International Journal of Education and Life Sciences (IJELS)
Vol. 4, No. 7 July 2026: 933-946

Classification Analysis of 2 K-Nearest Neighbor (KNN) and Decision Tree Algorithms Using Rapidminer in Breast Cancer

Asri Liya Astuti
Universitas Pelita Bangsa, Bekasi Regency
Corresponding Author: Asri Liya Astuti: asriliyaastuti@gmail.com

A R T I C L E I N F O
Keywords: Breast cancer, K-Nearest Neighbor (KNN, Decision Tree C4.5, ROC Curve, Matrix Confusion

Received : 27, May
Revised : 20, June
Accepted: 8, July

A B S T R A C T
Cancer remains the primary cause of mortality in developed nations and ranks second in developing countries. According to 2018 data from the Global Cancer Observatory, Indonesia recorded a cancer incidence rate of 136.2 per 100,000 people, placing it 8th in Southeast Asia and 23rd across Asia. Among women, breast cancer had the highest incidence rate at 42.1 per 100,000, with a mortality rate averaging 17 per 100,000, while cervical cancer followed at 23.4 per 100,000 incidence and 13.9 per 100,000 mortality. Many patients assume that once breast cancer is no longer detectable in the body, it will not return yet recurrence is a real possibility that remains poorly understood by the public. Breast cancer is among the most common cancers affecting Indonesians, especially women, and is generally classified into two forms: benign (non-cancerous) and malignant (cancerous). This study aims to identify an accurate algorithm for analyzing breast cancer datasets, generate insights into patterns or models that support early detection, and compare the performance of two classification algorithms Decision Tree C4.5 and K-Nearest Neighbor (KNN) in classifying breast cancer cases. The research follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The central question guiding this study is which of the two algorithms delivers higher classification accuracy and greater utility for identifying patterns relevant to early breast cancer detection. Findings from the CRISP-DM analysis revealed that KNN outperformed the Decision Tree C4.5 model, achieving an accuracy of 97.14% and an AUC score of 0.976, indicating excellent classification performance (AUC values between 0.90 and 1.00 are considered very good).

DOI: https://doi.org/10.59890/ijels.v4i7.44
ISSN : 3026-2429
https://ntlmultitechpublisher.my.id/index.php/ijels

INTRODUCTION

In addition to ranking first in terms of the number of cancer cases in Indonesia, breast cancer also ranks first in terms of the number of deaths due to cancer. One of the most common types of cancer suffered by Indonesians, especially women, is breast cancer. An uncontrolled cell growth disease known as cancer can spread to other organs in the body. This disease can cause death because it is considered a dangerous level. [1]

According to Globocan data in 2020, there were 65,858 new cases of breast cancer in Indonesia, which is 16.6% of the total 396,914 new cases of cancer. However, the number of cases reached more than 22 thousand deaths. One of the most common cancers in women around the world is breast cancer. Breast cancer is usually divided into two types: benign, or commonly called benign, and malignant, or commonly called malignant. Benign breast cancer is usually characterized by small, round, soft lumps that are not cancerous. Although it can be detected, this cancer will not spread and damage the surrounding tissues. An asymmetrical, rough, and painful shape is a sign of malignant breast cancer. Breast cancer usually spreads. [2]

Cancer is the leading cause of death in countries with advanced economies and the second cause of death in countries with developing economies. Data from the Global Cancer Observatory in 2018 shows that the incidence rate of cancer in Indonesia (136.2/100,000 population) is 8th in Southeast Asia and 23rd in Asia. Breast cancer, with an incidence rate of 42.1 per 100,000 population and an average mortality rate of 17 per 100,000 population, is followed by cervical cancer, with an incidence rate of 23.4 per 100,000 population. Many people don't realize that breast cancer can recur again because it has disappeared from the sufferer's body. After treatment, usually only a few cancer cells remain alive. But at small doses, the cells are more serious problems. The disease can recur in the chest, chest, or other parts of the body, such as bones and liver. [3]

According to data from the Global Cancer Observatory, breast cancer accounted for 30.8% of all cancer-related deaths among women across all age groups in 2020. This study utilizes breast cancer datasets with the goal of raising awareness and deepening understanding of the disease, as public knowledge of breast cancer is essential and should be widely accessible information [4]. Breast cancer originates in the cells of breast tissue and stands as both the most frequently diagnosed cancer among women and the foremost cause of cancer-related deaths in women globally. [5]

Data mining has become a widely used and validated approach for classification analysis across various fields, including medicine, finance, marketing, and social science. In the medical domain, for instance, machine learning techniques are applied to analyze tumor behavior in breast cancer patients.

While various data mining techniques are available, the choice of method must align with the specific objective of the analysis. Common examples include

K-Nearest Neighbor, Support Vector Machine, Naive Bayes, and C4.5. The K-Nearest Neighbor (KNN) algorithm, in particular, enables the calculation of both accuracy and error values. Accuracy reflects the proportion of correctly classified instances relative to the total sample size, whereas the error value indicates the extent of misclassification measured against the same total number of samples. [6]

A prior study conducted by Fahrurrozi and Wasilah demonstrated that the Decision Tree classifier algorithm is capable of handling large databases effectively for feature selection purposes. Additionally, their findings indicated that the K-Nearest Neighbors (KNN) algorithm recognized as a suitable and effective method for cancer analysis and diagnosis achieved an accuracy rate of 94.73%. [2].

Based on the description above, this study with the title "Classification Analysis 2 K-Nearest Neighbor (Knn) Algorithm and Decision Tree Using Rapidminer in Breast Cancer" will compare the K-Nearest Neighbor (KNN) and Decision Tree methods to classify breast cancer.

METHOD

This study used the Cross-Industry Standard Process for Data Mining (CRISP-DM) as its research methodology, a framework comprising six distinct phases, [30]namely:

1. Research Understanding Phase
2. Data Understanding
3. Data Prepation
4. Modeling
5. Evaluation
6. Deportment

Research Understanding

The dataset used in this research is the Breast Cancer Wisconsin (Original) dataset, obtained from the UCI Machine Learning Repository and accessed on July 16, 2024, via the following page: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original. The dataset comprises 699 records across 11 attributes, of which 241 instances are classified as malignant breast cancer cases. Given the challenge posed by insufficient accurate analytical methods for this data, the present study applies both the K-Nearest Neighbor (KNN) and Decision Tree algorithms to perform classification.

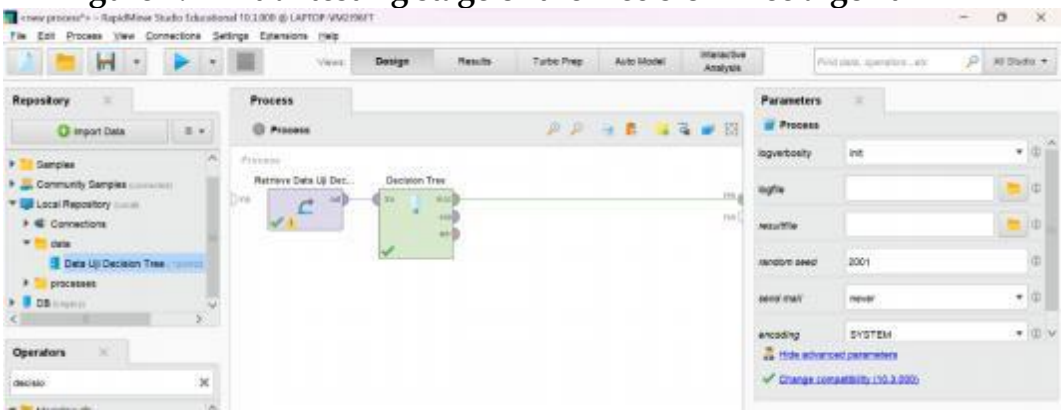
RESULTS

Research Results and Evaluation

As discussed in the second chapter, researchers can utilize a wide variety of techniques to build classification models. The focus is on applying the K-Nearest Neighbor and Decision Tree C4.5 algorithms. While the evaluation of a classification process can involve multiple criteria including interpretability, scalability, reliability, speed, and accuracy the primary objective here is to compare the two selected methods to identify which yields the highest precision. The main goal of this research is to measure how accurately breast cancer data can be analyzed using both the K-Nearest Neighbor and Decision Tree C4.5 algorithms. Consequently, a direct comparison was performed between these two machine learning models to thoroughly assess their respective predictive accuracies.

During this phase, the Decision Tree methodology is tested and implemented utilizing the RapidMiner software. The systematic procedure for the Decision Tree C4.5 model requires inserting a "read" operator to import data from Microsoft Excel, adding a "Decision Tree" operator to link the entire workflow together, and finally executing the process by pressing the run button. This visual setup is illustrated in Figure 4.1.

Figure 4.1 Initial testing stage of the Decision Tree algorithm



This resulted in the C4.5 Decision Tree Algorithm model. It can be seen in Figure 4.2.

Figure 4.2 Decision Tree C4.5 Algorithm Model

Figure 4.3 displays the process of extracting accuracy, precision, and recall measurements through the addition of the "apply model" and "performance" operators.

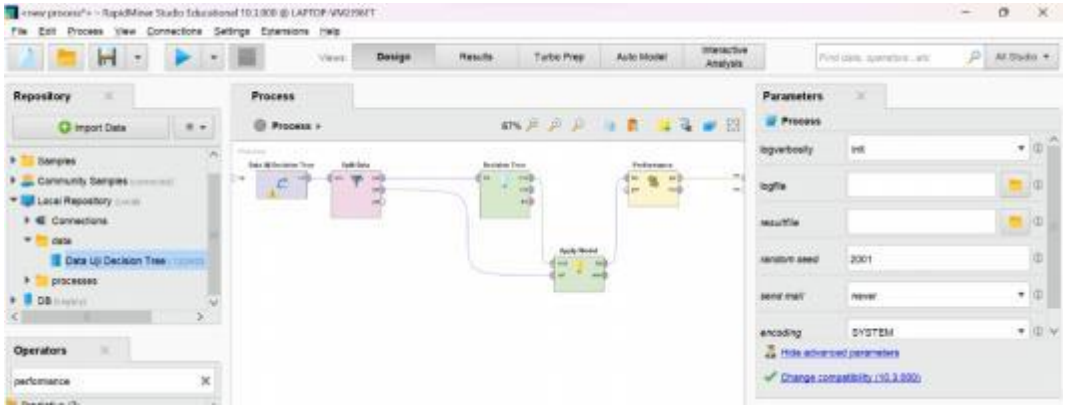


Figure 4.3 The final testing step for the Decision Tree algorithm

During this phase, the results are evaluated using performance measurement tools that generate a confusion matrix, which displays the accuracy, precision, and recall values. The primary focus here is to determine the accuracy achieved by each output.

Accuracy refers to how closely the predicted results align with the actual outcomes, while precision measures how well the system's responses match the information the user is seeking. Recall, meanwhile, reflects the system's success rate in retrieving relevant information.

accuracy: 95.71%

	true Benign	true Malignant	class precision
pred. Benign	44	1	97.78%
pred. Malignant	2	23	92.00%
class recall	95.65%	95.83%	

Figure 4.4 Final testing stage of the Decision Tree algorithm

Figure 4.4 presents the results obtained by splitting the dataset into 90% training data and 10% testing data. Out of the 70 testing samples, the C4.5 method correctly classified 44 instances as Benign. However, 1 instance predicted as benign was actually malignant, while 23 malignant instances were correctly identified, and 2 instances predicted as malignant were, in fact, benign.

ROC Decision Tree Curve

Figure 4. 5 ROC Decision Tree Algorithm

The ROC curve shown above illustrates the confusion matrix presented in Figure 4.5, where the horizontal axis represents the false positive rate and the vertical axis represents the true positive rate. To facilitate the interpretation of the breast cancer data, performance evaluation tools were incorporated to calculate the Root Mean Squared Error and Squared Error, with the corresponding results displayed in Figure 4.6.

Figure 4.6 Performance Vector Results

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.998 +/- 0.000
squared_error: 0.996 +/- 0.707
```

Subsequently, the Decision Tree approach is executed within the RapidMiner environment by adhering to the following steps:

Process the testing dataset in RapidMiner to produce confidence metrics for every assigned class.

Evaluate the model's effectiveness by computing both the Root Mean Squared Error and the Squared Error.

The split validation framework is separated into two distinct phases. The first is the training phase, which is responsible for managing the chosen classification algorithm. The second is the testing phase, where the "Apply Model" function is utilized to evaluate the test data, and the "Performance" function is deployed to generate the final Root Mean Squared Error and Squared Error metrics.

K-Nearest Neighbor (KNN)

During this phase, the K-Nearest Neighbor (KNN) methodology is tested and deployed via RapidMiner. Constructing the KNN model requires a systematic workflow: initiating data import via the "Read Excel" operator, adding

the K-Nearest Neighbor operator, and subsequently linking all components together. After configuring these connected operators, the analysis is executed by pressing the run button, a process visually depicted in Figure 4.7

Figure 4.7 Preliminary evaluation phase of the K-Nearest Neighbor (KNN) method.

This procedure subsequently generates the final K-Nearest Neighbor (KNN) model, which is illustrated in Figure 4.8.

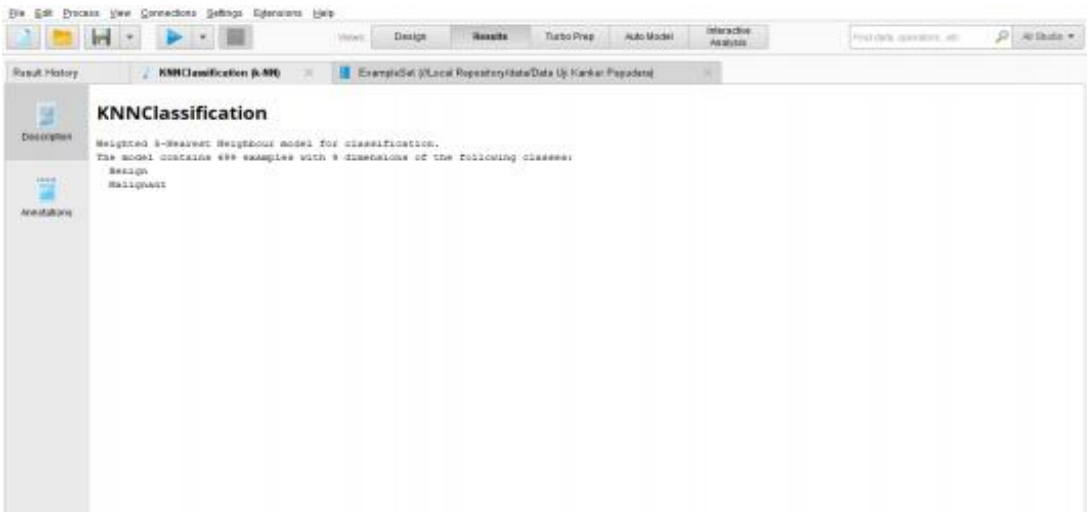


Figure 4.8 Decision Tree C4.5. Algorithm Model

Additionally, the accuracy, precision, and recall values can be obtained by incorporating the "apply model" and "performance" operators, as demonstrated in Figure 4.9.

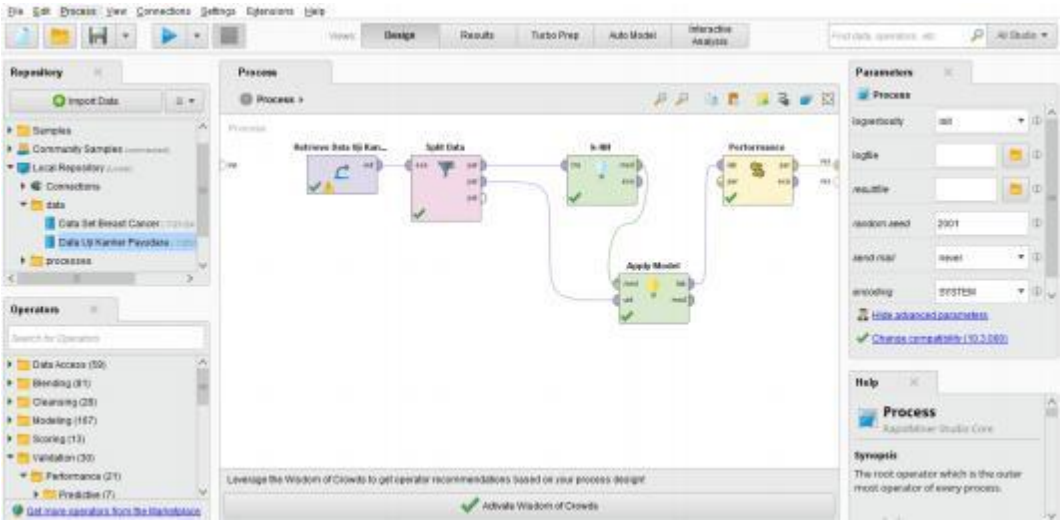


Figure 4.9 Concluding evaluation phase for the K-Nearest Neighbor (KNN) model.

During this step, performance measurement features are utilized to construct a confusion matrix, which displays the metrics for accuracy, precision, and recall. The primary goal of this assessment is to establish the specific accuracy and precision rates. In this context, accuracy indicates how well the system's output matches the user's requested information, whereas precision measures the exactness of the predicted classifications compared to the true data. Additionally, recall demonstrates the algorithm's effectiveness in accurately finding and returning pertinent data.

accuracy: 97.14%

	true Benign	true Malignant	class precision
pred. Benign	44	0	100.00%
pred. Malignant	2	24	92.31%
class recall	95.65%	100.00%	

Figure 4.10 Final testing stage of the K-Nearest Neighbor (KNN) algorithm

Figure 4.10 presents the results derived from splitting the dataset into 90% training data and 10% testing data. Among the 70 testing samples, the C4.5 method correctly classified 44 instances as Benign. However, 1 instance predicted as benign was actually malignant, 23 malignant instances were accurately predicted, and 2 instances predicted as malignant turned out to be benign.

ROCK-Nearest Neighbor (KNN) Curve

Figure 4.11 ROC K-Nearest Neighbor (KNN) Curve

The preceding ROC curve visually maps the confusion matrix detailed in Figure 4.11, positioning the false positive rate along the horizontal x-axis and the true positive rate on the vertical y-axis. To make the breast cancer dataset easier to analyze, performance evaluation operators were utilized to compute both the Squared Error and Root Mean Squared Error, the outcomes of which are shown in Figure 4.12.

PerformanceVector

```
PerformanceVector:
accuracy: 97.14%
ConfusionMatrix:
True:   Benign  Malignant
Benign: 44      0
Malignant:      2       24
precision: 92.31% (positive class: Malignant)
ConfusionMatrix:
True:   Benign  Malignant
Benign: 44      0
Malignant:      2       24
recall: 100.00% (positive class: Malignant)
ConfusionMatrix:
True:   Benign  Malignant
Benign: 44      0
Malignant:      2       24
AUC (optimistic): 0.995 (positive class: Malignant)
AUC: 0.976 (positive class: Malignant)
AUC (pessimistic): 0.976 (positive class: Malignant)
```

Figure 4.12 Evaluation Outcomes for Root Mean Square Error and Squared Error.

Following this, a linear regression algorithm is implemented utilizing the RapidMiner application. The steps required to execute this linear regression approach are as follows:

Classify the testing dataset within RapidMiner to yield confidence levels for each categorized instance.

Configure the performance metrics to extract the Squared Error and Root Mean Squared Error.

In the context of the split validation framework, the workflow is divided into two main segments. The training segment is responsible for processing the classification algorithm. Conversely, the testing segment employs the "Apply Model" tool to evaluate the model against the test dataset and uses the "Performance" tool to reveal the resulting Root Mean Squared Error and Squared Error figures.

RESULT

The resulting algorithm models were measured using the split validation technique. To discover the most superior model, researchers compared key metrics including accuracy, recall, precision, and the ROC curve.

A comparative analysis of the AUC and Accuracy metrics for both algorithms is shown in Table 4.1. The findings demonstrate that the K-Nearest Neighbor (KNN) approach yielded the highest scores across both metrics, although the Decision Tree C4.5 also delivered competitive outcomes. Drawing from the data in Table 4.1, it is evident that both the KNN and Decision Tree C4.5 models are highly effective for classification tasks, given that they both recorded AUC scores in the optimal 0.90 to 1.00 bracket.

Research Implications

The assessment indicates that the K-Nearest Neighbor (KNN) model surpassed the Decision Tree C4.5 in terms of predictive accuracy. By implementing these two classification methods on the breast cancer dataset, it became clear that KNN is highly proficient at correctly categorizing the information. Consequently, the KNN approach can be considered a reliable and practical tool for distinguishing between malignant and benign breast cancer cases.

Evaluation and Validation

While a wide array of algorithms can be utilized to construct classification models, as noted in Chapter 2, this research focuses strictly on the K-Nearest Neighbor (KNN) and Decision Tree C4.5 methods. These two approaches were contrasted to identify the most accurate model, factoring in elements like interpretability, scalability, reliability, speed, and overall accuracy during the classification process. The primary objective of this study is to evaluate and directly compare the performance and accuracy of the KNN and Decision Tree C4.5 algorithms when analyzing breast cancer data.

CONCLUSION

An evaluation of the two algorithms demonstrated that the K-Nearest Neighbor (KNN) method reached an accuracy of 97.14%, outperforming the Decision Tree approach, which recorded an accuracy of 95.71%. Additionally, testing procedures involving ROC curves, Confusion Matrices, and Split Validation indicated that the KNN model secured an AUC score of 0.976, while the Decision Tree model achieved a 0.957 AUC score. Both of these scores successfully fall into the "Excellent Classification" tier, though the ROC curve analysis formally validates that the KNN method was superior. Ultimately, these

results prove that the KNN algorithm is highly capable of data classification, making it a suitable option for diagnosing breast cancer tumors as either malignant or benign.

ADVICE

Drawing from the study's findings, the author proposes the following recommendations:

1. Incorporate additional and more precise features to better distinguish between the different categories of breast cancer.
2. Future studies should test alternative machine learning algorithms and methods to discover even higher accuracy rates for breast cancer identification.
3. Expand the sample size and include more datasets in future research to further enhance the accuracy of the classification models.

REFERENCES

Madyaningrum and Sulastri, "Analysis of Breast Cancer Recurrence Prediction Using K-Nearest Neighbor," *ProceedingSINTAK 2019*, pp. 180–185, 2019.

Fahrurrozi and Wasilah, "Early Detection of Breast Cancer Using K-Nearest Neighbor (KNN) and Decision Tree C-45 Algorithms," *Teknika*, vol. 17, no. 2, pp. 427–434, 2023, [Online]. Available: https://jurnal.polsri.ac.id/index.php/teknika/article/view/7565

Farahdiba, D. Yusuf, and S. Nugroho, "Classification of Breast Cancer Using the Gain Ratio Algorithm."

Angkasa and J. J. Pangaribuan, "Information System Development Compares the Accuracy Level of Random Forest and Knn to Diagnose Breast Cancer," *J. Inf. Syst. Dev.*, vol. 7, no. 1, pp. 37–38, 2022, [Online]. Available: http://dx.doi.org/10.19166/xxxx

Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in *Communications in Computer and Information Science*, Springer, 2020, pp. 108–117. doi: 10.1007/978-981-15-7205-0_ 10.

Findawati, I. R. I. Astutik, A. S. Fitroni, I. Indrawati, and N. Yuniasih, "Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast," *J. Phys. Conf. Ser.*, vol. 1402, no. 6, 2019, doi: 10. 1088/1742-6596/1402/6/066046.

Derisma and F. Febrian, "Comparison of Neural Network Classification Techniques, Support Vector Machine, and Naive Bayes in Detecting Breast Cancer," *Bina Insa. Ict J.*, vol. 7, no. 1, p. 53, 2020, doi: 10.51211/biict.v7i1.1343.

Abdul Jabbar, E. Hasmin, C. Susanto, W. Musu, and I. Article, "Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification," *October*, vol. 14, no. 3, pp. 258–270, 2022, [Online]. Available: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

Hidayati, F. S. Rahmat Suwandi, D. Ediana, and F. Nursing and Public Health, Universitas Prima Nusantara Bukittinggi, West Sumatra, "The Experience of Patients Diagnosed with Lung Cancer for the First Time Reviewed from the Theory of the Five Stages of Grieving Article Information a B S T R a K," vol. 14, pp. 70–073, 2023, [Online]. Available: http://ejurnal.stikesprimanusantara.ac.id/

R. Wulandari, W. Wijayanti, E. Hapsari, D. Widyastutik, and S. Putri H, "Efforts to Improve Cadre Skills in Early Detection of Breast Cancer with Self-Breast Examination (SADARI) at Posyandu Tanggul Asri RW 10 Kadipiro Village, Banjarsari District, Surakarta," *J. Healthy greetings Masy.*, vol. 3, no. 2, pp. 47–52, 2022, doi: 10.22437/jssm.v3i2.18171.

Aini Silvi Astuti, Yunia Renny Andhikatias, "The Effectiveness of Conscious Health Education on the Level of Knowledge of Adolescent Girls About Early Detection of Breast Cancer in Tegalsari Dam," *Angew. Chemistry Int. Ed. 6(11), 951–952.*, vol. 2, 2019.

Destria, "Decision Support System for Companies that Excel in the Industrial Sector with the Weighted Product Method," *J. Ris. Sist. Inf. and Technology. Inf.*, vol. 3, no. 2, pp. 1–11, 2021, doi: 10.52005/jursistekni.v3i2.88.

Nawangsih, I. Melani, S. Fauziah, and A. I. Article, "Pelita Technology Predicts Employee Appointment Using the C5.0 Algorithm Method (Case Study of Pt. Mataram Cakra Buana Agung," *J. Pelita Teknol.*, vol. 16, no. 2, pp. 24–33, 2021.

Kirsten, *Prediction (of metaphor)*. 2021.

Polo and H. A. Miot, "Applications of the ROC curve in clinical and experimental studies," *J. Vasc. Arms.*, vol. 19, pp. 13–16, 2020, doi: 10. 1590/1677-5449.200186.

Erwansyah, "Implementation of Data Mining to Analyze the Relationship of Chemical Product Sales Data to Inventory of Goods Using FP (Frequent Pattern) Growth Algorithm at PT . Grand Multi Chemicals," *J. Teknol. Sist. Inf. and Sist. Computer. TGD (J-SISKO TECH), VOL*. 2, no. 2, pp. 30–40, 2019.

Fig. 1 and P. S. Hasugian2, "DATA MINING OF OFFICE INVENTORY ORDER LEVELS USING A PRIORI ALGORITHM IN THE NORTH SUMATRA REGIONAL POLICE," 2019.

Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analysis of Predicted Student Graduation on Time Using the Naïve Bayes Data Mining Method: Systematic Review," *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020, doi: 10.30998/faktorexacta.v13i1.5548

Nasrullah, "Implementation of Decision Tree Algorithm for Best-Selling Product Classification," *J. Ilm. Computer Science.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

Fadrial, "Naive Bayes Algorithm for Finding Student Estimated Time Students," *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 20–29, 2021.

Hermanto and A. Jaelani, "APPLICATION OF DATA MINING FOR PREDICTION OF NON-CASH FOOD ASSISTANCE RECIPIENTS (BPNT) IN WANACALA VILLAGE USING THE BAYES METHOD."

P. Putra, A. M. H. Pardede, and S. Syahputra, "Analysis of K-Nearest Neighbour (Knn) Method in Classification of Iris Bunga Data," *J. Tek. Inform. Kaputama*, vol. 6, no. 1, pp. 297–305, 2022, [Online]. Available: https://garuda.kemdikbud.go.id/documents/detail/2458300

Mustakim, R. Hastarimasuci, P. Papilo, Zarkasih, Olive, and A. Nazir, "Variable Selection to Determine Majors of Students using K-Nearest Neighbor and Naïve Bayes Classifier Algorithm," *J. Phys. Conf. Ser.*, vol. 1363, no. 1, 2019, doi: 10. 1088/1742-6596/1363/1/012057.

M. Reza Noviansyah, T. Rismawan, D. Marisa Midyanti, J. Computer Systems, and F. H. MIPA Tanjungpura University Jl Hadari Nawawi, "The Application of Data Mining Using the K-Nearest Neighbor Method for the Classification of Fire Weather Index Based on AWS (Automatic Weather Station) Data (Case Study: Kubu Raya Regency)," *J. Coding, Sist. Computer. Untan*, vol. 06, no. 2, pp. 48–56, 2018.

D. Setiawati, I. Taufik, J. Friday, and W. B. Zulfikar, "Classification of Quranic Verses Translation of Science Using Mobile-Based Decision Tree Algorithm," *J. Online Inform.*, vol. 1, no. 1, p. 24, 2016, doi: 10.15575/join.v1i1.7.

K. P. Decision, "Decision Tree 3 Reference," vol. 11, no. November, pp. 243–257, 2020.

R. Rustam, S. Rahmatullah, S. Supriyato, and S. Wahyuni, "The Application of Data Mining for the Prediction of Plywood Product Sales at Pt Puncak Menara Hijau Mas," *J. Inf. and Compost.*, vol. 8, no. 2, pp. 75–86, 2020, doi: 10.35959/jik.v8i2.186.

B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analysis of Netflix Data Mining Using the Rapid Miner Application," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021, doi: 10.30813/jbase.v4i1.2729.

S. James and C. Alley, "Working Paper Series by," vol. 55, no. 97, pp. 1023– 1038, 2007.

S. Haryati, A. Sudarsono, and E. Suryana, "The Implementation of Data Mining to Predict the Study Period of Students Using the C4.5 Algorithm (Case Study: Dehasen University Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.