Emerging Information Science and Technology
Vol. 3, No. 2, (2022), pp. 78-84

Classification of Student Understanding on Covid-19 Booster Vaccine Using Machine Learning

Cahya Damarjati*, Slamet Riyadi, Ricki Irawan

Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia
**Corresponding author: cahya.damarjati@umy.ac.id*

Abstract

The outbreak of COVID-19 has been declared a global pandemic by the World Health Organization (WHO). Developing a vaccine is one of the best ways to reduce the virus's impact. Nevertheless, the development of virus mutations produces new variants that diminish the efficacy of the previous vaccine. Booster doses of the Covid-19 vaccine is still a matter of debate among the public, particularly among students, as evidenced by the low rate of booster vaccinations in the community, which is a result of a lack of knowledge about booster vaccines. The purpose of this study is to assess the level of understanding among Universitas Muhammadiyah Yogyakarta (UMY) students regarding booster vaccinations, with the results subsequently serving as a factor or strategy for future government booster vaccination policy decisions. ANN and SVM algorithms could be used to predict the level of understanding of booster vaccinations among UMY students. However, the maximum level of precision in classifying the level of comprehension is not yet known. To determine which of the two methods, kernel and k-fold, provided the maximum level of accuracy, a comparative study was conducted between them. The research was conducted by disseminating questionnaires containing assessments of booster vaccinations to a total of 2095 respondents. Using randomized sampling type, this study yielded an accuracy of 88.45% for the ANN method and 89.93% for the SVM method in each scenario. In addition, the authors conduct feature efficiency, which aims to reduce the time and cost associated with data computation.

Keywords: *Artificial Neural Network, Feature, Prediction, Students, Support Vector Machine,Understanding, Vaccines*

1. Introduction

Vaccines are a crucial tool in the fight against infectious diseases, as they stimulate the immune system to recognize and combat specific pathogens. By containing antigens from microorganisms, vaccines train the body's defenses to mount a robust immune response when faced with actual infection. Over the years, vaccines have played a pivotal role in preventing the spread of various diseases and have significantly reduced the burden of illness worldwide [1], [2].

In the case of the Covid-19 pandemic, vaccines have emerged as a beacon of hope. Vaccines are biological products containing antigens, i.e., the microorganisms they produce, which are then processed to make them safe for human administration; when administered, vaccines increase specific immunity against certain diseases [3], [4]. The administration of Covid-19 vaccines has proven to be highly effective in reducing the risk of infections. Recent studies have shown that the overall effectiveness of booster vaccines against the Delta variant stands at an impressive 93%, while it remains at a commendable 75% for the Omicron variant [5]. Despite these encouraging statistics, there is still a notable lack of enthusiasm among the population, including students, for receiving booster vaccinations [6], [7].

The low interest of students in booster vaccinations can be attributed to various factors, with a lack of understanding about the need for and benefits of booster shots being a key issue. Many individuals may not be fully aware of the importance of boosters in reinforcing immunity and providing prolonged protection against evolving variants. To address this knowledge gap and better predict the level of understanding among students at the University of Muhammadiyah Yogyakarta (UMY), this study aims to leverage two powerful algorithms: the Artificial Neural Network (ANN) Algorithm and the Support Vector Machine (SVM) Algorithm. By employing these advanced computational techniques, the study aims to accurately assess and predict the level of comprehension regarding booster vaccinations among UMY students. Such insights can pave the way for targeted educational interventions and communication strategies to promote greater awareness and acceptance of booster vaccinations within this specific population.

2. Method

2.1 Data colletion

According to [8], [9], Data mining is the process of extracting valuable information from multiple databases using machine learning, artificial intelligence (AI), mathematics, and statistical techniques. To implement data mining, data collection step must be passed. The data collection process is very important because the implementation process and conclusions cannot be carried out if this step has not been realized. The data taken is data from UMY students through a questionnaire in the form of a google form. The questionnaire contains data about what factors affect the level of student understanding of the covid booster vaccine. The questions are listed below:

1. I have done the first and second dose of vaccination (1 not yet vaccine, first 2 doses, second 3 doses)
2. The side effects from the first and second doses of vaccination made me not hesitate to take the booster vaccination.
3. I understand that booster vaccinations can maintain and enhance the body's immune system so that it is not easily infected with the Corona Virus.
4. I understand that booster vaccination is a government step to achieve herd immunity.
5. I understand that booster vaccination can strengthen antibodies that have been built up before after the first and second doses.
6. I understand that booster vaccines were created to fight new corona variants such as Omicron.
7. I understand what a homologous or heterologous booster vaccine is.
8. I understand in choosing the type of booster vaccine between homologous or heterologous.
9. I had no difficulty in getting information about booster vaccinations in my area, this made me plan to take booster vaccinations.
10. I plan to take booster vaccinations after knowing the benefits.
11. My level of understanding of booster vaccinations

2.2 Pre-Processing Data

Data pre-processing is needed in selecting data that needs to be used as test material. As an example of data that is not used later in this study is the name of the respondent. The data collected will be default which needs to be sorted again according to the needs in this study.

2.3 Data Cleaning

This process focuses on data that has the same entity, so data that has the same entity will be deleted.

2.4 Data Filtering

In the data filtering process carried out by the researcher, the researcher selects data and attributes that are relevant to this study, therefore the unused attributes will be deleted by the researcher.

2.5 Data Transformation

The data transformation carried out in this study is transforming the initial question into the core of the question, making it shorter and easier to read when implementing it into Machine Learning and transforming each respondent's answer on a scale in the form of numbers 1 to 3.

2.6 Impementasi Algoritma

This is the core of data mining process. Machine learning algorithm are implemented to do the prediction. Prediction is the process of systematically estimating what is most likely to occur in the future based on historical and current data to minimize error. The data that has been obtained because of data processing will go through a classification process using the Artificial Neural Network (ANN) algorithm and Support Vector Machine (SVM) and RapidMiner software.

According to [10]–[12], ANN is non-linear statistical data models that mimic the function of natural neural networks. It has been applied to a variety of problems, including prediction of protein secondary structure, speech recognition, gene prediction, and cancer classification. SVM is one of the algorithms of a classification method that can generate a learning process in a classification problem, which is interpreted as a search for a line (hyperlane) to separate the two groups[13]. To see the performance between SVM and ANN, confusion matrix is used. According to [14], [15], the confusion matrix is a cross table that documents the number of occurrences between two raters, actual or true classification, and predicted classification.

3. Results and Discussion

Figure 1 illustrates the results of the distribution of data for each query, as determined by the research findings. There are 813 data points for Neutral labels, 538 for Understand labels, and 744 for Don't Understand labels. Without removing the accuracy feature, the ANN method yields an accuracy of 88.45%. It is 86.47% in the Neutral precision class, 86.51% in the Does Not Understand precision class, and 93.98% in the Understands precision class. As shown in Figure 2, the recall for the Neutral class was 84.13%, the recall for the Do Not Understand class was 87.90%, and the recall for the Understanding class was 95.72%.

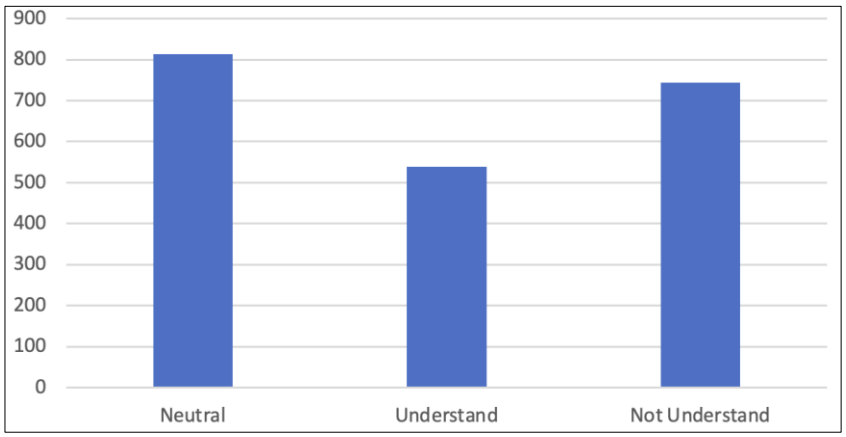

Figure 1. Data distribution of student's understanding on the covid booster vaccine

accuracy: 88.45% +/- 2.70% (micro average: 88.45%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	684	86	21	86.47%
pred. Not Understand	100	654	2	86.51%
pred. Understand	29	4	515	93.98%
class recall	84.13%	87.90%	95.72%	

Figure 2. Confusion matrix in ANN scenario

Without eliminating the accuracy feature, the SMV method achieves an accuracy of 89.93%. In the Neutral precision class of 86.51%, the Do not Understand precision class is 89.85% and the understand precision class is 95.34%. Figure 3 displays that the Neutral class recall was 88.31%, the Do Not Understand class recall was 88.04 %, and the Understanding class recall was 94.98%.

accuracy: 89.93% +/- 1.02% (micro average: 89.93%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	718	85	27	86.51%
pred. Not Understand	74	655	0	89.85%
pred. Understand	21	4	511	95.34%
class recall	88.31%	88.04%	94.98%	

Figure 3. Confusion matrix in SVM scenario

On the effectiveness of removing the seventh feature using the ANN method with two concealed layers and k-fold = eight, the obtained accuracy is 85.44 percent. It is 81.13 percent in the Neutral precision class, 85.77 percent in the Does Not Understand precision class, and 91.65 percent in the Understands precision class. As shown in Figure 4, the recall for the Neutral class was 83.03%, the recall for the Do Not Understand class was 83.47%, and the recall for the Understanding class was 91.82%.

accuracy: 85.44% +/- 0.98% (micro average: 85.44%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	675	117	40	81.13%
pred. Not Understand	99	621	4	85.77%
pred. Understand	39	6	494	91.65%
class recall	83.03%	83.47%	91.82%	

Figure 4. Confusion matrix in ANN with removal of feature number 7 scenario

The effectiveness of removing the seventh feature using the SVM kernel polynomial method with k-fold = 10 yields an accuracy of 86.45%. It is 82.22% in the Neutral precision class, 86.17% in the Does Not Understand precision class, and 93.33% in the Understands precision class. As shown in Figure 5, the recall for the Neutral class was 84.13%, the recall for the Do Not Understand class was 83.74%, and the recall for the Understanding class was 93.68%.

accuracy: 86.45% +/- 2.11% (micro average: 86.44%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	684	116	32	82.21%
pred. Not Understand	98	623	2	86.17%
pred. Understand	31	5	504	93.33%
class recall	84.13%	83.74%	93.68%	

Figure 5. Confusion matrix in SVM with removal of feature number 7 scenario

Based on the effectiveness of removing feature 9 using the ANN method with two concealed layers and k-fold = 8, 85.30% accuracy is obtained. It is 81.58% in the Neutral precision class, 84.54% in the Does Not Understand precision class, and 91.88% in the Understands precision class. As shown in Figure 6, the Neutral class recall was 81.18 percent, the Do Not Understand class recall was 84.5 percent, and the Understanding class recall was 92.5 percent.

accuracy: 85.30% +/- 1.91% (micro average: 85.30%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	660	110	39	81.58%
pred. Not Understand	114	629	1	84.54%
pred. Understand	39	5	498	91.88%
class recall	81.18%	84.54%	92.57%	

Figure 6. Confusion matrix in ANN with removal of feature number 9 scenario

Based on the effectiveness of feature removal feature no. 9 using the linear kernel SVM method with k-fold = 10, the obtained accuracy is 87.16 percent. In the Neutral precision class of 82.46%, the Do not comprehend precision class is 87.93% and the comprehend precision class is 93.58%. As shown in Figure 7, the Neutral class recall was 85.61 percent, the Do Not Understand class recall was 85.22 percent, and the Understanding class recall was 92.19%.

accuracy: 87.16% +/- 2.60% (micro average: 87.16%)

	true Neutral	true Not Understand	true Understand	class precision
pred. Neutral	696	107	41	82.46%
pred. Not Understand	86	634	1	87.93%
pred. Understand	31	3	496	93.58%
class recall	85.61%	85.22%	92.19%	

Figure 7. Confusion matrix in SVM with removal of feature number 9 scenario

4. Conclusion

In the feature efficiency scenario, it can be concluded that the accuracy of the SVM method for removing feature number 7 decreased by 3.24 percentage points, while the SVM method for removing feature number 9 decreased by 2.53 percentage points, and that the accuracy of the ANN method for removing feature number 7 decreased by 3.01 percentage points, while the ANN method for removing feature number 7 9 decreased by 3.15 percentage points. In the meantime, the pace of feature removal for the SVM method is between 0 and 1 second, while for the ANN method it is between 6-7 seconds. In this study, the authors prefer to maintain accuracy over feature efficiency by eliminating one of the features, as the accuracy decreases significantly with both methods, whereas the speed of feature removal is only a fraction of a second.

References

[1] S. B. Omer, D. A. Salmon, W. A. Orenstein, M. P. Dehart, and N. Halsey, "Vaccine refusal, mandatory immunization, and the risks of vaccine-preventable diseases," *New England Journal of Medicine*, vol. 360, no. 19, pp. 1981–1988, 2009.

[2] J. M. Caldwell *et al.*, "Vaccines and variants: Modelling insights into emerging issues in COVID-19 epidemiology," *Paediatric Respiratory Reviews*, vol. 39, pp. 32–39, 2021.

[3] A. J. Pollard, A. Finn, and N. Curtis, "Non-specific effects of vaccines: plausible and potentially important, but implications uncertain," *Archives of disease in childhood*, vol. 102, no. 11, pp. 1077–1081, 2017.

[4] D. M. Klinman, S. Klaschik, T. Sato, and D. Tross, "CpG oligonucleotides as adjuvants for vaccines targeting infectious diseases," *Advanced drug delivery reviews*, vol. 61, no. 3, pp. 248–255, 2009.

[5] K. Xu *et al.*, "Protective prototype-Beta and Delta-Omicron chimeric RBD-dimer vaccines against SARS-CoV-2," *Cell*, vol. 185, no. 13, pp. 2265–2278, 2022.

[6] M. Ghojoghdordi and A. Charkazi, "Barriers toward Getting Booster Dose of COVID-19 Vaccination among Turkmen people: A Content Analysis Study," 2022.

[7] F. Ciardi *et al.*, "Knowledge, attitudes and perceptions of COVID-19 vaccination among healthcare workers of an inner-city hospital in New York," *Vaccines*, vol. 9, no. 5, p. 516, 2021.

[8] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[9] L. Marlina, M. Muslim, A. U. Siahaan, and P. Utama, "Data Mining Classification Comparison (Naïve Bayes and C4. 5 Algorithms)," *Int. J. Eng. Trends Technol*, vol. 38, no. 7, pp. 380–383, 2016.

[10] O. I. Abiodun *et al.*, "Comprehensive review of artificial neural network applications to pattern recognition," *IEEE Access*, vol. 7, pp. 158820–158846, 2019.

[11] A. Krogh, "What are artificial neural networks?," *Nature biotechnology*, vol. 26, no. 2, pp. 195–197, 2008.

[12] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[13] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[14] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022.

[15] E. Beauxis-Aussalet and L. Hardman, "Visualization of confusion matrix for non-expert users," in *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*, 2014, pp. 1–2.