortunately, this cluster has the lowest awareness on posts feature compared to the other clusters. All clusters except cluster 4 have vulnerabilities in password security.

Based on the consideration of the feature area, cluster 1 and cluster 3 have the vulnerability to unsafe access features. It is necessary to reinforce the perspective of being careful in accessing adult sites, gambling sites, torrent-based download sites, or other similar sites. Vulnerability in the consequences feature is related to the impact of uploads on social media, especially in cluster 2. Based on the behavior of Clusters 1, 2, and especially 3, it can be noticed that there is a vulnerability to email attachments, so the participant should be aware of it. It is necessary to guide them in treating email attachments. Cluster 2 is more concerned with downloaded emails but lacks of awareness to open email attachments. Fig. 10 shows some parallel coordinates of 4 clusters that use a particular color to differentiate each cluster.

Agglomerative

In this research, the hierarchical structure to select the number is using a dendrogram representation (Fig. 11). It uses a euclidean distance of 8 so that the result cluster number is 3. In the visualization of model results of agglomerative algorithm the dataset is divided into three clusters as shown in Fig.12.

In this model, cluster 1 generally is considered a group with a better ISA than other groups. Cluster 1 has a vulnerability to email the attachment access. However, information security awareness in cluster 1 has a good value, such as similar passwords, and shoulder surfing. It shows that cluster 1 knows the way to avoid the use of the same password on several accounts, especially for a finance-related account. It also aware of the threat of the through shoulder surfing so that they feel uncomfortable when using gadgets while being noticed by strangers.

Fig. 13 shows that cluster 2 has a significant difference in each feature in knowledge, attitude, and behavior areas. Cluster 2 has similar vulnerabilities to cluster 2 of the K-Means model, i.e. secure password, unsafe access, and email attachments. The information security awareness is already owned by cluster 2 which includes device security, shoulder surfing, and file download. Cluster 3 on knowledge and attitude area is always in the middle between clusters 1 and 2. On the behavior side, cluster 3 is often under the others. It shows that those in cluster 3 have a good knowledge and attitude towards information security awareness. However, this is not being consistent in their behavior.

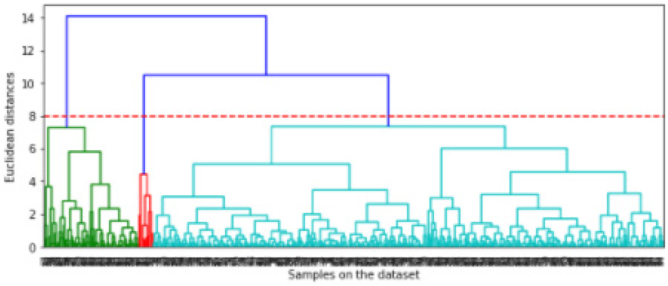


Figure 11: The dendrogram of dataset

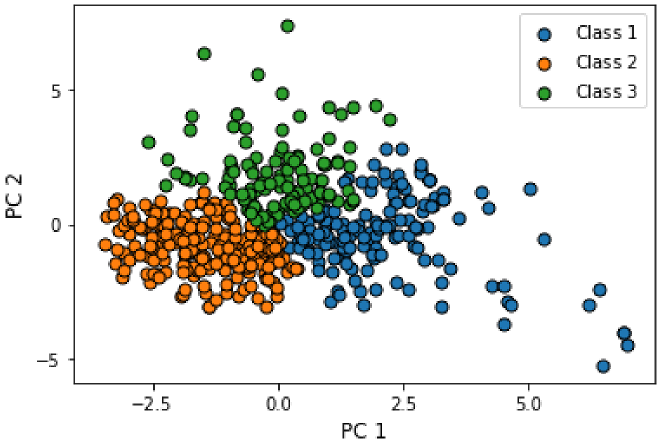


Figure 12: Agglomerative model visualization

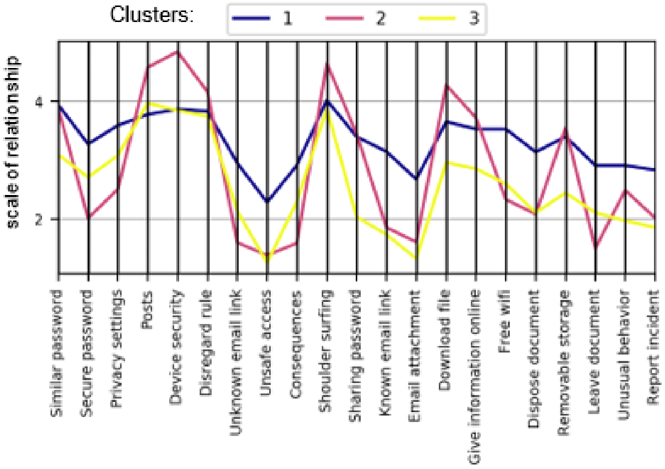


Figure 13: Parallel coordinates Agglomerative model

Density-based spatial clustering of applications with noise

Parameters that play a role in Density-Based Spatial Clustering Of Applications With Noise (DBSCAN) consist of epsilon (eps). The epsilon indicate the distance between the samples from the core sample and the minimum number of samples to form a cluster. Algorithm DBSCAN is implemented using a library sklearn.cluster from scikit-learn. It used the min_samples that specify the minimum number of samples to form a cluster. This research uses min_samples with a value of 5 as the minimum number of samples in a cluster. The following algorithm (Fig. 14) is used to determine the eps value in this research:

1. Discover the distance of each sample in the dataset from point x = 0. It used the NearestNeighbors library

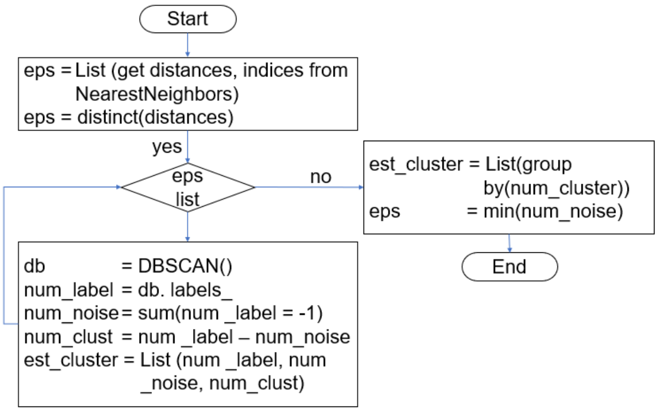

Start
eps = List (get distances, indices from NearestNeighbors)
eps = distinct(distances)
yes
eps list
no
est_cluster = List(group by(num_cluster))
eps = min(num_noise)
End
db = DBSCAN()
num_label = db. labels_
num_noise = sum(num _label = -1)
num_clust = num _label – num_noise
est_cluster = List (num _label, num _noise, num_clust)

Figure 14: Algorithm to determine of eps

with the parameter n_neighbors of 2 and auto as the value of the algorithm parameter.

2. Group the distance through the results of rounding off for each sample. Each group is stored in the variable eps.
3. Compile an estimate of the number of clusters and noise in each eps by implementing the DBSCAN algorithm in the Sklearn library.
4. Group the number of clusters and get the average eps.
5. Select the eps value with the minimum amount of noise criteria.

The algorithm has been applied to obtain eps resulted in a value of eps=0.46 with a noise of 55.33.

There are 5 clusters in the DBSCAN model result. A cluster in the model has the characteristics of noise. The DBSCAN noise in this research has a value of -1. Noise has been shown in Fig. 15. It emerges because some sample does not have the closest neighbors to be the members of a cluster or form a cluster with a minimum member.

Fig 16. illustrate, there are vulnerabilities in cluster 1 consist of unsafe access and email attachments features, but it has acceptable information security awareness on-device security features. Vulnerabilities in cluster 2 consist of secure passwords, unsafe access, sharing passwords, and the dispose documents features. Cluster 2 has some concerns related to Information security awareness in some features. It is such as device security, shoulder surfing, and file downloading. Cluster 3 generally has better scores than others. However, for a secure password feature, the lowest value is the vulnerability. The best information security awareness in cluster 3 is related to the posts feature, device security, and give online information. Vulnerabilities in cluster 4 consist of unsafe access and the email attachment features, while the best information security awareness is related to the shoulder surfing feature.

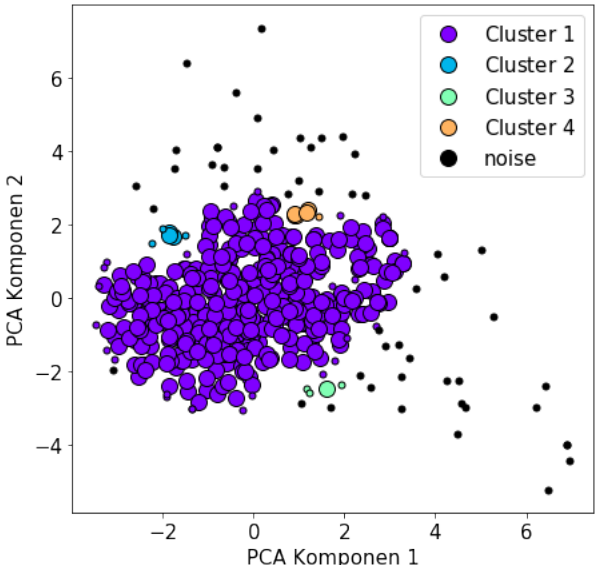


Figure 15: DBSCAN model visualization

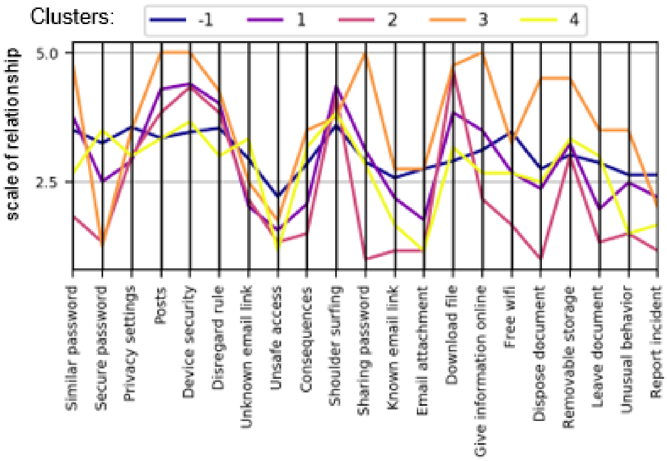


Figure 16: Parallel coordinates DBSCAN model

Clustering model analysis on information security awareness

ARI was previously described as a function to calculate the similarity between 2 labels. ARI results in DBSCAN produces values always close to 0 (Fig. 17). This shows that the labels have different characteristics. The value does not depend on the number of clusters and the eps value is consistently randomized in terms of its labeling so that it remains close to 0. Fig. 17 shows that the number of clusters used in the K-Means model and agglomerative is between 1 and 21. Thus, different in DBSCAN that uses eps values between 0.1 to 2.1 to determine the number of clusters. The calculation of the ARI value uses adjusted_rand_score in the scikit learn library.

Fig. 17 shows that the higher the number of clusters, the greater the ARI value in the k-Means and agglomerative algorithms. It shows that an increasing number of clusters has a high similarity with slight differences in the adjacent clusters thus the characteristics of the participant in one cluster with another cluster more similar. The ARI value on DBSCAN has a maximum value between the number of clusters of more than 11 to 13. Its value in the model of DBSCAN for the number of clusters becomes not effective.

The three models have a completeness value of 1 that indicates the model is successful in grouping each sample into one cluster. Homogeneity that shows clusters consists of a different sample (Table 8). The most reliable homogeneity was performed through DBSCAN with the lowest value.

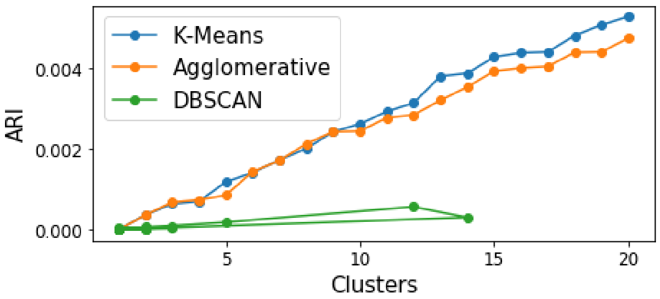
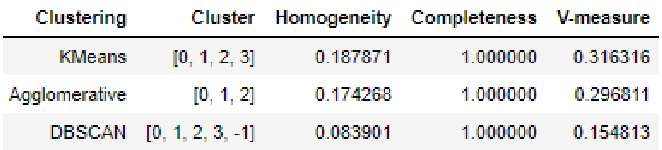


Figure 17: ARI clustering model

Table 8: Homogeneity, completeness, and v-measures



Clustering	Cluster	Homogeneity	Completeness	V-measure
KMeans	[0, 1, 2, 3]	0.187871	1.000000	0.316316
Agglomerative	[0, 1, 2]	0.174268	1.000000	0.296811
DBSCAN	[0, 1, 2, 3, -1]	0.083901	1.000000	0.154813

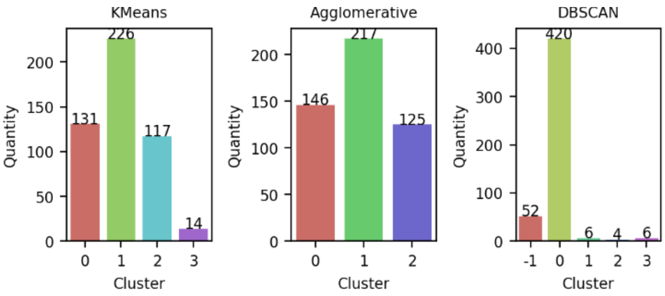


Figure 18: Cardinality clustering model

Cluster 1 in Fig. 18 presents the K-Means and the agglomerative models have higher samples number as cardinality cluster results. Cluster 1 on the k-mean model is 46% of the total sample in the dataset while in the agglomerative model it is 44%. Cluster 0 of the DBSCAN model consists of 420 samples with 52 noise samples. The samples included in the noise were not members of any group. This noise occurs because the samples do not have the closest neighbors to be a member of a cluster. The higher the min_samples parameter, the higher the number of samples in the noise.

CONCLUSIONS

In our research, there are different processes in building models using classification and clustering approaches. Implementation of the machine learning algorithms in this research is related to the human aspect of information security awareness. The selection of algorithms used is based on popularity and it has libraries support using the python programming language.

Models as the result of classification algorithms implementation were formed through a process examining the balance sample number in each class. Changes in sample distribution in each class affect the sample number in classes, the level of accuracy, and the misclassification rate of a model. The accuracy is influenced by the hyperparameter process. Classification models are evaluated using several methods.

Based on CV, the SVM algorithm has a higher value and the lowest deviation. The selection used CM that resulted in the SVM model based on high TP and TN values and low FP and FN values. Based on the ROC results, the SVM has a value of 1. The SVM model has an accuracy of 99.7% and an error rate of 0.3%. SVM is succeeded in predicting a sample of 262 of the 263 sample test data. The model generated using the SVM algorithm is superior to other models. This model shows the vulnerability such as leave documents, email attachments, consequences, and unsafe access to a website.

The model of the clustering algorithms involves PCA to handle multidimensional data. DBSCAN has the most suitable value based on the ARI value, while cluster number on K-Means and agglomerative increase, and consequently the ARI value will increase. The implementation of each model is used to analyze information security awareness in which cluster is suitable. The K-Means model produces 4 clusters, while agglomerative produces 3 clusters, and DBSCAN produces 4 clusters. Vulnerabilities in the K-Means model consist of the unsafe access feature as the cluster 4, the sharing password, and dispose documents as the cluster 2. The agglomerative and the DBSCAN models show a vulnerability in the unsafe access and email attachment features as the cluster 3. The agglomerative model has a sample with good information security awareness in cluster 2 while the DBSCAN model is in cluster 3. Vulnerability in the three clustering models is generally in an unsafe access feature.

The obstacles in the research process were the limited hardware and data separation. The hyperparameter process cannot be carried out optimally due to the limited Random Access Memory (RAM) in the research environment. Data separation in the research is divided into train data and test data therefore it does not have data validation.

The research can be enhanced through both classification and clustering approaches. The classification approach enhanced by building models using modern machine learning algorithms. Enhancement in the clustering approach can be done by exploring the characteristics of the information security awareness sample in the human aspect in more detail. Clusters of the clustering model can be used as a label in classification for future research and testing on a new dataset.

ACKNOWLEDGMENT

We thank the Ministry of Education and Culture of Republic of Indonesia for financial support for this research under the PTUPT Research Grant number NKB-356/UN2.RST/HKP.05.00/2020.

REFERENCES

1. B. P. Statistik, Statistik Telekomunikasi Indonesia 2017, Jakarta: Badan Pusat Statistik, 2018.
2. C. Easttom and W. Butler, "A Modified McCumber Cube as a Basis for a Taxonomy of Cyber Attacks," in IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2019.
3. R. v. Solms and J. v. Niekerk, "From information security to cyber security," Computers & Security, vol. 38, pp. 97-102, 2013.
4. J. McCumber, Assessing and Managing Security Risk in IT Systems: A Structured Methodology, USA: Auerbach Publications, 2004.

5. S. Kraemer, P. Carayon and J. Clem, "Human and organizational factors in computer and information security: Pathways to vulnerabilities," computers & security, vol. 28, pp. 509-520, 2009.
6. T. W. Edgar and D. O. Manz, "Chapter 6 - Machine Learning," in Research Methods for Cyber Security, United States, Syngress, 2017, pp. 153-173.
7. G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security," in 2018 10th International Conference on Cyber Conflict (CyCon), Tallinn, 2018.
8. M. Alohali, N. Clarke, S. Furnell and S. Albakri, "Information security behavior: Recognizing the influencers," in 2017 Computing Conference, London, 2017.
9. S. Bauer and E. W. Bernroider, "From Information Security Awareness to Reasoned Compliant Action: Analyzing Information Security Policy Compliance in a Large Banking Organization," ACM SIGMIS Database: the DATABASE for Advances in Information Systems, vol. 48, p. 44–68, 2017.
10. Y. Normandia, L. Kumaralalita, A. N. Hidayanto, W. S. Nugroho and M. R. Shihab, "Measurement of Employee Information Security Awareness Using Analytic Hierarchy Process (AHP): A Case Study of Foreign Affairs Ministry," in 2018 International Conference on Computing, Engineering, and Design (ICCED), Bangkok, Thailand, 2018.
11. A. Farooq, S. Alifov, S. Virtanen and J. Isoaho, "Towards comprehensive information security awareness: a systematic classification of concerns among university students," In Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18), p. 1–6, 2018.
12. A. Carella, M. Kotsoev and T. M. Truta, "Impact of security awareness training on phishing click-through rates," in 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017.
13. A. Cindana and Y. Ruldeviyani, "Measuring Information Security Awareness on Employee Using HAIS-Q: Case Study at XYZ Firm," International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 289 - 294, 2018.
14. M. G. Ikhsan and K. Ramli, "Measuring the Information Security Awareness Level of Government Employees," 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), 2019.
15. M. S. b. O. Mustafa, M. N. Kabir and F. Erna, "An Enhanced Model for Increasing Awareness of Vocational Students Against Phishing Attacks," in 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Selangor, Malaysia, 2019.
16. D. D. H. Wahyudiwan, Y. G. Sucahyo and A. Gandhi, "Information security awareness level measurement for employee: Case study at ministry of research, technology, and higher education," 3rd International Conference on Science in Information Technology (ICSITech), pp. 654 - 658, 2017.
17. S. Alelyani, J. Tang and H. Liu, "Feature Selection for Clustering: A Review," in Chapter 2, New York, Chapman and Hall/CRC, 2013.
18. A. L'Heureux, K. Grolinger, H. F. ElYamany and M. A. M. Capretz, "Machine Learning with Big Data: Challenges," IEEE Access, vol. 5, pp. 7776-7797, 2017.
19. V. S. Saridewi and R. F. Sari, "Feature Selection In The Human Aspect of Information Security Questionnaires Using Multicluster Feature Selection," International Journal of Advanced Science and Technology, vol. 29, no. 7, pp. 3484-3493, 2020.
20. M. Nieles, K. L. Dempsey and V. Y. Pillitteri, "An Introduction to Information Security," Special Publication (NIST SP), US, 2017.
21. K. Parsons, D. Calic, M. Pattinson and et.all, "The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies," Computers & Security, vol. 66, pp. 40-51, 2017.
22. H. Kruger and W. Kearney, "A prototype for assessing information security awareness," Computers & Security, vol. 25, no. 4, pp. 289-296, 2006.
23. E. Alpaydin, Introduction to Machine Learning, Second Edition, US: The MIT Press, 2009.
24. Google Developers, "Machine Learning Crash Course," Google Developers, [Online]. Available: https://developers.google.com/machine-learning/crash-course. [Accessed 11 4 2020].
25. I. Goodfellow, Y. Bengio and A. Courville, "Machine Learning Basics," in Deep Learning, The MIT Press, 2016, p. 98.
26. M. Swamynathan, Mastering Machine Learning with Python in Six Steps, Berkeley, CA: Apress, 2017.
27. W. Lee, Python® Machine Learning, Indianapolis: John Wiley & Sons, Inc., 2019.
28. A. K. Nandi and H. Ahmed, "Classification Algorithm Validation," in Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines, 307-319, 2019, pp. 307-319.
29. D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," Annals of Data Science, vol. 2, p. 165–193, 2015.
30. J. Wang, Y. Wu, H.-H. Hsu and Z. Cheng, "Spatial Big Data Analytics for Cellular Communication Systems," in Big Data Analytics for Sensor-Network Collected Intelligence, Academic Press, 2017, pp. 153-166.

31. D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vols. PAMI-1, no. 2, pp. 224-227, 1979.
32. scikit-learn.org, "scikit-learn," scikit-learn developers (BSD License), [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed 27 April 2020].
33. V. Kotu and B. Deshpande, "Chapter 7 - Clustering," in Data Science (Second Edition), Morgan Kaufmann, 2019, pp. 221 - 261.
34. M. C. Thomas and J. Romagnoli, "Extracting knowledge from historical databases for process monitoring using feature extraction and data clustering," Computer Aided Chemical Engineering, vol. 38, pp. 859-864, 2016.
35. V. Satopaa, J. Albrecht, D. Irwin and B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," in 31st International Conference on Distributed Computing Systems Workshops, Minneapolis, Minnesota, 2011.
36. R. J. Mejias, "An Integrative Model of Information Security Awareness for Assessing Information Systems Security Risk," in 2012 45th Hawaii International Conference on System Sciences, Maui, HI, 2012.
37. J. Brownlee, "Machine Learning Mastery," Machine Learning Mastery Pty. Ltd., 2020. [Online]. Available: https://machinelearningmastery.com/. [Accessed 23rd May 2020].
38. S. Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, eprint arXiv:1811.12808, 2018.

Paper submitted: 24.09.2020.
Paper accepted: 09.04.2021.