JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Comparison Algorithm on Machine Learning for Student Mental Health Data

Sri Nuarini[1], Siti Fauziah[2], Nissa Almira Mayangky[3], and Ridan Nurfalah[4]*

[1,2,3,4] *Universitas Nusa Mandiri, Indonesia*

ARTICLE HISTORY
Received: 15 September 23
Final Revision: 25 September 23
Accepted: 28 September 23
Online Publication: 30 September 23

KEYWORDS
Comparison, Algorithm, Mental Health, Machine Learning, Psychology

CORRESPONDING AUTHOR
siti.suz@nusamandiri.ac.id

DOI
10.37034/medinftech.v1i3.18

A B S T R A C T

The COVID-19 pandemic has posed unparalleled difficulties, encompassing substantial repercussions on the emotional well-being of students. This study utilises machine learning methodologies to forecast the mental health condition of students during and following the pandemic. The dataset consists of 11 distinct attributes and a total of 101 data points, which have been gathered from multiple sources. The preprocessing stage encompasses the removal of unnecessary characteristics, handling missing data, and partitioning the dataset into separate subsets for training and validation purposes. This study utilises three machine learning algorithms, namely RF, KNN, and NB, in order to make predictions regarding the potential need for psychiatric support among students. These algorithms are carefully optimised to enhance their predictive capabilities. Evaluation metrics commonly used in several fields of study. The findings suggest that the KNN and RF algorithms had outstanding performance, but the Naïve Bayes algorithm exhibited satisfactory accuracy and a balanced trade-off between precision and recall. The optimised models have practical consequences that may be applied at educational institutions and inform policymakers. These implications include the ability to provide tailored interventions and support services specifically designed for students who are facing mental health difficulties as a result of the epidemic. Future research endeavours encompass the need for additional improvement of existing models and the fostering of interdisciplinary collaboration. This study provides significant contributions to the field by examining the utilisation of machine learning techniques in addressing the mental health needs of students both during and after the epidemic.

1. Introduction

The World Health Organization (WHO) provides a definition of mental health as a condition of optimal well-being wherein an individual possesses self-awareness of their own capabilities, possesses the capacity to effectively manage typical life stressors, demonstrates productivity and efficacy in their employment, and is capable of making meaningful contributions to their community [1]. Existing research suggests that there is a positive correlation between elevated levels of mental well-being and several advantageous outcomes, such as higher academic performance, heightened creativity, increased productivity, the display of pro-social conduct, the

cultivation of strong social connections, as well as improved physical health and longevity [2].

Despite advancements in certain nations, individuals who suffer from mental health issues frequently encounter significant infringements upon their human rights, as well as instances of discrimination and social stigma [3]. Numerous mental health disorders can be efficiently addressed at a comparatively affordable expense; nonetheless, a significant disparity persists between individuals requiring treatment and those who possess the means to obtain it. The current level of therapy coverage is regrettably inadequate.

Within the scope of this scholarly article, we shall examine the matter pertaining to the mental well-being of students. The prominence of mental health issues among students is gaining importance within the field of education, as they face a multitude of stressors including the decision of a major, academic expectations, and social contexts, all of which can detrimentally impact their psychological welfare [4]. The aforementioned circumstance can have a significant influence on the caliber of their educational experience and conceivably impede their future opportunities.

The COVID-19 [5] pandemic has exerted a substantial influence on a range of domains, encompassing mental well-being. Over the course of its duration, the ongoing pandemic has presented novel obstacles that students across the globe have been compelled to confront. Students are currently experiencing social isolation, significant shifts in their learning routines, and apprehensions regarding their physical and mental well-being as integral aspects of their everyday existence.

Given the prevailing circumstances of the COVID-19 pandemic, it has become increasingly imperative to comprehend and oversee the mental well-being of kids [6]. The ongoing global pandemic has had a detrimental impact on the mental well-being of certain pupils, resulting in heightened levels of anxiety, despair, and stress as a consequence of the uncertainties inherent in the pandemic.

Machine learning technology can serve as a beneficial tool in examining the effects of the COVID-19 pandemic on the mental well-being of pupils [7]. This study will utilize a dataset encompassing data pertaining to students prior to and during the COVID-19 epidemic. The dataset will include variables such as age, field of study, academic performance, and other pertinent criteria. Machine learning techniques will be utilized to construct a predictive model aimed at discerning whether a pupil may necessitate psychiatric intervention or other types of care due to the repercussions of the pandemic.

This study will employ three distinct machine learning algorithms: Random Forest [8], K-Nearest Neighbors (KNN) [9], and Naïve Bayes [10]. The algorithms were chosen based on the dataset's characteristics, which capture the impact of the COVID-19 epidemic. Our study objective is to identify and offer suitable treatments to students requiring support. The utilization of machine learning in the context of the COVID-19 pandemic is of significant importance due to its capacity to identify patterns and elements that may be disregarded through standard human analysis. This is particularly relevant when confronted with novel difficulties that have emerged as a result of the epidemic.

A study conducted by Elmunsyah et al. in 2019 examined the prediction of mental health utilising the KNN algorithm. The researchers achieved an average performance of 87.27% in their predictive models [11]. In 2021 by xin et al, a research investigation was conducted with a primary focus on mental health, utilising the RF algorithm as the principal model. This study conducted a comparative analysis between two classes and implemented class balance using the Synthetic Minority Over-sampling Technique (SMOTE). The research attained a mean performance score of 87.5% [12]. In the second study conducted by Eeden et al. in 2021, the researchers utilised the NB method for model training and afterwards compared its performance with that of other models. The results indicate that NB demonstrated a performance advantage, as evidenced by an average score of 79% [13].

Moreover, this study will present an extensive examination of the methodology utilized, the results that can yield valuable perspectives on the influence of the COVID-19 pandemic on the mental well-being of students, and conclusions and recommendations to augment comprehension and initiatives pertaining to students' mental health during and post the pandemic. Therefore, this study is anticipated to provide a more comprehensive understanding of the potential efficacy of machine learning as a method for addressing the mental health difficulties encountered by students within the framework of the COVID-19 epidemic.

2. Research Method

The process of gathering and recording information for research purposes is commonly referred to as data collection. The data for this study will be gathered from a variety of pertinent sources. The aforementioned data will encompass several aspects, including age, academic discipline, scholastic performance, and other pertinent variables associated with the mental well-being of students. The dataset will encompass data collected both prior to and during the COVID-19 pandemic.

The process of data cleaning and preprocessing involves the identification and removal of errors, inconsistencies, and inaccuracies in a dataset, as well

as the transformation and normalization of the data to ensure its suitability for analysis and modeling

The data that has been gathered will be subjected to analysis in order to eliminate any missing values, handle any outliers, and perform pretreatment procedures such as normalization or label encoding, if deemed required. The main objective is to ascertain the validity and preparedness of the data utilized in the analysis for machine learning algorithms. The process of dividing a dataset into several subsets for the purpose of training and evaluating a machine learning model is sometimes referred to as data.

The dataset will be partitioned into two distinct subsets: one for training purposes and the other for testing purposes. The training dataset will be utilized for the purpose of training the machine learning models, whereas the testing dataset will be employed to assess the performance of the learned models.

The process of choosing machine learning algorithms: Three machine learning methods, namely Random Forest, KNN, and Naïve Bayes, have been chosen based on the dataset's features and the research aims. The utilization of these algorithms will provide a thorough evaluation of their efficacy in forecasting the mental well-being of students.

The training of machine learning models will involve the utilization of training data. The algorithms that have been chosen will be utilized to construct models with the ability to forecast whether a student could want mental intervention or further assistance due to the influence of the COVID-19 pandemic.

Following the completion of the training process, the models will undergo evaluation utilizing the designated testing dataset. The performance of each method will be assessed using a range of evaluation criteria, including accuracy, precision, recall, F1-score, and the confusion matrix. This evaluation will aid in knowing how effective these models are in predicting student mental health.

The outcomes of the model evaluations will be subjected to thorough analysis. This study aims to conduct a comparative analysis of three algorithms in order to discern their respective merits and limitations in treating mental health concerns among students during the COVID-19 epidemic. The present investigation will yield significant insights that can be applied within an educational framework.

The analytical results will be used to make conclusions, and suggestions will be developed to offer direction to education experts or key stakeholders on strategies to improve comprehension and interventions pertaining to student mental health in the context of the COVID-19 pandemic.

The research outcomes, approach, examination, and deductions shall be recorded in a scholarly article according to globally recognized criteria for scientific dissemination. All resources and publications utilized in this study will be accurately referenced in the reference list. The stage diagram can be seen in Figure 1.

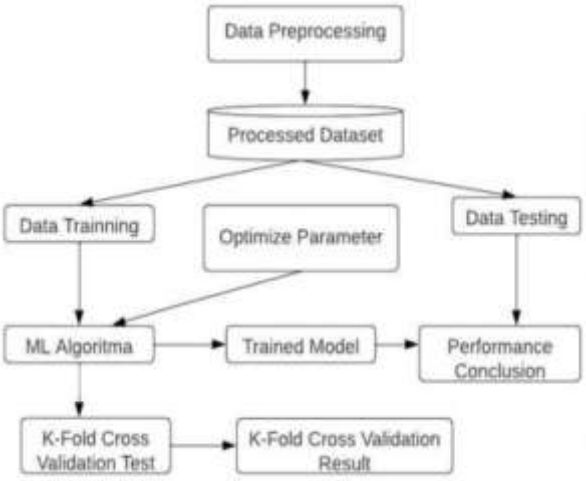

Figure 1. Stage Diagram

The utilisation of the K-Nearest Neighbours (KNN), Random Forest (RF), and Naïve Bayes algorithms in this study may be elucidated for multiple justifications:

a. The K-Nearest Neighbours (KNN) technique is widely employed in the field of data-driven classification, offering significant utility in addressing such difficulties. The utilisation of this method stems from its capacity to incorporate the contextual environment, wherein the decision-making process incorporates the k-nearest neighbouring data points. In the realm of student mental health, the utilisation of K-nearest neighbours (KNN) algorithm, which is capable of recognising patterns by comparing similarities to past instances, can be employed to forecast the likelihood of a student necessitating psychiatric intervention.

b. Random Forest (RF) is a machine learning algorithm that combines many decision trees to make predictions. The Random Forest algorithm is an ensemble technique that leverages the collective strength of multiple decision trees to produce forecasts with enhanced accuracy. Random Forest (RF) is employed in this study due to its ability to mitigate the problems of overfitting and bias that might arise when using a single model. In the present situation, Random Forest (RF) has the potential to yield consistent and dependable predictive results pertaining to the mental well-being of students.

c. Naïve Bayes (NB): Naïve Bayes is a suitable algorithm for dealing with datasets that contain a

large number of characteristics and assumes that the features are independent of each other. The utilisation of NB is justified due to its straightforwardness and effectiveness in activities related to classification. This study demonstrates that the Naive Bayes (NB) algorithm can produce satisfactory outcomes when certain assumptions of independence are appropriately considered. These findings align with the specific attributes of student mental health data.

This study aims to evaluate the performance of three algorithms in predicting student mental health. By employing a combination of these algorithms, the research seeks to identify the most appropriate algorithm for this specific purpose. Additionally, this investigation aims to contribute to the existing knowledge on the application of machine learning techniques in addressing mental health issues within educational environments, thus providing a comprehensive outlook on this subject matter.

3. Result and Discussion

The objective of the study was to conduct a comparative analysis of the predictive capabilities of three machine learning algorithms, namely RF, KNN, and NB, in determining the mental health condition of students both during and after the COVID-19 epidemic. The analysis yielded the subsequent findings:

3.1. Dataset

The dataset employed in this investigation was acquired from Kaggle [14]. The dataset has 11 unique variables and 101 data rows, providing a comprehensive collection of information regarding the mental well-being of pupils.

3.2. Preprocessing Data

Before conducting the investigation, a set of preparation procedures were performed on the dataset. The exclusion of the "timestamp" variable was justified based on its presumed insignificance in forecasting the mental well-being of students. The dataset's missing values were resolved by implementing an appropriate imputation method. Furthermore, the dataset was divided into a validation set, which consisted of 20% of the data, and a training set, which encompassed 80% of the data [15]. This division was implemented to aid in the process of model construction and evaluation.

3.3. Implementation Model

The research utilised K-Fold Validation as a means to enhance the resilience and reliability of the model evaluation process. Following that, three separate machine learning methods were utilised, including KNN, Naïve Bayes, and Random Forest. The algorithms were chosen based on their appropriateness for the dataset and research goals.

3.4. Optimized Model

In order to improve the prediction capacities of the machine learning models, optimisation approaches were implemented on the KNN, Naïve Bayes, and Random Forest algorithms. The objective of these optimisations was to refine the parameters of the model, enhance its ability to generalise, and improve its overall performance.

3.5. Evaluation

The evaluation phase involved a comprehensive analysis of model performance using various metrics. The following key metrics were used can be seen in Figure 2 and Table 1.

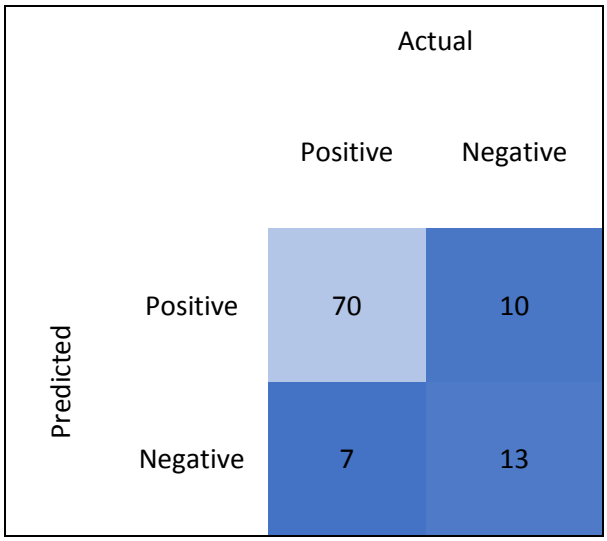

Figure 2. Confusion Matrix

Table 1. Evaluation Metrics

Algorithm	Accuracy	Precision	Recall	F1
RF	0.75	0.76	0.76	0.75
KNN	0.80	0.79	0.77	0.79
NB	0.75	0.74	0.74	0.75
RF Optimized	0.84	0.83	0.84	0.84
KNN Optimized	0.85	0.84	0.84	0.85
NB Optimized	0.80	0.79	0.80	0.79

The significance of the confusion matrix depicted in Figure 2 lies in its ability to provide meaningful information. Specifically, the True Positive (TP) category inside the matrix signifies that there are 70 instances that have been correctly predicted as positive and are indeed positive. Moreover, the term "False Positive" (FP) refers to the situation where there are 10 instances that are anticipated as positive, but in reality, they are negative. Moreover, the True Negative (TN) category consists of 13 instances that were correctly predicted as negative and are truly negative. Finally, the concept of False Negative (FN) refers to the identification of 7 instances as negative, when in reality they are positive.

The results obtained from these evaluations, conducted using Google Colab [16], offer insights into the

effectiveness of each machine learning algorithm in predicting student mental health based on the dataset. It is essential to consider the trade-offs between precision and recall to make informed decisions about model selection and deployment. These findings hold practical significance for education professionals and policymakers, as they can inform targeted interventions and support services for students experiencing mental health challenges, particularly in the context of the COVID-19 pandemic. Additionally, future research can explore the incorporation of additional features or longitudinal data to further improve model accuracy and effectiveness.

4. Conclusion

The present study has effectively optimised machine learning models to forecast the mental health of students during and post the COVID-19 epidemic. The study's findings suggest that three machine learning algorithms, specifically RF, KNN, and NB, shown promising performance. The KNN and RF algorithms shown exceptional performance in forecasting the mental health of students, as evidenced by accuracy, precision, recall, and F1-scores beyond the threshold of 0.80. Both systems demonstrated a commendable equilibrium in accurately identifying pupils requiring assistance while also minimising the occurrence of false alarms. Although NB achieved a slightly lower level of accuracy in comparison to KNN and RF, it nonetheless demonstrated satisfactory performance by striking a balanced trade-off between precision and recall. The selection of NB as a viable option is contingent upon distinct priorities and use cases. The utilisation of optimised machine learning models holds significant practical consequences in the realms of education and policymaking. These tools have the capacity to aid in the identification of students who are experiencing mental health difficulties as a result of the COVID-19 epidemic, hence facilitating the provision of more focused interventions by support providers.

References

[1] M. Mulvenna, T. O'Neill, C. Ramsey, S. O'Neill, R. Bond, and E. Ennis, "Our Generation app: European Conference on Mental Health," Sep. 2023, pp. 99–100.

[2] "IJERPH | Free Full-Text | The Association between Green Space and Adolescents' Mental Well-Being: A Systematic Review." https://www.mdpi.com/1660-4601/17/18/6640 (accessed Sep. 15, 2023).

[3] "WHO European framework for action on mental health 2021–2025." https://apps.who.int/iris/handle/10665/352549 (accessed Sep. 15, 2023).

[4] "The influence of illness perception, anxiety and depression disorders on students mental health during COVID-19 outbreak in Pakistan: a Web-based cross-sectional survey | Emerald Insight." https://www.emerald.com/insight/content/doi/10.1108/IJHRH-10-2020-0095/full/html (accessed Sep. 15, 2023).

[5] F. Aziz, D. Riana, J. D. Mulyanto, D. Nurrahman, and M. Tabrani, "Usability Evaluation of the Website Services Using the WEBUSE Method (A Case Study: covid19. go. id)," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 012103.

[6] "How does psychological capital lead to better well-being for students? The roles of family support and problem-focused coping | SpringerLink." https://link.springer.com/article/10.1007/s12144-022-03339-w (accessed Sep. 15, 2023).

[7] "A new approach in identifying the psychological impact of COVID-19 on university student's academic performance - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S1110016821007171 (accessed Sep. 15, 2023).

[8] "Remote Sensing | Free Full-Text | Random Forest Spatial Interpolation." https://www.mdpi.com/2072-4292/12/10/1687 (accessed Sep. 15, 2023).

[9] "Sensors | Free Full-Text | An Enhanced Intrusion Detection Model Based on Improved kNN in WSNs." https://www.mdpi.com/1424-8220/22/4/1407 (accessed Sep. 15, 2023).

[10] "A novel selective naïve Bayes algorithm - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0950705119306185 (accessed Sep. 15, 2023).

[11] H. Elmunsyah, R. Mu'awanah, T. Widiyaningtyas, I. A. E. Zaeni, and F. A. Dwiyanto, "Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm," in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Oct. 2019, pp. 211–215. doi: 10.1109/ICEEIE47180.2019.8981418.

[12] L. K. Xin and N. binti A. Rashid, "Prediction of Depression among Women Using Random Oversampling and Random Forest," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, Mar. 2021, pp. 1–5. doi: 10.1109/WiDSTaif52235.2021.9430215.

[13] W. A. van Eeden *et al.*, "Predicting the 9-year course of mood and anxiety disorders with automated machine learning: A comparison between auto-sklearn, naïve Bayes classifier, and traditional logistic regression," *Psychiatry Res.*, vol. 299, p. 113823, May 2021, doi: 10.1016/j.psychres.2021.113823.

[14] "Student Mental health." https://www.kaggle.com/datasets/shariful07/student-mental-health (accessed Sep. 15, 2023).

[15] T. Hidayat, D. U. E. Saputri, and F. Aziz, "MEAT IMAGE CLASSIFICATION USING DEEP LEARNING WITH RESNET152V2 ARCHITECTURE," *J. Techno Nusa Mandiri*, vol. 19, no. 2, pp. 131–140, 2022.

[16] "Google Colaboratory | SpringerLink." https://link.springer.com/chapter/10.1007/978-1-4842-4470-8_7 (accessed Sep. 15, 2023).