Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 2, April 2022, hlm. 367-374

DOI: https://doi.org/10.20884/1.jutif.2022.3.2.206
p-ISSN: 2723-3863
e-ISSN: 2723-3871

LINEAR REGRESSION FOR PREDICTION OF EXCESSIVE PERMISSIONS DATABASE ACCOUNT TRAFFIC

Agus Pamuji[*1], Heri Satria Setiawan[2]

[1]Bimbingan Konseling Islam, Fakultas Ushuluddin Adab dan Dakwah, IAIN Syekh Nurjati Cirebon, Indonesia
[2]Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Indraprasta PGRI, Indonesia
Email: [1]agus.pamuji@syekhnurjati.ac.id, [2]herisatria@unindra.ac.id

(Naskah masuk: 10 Maret 2022, Revisi: 23 Maret 2022, Diterbitkan: 25 April 2022)

Abstract

Today, the security of information and data is an important asset for everyone in protecting data. Data and information become critical when weaknesses and threats come. In this study, an estimation of the observed variables will be carried out. The application of data mining, especially with the estimation method by using linear regression techniques. The next stage is data preparation by referring to the dataset recorded in the user activity log. Data preparation takes a lot of time because you have to make sure the data fits the needs of data mining analysis. The analysis technique with linear regression involves three independent variables as Type Permissions, Type of User Account, Status, and the dependent variable, namely User Actions. The strongest effect was found in type_permissions and state when together on user_actions. The type_permissions variable keeps increasing when the state on the user is active. The status attribute also suffers from the same condition. Accrording to the results, our findings in root mean squared error is 37.614 and absolute error is 31.058, and mean absolute percentage about 23%. Furthermore, User_action as an estimated variable gives two data opportunities whether it is allowed or not. Therefore, in future research, it is necessary to map users of the database system still in the context of data mining when digging for information on excessive permissions.

Keywords: *Estimation, Excessive Permissions, User Accounts, Database Data Mining.*

1. INTRODUCTION

Today, the security of information and data is an important asset for everyone in protecting data[1]. Data and information become critical when weaknesses and threats come. Thus, there are challenges to be faced when applying security to data and information. The essence of data and information security is confidentiality, integrity and availability. We state that these three aspects (CIA) are security principles that are adhered to when protecting data and information. Therefore, data mining has become a scientific discipline starting with inferring data and developing algorithms to help academics, researchers, practitioners solve complex problems[2].

Database security is related to the problem of malicious activity when found based on log history by the admin[3]. To overcome these threats, an integrated intruder detection system is proposed with certain applications. The intruder detection system performs its performance by recording all the information. The first step begins with user authentication data when logging in. Thus, security is implemented to ensure safety at the logic level[4]. Different users, administrators and other users have different access. Malicious database system transactions relate to security attacks carried out by individuals or organizations internally or externally. External attacks are carried out by attempting to damage or access private data or are executed by unauthorized users [30]. There are three types of attacks and intrusions include application-based, network and host-based intrusion systems[5]. Filtering against spam includes an intruder detection system by detecting specific objects. Thus, there are two types of attacks in the concept of the intruder detection approach, namely signature-based detection systems and anomaly detection systems [6].

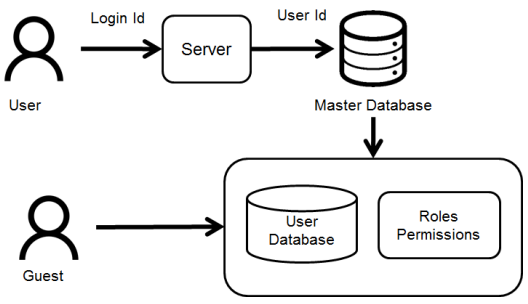

Figure 1. Simple database access system

The database system simply has the system architecture in figure 1. The user authenticates by matching the data to the master database based on the user id. If the user is not registered, it must be

registered. The permissions procedure applies to all database users and accounts

Data models can be presented when they are generated from data-driven discoveries. Data mining has almost the same technique when tracing (retrieving) information from various data sources. It is the main topic of research in the field of artificial intelligence and data processing [7]. Not only in the field of intelligence but ushering in the presence of recognition technology, automatic learning in addition to being faced with big data and statistics. Data mining can also be understood advantages when analyzing data automatically, finding hidden patterns or potential decision support, reducing the impact of risk, and predicting accuracy[8]. The concept of data mining can be considered more familiar with machine learning because of some of the techniques applied to adopt machine learning algorithms[9]. The goal of this algorithm is to find abnormal events hidden in large data volumes[10]. Utilization of raw data as datasets in data mining obtained from various sources. The data consists of structured data included in relational database systems and semi-structured data. Ways to explore knowledge from the information and data obtained include induction, deduction, and mathematical models. The final result of data mining shows that knowledge can be extracted with the need for information retrieval, rapid inquiry and process control. Thus, data mining as a hybrid technology when exploiting and decomposing complex data had reduced to a simple one[11].

At the same time, data mining has the advantage of revealing information and presenting information into knowledge. The rest, the data mining method is the key to how it is applied depending on the type of dataset and the case being investigated. Not only on datasets but methods, one of which can be estimated, such as linear regression analysis. This method exploits various characteristics of data sources. It maps data categories into predictive variable functions. In addition, conducting explorations to build relationships between variables adopted from mathematical theory.

The concept in this paper becomes hard when receiving information from various sources for the intruder detection system. Normal and abnormal patterns were analyzed by the data mining approach is essentially in its role. Thus, it can describe the data mining techniques which are applied when measuring database security problems[12]. an example is estimating or predicting the danger of malicious activity found in the intruder detection system[13].

Most database security builds a detection system when an intruder is present[14]. Intruders can be considered as foreign parties wanting to enter the system or work environment. Usually, intruders are included in illegal users. The weakness of intruder detection only focuses on finding anomalous data, especially on activity logs[15]. Intruder detection integrated with the application is usually equipped with encryption algorithms to protect sensitive data[16]. The performance of the encryption algorithm also needs to be analyzed and considered, although it is still were considered significant in reducing the risk of intruder harm. Thus, the encryption algorithm and intruder detection system can only work when there is log data entered in the database but not as normal[17]. Some intruder detection is implemented in computer networks. The essence of an intruder detection system is to identify malicious activities carried out by users internally and externally.

With the increase in malicious activities that often occur, the motivation in this study is to investigate and estimate the dangers of excessive permissions[18]. To support performance, estimation methods on the concept of data mining will be applied to analyze the forms and patterns of activities carried out by illegal users. Furthermore, the dataset will be derived, namely legal and illegal user data and even redistribute with normal and abnormal data.

2. RESEARCH METHODOLOGY

A. Data Mining Method

Performance in data mining methods is how to browse data, present data to be able to provide knowledge. Several techniques that can be used to reveal the data are classification, clustering, association, and prediction or estimation. One of them is estimating or predicting events based on data as real objects. With the case raised in this paper, it is possible to estimate the population value at the same time as sampling on a dataset containing database activity logs. The reason is that estimation techniques can provide opportunities for decision support, activity scheduling and so on. Therefore, one of the algorithms that can model the estimation of excessive permissions is a linear regression algorithm[31].

B. Linear Regression

Estimation of malicious activity in the form of excessive permissions was carried out using linear regression analysis[19]. This estimation is included in the concept of data mining as an effort to extract patterns from large amounts of data in the activity log in the database system[20]. Thus data mining can be a process of generating valuable information from a set of data or processing data into information presenting knowledge. Method for estimating excessive permissions in multiple linear regression[21]. With regression, there is an independent variable with one dependent variable in the form of a straight line, it can also be considered a linear relationship with the equation below.

$$Y = A + BX$$

This linear regression model relates between two or more independent variables (X1, X2, Xn) with the dependent variable (Y).

$$Yt = a + B1.X1 + B2.X2 + B3.X3 \ldots\ldots\ldots + Bn.Xn$$

Regression coefficients b1, b2, b3 and b can be obtained by the equation below.

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2 + b_3 \sum X_3$$

$$\sum X_{1Y} = a \sum X_1 + b_1 \sum X_{1^2} + b_2 \sum X_1 X_2 + b_3 \sum X_1 X_3$$

$$\sum X_{2Y} = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + b_3 \sum X_2 X_3$$

$$\sum X_{3Y} = a \sum X_3 + b_1 \sum X_1 X_3 + b_2 \sum X_2 X_3 + b_3 \sum X_3^2$$

Thus, there is a search for the constant value and the regression variable for each variable can be determined by the determina

C. *Problem Statement*

Permissions had described as the type of access to the database include the owner of the data is considered the owner, then derived into grant and full-control. Grant and full control types together have full power access to all resources[23]. The difference is that full-control is oriented towards monitoring during the activities carried out by the user while Grant is not. In addition, full control can delete data along with reading or opening up data modifications. Furthermore, the Read type is defined as the type of user who can only open data through certain SQL query while Write is defined as an access type that provides the opportunity to modify and create data.

The sequence of data mining work can be presented in table 1. There are four stages of data mining including business planning and understanding, data understanding, modeling and evaluation. The stage of understanding data almost takes a long time during data mining analysis.

Table 1. Data Mining Steps

Process Stage	Process Description
Planning and Business Understanding	- Define problem objectives - Translate into Data Mining objectives - Define Analysis Approach
Understanding Data	- Data Requirements - Collecting Data from Multiple Source - Combining Data into Single Source - Cleaning Data, Selecting - Transforming Data
Modelling	- Predictive Data Mining - Descriptive Data Mining
Evaluation	- Evaluate Quality - Accuracy, Precision, F-Measure

Identification of attribute problems is shown in table 2. The description of the process shows an explanation of each stage of data mining. Thus, the stages of data mining become transparent.

Table 2. Identify Problem Attributes

Process Stage	Process Description
Planning and Business Understanding	Excessive Permissions and factors were observed
Understanding Data	3 independent variables and 1 dependent variable.
Modelling	- Linear Regression
Evaluation	- Cross-validation - MAE, RMSE, MSE

Referring to the main concept, in cases of excessive permissions, there are 3 independent variables and 1 dependent variable being observed. The figure below presents a research framework on how to estimate the dangers of excessive permissions using data mining, especially in the estimation or prediction method[24].

The three independent variables are Type Permissions, Type of User Account, Status, and User Actions. The data set contains 3 independent variables such as the permissions type consisting of Owner, Grant, full control, read, and write, then the type of user account consists of a system account, superuser account, regular user account, guest user account, and anonymous user. accounts. The status attribute consists of Active, Blocked, and Closed. Finally, user actions consist of Allowing and Disallowing. At the proposal stage, through the estimation method, the datasets used are private as primary data and additional data through secondary or public data are adapted to raise certain cases.

D. *Proposed Framework*

Various sectors have taken advantage of data mining and academics and researchers. The data extraction process continues to drastically improve in addition to its effectiveness[25]. Thus, data mining can be relied on in analyzing data with large, fast and always available data demands. Some steps, in general, determine the pattern in data mining as follows:

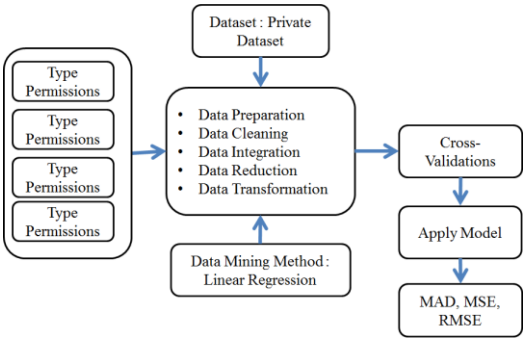

Figure 2 Estimation with a linear regression model

The estimation modeling with linear regression can be shown in figure 2. Observation variables consist of type permissions, type of user account, status and user actions. Data preprocessing and data mining methods are applied together to implement the model. The last stage is measuring model performance with MAD, MSE.

E. Data and Dataset Preparation

Data preparation is an important stage before analyzing the dataset. Data cleansing, data integration, data reduction, and data transformation can be part of the data preparation process. The data preparation process is shown in Figure 3. With this concept, it explains how the data can be used as an analysis of certain cases.. Pattern selection, data representation combined in data mining. Thus, the data extraction process is divided into two parts, namely preprocessing and data extraction[26]. The concept in this paper confirms that data mining methods require large volumes of data. The data mining method can hardly be completed in one stage. In other words, data preparation can be said as a tool for handling raw datasets. Furthermore, the data preparation stage takes a long time apart from being considered a long-term activity[27]. It is caused by reformatting the data, the data must be corrected, and repairing the data by fusion techniques. Data must be entered in a context-appropriate format as a prerequisite when making observations and eliminating distortions.

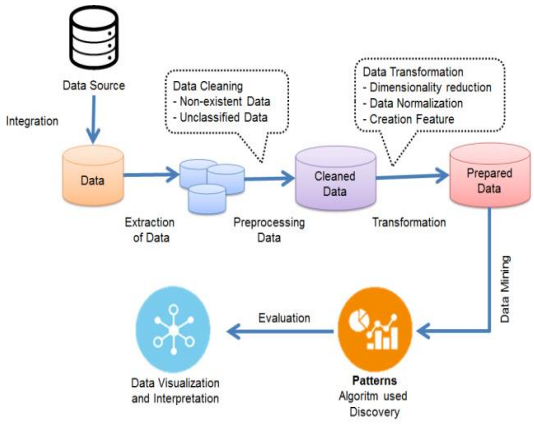

Figure 3 Data Preparation flow

The estimation method is different from classification, clustering and its friends in data mining. The dataset contains a collection of collected data containing activity logs in the database. The data type is almost numeric when executed before entering the data preparation stage. Initially, the existing dataset contained non-numeric content so it had to be converted into another format. Using estimation methods such as linear regression does not match the non-numeric data types.

F. Validation

We use model validation techniques to select how the analysis results and generalize to the data set. This technique has been proven by its reliability, namely Cross-validation. The data analyzed is the independent variable and the dependent variable. The estimation method is very suitable when using cross-validation techniques such as making predictions. In addition, it can estimate the degree of accuracy in the type of predictive model. In this case, adopting a method with estimation capability, then one of the techniques is k-fold cross-validation. The dataset containing the database activity logs will be broken down into k parts of the dataset with the adjusted size. Bias in the data will appear when validating so that this technique can overcome it by eliminating some of it. There is a process of training and testing during validation which is carried out for k iterations of the experiment. Several assumptions and provisions for the distribution of validation are described in the literature, 80% training data and 20% testing data. others, configure data validation to be 70% training data and 30% testing data and can be customized according to the case and data pattern. Thus, in this case, we will use the first ratio of 80%: 20%. In the figure, it can be determined using k fold cross-validation with 10 folds. This 10 step fetch will be applied to the dataset[28].

3. RESULTS AND DISCUSSION

A. Data Preparation

The estimation method with regression technique is different from other data mining methods. The proof is in the dataset used when in the estimation method almost all of the data is numeric with integer or real type. If the classification method such as decision tree or k - Nearest Neighbor uses non-numeric type data, then the estimation method in linear regression becomes different in treating the data. In this dataset, the rapid miner application is used as a data processing tool. Data preparation has data cleaning, data integration, data reduction, and data transformation. In data cleaning to clean data that has not met the format in data mining, it can be confirmed to be valid. Data integration, where we collect various kinds of data related to user activities in the database system when dealing with the system. Almost all activities are recorded in the form of logs and various places are made into one container and combined. Existing datasets sometimes experience overcrowding in format and size, so with data reduction, we make the data more suitable for analysis needs. Finally, there is data in the activity history log of database system users, some data and information are converted into other formats by converting data values that do not have criteria.

A. *Model Validation*

The purpose of validating is to ensure that the model used is sufficiently effective. Although the model has a relative nature, it must be ensured that the accuracy value is measured by measuring accuracy. Thus, data validation in the model becomes an advantage in analyzing problems related to data. Cases raised on the dangers of excessive permissions will be of full attention by looking at the dataset. The dataset containing user logs related to real activities on the database system will be a description of how to analyze the dangers and threats to database security. Thus, the cross-validation method becomes a recommendation invalidating the data. Among the many pieces of kinds of literature, the research results do not present many data validation results related to the models used such as estimation, clustering classification and other data mining methods. It is clear, the rest of the validation test will be able to describe in confirm the results of the analysis that are related to the case of the danger of excessive permissions. The bottom line is that database security does not only focus on how to hide the database but also has to think about access patterns. The flow of data traffic that also involves many users and coupled with the culture of sharing data is also a consideration for the next.

The existing methods of validating facilitate the way of proof in terms of testing datasets. The same is the case for database security on excessive permissions. Thus, the data mining approach will be empowered to test and confirm the estimation of the danger of excessive permissions. Validation with the cross-validation technique responds to dataset testing due to its reliability when it has been widely applied to other research fields of study. In the cross-validation method, the user dataset is divided into 10 iterations with the results presented in the picture.

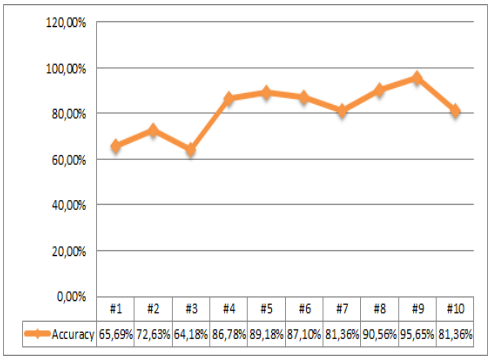

Figure 4. Cross-Validation Measurement

Initially, the accuracy value in the above model test was still considered quite good even though it was only 65.69%. With the addition of iterations on the test, it shows a change in value, especially in the fourth step to the seventh step. This condition has a stable value in maintaining the accuracy value so that it reaches the best point in step 9 which is presented in Figure 4. The largest value in iteration 9 is 95.65% which indicates that the certainty of the dataset remains effective when estimating cases of excessive permissions. The average accuracy of the results of the model testing phase is a consideration of how successful the linear regression method is in estimating the excessive permissions of the dataset[29]. The calculation is 81.45%, which indicates in this study, that the linear regression method in estimating includes good classification.

B. *Model Performance*

1) *Root Mean Square Error*: The dataset is tested by measuring the difference in the dataset containing the value in the estimate of the observed value. The attribute values tested are type_permissions, type_user_account, status and user_actions. The accuracy of the dataset can be considered good because the RMSE recommends a smaller value than a larger one. It must be admitted that the RMSE is very well applied to the estimation even in the case of database security on excessive permissions. The RMSE method provides information on the degree of accuracy in the resulting model. By looking at security conditions such as a database full of sensitive data, mitigation can be done from the estimation results. Not all estimated values can be accurate, but RMSE can provide a fairly clear description.

2) *Mean Absolute Error:* The value on the RMSE is still in the model quality analysis stage. By calculating the average difference as an absolute value with the estimated results. The motivation in implementing MAE is to measure the value of accuracy in the model in estimation or forecasting. MAE and MAPE (Mean Absolute Percentage Error) are almost the same, only have slight differences with the aim of time series analysis (Time Series).

The following table shows the accuracy values in the linear regression model.

Table 3 Performance Model

Performance Vector	**Accuracy Value**
root_mean_squared_error:	37.614 +/- 0.000
absolute_error	31.058 +/- 21.218
relative_error	290.17% +/- 842.78%
normalized_absolute_error:	1.242
root_relative_squared_error	1.303
Mean Absolute Percentage Error	23%

The results of the model performance are shown in table 3. The value of the performance vector is 37,614 while the absolute error of 31,058 shows that the performance of the model is quite good. Finally, the Mean Absolute Percentage is only 23% so that the model is considered reliable.

Linear regression is very suitable to be applied there is an estimation method into the concept of

data mining. The main performance is shown when it involves the dependent variable and the independent variable. These two variables can correlate if using a correlation test. All variables were tested for correlation to see how strong the degree of strength between variables was. If the correlation test does not produce a strong relationship pattern, the regression test cannot be run. Thus, it is confirmed that the regression analysis has been carried out with correlation test analysis.

This discussion, how the involvement of 3 independent variables and 1 dependent variable. The three independent variables as observation variables show different levels of significance. The strongest effect on type_permissions and state when together is on user_actions. The movement of the type_permissions value continues to increase when the state on the user is active. The upgrade condition indirectly becomes a warning for all, including the admin as full control of the data flow. The status attribute also suffers from the same condition. User_action as a predicted class can only give two data chances whether it is allowed or not. Users whose access status is active can be denied or vice versa. Thus, the level of the ratio between user state and type_permissions is a major concern in user management. Some Users are given the freedom to access data. Access overload occurs when there are multiple tasks to be done with the same user. The rest without any consideration in ensuring the user's supervision. Finally, users whose excessive activity is detected are considered an anomaly by the admin. Based on the dataset, it is clear that there are indications of activity anomalies even though the test only uses averages before further analysis. Thus, activity log data can be monitored regularly.

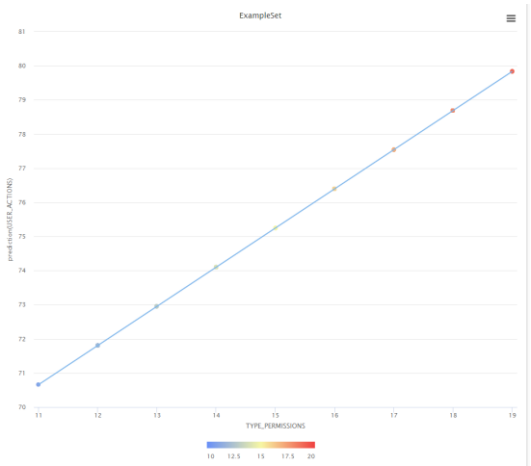

Figure 5. Estimation with a linear regression model

The results of the estimation model are presented in Figure 5. Figure 5 shows that there is an excessive increase in cases. The identification results are quite significant when viewed and compared with real data in the activity log.

Not only researchers, academics and computer practitioners but everyone in the database environment. All are aware by the characteristics of the data that it always moves quickly even in simple data transaction logs. The volume of database activity logs that have small anomalies can be a potential threat. Linear regression in addition to a technique that is considered well-known in several fields including science, but also has reliable predictive power only if it is supported by valid data. In the future, it may be necessary to pay attention to the data to be processed and the analysis techniques.

4. CONCLUSION

We need to analyze and in the case of excessive permissions, the estimation method, especially in the linear regression technique. The summary with address to the result, our findings in root mean squared error is 37.614 and absolute error is 31.058, and mean absolute percentage about 23%. Linear regression method was conducted with the RMSE technique to ensure this model is valid and have high reliability. The most important for analysis to Linear regression related to data simulation widely as statistic field. Today, unity for any field not only statistic but the computer science and other have compatibility to build analysis with big data. This technique has been widely applied because it is also easy to apply to the case of database security estimation. The linear regression technique used can generalize and extract the excessive permission dataset by determining the pattern in the processed data. Type_permissions is said to be the user type in access to database access to be the key when estimated. In addition, type_user_account also indicates a strong relationship with user_action. The prediction results also state that type_permission and type_user_account status are even separate ways to identify the potential danger of excessive permission. Thus, the final result of linear regression provides recommendations for future estimates and also follow-ups for mitigation and policy changes to user management. User management has a ranking of determining database security because data is a highly sensitive resource.

Although the linear regression technique is considered a superior method, it must still be criticized for its weaknesses. The disadvantage is that if the data is entered, the results will be the same. Regression techniques only run procedures with numerical data and produce estimated values. The rest, when there is a causal relationship, is also recorded during the experiments that have been carried out. Sometimes it is difficult for us to build a model and think about the pattern of cause-and-effect relationships by involving the dependent variable and independent variables, especially dealing with cases of excessive permissions. We also make observations in addition to the data that

has been collected. These observations are helpful when specifying variables. The case of excessive permissions is mainly aimed at the user characters that are revealed in the database. A person who has activities outside normal limits can be estimated to be a symptom of excessive permissions. Identification shows that three independent variables contribute and influence the dependent variable such as user_action. It is known, user_action itself is the determination of the user in accessing the database whether it is allowed or not. Efforts to grant permission are also the gateway to the mitigation and monitoring section of database system users.

REFERENCES

[1] P. Menard, G. J. Bott, and R. E. Crossler, "User Motivations in Protecting Information Security: Protection Motivation Theory Versus Self-Determination Theory," *J. Manag. Inf. Syst.*, vol. 34, no. 4, pp. 1203–1230, 2017, doi: 10.1080/07421222.2017.1394083.

[2] R. Diesch, M. Pfaff, and H. Krcmar, "A comprehensive model of information security factors for decision-makers," *Comput. Secur.*, vol. 92, no. 3, pp. 1–21, 2020, doi: 10.1016/j.cose.2020.101747.

[3] N. Miloslavskaya and A. Tolstoy, "Internet of Things: information security challenges and solutions," *Cluster Comput.*, vol. 22, no. 1, pp. 103–119, 2019, doi: 10.1007/s10586-018-2823-6.

[4] J. Norbekov, "Ensuring information security as an ideological problem," *Ment. Enlight. Sci. J.*, vol. 2020, no. 1, pp. 56–65, 2020.

[5] D. Tse, B. Zhang, Y. Yang, C. Cheng, and H. Mu, "Blockchain application in food supply information security," in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2018, vol. 2017-Decem, pp. 1357–1361, doi: 10.1109/IEEM.2017.8290114.

[6] M. D. McLaughlin and J. Gogan, "Challenges and best practices in information security management," *MIS Q. Exec.*, vol. 17, no. 3, pp. 237–262, 2018.

[7] E. Bertino, M. Kantarcioglu, C. G. Akcora, S. Samtani, S. Mittal, and M. Gupta, "AI for Security and Security for AI," *CODASPY 2021 - Proc. 11th ACM Conf. Data Appl. Secur. Priv.*, pp. 333–334, 2021, doi: 10.1145/3422337.3450357.

[8] M. Y. Alyousef and N. T. Abdelmajeed, "Dynamically detecting security threats and updating a signature-based intrusion detection system's database," in *Procedia Computer Science*, 2019, vol. 159, pp. 1507–1516, doi: 10.1016/j.procs.2019.09.321.

[9] Y. A. Basallo, "Artificial Intelligence Techniques for Information Security Risk Assessment," *IEEE Lat. Am. Trans.*, vol. 16, no. 3, pp. 3–7, 2018.

[10] J. Chevalier and D. Buckles, *Participatory Action Research*. 2019.

[11] S. S. Sarmah, "Database Security – Threats & Prevention," *Int. J. Comput. Trends Technol.*, vol. 67, no. 5, pp. 46–53, 2019.

[12] W. C. Alisawi, A. A. A. Hussain, and W. A. Alawsi, "Estimate new model of system management for database security," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1391–1394, 2019, doi: 10.11591/ijeecs.v14.i3.pp1391-1394.

[13] R. A. Teimoor, "A Review of Database Security Concepts, Risks, and Problems," *UHD J. Sci. Technol.*, vol. 5, no. 2, pp. 38–46, 2021, doi: 10.21928/uhdjst.v5n2y2021.pp38-46.

[14] K. Jonah, M. Elem, N. Elem, C. Obinna, and N. Chiemezuo, "Online Database Security Threats and Solutions: The NetFlix Incident," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 12, pp. 1–6, 2020, [Online]. Available: www.ijisrt.com.

[15] J. H. Park, S. M. Yoo, I. S. Kim, and D. H. Lee, "Security Architecture for a Secure Database on Android," *IEEE Access*, vol. 6, pp. 11482–11501, 2018, doi: 10.1109/ACCESS.2018.2799384.

[16] M. Miao, Y. Wang, J. Wang, and X. Huang, "Verifiable database supporting keyword searches with forward security," *Comput. Stand. Interfaces*, vol. 77, p. 103491, 2021, doi: 10.1016/j.csi.2020.103491.

[17] G. Chen, "Security precautionary technology for enterprise information resource database based on genetic algorithm in age of big data," *J. Comput. Methods Sci. Eng.*, vol. 20, no. 2, pp. 427–434, 2020, doi: 10.3233/JCM-193874.

[18] A. Motro, "A unified model for security and integrity in relational databases," *J. Comput. Secur.*, vol. 1, no. 2, pp. 189–213, 1992, doi: 10.3233/JCS-1992-1204.

[19] J. Pek, O. Wong, and C. M. Wong, "Data Transformations for Inference with Linear Regression: Clarifications and Recommendations," *Pract. Assessment, Res. Eval.*, vol. 22, no. 9, pp. 1–11, 2017.

[20] A. Jozaghi *et al.*, "Multi-model streamflow prediction using conditional bias-penalized multiple linear regression," *Stoch. Environ. Res. Risk Assess.*, vol. 35, no. 11, pp. 2355–2373, 2021, doi: 10.1007/s00477-021-02048-3.

[21] L. Zhou, Y. Zhu, and K. K. R. Choo, "Efficiently and securely harnessing cloud to solve linear regression and other matrix

operations," *Futur. Gener. Comput. Syst.*, vol. 81, no. 2, pp. 404–413, 2018, doi: 10.1016/j.future.2017.09.031.

[22] A. F. Schmidt and C. Finan, "Linear regression and the normality assumption," *J. Clin. Epidemiol.*, vol. 98, no. 10, pp. 146–151, 2018, doi: 10.1016/j.jclinepi.2017.12.006.

[23] Z. Jin, Y. Cui, and Z. Yan, "Survey of intrusion detection methods based on data mining algorithms," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, no. 2, pp. 98–106, 2019, doi: 10.1145/3341620.3341632.

[24] J. Zhao, S. Fang, and P. Jin, "Modeling and Quantifying User Acceptance of Personalized Business Modes Based on TAM , Trust and Attitude," *J. Sustain.*, vol. 10, no. 356, pp. 1–26, 2018, doi: 10.3390/su10020356.

[25] P. Zeng, G. Lin, L. Pan, Y. Tai, and J. Zhang, "Software vulnerability analysis and discovery using deep learning techniques: A survey," *IEEE Access*, vol. 8, no. 4, pp. 197158–197172, 2020, doi: 10.1109/ACCESS.2020.3034766.

[26] V. R. Balpande and R. D. Wajgi, "Prediction and severity estimation of diabetes using data mining technique," *IEEE Int. Conf. Innov. Mech. Ind. Appl. ICIMIA 2017 - Proc.*, no. Icimia, pp. 576–580, 2017, doi: 10.1109/ICIMIA.2017.7975526.

[27] M. A. Meena, "Data Mining Techniques Used in Cyber Security," *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol. 4, no. 11, pp. 11–19, 2018, [Online]. Available: http://www.ijfrcsce.org.

[28] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Comput. Stat. Data Anal.*, vol. 120, no. xxxx, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.

[29] R. C. Sharma, K. Hara, and H. Hirayama, "A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data," *Scientifica (Cairo).*, vol. 2017, 2017, doi: 10.1155/2017/9806479.

[30] A. A. Argasah and D. Gustian, "DATA MINING ANALYSIS TO DETERMINE EMPLOYEE SALARIES ACCORDING TO NEEDS BASED ON THE K-MEDOIDS CLUSTERING ALGORITHM", *J. Tek. Inform. (JUTIF)*, vol. 3, no. 1, pp. 29-36, Feb. 2022.

[31] Y. I. Kurniawan, A. Fatikasari, M. L. Hidayat, and M. Waluyo, "PREDICTION FOR COOPERATIVE CREDIT ELIGIBILITY USING DATA MINING CLASSIFICATION WITH C4.5 ALGORITHM ", *J. Tek. Inform. (JUTIF)*, vol. 2, no. 2, pp. 67-74, Mar. 2021