Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.9 No.1 January 2024, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Analytical Comparison of Lung Cancer Classification Using K-Nearest Neighbor and Naïve Bayes Algorithms

Yunisa Darmayanti[1*], Fitri Marisa[2], Aviv Yuniar Rahman[3]

[1,2,3]*Informatics Department, Universitas Widyagama, Malang, Indonesia*

[1]unisadarma199@gmail.com (*)

[2,3][fitrimarisa, aviv]@widyagama.ac.id

Received: 2023-12-20; Accepted: 2024-01-28; Published: 2024-02-08

Abstract— Lung cancer stands as a significant global contributor to human mortality, constituting 25% of all cancer-related deaths in 2021. Its elusive nature, often devoid of early symptoms in a quarter of diagnosed cases, poses a challenge for timely detection. Unlike some other cancers, lung cancer remains hidden from the naked eye, with its symptoms often masquerading as those of other ailments like bronchitis, asthma, or persistent coughs. Early diagnosis is pivotal for effective treatment and increased survival rates. In light of the pressing nature of the situation, this study investigates the prediction of lung cancer by using data mining tools. It is essential to conduct data mining, which is a process that involves searching for patterns and trends inside vast data repositories to discover valuable insights. Within this context, classification emerges as a fundamental aspect that discerns objects based on their distinctive characteristics. A comparative study is undertaken to address the complexities associated with lung cancer classification, focusing on the K-Nearest Neighbor (KNN) and Naïve Bayes Classifier (NBC) algorithms. Through the utilization of a dataset that contains one thousand instances and twenty-four criteria, the purpose of this study is to determine which algorithm is preferable in the categorization of lung cancer. Upon analysis, the study yields noteworthy results. The KNN algorithm exhibits an accuracy rate of 98.34%, surpassing the NBC algorithm's accuracy of 89.37%. Consequently, this research concludes that, in lung cancer classification, the KNN algorithm outperforms the Naïve Bayes algorithm. These findings promise to enhance the efficacy of early lung cancer detection, potentially saving numerous lives through improved classification methods.

Keywords— Lung Cancer Classification; K-Nearest Neighbor; Naïve Bayes Classifier.

I. INTRODUCTION

The most significant cause of death in humans is cancer [1], [2]. Cancer is an abnormal growth of cells that can grow and spread to other parts of the body [3]. Cancer is one of the diseases that can linger for a very long time. Patients who have cancer frequently experience pain, which may be brought on by the symptoms of the disease itself or by the procedures and treatments they undergo [4]. Lung cancer remains the most common type of cancer globally and accounts for nearly 25% of all cancer-related deaths in 2021. More than 80% of these cases can be directly attributed to smoking. In addition, about another 2.7% of deaths are caused by passive exposure to cigarette smoke [5], [6]. Lung cancer is characterized by uncontrolled cell growth in the lung tissue, especially the cells that line the respiratory system [7], [8]. Based on the research results, approximately a quarter of individuals diagnosed with early-stage lung cancer do not manifest any noticeable symptoms. In contrast to some other forms of cancer, lung cancer remains invisible to the unaided eye, and its indications frequently overlap with those of other illnesses like bronchitis, asthma, and cough. [9], [10]. The early classification of lung cancer is necessary to achieve a cure for the disease. Early diagnosis of lung disease has the potential to save a significant number of lives [1]. Using data mining tools makes it possible to make predictions regarding lung cancer [1].

Data mining is a process that utilizes statistical, mathematical, artificial intelligence, and machine learning techniques to discover and extract valuable information and pertinent knowledge from a variety of interconnected databases. [11]. Data mining also aims to find new similarities, patterns, and trends that have meaning by categorizing extensive data in repositories. This is done through applying pattern recognition technology and statistical and mathematical techniques [12]. This process can involve several stages, including data preprocessing, exploration data, modeling, and evaluation [13]. One of the most essential things in data mining is classification [14]. Classification is a way to distinguish objects based on their characteristics [15]. Classification is a process that consists of two stages, namely, the learning stage and the classification stage. A classification algorithm will build a classification model in the learning stage by analyzing training data. The learning stage can also be viewed as a function or mapping stage $Y=F(X)$, where Y is the predicted class, and X is the tuple whose class will be indicated. Classification is the process of finding a set of models that describe and distinguish data classes with the aim that the model can be used to predict the class of an object whose class is unknown [16]. Classification uses technology with several algorithms, such as Naive Bayes, Decision Tree, and K-Nearest Neighbor [1], [17].

In previous research, the Naïve Bayes algorithm has been applied to classify lung cancer. The study used 309 data in the form of CSV, divided into 70% training data and 30% testing

DOI : https://doi.org/10.25139/inform.v9i1.7592

data with 16 attributes. This research resulted in an accuracy of 94.62% [18].

The following study also used Naïve Bayes for lung cancer classification. The study used 134 data, which were divided into 100 training data and 34 testing data, with the criteria of shortness of breath, cough, bloody cough, phlegm, fever, weakness, decreased appetite, nausea, vomiting, stool, tub, history of asthma, history of stroke, history of tuberculosis, headache, heartburn, chest pain, weight loss, and night sweats. The study resulted in an accuracy of 97.06% [14].

Previous research has compared SVM and KNN algorithms for lung cancer classification. The study used 309 data divided into 70% training data and 30% testing data with 16 attributes. The results showed that the SVM algorithm has better accuracy than the KNN algorithm, which is 92.61% compared to 89.65%. [19].

Previous research has implemented classification using the Random Forest algorithm. Tests were conducted using Confusion Matrix and K-fold Cross Validation. The Confusion Matrix test results showed the highest accuracy of 0.904 and an average accuracy of 0.813. Test results using K-fold cross-validation show the highest average accuracy of 0.889 when using 5-fold cross-validation [20].

The objective is to solve this issue by conducting comparative research between the K-Nearest Neighbour and Naïve Bayes algorithms for lung classification. This research will use 1000 datasets with 24 criteria. The results are expected to provide information on which algorithm is better for lung classification.

II. RESEARCH METHODOLOGY

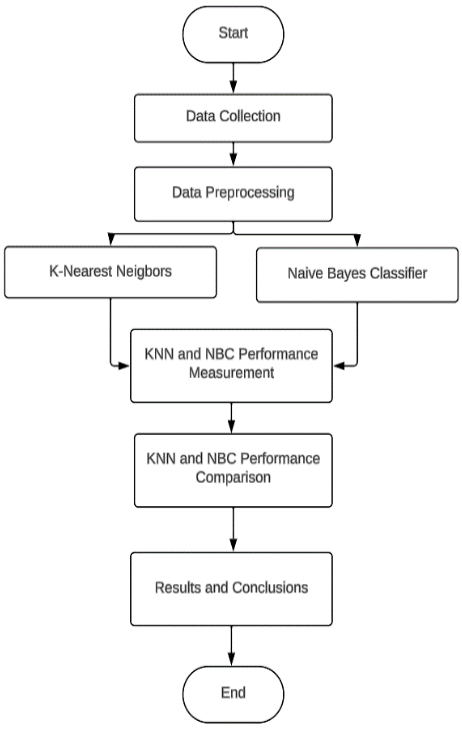

Figure 1. Research Stages

The study presents a visual representation in Figure 1, illustrating the comparison between the KNN and Naïve Bayes algorithms in the context of lung cancer classification. They started with preparing the dataset and then preprocessing it. Then, the data is divided into two: training and testing. Then, calculations are carried out using the KNN and Naïve Bayes algorithms using RapidMiner toll. Following the completion of algorithmic computations, the accuracy is subsequently assessed. The accuracy results are then compared to which one is more suitable for classifying lung cancer.

A. K- Nearest Neighbor

The KNN algorithm is a classification technique that identifies the k nearest neighbors of a given data point slated for classification. [21]. The nearest neighbor is another data point with the closest distance to the data point being classified [21]. It is possible to determine the distance between two data points by employing a variety of distance metrics, such as the Euclidean distance, the Manhattan distance, and the Minkowski distance [21]. This research employs Euclidean distance to compute the distance between two data points. The Euclidean distance formula is in the Equation (1)[21]

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

The following symbols have a unique role in measuring the distance between data in a system. Expressed as d, the variable distance (d) indicates the distance between two data in the system. The variable (x) represents training data, the data set used to train the model or system. Meanwhile, variable (y) refers to test data, which is the data used to test the performance of the trained model or system. The variable (n), or data dimension, characterizes the number of dimensions or attributes the data owns. Finally, the variable (i) signifies the data variable, which can refer to a particular attribute or dimension of the data. KNN Algorithm modeling flow is presented in Figure 2.

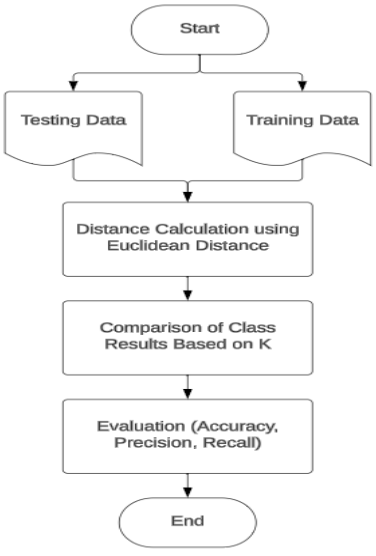

Figure 2. Testing Flow of KNN Algorithm

DOI : https://doi.org/10.25139/inform.v9i1.7592

B. Naïve Bayes

The Naïve Bayes algorithm functions as a straightforward probability classifier, determining a set of probabilities by tallying frequencies and combinations of values within a provided dataset. Grounded in Bayes' theorem, the algorithm operates under the assumption of independence among all variables, considering the class variable's value. This presumption of conditional independence is often not met in practical, real-world scenarios. Consequently, it is termed "Naïve." However, the algorithm demonstrates swift learning capabilities across controlled classification problems despite this simplification. [22]. Naïve Bayes can be calculated using Bayes Theorem Equation (2) [23].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

Where, *P(H)* is the prior probability value of the hypothesis on a sample, commonly referred to as a priori. The *P(X)* variable is the evidence of the training data probability. The *P(H|X)* variable is the probability value of *H* affecting *X* (posterior density), while *P(X|H)* is the probability of x to h called the likelihood. The NBC Algorithm modelling flow is presented in Figure 3.

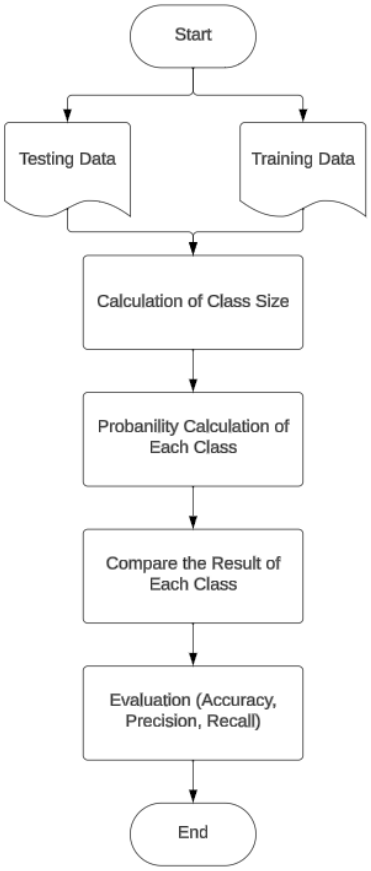

Figure 3. Testing Flow of NBC Algorithm

C. Evaluation

Following the completion of classification using the Naïve Bayes and KNN algorithms, a comparison will be conducted between the accuracy values obtained from Equation (3), the precision values obtained from Equation (4), and the recall values obtained from Equation (5) in the context of predicting lung cancer disease. Accuracy entails the proportion of accurate predictions concerning the total predictions made. Precision represents the ratio of correct optimistic predictions to the overall positive predictions. Recall delineates the ratio of correct optimistic predictions to the total actual positive data [24].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (3)$$

$$Precision = \frac{TP}{FP+TP} x100\% \quad (4)$$

$$Recall = \frac{TP}{FN+TP} x100\% \quad (5)$$

In analyzing the performance of a model or classification system's performance, four important metrics reflect the extent to which the model's predictions are accurate. True positive (TP) indicates the true-positive data was correctly predicted as positive. True-negative (TN) refers to the true-negative data correctly predicted as negative. Meanwhile, false negative (FN) is the amount of data that is actually positive but was mistakenly predicted as negative. On the other hand, false positive (FP) describes the amount of data that is actually negative but mistakenly predicted as positive. These four metrics help provide a comprehensive picture of the model's ability to recognize and predict the data class [25].

III. Result and Discussion

This study employs data obtained from the Kaggle repository, presented in the format of CSV files, comprising a total of 1000 entries. Out of the complete dataset, 700 entries were utilized for the training phase, while the remaining 300 were reserved for the testing phase. Twenty-four criteria become research variables, including individual characteristics such as Genetic Risk, Obesity, Wheezing, Alcohol use, Chronic Lung Disease, Dust Allergy, Passive Smoking, Snoring, Chest Pain, Coughing of Blood, Air Pollution, Smoking, Weight Loss, Shortness of Breath, Fatigue, Balanced Diet, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, Age, Gender, Occupational Hazards and Level as outcome criteria. The sample of the dataset used is presented (Table I).

Table I
Lung Cancer Data

Index	Patient Id	Age	…	Gender	Snoring	Level
0	P1	33	…	1	4	Low
1	P2	17	…	1	2	Medium
2	P3	35	…	1	2	High
3	P4	37	…	1	5	High
4	P5	46	…	1	3	High

DOI : https://doi.org/10.25139/inform.v9i1.7592

Index	Patient Id	Age	…	Gender	Snoring	Level
5	P6	35	…	1	2	High
…	…	…	…	…	…	…
994	P995	33	…	1	2	High
995	P996	44	…	1	3	High
996	P997	37	…	2	4	High
997	P998	25	…	2	2	High
998	P999	18	…	2	3	High
999	P1000	47	…	1	2	High

A. Data Preprocessing

In the preprocessing stage, data transformation is a process in which data patterns can be changed through the application of data mining techniques. In this context, data mining techniques are used to identify, explore, and analyze hidden or not directly visible patterns in the dataset. The primary purpose of data transformation is to change the data's structure or distribution to better suit the needs of further analysis or processing [26]. The sample of the dataset after preprocessing is presented in Table II.

TABLE II
LUNG CANCER DATA AFTER PRE-PROCESSING

Index	Patient Id	Age	…	Gender	Snoring	Level
0	P1	Dewasa Muda	..	1	4	Low
1	P2	Remaja	..	1	2	Medium
2	P3	Dewasa Muda	..	1	2	High
3	P4	Dewasa	..	1	5	High
4	P5	Dewasa Lanjut	..	1	3	High
5	P6	Dewasa Muda	..	1	2	High
…	…	…	..	…	…	…
994	P995	Dewasa Muda	..	1	2	High
995	P996	Dewasa	..	1	3	High
996	P997	Dewasa	..	2	4	High
997	P998	Remaja	..	2	2	High
998	P999	Remaja	..	2	3	High
999	P1000	Dewasa Lanjut	..	1	2	High

B. Classification using KNN and Naïve Bayes Algorithms

The classification results using the K-Nearest Neighbors (KNN) algorithm on the RapidMiner tool with a split ratio of 70:30 and a value of k = 11 showed an accuracy rate of 98.34%. Meanwhile, in the classification using the Naïve Bayes Classification (NBC) algorithm with the same split ratio, an accuracy rate of 89.37% was obtained. Details of the classification results for both algorithms can be found in Table III.

TABLE III
KNN AND NBC TESTING RESULTS

	True Low		True Medium		True High	
	KNN	NBC	KNN	NBC	KNN	NBC
Pred. Low	91	85	0	0	0	0
Pred. Medium	0	4	95	81	0	7
Pred. High	0	2	5	19	110	103

Furthermore, to offer a more comprehensive comprehension of the performance of the two algorithms, the results of the K-Nearest Neighbours and Naïve Bayes Classifier tests can be elucidated through the Confusion Matrix. This matrix encompasses the average precision, recall, and accuracy values presented in Table IV.

TABLE IV
KNN AND NBC MODEL PERFORMANCE

Model Performance	KNN	NBC
Precission	98,55%	90,36%
Recall	98,33%	89,35%
Accuracy	98,34%	89,37%

C. Performance Comparison of KNN and Naïve Bayes Algorithms

The classification testing process that has been carried out on the KNN and NBC algorithms, the performance of each algorithm. Figure 4 shows the precision, recall, and accuracy values for the KNN and Naïve Bayes algorithms. In the KNN algorithm, the precision, recall, and accuracy values are 98.55%, 98.33%, and 98.34%, respectively. While in the Naïve Bayes algorithm, the precision, recall, and accuracy values are 90.36%, 89.35%, and 89.37%, respectively. This study shows that the KNN algorithm has better performance than Naïve Bayes in lung cancer classification.

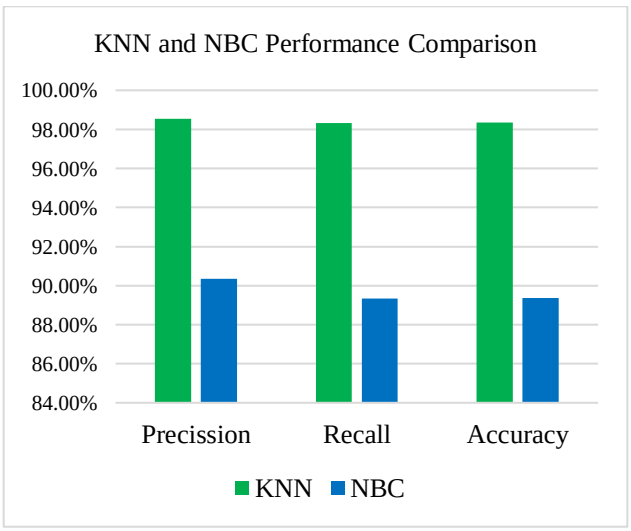

Figure 4. Performance Comparison of Algorithms

DOI : https://doi.org/10.25139/inform.v9i1.7592

IV. CONCLUSION

In this investigation, data related to lung cancer was extracted from the Kaggle repository, presenting a dataset of 1000 entries formatted in CSV. The dataset was divided into two subsets, with 700 entries designated for the training phase and 300 for the subsequent testing process. A comprehensive analysis encompassed 24 criteria to delve into various aspects of lung cancer.

The testing phase was executed using the powerful RapidMiner software, facilitating a meticulous examination of the accuracy values yielded by the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms. The obtained results delineated a noteworthy disparity in accuracy between the two algorithms. Specifically, the KNN algorithm showcased an impressive accuracy of 98.34%, while the Naïve Bayes algorithm trailed with a slightly lower accuracy value of 89.37%.

This discernible difference in accuracy strongly suggests that the KNN algorithm outperforms its Naïve Bayes counterpart in the classification of lung cancer based on the dataset employed in this study. The findings underscore the KNN algorithm's capacity to provide more precise and effective results in lung disease classification. This conclusion carries valuable implications for advancing lung cancer disease detection and classification methodologies utilizing data mining techniques, presenting a promising stride toward enhancing diagnostic accuracy and patient outcomes in the field of pulmonary health.

REFERENCES

[1] M. Ismail, "Lung Cancer Prediction using Data Mining Techniques," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, pp. 2277–3878, Nov. 2019, doi: 10.35940/ijrte.D9914.118419.

[2] N. Kalaivani, N. Manimaran, Dr. S. Sophia, and D. D Devi, "Deep Learning Based Lung Cancer Detection and Classification," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 994, no. 1, p. 012026, Dec. 2020, doi: 10.1088/1757-899X/994/1/012026.

[3] L. Rahayuwati, I. A. Rizal, T. Pahria, M. Lukman, and N. Juniarti, "Pendidikan Kesehatan tentang Pencegahan Penyakit Kanker dan Menjaga Kualitas Kesehatan," *Media Karya Kesehatan*, vol. 3, no. 1, Art. no. 1, Apr. 2020, doi: 10.24198/mkk.v3i1.26629.

[4] A. Rahman, D. Gayatri, and A. Waluyo, "Dukungan Sosial terhadap Kualitas Hidup Pasien Kanker," *Journal of Telenursing (JOTING)*, vol. 5, no. 1, Art. no. 1, Jun. 2023, doi: 10.31539/joting.v5i1.5770.

[5] Y. Cheng, T. Zhang, and Q. Xu, "Therapeutic advances in non-small cell lung cancer: Focus on clinical development of targeted therapy and immunotherapy," *MedComm*, vol. 2, no. 4, pp. 692–729, 2021, doi: 10.1002/mco2.105.

[6] M. Vedaraj, C. S. Anita, A. Muralidhar, V. Lavanya, K. Balasaranya, and P. Jagadeesan, "Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 6s, Art. no. 6s, May 2023.

[7] J. Amalia, "ANALISIS MODEL REGRESI COX PROPORTIONAL HAZARD PADA STUDI KASUS PASIEN KANKER PARU-PARU," *JURNAL ILMIAH SIMANTEK*, vol. 4, no. 1, Art. no. 1, Feb. 2020.

[8] A. M. T. I. S. Ua *et al.*, "Penggunaan Bahasa Pemrograman Python Dalam Analisis Faktor Penyebab Kanker Paru-Paru," *Jurnal Publikasi Teknik Informatika*, vol. 2, no. 2, Art. no. 2, Jul. 2023, doi: 10.55606/jupti.v2i2.1742.

[9] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Comput & Applic*, vol. 31, no. 10, pp. 6863–6877, Oct. 2019, doi: 10.1007/s00521-018-3518-x.

[10] F. Taher, N. Prakash, A. Shaffie, A. Soliman, and A. El-Baz, "An Overview of Lung Cancer Classification Algorithms and their Performances," vol. 48, no. 4, 2021.

[11] M. Nabeel, S. Majeed, M. Awan, H. Muslih-Ud-Din, M. Wasique, and R. Nasir, "Review on Effective Disease Prediction through Data Mining Techniques," *International Journal on Electrical Engineering and Informatics*, vol. 13, Sep. 2021, doi: 10.15676/ijeei.2021.13.3.13.

[12] S. K. P. Loka and A. Marsal, "Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier untuk Klasifikasi Status Gizi Pada Balita: Comparison Algorithm of K-Nearest Neighbor and Naïve Bayes Classifier for Classifying Nutritional Status in Toddlers," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, Art. no. 1, May 2023, doi: 10.57152/malcom.v3i1.474.

[13] M. A. R. Wahid, A. Nugroho, and A. H. Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, Art. no. 1, Mar. 2023, doi: 10.47065/bit.v4i1.501.

[14] M. Y. Haffandi, E. Haerani, F. Syafria, and L. Oktavia, "KLASIFIKASI PENYAKIT PARU-PARU DENGAN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER," *Jurnal Tekinkom (Teknik Informasi dan Komputer)*, vol. 5, no. 2, Art. no. 2, Dec. 2022, doi: 10.37600/tekinkom.v5i2.649.

[15] Y. A. Suwitono and F. J. Kaunang, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras," *1*, vol. 6, no. 2, Art. no. 2, Nov. 2022, doi: 10.31603/komtika.v6i2.8054.

[16] H. Susana, "PENERAPAN MODEL KLASIFIKASI METODE NAIVE BAYES TERHADAP PENGGUNAAN AKSES INTERNET," *Jurnal Riset Sistem Informasi dan Teknologi Informasi (JURSISTEKNI)*, vol. 4, no. 1, Art. no. 1, Feb. 2022, doi: 10.52005/jursistekni.v4i1.96.

DOI : https://doi.org/10.25139/inform.v9i1.7592

[17] K. Hayati and R. Habibi, "KLASIFIKASI KELAYAKAN MAHASISWA MASUK PROGRAM MSIB KAMPUS MERDEKA:," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 3, Art. no. 3, Nov. 2023, doi: 10.36040/jati.v7i3.6882.

[18] E. Wulandari, "KLASIFIKASI KANKER PARU-PARU MENGGUNAKAN METODE NAIVE BAYES," *International Research on Big-Data and Computer Technology: I-Robot*, vol. 6, no. 2, Art. no. 2, Sep. 2022, doi: 10.53514/ir.v6i2.325.

[19] A. Desiani *et al.*, "Perbandingan Klasifikasi Penyakit Kanker Paru-Paru menggunakan Support Vector Machine dan K-Nearest Neighbor," *Jurnal PROCESSOR*, vol. 18, no. 1, Art. no. 1, Apr. 2023, doi: 10.33998/processor.2023.18.1.700.

[20] R. D. Marzuq, S. A. Wicaksono, and N. Y. Setiawan, "Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 14, Art. no. 14, Oct. 2023, Accessed: Dec. 19, 2023. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12964

[21] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[22] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, Art. no. 2, Jun. 2019, doi: 10.18201//ijisae.2019252786.

[23] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 1, pp. 59–66, Jan. 2022, doi: 10.34010/komputika.v11i1.4350.

[24] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *Jurnal Informatika*, vol. 9, no. 1, Art. no. 1, Apr. 2022, doi: 10.31294/inf.v9i1.11931.

[25] I. H. Kusuma and N. Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 3, Art. no. 3, Sep. 2023, doi: 10.30591/jpit.v8i3.5734.

[26] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science (JINACS)*, vol. 5, no. 01, pp. 97–100, Jul. 2023.

DOI : https://doi.org/10.25139/inform.v9i1.7592