Journal of Positive School Psychology
2022, Vol. 6, No. 3, 8442–8451

Implementation of SARIMA Algorithm in Understanding Cybersecurity Threats in University Network

Norkhushaini Awang[1] , Ganthan A/L Narayana Samy[2] , Noor Hafizah Hassan[3] , Nurazean Maarop[4] and Sundresan Perumal[5]

[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia*
[2,3,4]*Razak Faculty of Technology and Informatics, UniversitiTeknologi Malaysia (UTM), Malaysia*
[5]*Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Malaysia*
Email: [1]shaini@tmsk.uitm.edu.my, [2]ganthan.kl@utm.my, [3]noorhafizah.kl@utm.my, [4]nurazean.kl@utm.my, [5]sundresan.p@usim.edu.my

Abstract

Currently, there are few studies on cybersecurity threats risk assessment in the university network. This research aims to fill the gaps by identifying cybersecurity threats through a quantitative study conducted in selected university. This study aims to investigate the use of predictive analysis by looking at network packets in the university network. We proposed a prediction model using Time Series Analysis (TSA) whereby data is gathered from the selected university network firewall. Conducting a risk assessment is a very important activity in an organization. The results of the risk assessment process can help network admin to make decisions in managing risks. The risk assessment also important because university network have many systems to be protected from cybersecurity threats. With risk assessment, network admin can manage the risks and prevent before cybersecurity threats interrupt the whole system in the university network. In this research, we adapted quantitative methods to analyze risk. Moreover, there are few studies on risk assessment prediction at university network. In building predictive models, we implemented the Seasonal Autoregressive Integrated Moving Average (SARIMA) method for time series forecasting with univariate data containing trends and seasonality. SARIMA has built a predictive model by looking at several variables namely seasonal autoregressive order, seasonal difference order, seasonal moving average order and also the number of time steps for a single seasonal period. The conclusion from this study shows that by using SARIMA algorithm, the researcher get the best prediction value in order to get a small Root Mean Squared Error (RMSE) value

Keywords: cybersecurity threats, risk assessment, time series analysis, sarıma algorithmn, predictive analytics

I. INTRODUCTION

Cybersecurity is a practical way to protect network infrastructure and data from malicious attacks. Nowadays, it becomes an important thing to manage cybersecurity from threats as computer networks are now increasingly complex. The discussion from (Knight & Nurse, 2020) explains that major cybersecurity incidents can represent a cyber crisis for organizations, primarily because of the risks associated with reputational damage. The effects of the cyber crisis such as data breaches affect the reputation and competitiveness of companies. This is a critical problem and research should be done to provide a solution to this problem. Other researchers suggested the use of a framework and proper methodology

can help organization understand cybersecurity threats and in turn can understand attack patterns in their study (Sheehan, Murphy, Kia, & Kiely, 2021). The right methodology gave an initial picture of how organizations can manage cybersecurity in their organizations. In Malaysia, the Information Security Management Plan Report is referred in managing public sector cybersecurity network (CyberSecurity Malaysia, MAMPU, MIMOS, & Chief Government Security Office, 2016). In this report stated that the information security in relation to system architecture, technology and security controls. These can be classified into several categories which are personal computing devices, network devices, applications, servers and physical environment. This report views cybersecurity as necessary in the government sector in Malaysia as government sector is now done in cyber networks.

II. RELATED STUDY

In understanding the threats that occur in the university network, we look at the categories of threats that may occur in the network. As discussed by (Singh & Joshi, 2017) stated that in university network, they found that SQL injection, weak password and cross-site request forgery attack is the most attack capture from their analysis. Another researcher (Guo, 2019) mention that data transmission and storage increase the risks of data theft and virus infection in university network. According to (Lazar, Cohen, Freund, Bartik, & Ron, 2021) mention about their studied performed from server domains found that cybersecurity threats occur on networks by masquerading as legitimate websites known as phishing attacks. This attack occurs by taking the vulnerabilities found in the user's web browser. Researchers have also found that there are malicious domains for which it is a well-known problem in cybersecurity systems. Research on cybersecurity threats on the university network were also conducted by (Yevseiev et al., 2019) mention that in their research, they focusing on web infrastructure where the findings from this study found that the university network is vulnerable to cybersecurity threats through the web as it is access to applications in the university today.

The research about cybersecurity threats in university network also been discussed by (Roberts, 2013) in the report. The author emphasizes that many university users do not comprehend the risks to basic information security. Users of university ignore a compromised system that can be used by computer networks to target another system. In the report, author list seven (7) cybersecurity threats that can be found in university network. The threats that have been reported in university network are key loggers, viruses, worms and Trojans, denial of service attacks, sniffers, wireless sniffing, file sharing threats and abundance of bandwidth. All of these threats occur through the use of applications on the web. In 2002, University of California banned Windows and Windows NT from being used by all campus users. This policy is made because numerous security threats of viruses, worms and denial-of-service attacks were reported and caused many outages of the university network. Support staff argued that it was very difficult to maintain student computers installed in a stable way. A research conducted by (Georgetown University, 2017) listed ten (10) types of threats which found in university network. The threats are includes technology with weak security, social media attacks, mobile malware, third-party entry, neglecting proper configuration, outdated security software, social engineering, lack of encryption, corporate data on personal devices and inadequate security technology. From these ten threats, most threats occur within the network and the rest occur on software.

According to (Roberts, 2013) mention that the Computer Emergency Response Team reported that the number of network incidents in university campus in 2001 was 52,658, which jumped from 21,756 the previous year. The increasing of threats in today’s universities because of the lead of technological advancement (Joshi, 2016). The author also

discuss about accessing to technology in the university campus results in vulnerable computing environment with more security threats. University is a place provided with various technologies in preparing for student learning including Wi-Fi technology facilities, online learning, digital library, and web conferencing. The widespread use of the internet within university networks leaves universities vulnerable to cybersecurity threats. Attacks that took place on the university network were also discussed by (Naagas & Palaoag, 2018) mention about twenty six (26) threats that might happen in university network which are spoofing, sniffing, session hijacking, denial of service, viruses, foot printing, password cracking, arbitrary code execution, buffer overflow, cross-site scripting, SQL injection, network eavesdropping, elevation of privilege, brute force attacks, dictionary attacks, man in the middle, information disclosure, attacker exploits, war driving and wireless attack. In their study, random black box penetration testing was implemented in assessing the university network. From the 26 threats identified by researchers, we can classify these threats into 3 categories, which are threats from viruses, from web applications and also intrusion into the network.

This discussion is continued by (Joshi, 2016) on cybersecurity threats in university network where authors explained that university campus mainly suffers following security threats as groups such as phishing, ransomware, and malware, viruses spreading through social media, mobile devices operating system vulnerability and embedded devices connectivity. All cyber threats obtained from previous studies point to some general classifications. We can conclude that cyber security threats can be categorized into network intrusion, malware and web application threats. The threats might have the same with others researchers found or the new threats depends on the campus users activities using the information system in campus network.

III. METHODOLOGY

A. Proposed Risk Assessment

Risk analysis is performed by looking at, the availability and reliability of information and available resources. In this research, Time Series Analysis is adapted to forecast cyber threats. In this study, we have collected data from firewall logs to study cybersecurity threats. This firewall log is taken from selected university from 2019 to 2021. The logs taken include information on threats from web filtering, malware filtering and also intrusion systems.

The previous discussion has explained, at this phase, we used quantitative methods in conducting experiments. In this study, we used machine learning as an application of artificial intelligence in analyzing the collected data. A study from (Paltrinieri, Comfort, & Reniers, 2019) stated that researchers used a machine learning approach to address challenges in making risk assessments. This approach facilitated researchers get suggestions answers in the study conducted. Past studies from (Boutaba et al., 2018) (Salim et al., 2021) (Syed Nor, Ismail, & Yap, 2019) also mentioned that machine learning simplified in their study by examining the data and represented conclusions of knowledge. In using this machine learning approach, the programming language we used is python language in developing prediction algorithms as well as data exploring.

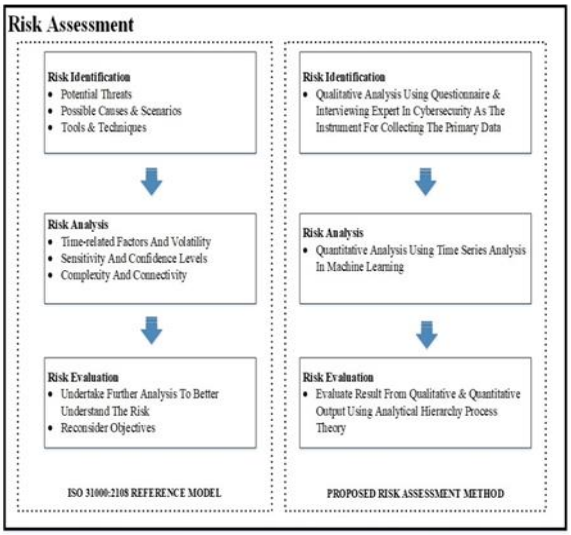

Figure 1. Proposed Risk Assesment Framework

Figure 1 shows the proposed risk assessment framework used in this study. In this study, we adapted the ISO 31000: 2018 standard as a reference model in conducting risk assessment. We applied a mix-method approach in conducting this research. Qualitative and quantitative methods are used in our study.The discussion in this paper covered the use of SARIMA algorithm in developed cybersecurity threat prediction that occurs in university network.

B. Time Series Analysis

Cybersecurity threats included computer malware, data breach, Denial of Service (DoS) attacks and other attack vectors included in our studies that can be identified by the firewall. In this study, we implemented Time Series Analysis (TSA) algorithm in conducting a cybersecurity threat analysis. Time Series Analysis is an important field of machine learning since there are several problems involving time components to make predictions. A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. An observed time series can be decomposed into three components which are the trend a long term direction, the seasonal a systematic, calendar related movements and the irregular an unsystematic, short term fluctuations. We used this firewall log in conducting the study. This raw data is unstructured data where we have to do a data massage to understand and explore the data.

According to (Farsi et al., 2021) mention that time series analysis is widely used in gathering and analyze past series observations to extract hidden structure, meaningful statistics, and other features to deduce acceptable models. Explanation from their study elaborate that Autoregressive Integrated Moving Average (ARIMA) is the most commonly used time series model, well-known for statistical forecasting, with a notable forecasting accuracy and efficiency. In a study conducted by (Liu, Loo, & Pasupa, 2021) mention that they used machine learning to generate a time series prediction model. Researchers also mention that time series analysis is necessary for a continuous stream of data, such as time.

For other researchers, data-driven inventory found to be useful for time-series forecasting (Punia, Singh, & Madaan, 2020). Researchers also used machine learning in producing an estimation models to leverage the vast amount of data generated by a large number of products in a retail environment to generate accurate demand estimates. A research about the time series were also discussed by (Dimri, Ahmad, & Sharif, 2020) mention that analysis of time series invariably involves the evaluation of trends and seasonality in the data. In their research, they look at dynamic climate structures as well as temperature changes using time series analysis. From previous studies, it shows that time series analysis are used to examine how changes related to selected data points are compared to other variables over the same time period. In this study, we used time series analysis with applications to machine learning on data network packets that have been filtered by the firewall. Our study later discussed how time series analysis develop a forecasting threats model that occurred so that network management can be prepared in advance in preventing cyber threats from occurring.

IV. FINDINGS

In conducting this study, we have adapted the data analytics approach using machine learning. Data analytics process goes through several processes that start with data collection. Next, the processes involved in this data analysis include exploratory data analysis, data preparation, feature engineering, model building and model evaluation.

A. Exploratory Data Analysis

Exploratory data is done by looking at the raw data and discarding data that contains null values. In order to understand the firewall log dataset, we have categorized cybersecurity threats into 3 main groups, namely web application attack, malware attack and network

intrusion attack. Figure 2,3,4 shows the data obtained from the firewall for the 3 categories of attack. Data taken from 2019 -2021 in university network. We identified from firewall log, data in web application attack are equal to each other where values are identical and gives a standard deviation value = 0. We have combined all three categories of cybersecurity threats to allow us to perform data analysis later. This can be seen from Figure 5.

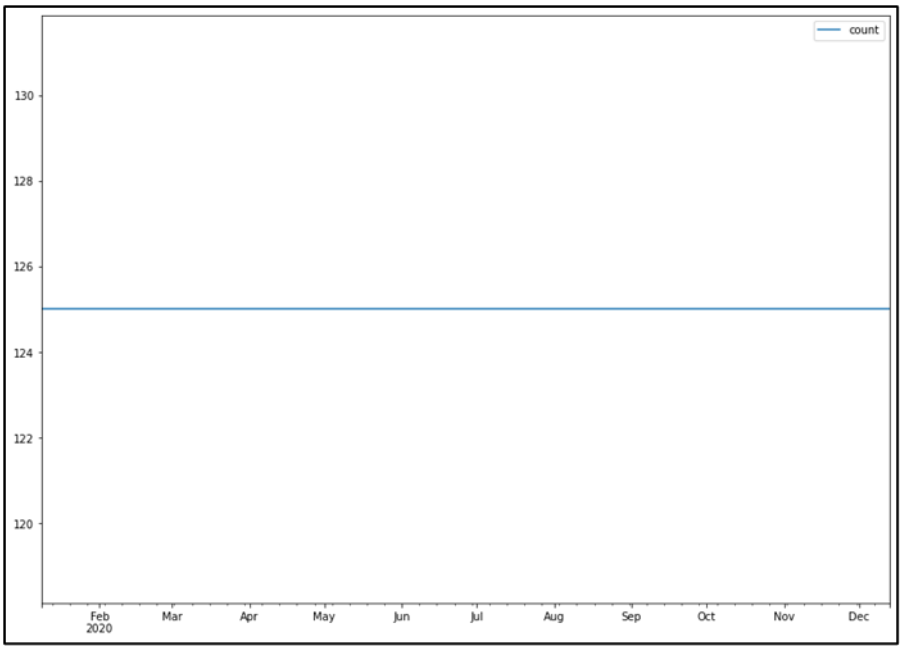

Figure 2. Firewall Log for Web Application Attack

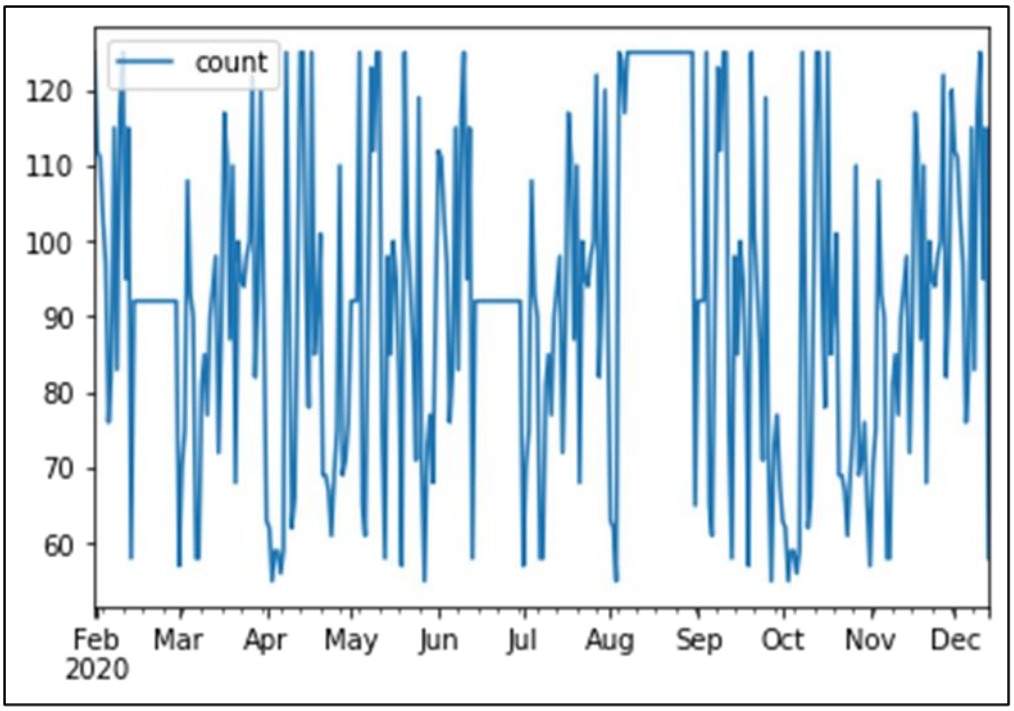

Figure 3. Firewall Log for Malware Attack

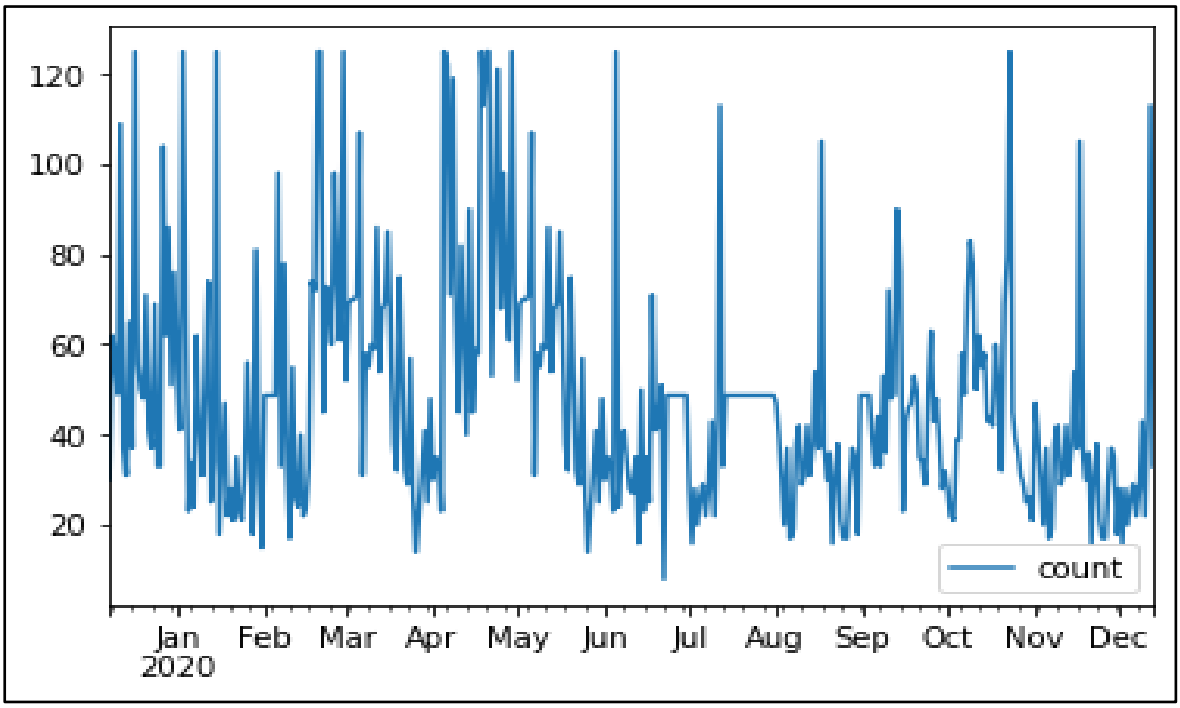

Figure 4. Firewall Log for Intrusion Attack

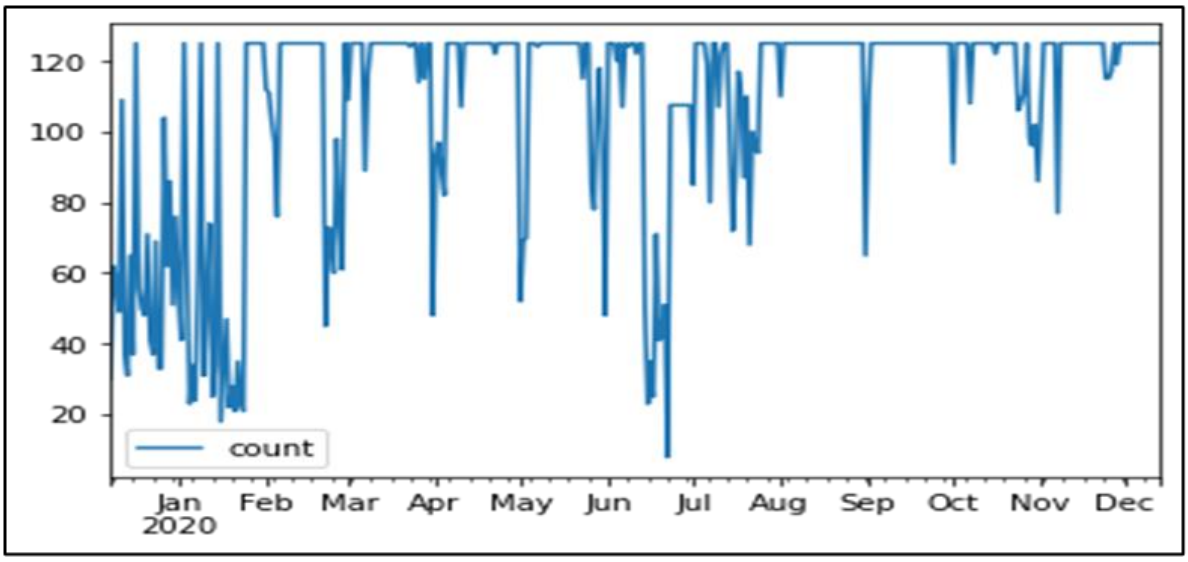

Figure 5. Combined 3 Categories of Cybersecurity Attack

In the next phase, we conducted a test after taking out the outliers to learn about the current data. We used Dickey Fuller Augmented (ADF) test to the firewall log. The ADF test is performed to see if a particular time series data is stationary data. This is one of the most commonly used statistical tests when analyzing series stationary.

B. SARIMA Algorithm

In this study, we have proposed to use SARIMA as a time series analysis algorithm on firewall log data taken from selected university. SARIMA is Seasonal ARIMA where ARIMA is a statistical analysis model that uses time-series data to either better understand the data set or to predict future trends. We have decomposed the time series data into examples of trend and seasonality and further described the observed time series, trends, seasons, and residues. From the observation through Figure 6, we can see that the trends and seasonal information taken from this firewall log data give an initial picture of cybersecurity attacks for the first 3 years. The results from the decomposed the time series data provides an explanation mainly for the analysis of time

series, to tell in general the problems that may occur to the forecasting model that will be done. We can see that time series cybersecurity attacks are not seasonal and not stationary with no consistency and no recurrence described and modeled as in Figure 6. This may be because the number of users using the network at the university network in year 2020 are less because students, management staff and lecturers work from home. This shows that with the increase in the number of users in the network is vulnerable to cybersecurity threats.

Next we have implemented SARIMA algorithm on the time series data and obtained the diagram as in Figure 7. From the results obtained shows that the predictions issued a lot of noise towards the time series data. According to (Ramli, Ab Mutalib, & Mohamad, 2018) (Mutalib, Ramli, & Mohamad, 2018) (Ab Mutalib, Ramli, & Mohamad, 2017) have discussed the value of RMSE in making predictions in time series. A small RMSE value indicates that the proposed model is very good. However, in our study, the RMSE value obtained is 10 where the proposed model does not provide the best fit to the dataset. From the results of this study show that the use of SARIMA algorithm is a good method in looking at forecasting attacks that may occur in the future.

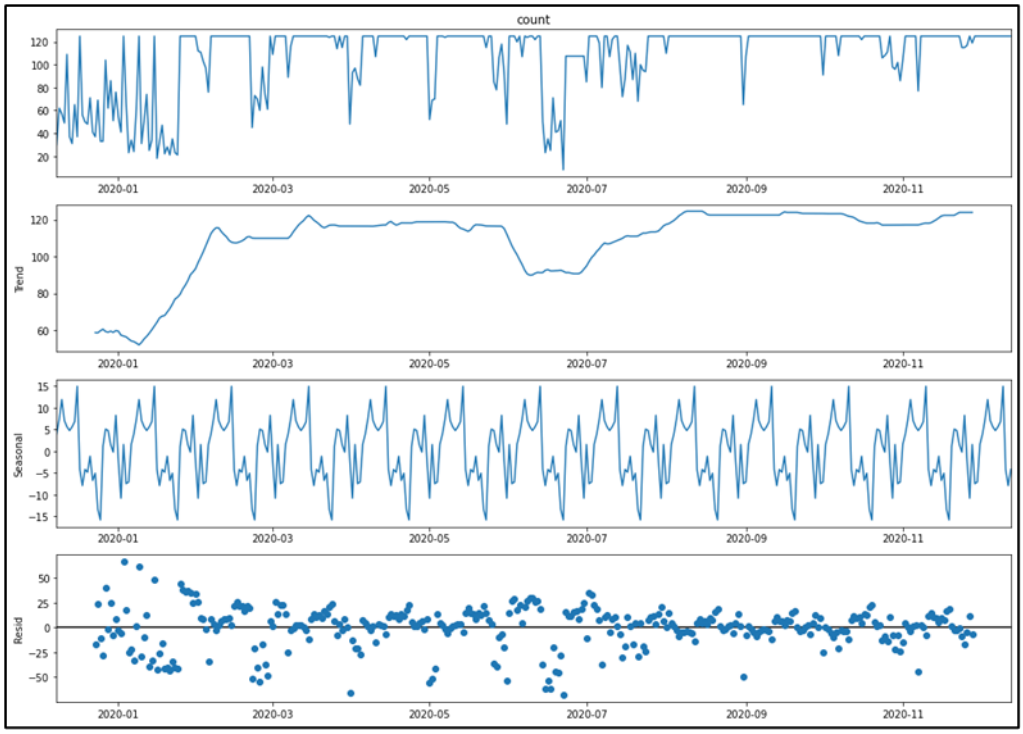

Figure 6. Decomposed Time Series Data

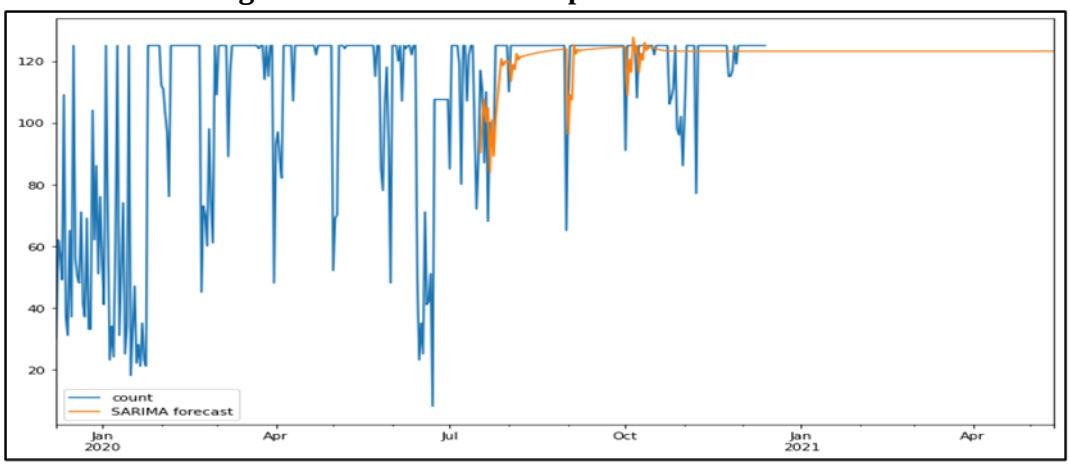

Figure 7. SARIMA Algorithmn to Time Series Data

Experiments need to be done on time series data by reducing the value of the RMSE value in future work.

V. DISCUSSION AND CONCLUSION

As discussed earlier, using predictive methods can produce results in analyzing risks. Researchers used machine learning method in conducting a predictive analysis, where the data used is trained through statistical models to understand the data. To perform predictive analysis, extensive data is explored, and this process is called data mining. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. For us to carry out data analysis, the process of data mining needs to be done in understanding the data we have. According to (Pokhrel, Rodrigo, & Tsokos, 2017), vulnerabilities are always present on computer networks and cause risks in cyber security. To find vulnerabilities for network systems, analytics methods can be done.

Recently, university databases have been exposed to security threats that have prompted university management to seek tools in better risk monitoring and modelling. According to (Giannopoulos, Filippini, & Schimmer, 2012) explained that by using data statistics techniques, it helps network administrators to analyse and display data coherently and structurally. The authors also stated that it is very important to conduct an accurate risk assessment using appropriate methodologies in finding potential network security threats. In this research, we have recommended using predictive analysis, where with this forecasting model can help management make decisions. There are a few data mining algorithm in predictive modelling which are logistic regression, time series analysis, decision tree and neural network. From a study, (Ahmed, Calders, Lu, & Pedersen, 2014) mention how they apply the risk prediction analysis to predict risk in real time. The model is used to forecast an outcome at desired future state or time based refer from data inputs. From the authors view, analyzing from data input is much faster after doing sampling activities before choosing the right model.

It can be concluded that by doing predictive analysis, this allows the organization to be more proactive in making predictions before an event occurs and this can help them to make better decisions.

VI. ACKNOWLEDGMENT

The authors would like to thank the Faculty of Computer and Mathematical Sciences (FSKM), Universiti Teknologi Mara (UiTM) for sponsoring this research.

BIBLIOGRAPHY

1. Ab Mutalib, S. M., Ramli, N., & Mohamad, D. (2017). Forecasting unemployment based on fuzzy time series with different degree of confidence. Journal of Telecommunication, Electronic and Computer Engineering, 9(1–4), 21–24.
2. Ahmed, T., Calders, T., Lu, H., & Pedersen, T. B. (2014). Risk Detection and Prediction from Indoor Tracking Data, 11–18.
3. Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. Journal of Internet Services and Applications, 9(1). https://doi.org/10.1186/s13174-018-0087-2
4. CyberSecurity Malaysia, MAMPU, MIMOS, & Chief Government Security Office. (2016). Rangka Kerja Keselamatan Siber Sektor Awam (Vol. versi 1.0). Retrieved from http://www.mampu.gov.my/images/suara_anda/RAKKSSA-VERSI-1-APRIL-2016-BM.pdf%0Ahttps://www.malaysia.gov.my/portal/content/30090

5. Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. Journal of Earth System Science, 129(1). https://doi.org/10.1007/s12040-020-01408-x
6. Farsi, M., Hosahalli, D., Manjunatha, B. R., Gad, I., Atlam, E. S., Ahmed, A., … Ghoneim, O. A. (2021). Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data. Alexandria Engineering Journal, 60(1), 1299–1316. https://doi.org/10.1016/j.aej.2020.10.052
7. Georgetown University. (2017). Top 10 Threats to Information Security.
8. Giannopoulos, G., Filippini, R., & Schimmer, M. (2012). Risk assessment methodologies for Critical Infrastructure Protection. Part I: A state of the art. European Commission JRC (Joint Research Center) Technical notes. https://doi.org/10.2788/22260
9. Guo, J. (2019). Big Data Security and Privacy Protection in Colleges and Universities. Application of Intelligent Systems in Multi-Modal Information Analytics, 929, 727–735. https://doi.org/10.1007/978-3-030-15740-1
10. Joshi, C. (2016). Quantitative Information Security Risk Assessment Model for University Computing Environment. In International Conference on Information Technology (pp. 69–74). https://doi.org/10.1109/ICIT.2016.11
11. Knight, R., & Nurse, J. R. C. (2020). A Framework for Effective Corporate Communication after Cyber Security Incidents. Computers & Security Journal, 40(2), 366–374.
12. Lazar, D., Cohen, K., Freund, A., Bartik, A., & Ron, A. (2021). IMDoC: Identification of Malicious Domain Campaigns via DNS and Communicating Files. IEEE Access, 9, 45242–45258. https://doi.org/10.1109/ACCESS.2021.3066957
13. Liu, Z., Loo, C. K., & Pasupa, K. (2021). A novel error-output recurrent two-layer extreme learning machine for multi-step time series prediction. Sustainable Cities and Society, 66, 102613. https://doi.org/10.1016/j.scs.2020.102613
14. Mutalib, S. M. A., Ramli, N., & Mohamad, D. (2018). Forecasting fuzzy time series model based on trapezoidal fuzzy numbers with area and height similarity measure concept. AIP Conference Proceedings, 1974, 1–8. https://doi.org/10.1063/1.5041571
15. Naagas, M. A., & Palaoag, T. D. (2018). A Threat-Driven Approach to Modeling a Campus Network Security. In International Conference on Communications and Broadband Networking (pp. 1–7). https://doi.org/10.1145/3193092.3193096
16. Paltrinieri, N., Comfort, L., & Reniers, G. (2019). Learning about risk: Machine learning for risk assessment. Safety Science, 118(June), 475–486. https://doi.org/10.1016/j.ssci.2019.06.001
17. Pokhrel, N. R., Rodrigo, H., & Tsokos, C. P. (2017). Cybersecurity: Time Series Predictive Modeling of Vulnerabilities of Desktop Operating System Using Linear and Non-Linear Approach. Journal of Information Security, 08(04), 362–382. https://doi.org/10.4236/jis.2017.84023
18. Punia, S., Singh, S. P., & Madaan, J. K. (2020). From predictive to prescriptive analytics: A data-driven multi-item newsvendor model. Decision Support Systems, 136(May), 113340.

https://doi.org/10.1016/j.dss.2020.113340

19. Ramli, N., Ab Mutalib, S. M., & Mohamad, D. (2018). Fuzzy Time Series Forecasting Model based on Centre of Gravity Similarity Measure. Journal of Computer Science & Computational Mathematics, 8(4), 121–124. https://doi.org/10.20967/jcscm.2018.04.010
20. Roberts, T. L. (2013). Information Security in Higher Education: Threats & Response. Global Information Assurance Certification Paper. SANS Institute. Retrieved from http://zma.es/Incident Handler/real-world-arp-spoofing/real-world-arp-spoofing_487.pdf
21. Salim, N. A. M., Wah, Y. B., Reeves, C., Smith, M., Yaacob, W. F. W., Mudin, R. N., … Haque, U. (2021). Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Scientific Reports, 11(1), 1–9. https://doi.org/10.1038/s41598-020-79193-2
22. Sheehan, B., Murphy, F., Kia, A. N., & Kiely, R. (2021). A quantitative bow-tie cyber risk classification and assessment framework. Journal of Risk Research, 0(0), 1–20. https://doi.org/10.1080/13669877.2021.1900337
23. Singh, U. K., & Joshi, C. (2017). Information security risk management framework for University computing environment. International Journal of Network Security, 19(5), 742–751. https://doi.org/10.6633/IJNS.201709.19(5).12
24. Syed Nor, S. H., Ismail, S., & Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model. Journal of Economics, Finance and Administrative Science, 24(47), 157–170. https://doi.org/10.1108/JEFAS-08-2018-0076
25. Yevseiev, S., Aleksiyev, V., Balakireva, S., Peleshok, Y., Milov, O., Petrov, O., … Shmatko, O. (2019). Development of a methodology for building an information security system in the corporate research and education system in the context of university autonomy. Eastern-European Journal of Enterprise Technologies, 3(9–99), 49–63. https://doi.org/10.15587/1729-4061.2019.169527