Health Maintenance In The New Normal: An Analysis Of University Students' Awareness Using Clustering And Naïve Bayes

Fiona Tanadi[1], Raymond Sunardi Oetama[2*]

[1,2] Information System Department, Universitas Multimedia Nusantara
Tangerang, Banten 18511, Indonesia.
[*]Corresponding Author:
Email: raymond@umn.ac.id

Abstract.

The new normal has posed significant challenges for students in maintaining their health. To address this issue, this study aimed to examine the awareness of university students in maintaining their health during this period using clustering and Naïve Bayes methods. The results of the study revealed two distinct clusters, "Aware" and "Not Aware," indicating that Indonesian students have varying levels of awareness regarding health precautions during the new normal. Most students are in the "Aware" cluster, suggesting a positive trend in health-conscious behavior among students in Indonesia. The Naive Bayes method was used to validate the cluster analysis, which resulted in an overall accuracy of approximately 96.25%. The study also found that the students are generally trying to maintain their health during the new normal, with most scoring between moderate and good. However, there are differences in the levels of effort put in by different age groups, with those aged 22 scoring lower. Moreover, students are generally making a good effort to maintain their health in public places during the new normal, but again, those aged 22 had the lowest scores. In terms of healthy food consumption, most students across all age groups scored a level of moderate consumption. The purpose of these findings is to inform policymakers and healthcare providers to develop sustainable programs for improving health awareness among students and addressing the challenges presented by the new normal.

Keywords: *Clustering, health awareness, naïve Bayes, and new normal.*

I. INTRODUCTION

The New Normal has brought about significant changes in how students live their lives, and the impact of these changes will likely continue even after the pandemic is over [1]. Although students' activities may gradually resume, they may be different from before, as students have learned the importance of being adaptable and resilient in the face of challenges. Studying from home may become more common, reducing commuting and travel, and leading to positive impacts on the environment and quality of life. Precautions such as wearing masks, practicing social distancing, and washing hands regularly when going out for activities such as shopping, dining, or socializing, may continue to be part of the new normal, especially with the emergence of new variants of the virus [2]. These changes highlight the importance of taking care of their health and being mindful of their actions.This pandemic has brought health to the forefront of students' minds, and it is crucial to maintain good health to carry out daily activities optimally [3].

By adhering to this awareness, students can not only protect themselves but also help break the chain of transmission of COVID-19, safeguarding the health of their loved ones and those around them [4]. The pandemic has highlighted the importance of collective action in mitigating its impact, and the initiatives taken by students demonstrate their ability to adapt and respond to disasters. Such experiences can serve as valuable social capital for students to face future health crises with readiness and resilience [5].Previous studies have examined Health Awareness using different methods. For example, one study utilized QRIS [6], while another employed Latent Dirichlet Allocation [7]. Additionally, some previous studies have used clustering and naive Bayes, but their focus was on sentiment analysis [8-9]. In contrast, this research aims to determine the clustering of health awareness using data mining techniques. The analysis is carried out using the K-Means Clustering Algorithm. The output produced is a group of students who have implemented health awareness. Afterward, there is the Naïve Bayes algorithm, which is also used in this study as a validation of the labeling data from the results of the Clustering process. The purpose of this study is to classify public health awareness based on age among Indonesian students.

II. METHODS

2.1 Data Collection Method

The object of this research is to several aspects of public health viewed from how aware the public is about the importance of health, which has a significant impact during the new normal period. The motivation for choosing this object is to obtain a clearer picture of the public's awareness of their health especially among university students. Population refers to a group of individuals or objects that share certain characteristics, while a sample represents a subset of the population that possesses the same characteristics and features [10]. This study is approached as a case study in the Information System Study Program at Universitas Multimedia Nusantara. The total number of active students in this study program is about 400 students as the population. The sample size is taken according to the Slovin formula [11]:

$$n = \frac{N}{1 + Ne^2} \quad (1)$$

Where n is the sample size, N is the number of populations, and e is the error margin. By taking N=400 and e=10% the minimum sample that must be taken is 80.

2.2 Research Flows

In practice, data mining is often referred to as Knowledge Discovery in Database. This approach involves a series of steps that includes data selection, Preprocessing, Transformation, Data Mining, and Pattern Evaluation [12]. The step-by-step process of Knowledge Discovery in the Database is shown in Fig. 1.

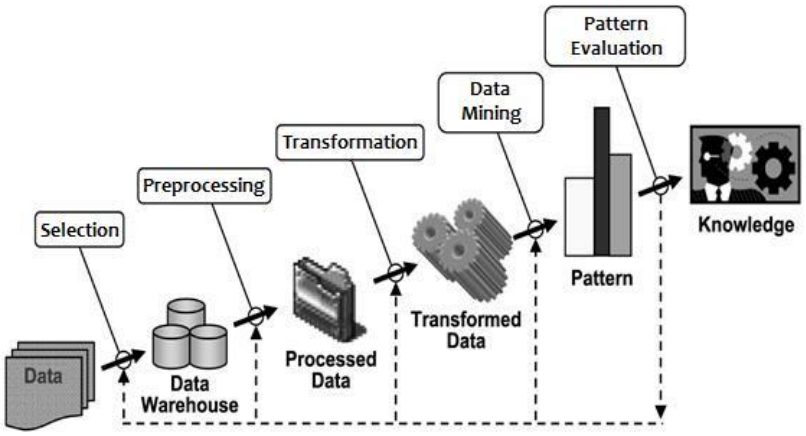

Fig 1. Knowledge Discovery in Database Processes [13]

The data source for this study is obtained from the results of a survey containing several general questions that have been distributed previously, for 100 responses selected randomly to answer the questions asked in the Google Form. After collecting the data, it is then pre-processed to be able to be analyzed in data analysis, because in data analysis, incomplete attributes cannot be systematically handled by algorithms. As a result, only 80 out of 100 respondents are valid. However, it still accomplishes the minimum sample size. Data cleansing by removing data that is deemed unnecessary or cannot be used, then transforming the cleansed data into the data frame in the R program. Data transformation involves converting the raw data into a format that is more suitable for Clustering.After transformation, the next step is to perform data mining. Data mining involves using clustering models such as K-Means clustering to discover patterns and relationships in the data. K-Means clustering groups similar data points together by partitioning the dataset into k clusters, assigning data points to the nearest centroid, and recalculating centroids until convergence [14].

In addition, the silhouette method evaluates clustering quality by calculating a silhouette coefficient for each data point, which measures the similarity of an object to its cluster compared to other clusters [15]. One of the main purposes of the silhouette method is to determine the optimal number of clusters in a dataset

by calculating the silhouette coefficient for different values of k number of clusters and selecting the value of k that gives the highest average silhouette coefficient. This can help identify the most appropriate number of clusters to use for a given dataset and improve the accuracy and efficiency of clustering algorithms.However, one of the weaknesses of clustering methods is the inability to validate the results due to the absence of labels on the data as Clustering is an unsupervised method [16]. The naive Bayes method is used to validate the effectiveness of clustering models created with the K-Means algorithm. It assigns each data point to a cluster based on its probability of belonging to that cluster, ensuring that the clusters represent the underlying patterns in the data and providing reliable insights. [17]. At this stage, the clustering result is labeled as "Aware" and "Not Aware" based on the assumption of only two clusters. In case of more than two clusters, the labels will be extended up to a maximum of k. Thus, the variable settings used for validating the results are shown in Table 1 as followings:

a. Self-health maintenance involves the practices and behaviors that students adopt to promote their physical, mental, and emotional well-being. Examples of self-health maintenance include eating a healthy diet, engaging in regular exercise, getting enough sleep, managing stress, and seeking preventive healthcare services. By engaging in these activities, students can reduce their risk of developing chronic diseases, improve their overall health, and enhance their quality of life. Self-health maintenance is an important aspect of overall health and well-being, as it empowers students to take an active role in their health and make choices that promote their well-being [18].
b. Health maintenance in public areas involves practices and measures when maintaining public health and safety in shared spaces, such as schools, workplaces, public transportation, and other community settings. These practices can include regular cleaning and sanitizing, promoting proper hand hygiene, enforcing social distancing measures, and providing access to healthcare resources. The responsibility for health maintenance in public areas often falls on government agencies, public health organizations, and community leaders who work to develop and implement policies and guidelines that promote public health and safety. Health maintenance in public areas is essential for promoting and protecting public health and safety, preventing the spread of infectious diseases, reducing the risk of injury and illness, and promoting a safe and healthy community for everyone [19].
c. Healthy food consumption is the practice of eating nutrient-dense foods that provide the body with essential vitamins, minerals, and macronutrients needed for optimal health and well-being. Examples of healthy foods include fruits, vegetables, whole grains, lean proteins, and healthy fats. Consuming healthy foods is important for maintaining a healthy weight, reducing the risk of chronic diseases such as heart disease and type 2 diabetes, and improving overall health and well-being. It is essential to consume a balanced and varied diet, in appropriate portions, to ensure the body receives essential nutrients that support immune function, bone health, and cognitive function. Overall, healthy food consumption is a critical aspect of promoting a healthy lifestyle and preventing chronic diseases [20].

Table 1. Dependent and Independent Variables

Variable type	Variable Names	Values
Dependent Variable (Target)	Y= Health Awareness	1= Aware 0= Not Aware
Independent Variables (Predictors)	X_1=Self-Health Maintenance	1=Very poor 2=Poor 3=Enough 4=Good 5=Very good
	X_2=Health Maintenance in Public Areas	1= Very poor 2=Poor 3=Enough 4=Good 5=Very good
	X_3=Healthy Food Consumption	1=Frequently 2=Moderately 3=Rarely

The final step is to evaluate the accuracy of the validation results. This is typically done by comparing the classified cluster assignments generated by the naive Bayes method to the actual cluster assignments of the data. The accuracy is calculated from the confusion matrix using 5-fold validation since it is a common validation technique used in machine learning and data analysis to assess the performance of a Naïve Bayes algorithm and it can be particularly useful when working with small datasets where every data point is important [21].Accuracy is a fundamental metric that evaluates the overall performance of a classification model. It reflects the ratio of correctly predicted instances to the total number of instances. True Positive (TP) stands for the number of positive instances that have been correctly classified, True Negative (TN) refers to the number of negative instances that have been correctly classified, False Positive (FP) indicates the number of negative instances that have been mistakenly predicted as positive, and False Negative (FN) is the number of positive instances that have been erroneously predicted as negative. The accuracy formula is derived by dividing the sum of TP and TN by the total number of instances (TP+TN+FP+FN) [22].

III. RESULTS AND DISCUSSION

Fig. 2 illustrates the distribution of students' efforts to maintain their health during the new normal, displayed as a box plot. The effort level is rated on a scale ranging from very poor (1) to very good (5) along the vertical axis, while age is plotted along the horizontal axis. Based on the box plot, it appears that overall, students across different age groups are making a reasonable effort to maintain their health during the new normal. The median effort ranges from 2 to 4, with most students scoring between 3 and 4.5. However, there is some variation in the spread of ratings across different age groups. The 17-year-old and 20-year-old groups have a narrower spread of ratings, while the 18-year-old and 19-year-old groups have a wider spread. The 21-year-old group also has a relatively narrow spread but with a lower median rating than some of the other age groups. The effort scores for age 22 are noticeably lower than the other age groups, with a median effort score of 2, indicating that students in this age group are putting in significantly less effort to maintain their health during the new normal.

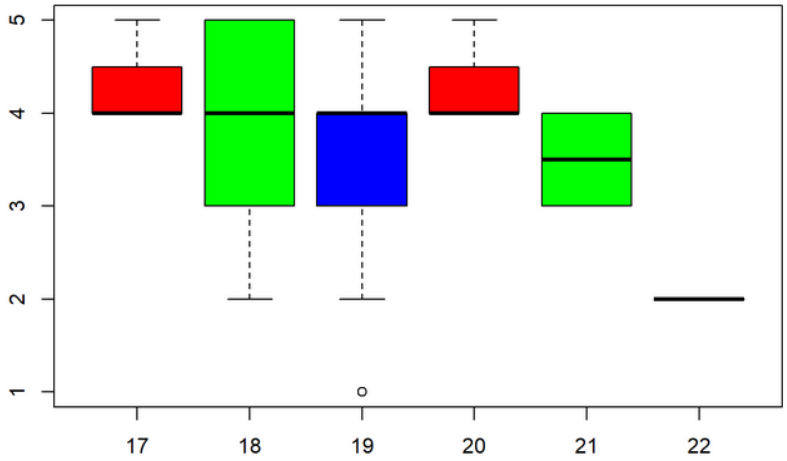

Fig 2. The Boxplot Diagram of Self-Health Maintenance Based on Age

Capturing the essence of students' efforts to maintain their health in public places during the new normal, Fig. 3 showcases a box plot analysis that unravels intriguing insights based on age. The vertical axis represents the effort rating scale, ranging from very poor (1) to very good (5), while the horizontal axis displays age. Students are making a good effort to maintain their health in public areas, with a median score of 4 out of 5. However, there are some variations in self-health maintenance behaviors based on age. Specifically, the age group between 17 and 21 shows the highest median score of 4, indicating that young adults in this range tend to prioritize their health more than other age groups in public areas. On the other hand, the age group of 22 shows the lowest median score of 2, suggesting that individuals in this age range may not be putting in as much effort as others.

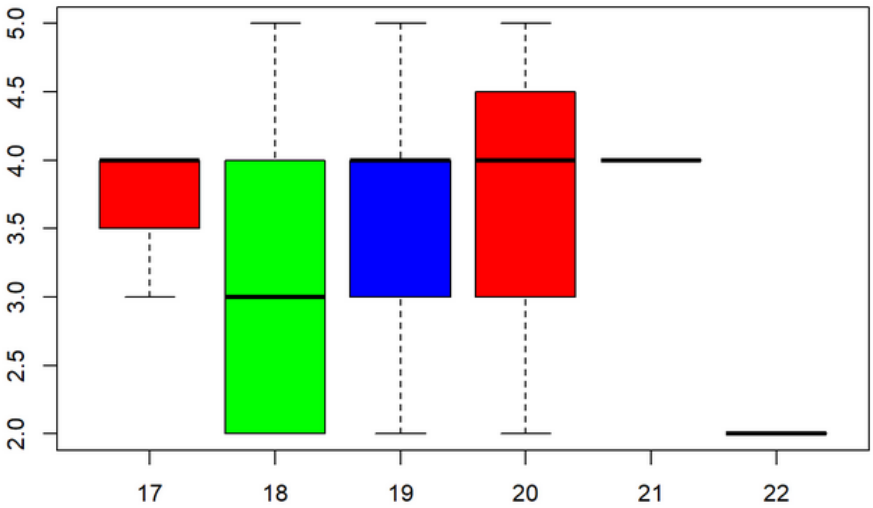

Fig 3. Box Plot of Health Maintenance in Public Areas by Age

Fig. 4 showcases a box plot representing the level of consumption of healthy food during the new normal, based on age. The frequency of ratings, ranging from frequent (1) to rare (3), is displayed along the vertical axis, while the horizontal axis elegantly depicts age. The consumption of healthy food is moderate (level 2) in most individuals across all age groups. The variability in consumption levels increases in ages 18 and 19 compared to age 17, indicating that some individuals consume healthy food more frequently or rarely. The median consumption level at age 20 moves to a level between moderate and frequent (level 1.5), with a wider range of consumption levels. At age 21, the median consumption level moves to the frequent level, indicating that most individuals consume healthy food more often than in the previous age groups. However, at age 22, the median consumption level moves back to moderate (level 2), which is like age 17.

There could be various reasons why some students consume healthy food in moderation. Some may find healthy food to be less tasty or satisfying than other types of food [23]. Unhealthy foods are often highly processed, containing sugar, salt, and unhealthy fats, leading to pleasure and satisfaction. Whilst healthy foods are often less processed and may not contain these elements, making them less appealing to some students. Afterward, some students may simply not have access to a wide range of healthy food options [24]. Limited access to fresh and healthy food options may occur due to factors such as geographic location, income level, and transportation barriers. Moreover, factors such as cost may also impact one's ability to regularly consume healthy food [25]. Healthy foods can be expensive, while unhealthy foods are often more affordable, which can make them a more convenient option for students on a tight budget or with limited access to healthy food options.

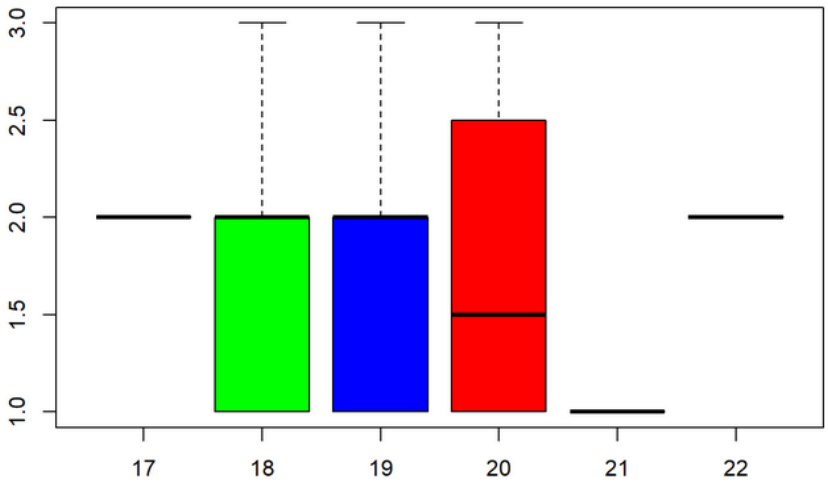

Fig 4. Box Plot of Healthy Food Consumption by Age

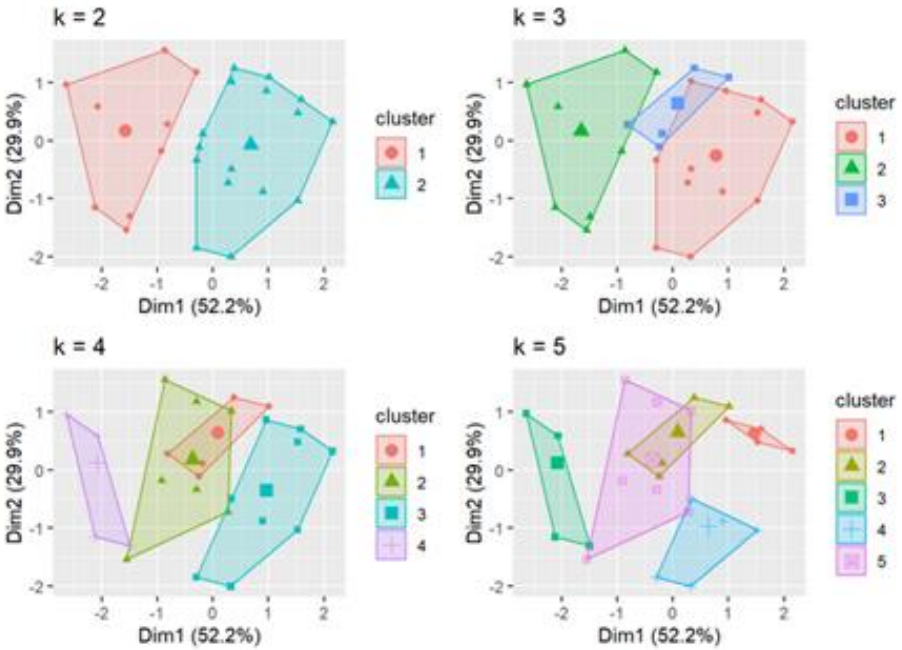

Fig 5. Clustering Comparison from k=2 to k=5

By comparing clustering results as shown in Fig. 5, the k=2 shows the best results. The two data sets have little similarity, thus forming closely adjacent patterns, but not touching each other. While others show overlapping between their clusters. By doubling crosscheck with the silhouette method as shown in Fig. 6. The vertical line that points out k=2 in the silhouette plot indicates that the dataset is best clustered into two distinct groups, based on the similarity between the data points.

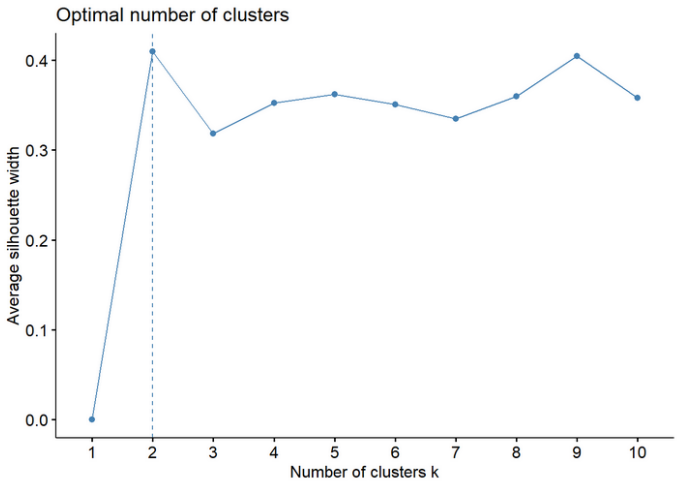

Fig 6. Silhouette method shows k=2

The analysis performed on the Indonesian Students' Awareness of Maintaining Health led to the identification of two distinct clusters. Most of the students appear to be aware of maintaining their health, with 80% of students falling into the aware group. The remaining 20% were classified as not aware based on their responses and behaviors related to health maintenance. The findings from the analysis of the student's awareness of maintaining health reveal both positive and negative aspects. On the positive side, most of the students appear to have a good level of health knowledge and behaviors, indicating a strong foundation for promoting and maintaining health. However, the negative aspect is that the remaining 20% of the students are not aware of health maintenance practices, which places them at risk of experiencing negative health outcomes. Addressing this issue would require targeted interventions and educational programs that can improve the health awareness and behaviors of this group.

Table 2. 5-Fold Validation

Fold	True Positive	True Negative	False Positive	False Negative
1	13	2	1	0
2	13	2	1	0
3	13	3	0	0
4	13	3	0	0
5	12	3	1	0
Total	64	13	3	0

As shown in Table 2, the results of 5-fold validation indicate that TP=64, TN=13. This shows that the total correct classification of Naïve Bayes classification is 77, with a total of 3 misclassifications out of the total 80 data points. Therefore, the accuracy is 96.25%. Clustering groups similar data points, and Naive Bayes can be used to build a classification model on the clustered data to measure its effectiveness. A high accuracy result, such as 96.25%, suggests that the clustering algorithm is performing well in creating meaningful and well-defined clusters. It also indicates that the Naive Bayes model is effective in correctly classifying the new data points based on their cluster membership.

IV. CONCLUSION

In conclusion, the study examined the awareness of Indonesian students in maintaining their health during the new normal using clustering and Naïve Bayes methods. The findings revealed two distinct clusters, "Aware" and "Not Aware," indicating that Indonesian students have varying levels of awareness regarding health precautions during the new normal. Notably, the results showed that most students fell under the "Aware" cluster, suggesting a positive trend in health-conscious behavior among the youth population in Indonesia. The study used the Naive Bayes method to validate the cluster analysis resulting in an overall accuracy of approximately 96.25%. This high accuracy shows that the clustering algorithm has created meaningful and well-defined clusters. In addition, students are making a reasonable effort to maintain their health during the new normal.

Most students scored between moderate and good. However, there is some variation in the spread of effort ratings across different age groups, with the age group of 22 showing a significantly lower median score. Afterward, this study also found that students generally put in a good effort to maintain their health in public places during the new normal. However, the age group of 22 showed the lowest median score. Finally, the consumption of healthy food during the new normal is generally moderate across all age groups.Students aged 22 have been identified as a group that needs more attention when it comes to promoting healthy habits, based on the findings of this study. It is recommended that any health improvement program aimed at university students should target this age group specifically, as this could lead to an improvement in overall self-health maintenance behaviors and the promotion of healthier habits that could have long-term benefits for both individuals and society.

V. ACKNOWLEDGMENTS

We would like to express our gratitude to Universitas Multimedia Nusantara for their exceptional services and support throughout this study, particularly in the context of the new normal. Their contributions have been invaluable to the success of this research.

REFERENCES

[1] N. Donthu, A. Gustafsson, *Effects of COVID-19 on business and research*, ***Journal of business research, 117***, 2020, pp.284-289.

[2] D. S. Vieira de Jesus, D. Kamlot, V. J. Correia Dubeux, *Innovation in the New Normal'Interactions, the Urban Space, and the Low Touch Economy: The Case of Rio de Janeiro in the Context of the COVID-19 Pandemic*, ***Int'l J. Soc. Sci. Stud., 8***, 2020.

[3] G. R. Tortella, A. B. Seabra, J. Padrão, R. Diaz-San Juan, *Mindfulness and other simple neuroscience-based proposals to promote the learning performance and mental health of students during the COVID-19 pandemic*, ***Brain sciences, 11***, 2020.

[4] S. Sinurat, I. S. Saragih, M. F. Larosa, *Correlation of Public Self-Awareness with Behaviour in Suppressing the Spread of COVID-19 at Parombunan Sub District Zone VI Sibolga City in 2021*, ***Jurnal Kesehatan LLDikti Wilayah 1 (JUKES), 1***, 2021, pp.51-59.

[5] S. Panday, S. Rushton, J. Karki, J. Balen, A. Barnes, *The role of social capital in disaster resilience in remote communities after the 2015 Nepal earthquake*, ***International Journal of Disaster Risk Reduction, 55***, 2021.

[6] N. P. A. Karniawati, G. S. Darma, L. P. Mahyuni, & I. G. Sanica, *Community Perception of Using QR Code Payment in Era New Normal*, ***PalArch's Journal of Archaeology of Egypt/Egyptology, 18***, 2021, pp.3986-3999.

[7] D. S. A. Maylawati, M. A. Ramdhani, *Indonesian Citizens' Health Behavior in a Pandemics: Twitter Conversation Analysis using Latent Dirichlet Allocation*, ***IEEE In 2022 8th International Conference on Wireless and Telematics (ICWT),*** 2022, pp.1-6.

[8] R. Pradipta, R. Jayadi, *The Sentiment Analysis Of The Indonesian Palm Oil Industry In Social Media Using A Machine Learning Model*, ***Journal of Theoretical and Applied Information Technology, 100***, 2022.

[9] K. Ponmani, M. Thangaraj, *Clustering-based sentiment analysis on Twitter data for COVID-19 vaccines in India*, ***Int. J. Health Sci***, 2022, pp.4732-4748.

[10] E. Istanti, B. K. Negoro, A. D. GS, *The Effect of Job Stress and Financial Compensation toward OCB and Employee Performance:(Case Study in PT. MENTARI SEJATI PERKASA Private Companies in Surabaya).* ***Media Mahardhika, 19***, 2021, pp. 560-570.

[11] T. A. Omang, P. U. Angioha, *Assessing the impact of the COVID-19 pandemic on the educational development of secondary school students.* ***JINAV: Journal of Information and Visualization, 2***, 2021, pp.25-32.

[12] A. D. Pramesti, M. Jajuli, B. N. Sari, *Implementasi Metode Double Exponential Smoothing dalam Memprediksi Pertambahan Jumlah Penduduk di Wilayah Kabupaten Karawang*, ***Ultimatics: Jurnal Teknik Informatika, 12***, 2020, pp.95-103.

[13] M. S. Suryono, R. Oetama, *Peramalan terhadap Forex dengan Metode ARIMA Studi Kasus GBP/USD*, ***Ultimatics: Jurnal Teknik Informatika, 11***, 2019, pp.6-10.

[14] P. Govender, V. Sivakumar, *Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)*, ***Atmospheric pollution research, 11***, 2020, pp.40-56.

[15] K. Tian, J. Li, J. Zeng, A. Evans, L. Zhang, *Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm*, ***Computers, and Electronics in Agriculture, 165***, 2019.

[16] C. Hansen, J. G. Simonsen, S. Alstrup, C. Lioma, *Unsupervised neural generative semantic hashing*, ***In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval,*** 2019, pp. 735-744.

[17] A. R. Braga, D. G. Gomes, B. M. Freitas, J. A. Cazier, *A cluster-classification method for accurate mining of seasonal honey bee patterns*, ***Ecological Informatics, 59***, 2020.

[18] H. Hah, J. Khuntia, A. Kathuria, J. Tan, *Rationalizing personal health management (PHM) policy: identifying health IT use patterns via observations of daily living (ODLs) data*, ***Health Policy and Technology, 9***, 2020, 185-193.

[19] J. Ali, S. Singh, W. Khan, *Health awareness of rural households towards COVID-19 pandemic in India: Evidence from Rural Impact Survey of the World Bank*, ***Journal of Public Affairs, 23***, 2023.

[20] T. Ali, J. Ali, *Factors affecting the consumers' willingness to pay for health and wellness food products*, ***Journal of Agriculture and Food Research, 2***, 2020.

[21] M. Ra, B. Ab, S. Kc, *COVID-19 outbreak: tweet based analysis and visualization towards the influence of coronavirus in the world*, ***Gedrag. Organ. Rev, 33***, 2020.

[22] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, J. M. Moguerza, *General Performance Score for classification problems.* ***Applied Intelligence, 52***, 2022, pp.12049-12063.

[23] L. Hagen, *Pretty healthy food: How and when aesthetics enhances perceived healthiness*, ***Journal of Marketing, 85***, 2021, 129-145.

[24] C. Li, M. Mirosa, P. Bremer, *Review of online food delivery platforms and their impacts on sustainability.* ***Sustainability, 12***, 2020.

[25] D. Fróna, J. Szenderák, M. Harangi-Rákos, *the challenge of feeding the world*, ***Sustainability, 11***, 2019.