International Journal of Applied Research and Sustainable Sciences (IJARSS)
Vol. 2 No. 12, 2024: 1039 - 1060

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

Ema Rosida[1], Andri Firmansyah[2], Suherman[3]
Universitas Pelita Bangsa
Corresponding Author: Ema Rosida emarosida71@gmail.com

A R T I C L E I N F O

Keywords: Breast Cancer, K-Nearest Neighbor (KNN), Decision Tree, Data Mining, ROC Curve, RapidMiner

Received : 18 October
Revised : 20 November
Accepted: 25 December

A B S T R A C T

Breast cancer is the leading cause of cancer-related deaths among women in Indonesia and worldwide. Early detection is critical for improving survival rates, yet many cases are diagnosed in late stages due to inadequate awareness and diagnostic tools. This study compares the performance of K-Nearest Neighbor (KNN) and Decision Tree algorithms for breast cancer classification using the Wisconsin Breast Cancer dataset. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework was applied, consisting of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment phases. The results indicate that KNN achieved the highest accuracy (97.14%) and Area Under the Curve (AUC) value (0.976), outperforming the Decision Tree algorithm (accuracy: 96.49%, AUC: 0.965). These findings highlight the potential of data mining techniques for enhancing early breast cancer detection and improving clinical decision-making.

1. INTRODUCTION

One of the most common types of cancer, breast cancer poses serious health dangers to people all over the world. Reducing mortality rates and improving treatment outcomes depend heavily on early detection and precise diagnosis. Machine learning (ML) has emerged as a valuable tool in medical diagnostics, enabling the classification of diseases based on patient data. However, selecting the most effective ML algorithm is vital for ensuring reliable predictions.

This study aims to compare the performance of two popular ML algorithms, K-Nearest Neighbor (KNN) and Decision Tree, in diagnosing breast cancer. Using the Wisconsin Breast Cancer dataset and RapidMiner software, the research seeks to identify the algorithm that provides superior accuracy in classifying malignant and benign cases. The findings can guide healthcare practitioners in adopting data-driven approaches for breast cancer detection.

2. LITERATURE REVIEW

2.1 Cancer

Cancer encompasses more than a hundred groups of diseases that start with uncontrolled cell growth. Normal cells cannot invade other tissues like these cancer cells. Cells turn cancerous when their DNA (deoxyribonucleic acid) changes [9].

Public health faces the problem of non-communicable diseases such as cancer worldwide. Cancer is a disease characterized by the presence of abnormal cells that can develop without control and invade particles moving between cells and body tissues. The World Health Organization (WHO) has released the Global Burden of Cancer (GLOBOCAN) data, which shows that the number of cases and deaths due to cancer reached 18.1 million cases and 9.6 million deaths in 2018 [10].

2.2 Breast Cancer

Breast cancer is a malignancy in breast tissue that can originate from the epithelium of the ducts or lobules. Breast cancer is one of the most common types of cancer in Indonesia [10].

Generally, the signs and symptoms of breast cancer are a lump in one or both breasts, tenderness or pain, discharge from the nipple even though it is not in a breastfeeding condition, there are skin abnormalities on the breast (dumpling, peau d'orange, redness, ulceration), enlarged lymph nodes or signs of distant metastasis [11]. Breast cancer can be prevented, treated and cured so that it has a high cure rate if known as early as possible signs and symptoms of cancer [10].

2.3 *Confusion Matrix*

A highly helpful tool for evaluating a classifier's ability to identify tuples from various classes is the confusion matrix. Values for accuracy, precision, and recall will be obtained by evaluation using the *confusion matrix* function. Confusion matrix is a matrix table consisting of two classes, namely classes that are considered positive and classes that are considered negative [12].

A method for assessing categorization models that determine whether an object is accurate or incorrect is the confusion matrix. Information about real values and classifications is contained in a matrix of classifications that will be compared to the original input class. [13].

2.4 ROC *Curve*

The receiver operating characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It depicts the trade-off between true positive rate (sensitivity) and false positive rate (specificity 1) for different threshold values. The ROC curve is a useful tool for evaluating the discriminative ability of a classification model and assessing its performance in distinguishing sick and non-diseased individuals [14].

Receiver Operating Characteristic Curve is a method used to analyze the performance of classification models and compare the effectiveness of different classification systems. ROC curves are also used to calculate the area under the curve (AUC) and interpret it. In addition, ROC curves can also be extended to evaluate multivariate models and ordinal classifications. It is important to accurately describe the parameters used when using ROC curves in research. ROC curves are commonly used in clinical and experimental studies [15].

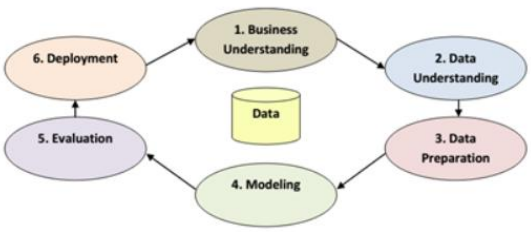
2.5 *Data Mining*

Data mining is a set of procedures used to unearth previously unknown information from databases in order to extract more value. By identifying and extracting significant or intriguing patterns from the database's data, the resulting information is produced. Data mining is frequently referred to as Knowledge Discovery Databases (KDD) because its primary purpose is to extract knowledge from massive databases. [16].

According to Han and Kamber (2011), data mining is the process of finding interesting patterns and knowledge from large amounts of data (Septiani, 2017). Meanwhile, according to Berry & Linoff (1997), data mining is a search and analysis of very large amounts of data and aims to find the meaning of patterns and rules (Chien, Wang, & Cheng, 2007). Then according to Connolly (2005). Data mining is a process of extracting or extracting previously unknown, but understandable and useful data from large databases and is used to make very important business decisions (Wulandari, Jatnika, & Purwanto, 2015). From several theories described by the experts above, an outline can be drawn. The process of searching and analyzing a database to identify an intriguing pattern in order to extract relevant and potentially helpful information and knowledge that can be comprehended and applied to decision-making is known as data mining [18].

2.6 Data Mining Stages Process

The Cross-Industry Standard Process for Data Mining, or CRISP-DM, is a standard technique for data mining research.. CRISP-DM is the result of



collaboration from several companies, including Daimler-Benz, OHRA, NCR Corp. and SPSS Inc. which began in 1999 [19].

Figure 1 CRISP-DM Cycle

2.7 Klasifikasi

To classify classes of items whose class labels are unknown, models or functions that define and differentiate between classes of data or concepts are sought after. [16].

Decision/classification trees, Bayesian/Naïve Bayes classifiers, neural networks, statistical analysis, genetic algorithms, rough sets, k-nearest neighbor, rule-based techniques, memory-based reasoning, and support vector machines (SVM) are examples of commonly used classification algorithms. [20].

2.8 *Split Validation*

RapidMiner has a layered operator called split validation. There are two subprocesses in split validation: the training subprocess and the testing subprocess. The training subprocess is used for learning or model building, and the made model is applied in the testing subprocess. The testing phase also measures the performance of the built model. [21].

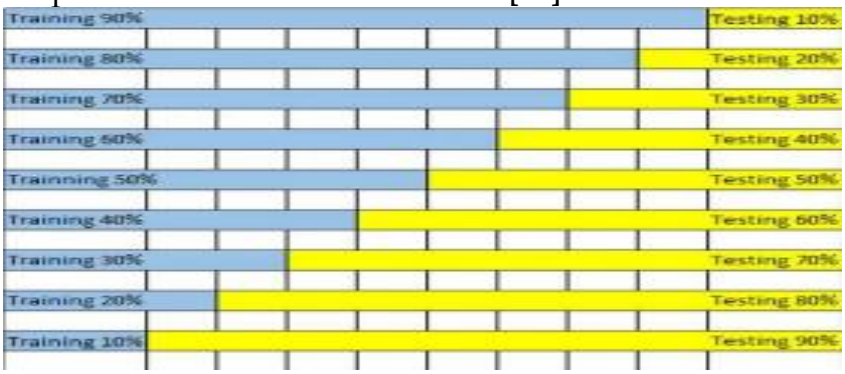


Figure 2 Illustration of Validation Table

2.9 *K-Nearest Neighbor*

KNN classification is a straightforward non-parametric classification technique. Even though the algorithm is straightforward, it performs exceptionally well and is a significant benchmarking technique. A metric and a positive integer are necessary for KNN classification. [22].

Finding the K groups of items in the training data that are nearest to the objects in the new data or test data is how KNN, a data clustering technique, determines categories based on the majority of categories in K-Nearest Neighbor [23]. [24].

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

2.10 *Decision Tree*

A *Decision tree* is a tree with a choice among several options displayed on each branch, and the chosen option displayed on each leaf [25]. Decision tree is commonly used to obtain information for the purpose of making a decision. Decision Tree is used to study the classification and classification of patterns

from data and describe the relationship of attribute variable x and target variable y in the form of a tree [19]. nodes, the selected features will distinguish a criterion compared to other criteria in the same node.

2.11 RapidMiner

RapidMiner is software created by Dr. Markus Hofmann from the Institute of Technology Blanchardstown and Ralf Klinkenberg from rapid-i.com with a GUI (Graphical User Interface) display making it easier for users to use this software. This software is open source and created using the Java program under the GNU. RapidMiner is specialized for the use of data mining. The models provided are also quite numerous and complete, such as Bayesian Models, Modeling, Tree Induction, Neural Networks and others. Many methods are provided by RapidMiner ranging from classification, clustering, association and others. If there is no model or algorithm model that does not exist in weka, users can add other modules, because weka is open source, so anyone can participate in developing this software [28].

3. METHODOLOGY

The study used the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which consists of six steps: Deployment, Modeling, Evaluation, Data Preparation, Business Understanding, and Data Understanding.

1. **Dataset**: The Wisconsin Breast Cancer dataset was obtained from the UCI Machine Learning Repository. It contains 699 instances with attributes describing cell nucleus features, including clump thickness, uniformity of cell size, and mitosis.
2. **Data Preprocessing**:
    - o The mean of the corresponding characteristics was used to fill in the missing values.
    - o To improve model performance, attribute ranges were standardized through data normalization.
3. **Algorithms**:
    - **K-Nearest Neighbor (KNN)**: A distance-based algorithm where classification depends on the majority class among k-nearest data points.
    - **Decision Tree**: A rule-based algorithm that builds a tree-like structure to make decisions based on feature splits.
4. **Modeling and Evaluation**:
    - RapidMiner was used to train and evaluate both algorithms.
    - The dataset was split into training (70%) and testing (30%) subsets.
    - Accuracy was the primary metric for performance evaluation.

4. RESULTS AND RESEARCH

4.1 *Decision Tree*

Training data is to determine whether a person has Benign or Malignant breast cancer. The following will discuss the prediction of whether a person has Benign or Malignant breast cancer, using the classification method.

Calculate the Entropy value. It is known that there are 699 cases that belong to the Benign class, 458 records and 241 records.

$$\text{Entropy (S)} = \sum_{i=1}^{n} -\text{pi} * \log_2 \text{pi}$$

$$= \left(-\frac{458}{699} * log_2\left(\frac{458}{699}\right)\right) + (-\frac{241}{699} * log_2\left(\frac{241}{699}\right))$$

$$= 0{,}929$$

Then calculate the Entropy value for each attribute, for example below calculate the Entropy value for the Clump Thickness attribute >5 with the number of patients 513 with benign class 437 and malignant class 76:

$$E_{Clump\ Thickness}(\leq 5) = \left(-\frac{437}{513} * log_2\left(\frac{437}{513}\right)\right) + (-\frac{76}{513} * log_2\left(\frac{76}{513}\right))$$

$$= 0,605$$

Clump Thickness <5 with the number of patients 186 with benign class 21 and malignant class 165:

$$E_{Clump\ Thickness}(> 5) = \left(-\frac{21}{186} * log_2\left(\frac{21}{186}\right)\right) + (-\frac{165}{186} * log_2\left(\frac{165}{186}\right))$$

$$= 0{,}509$$

After obtaining the entropy value of each attribute, for example, the entropy reduction value for the Clump Thickness attribute >5 with the number of patients is 513 with a benign class of 437 and a malignant class of 165.

$$I_{res\ Clump\ Thickness} = \sum p(v)I(v)$$

$$= \frac{513}{699} * 0{,}605 + \frac{437}{699} * 0{,}509$$

$$= 0{,}579$$

Then calculate the Gain for the above attributes as follows:

$$Gain_{Clump\ Thickness} = -\left((\frac{513}{699}) * 0{,}605\right) + \left((\frac{186}{699}) * 0{,}509\right)$$

$$= 0{,}350$$

After getting the Gain on each attribute, the highest attribute Gain result will be selected. So the highest attribute gain is Bare Nuclei.

Table 4. 1 Calculation of Entropy and Gain Values

Node	Case	*Benign* (i)	*Malignant* (j)	*Entropy*	*Entropy Res* (A)	*Gain* (A)

	699	**458**	**241**	**0,929**		
Clump Thickness						
<=5	513	437	76	0,605	0,579	0,350
>5	186	21	165	0,509		
Uniformity of Cell Size						
<=5	551	453	98	0,675	0,577	0,352
>5	148	5	143	0,213		
Uniformity of Cell Shape						
<=5	546	407	139	0,818	0,692	0,238
>5	153	6	147	0,239		
Marginal Adhesion						
<=5	579	452	127	0,759	0,678	0,251
>5	120	6	114	0,286		
Single Epithelial Cell Size						
<=5	592	450	142	0,795	0,690	0,240
>5	41	2	39	0,281		
Bare Nuclei						
<=5	**525**	**452**	**73**	**0,582**	**0,491**	**0,439**
>5	**174**	**6**	**168**	**0,216**		
Bland Chromatin						
<=5	557	450	107	0,706	0,626	0,303
>5	142	8	134	0,313		
Normal Nucleoli						
<=5	560	447	113	0,725	0,661	0,269
>5	139	11	128	0,399		
Mitoses						
<=5	665	456	209	0,898	0,870	0,059
>5	34	2	32	0,323		

Then calculate the Information Gain Ratio for the attributes above, for example below calculate the Information Gain Ratio as follows:

$$Gain\ Ratio_{Clump\ Thickness} = \frac{I - I_{res}(A)}{I(A)}$$
$$= \frac{0,929-0,579}{\sum_v p(V) log_2(p(v))}$$

$$= \frac{0{,}929 - 0{,}579}{0{,}836}$$
$$= 0{,}419$$

Here are the results of the Gain Ratio calculation of all attributes.

Table 4. 2 Calculation of Gain Ratio

Node	*Gain Ratio* (A)		Knot	*Gain Ratio* (A)
Clump Thickness			*Bare Nuclei*	
<=5	0,419		<=5	0,893
>5			>5	
Uniformity of Cell Size			*Bland Chromatin*	
<=5	0,609		<=5	0,485
>5			>5	
Uniformity of Cell Shape			*Normal Nucleoli*	
<=5	0,344		<=5	0,407
>5			>5	
Marginal Adhesion			*Mitoses*	
<=5	0,371		<=5	0,068
>5			>5	
Single Epithelial Cell Size				
<=5	0,348			
>5				

Then calculate Gini for the attributes above, for example below calculate Gini as follows:

$$Gini = \sum_{i \neq j} p(i)p(j)$$
$$= \frac{458}{699} x \frac{241}{699}$$
$$= 0{,}226$$

After calculating Gini, then the calculation of Gini Attributes, for example below calculates Gini Attributes as follows:

$$Gini_{Clump\ Thickness} = \sum_{v} p(v) \sum_{i \neq j} p\,(i|v)p(j|v)$$
$$= \frac{513}{699} x \left(\frac{437}{513} x \frac{76}{513}\right) + \frac{186}{699} x \left(\frac{21}{186} x \frac{165}{186}\right)$$
$$= 0{,}119$$

The last is to calculate the Gini Gain, as an example below calculate the Gini Gain as follows:

$$Gini\ Gain_{Clump\ Thickness} = Gini - Gini\ (A)$$
$$= 0{,}226 - 0{,}119$$
$$= 0{,}107$$

Here are the results of the calculation of Gini Gain for all attributes.

Table 4. 3 Calculation of Gain Value to Determine the Root Node

Node	*Gain* (A)	*Gain Rasio* (A)	*Gini Gain* (A)
Clump Thickness <=5 >5	0,350	0,419	0,107
Uniformity of Cell Size <=5 >5	0,352	0,609	0,104
Uniformity of Cell Shape <=5 >5	0,238	0,344	0,069
Marginal Adhesion <=5 >5	0,251	0,371	0,076
Single Epithelial Cell Size <=5 >5	0,240	0,348	0,069
Bare Nuclei <=5 >5	**0,439**	**0,893**	**0,128**
Bland Chromatin <=5 >5	0,303	0,485	0,091
Normal Nucleoli <=5 >5	0,269	0,407	0,082
Mitoses <=5 >5	0,059	0,068	0,018

From the results of the Entropy calculation in table 4.2 and the gain calculation in table 4.3, it can be seen that the attributes of Bare Nuclei have the highest gain value, namely gain 0.439, gain ratio 0.893, Gini gain 0.128. Therefore, Bare Nuclei is the root node in the decision tree. To determine the

next node, i.e. node 1.1, entropy and gain calculations are carried out again based on the attributes of the Bare Nuclei. The number of cases counted is the number of cases with the value of the root node (Bare Nuclei).

4.2 *K-Nearest Neighbor* (KNN)

The implementation of the KNN algorithm at this stage is an example of manual calculations starting from determining training data and testing data, implementing the KNN algorithm. The following 72 sample data that will be applied for manual calculations are shown in Table 4.4 and 4.5, showing 72 sample data that will be applied for manual calculations, 70 training data and 2 testing data.

Table 4. 4 *Training Data*

ID	*Clump Thickness*	*Uniformity of Cell Size*	*Uniformity of Cell Shape*	*Marginal Adhesion*	*Single Epithelial Cell Size*	*Bare Nuclei*	*Bland Chromatin*	*Normal Nucleoli*	*Mitoses*	*Class*
1000025	5	1	1	1	2	1	3	1	1	Benign
1002945	5	4	4	5	7	10	3	2	1	Beningn
1015425	3	1	1	1	2	2	3	1	1	Beningn
1016277	6	8	8	1	3	4	3	7	1	Beningn
1017023	4	1	1	3	2	1	3	1	1	Beningn
1017122	8	10	10	8	7	10	9	7	1	Malignant
1018099	1	1	1	1	2	10	3	1	1	Beningn
1018561	2	1	2	1	2	1	3	1	1	Beningn
1033078	2	1	1	1	2	1	1	1	5	Beningn

Table 4. 5 Data Testing

Clump Thickness	*Uniformity of Cell Size*	*Uniformity of Cell Shape*	*Marginal Adhesion*	*Single Epithelial Cell Size*	*Bare Nuclei*	*Bland Chromatin*	*Normal Nucleoli*	*Mitoses*	*Class*
5	10	8	10	8	10	3	6	3	?
3	1	1	1	2	1	2	2	1	?

Table 4.4 shows the dataset used where 70 data is used as training data 2 as testing data. After determining the testing data, the next stage is the process of implementing the KNN method using equation 1.

Tables 4.6 and 4.7 show the results of the first testing data calculation. Where every one testing data will be calculated for all training data. The following calculation is the outcome of comparing each attribute to the training data attribute using the Manhattan KNN formula. The results are then sorted based on a predefined number K, which indicates which class of results are the most dominant. Matching TP, TN, FP, and FN on the Confusion matrix serves as a benchmark for evaluating performance.

Table 4.6 Data Testing Attributes 1

Clump Thickness	*Uniformity of Cell Size*	*Uniformity of Cell Shape*	*Marginal Adhesion*	*Single Epithelial Cell Size*	*Bare Nuclei*	*Bland Chromatin*	*Normal Nucleoli*	*Mitoses*	*Class*
5	10	8	10	8	10	3	6	3	?

Table 4. 7 *Data Testing*

Num.	Distance Calculation Results	Nearest Order	Actual Label
1	5,916079783	1	*Malignant*
2	7,681145748	2	*Malignant*
3	8,717797887	3	*Malignant*
4	9,327379053	4	*Malignant*
5	9,695359715	5	*Malignant*
6	9,695359715	6	*Malignant*
7	9,848857802	7	*Malignant*
8	9,899494937	8	*Benign*
9	9,899494937	9	*Malignant*
10	10,34408043	10	*Malignant*

Tables 4.8 and 4.9 show the results of the second test data calculation. Where each data testing will be calculated all training data. The following calculation shows the outcome of comparing each attribute with the training data attribute using the Eucludien KNN formula. The results are then sorted by the number of K that has been established, where the sequence's most dominant class is identified. The reference for performance measurement is derived from this result by matching TP, TN, FP, and FN on the Confusion matrix.

Table 4.8 *Data Testing Attributes 2*

Clump Thickness	*Uniformity of Cell Size*	*Uniformity of Cell Shape*	*Marginal Adhesion*	*Single Epithelial Cell Size*	*Bare Nuclei*	*Bland Chromatin*	*Normal Nucleoli*	*Mitoses*	*Class*
3	1	1	1	2	1	2	2	1	?

Tabel 4.9 Perhitungan Data *Testing* 2

No.	Hasil Perhitungan Jarak	Urutan Terdekat	Label Aktual
1	1	1	*Benign*
2	1,414214	2	*Benign*
3	1,414214	3	*Benign*
4	1,414214	4	*Benign*
5	1,414214	5	*Benign*
6	1,414214	6	*Benign*
7	1,414214	7	*Benign*
8	1,732051	8	*Malignant*
9	1,732051	9	*Benign*
10	1,732051	10	*Benign*

4.3 Evaluation and Validation Results

As mentioned in chapter 2, many methods can be used to create a classification model. In this paper, for example, the methods used are Decision Tree and K-Nearest Neighbor algorithms. Then a comparison is made between the two and measure which method is the most accurate. Classification methods can be evaluated such as accuracy, speed, reliability, scalability and interpretability.

This research aims to see the accuracy of breast cancer analysis using the Decision Tree and K-Nearest Neighbor algorithms. Then analyze the accuracy by comparing the two algorithms.

4.4 Model Testing

After manual calculations, the next process to classify breast cancer in this test uses the RapidMiner tool.

4.4.1 *Decision Tree*

Using the RapidMiner tool, the decision tree procedure is implemented in order to conduct testing in this phase. In the RapidMiner structured release process, the decision tree algorithm model is displayed along with the operator input, Excel input, decision tree operator input, connection of all operators, and Run button clicks. It can be seen in Figure 4.1.

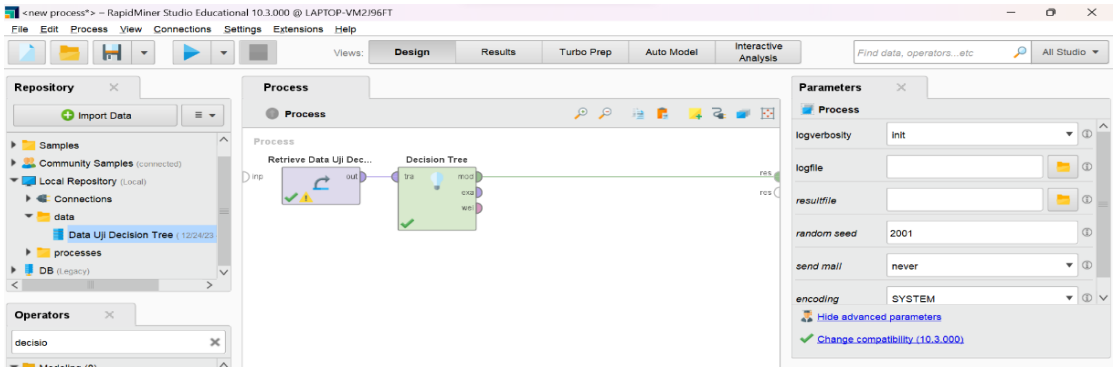


Figure 4.1 Initial Testing Stage of the Decision Tree Algorithm

Thus producing a Decision Tree Algorithm model. .

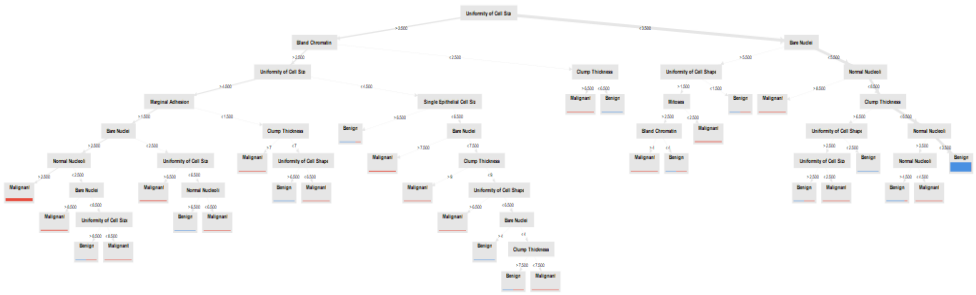


Figure 4. 2 *Decision Tree* Algorithm Model

Furthermore, to get acuracy, precision, and recall values by adding operators, apply models, and performance can be seen in the image.

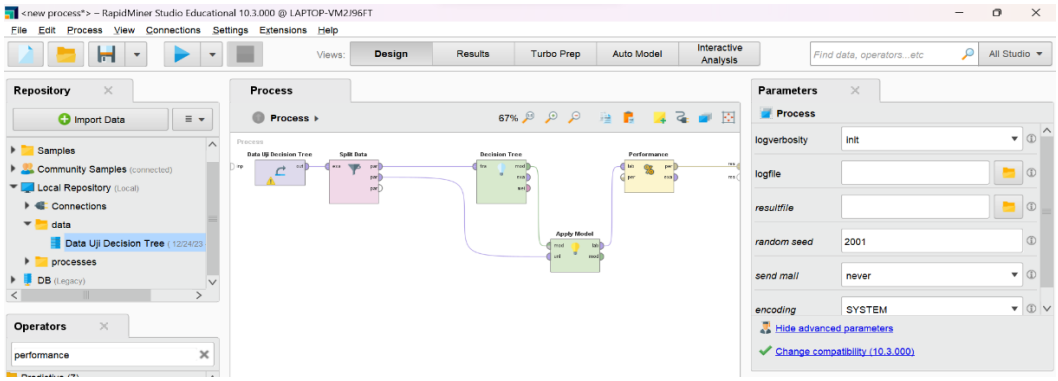


Figure 4.3 Final Testing Stage of *Decision Tree* Algorithm

At this stage, the results are assessed using performance tools shown to display the confusion table, which is used to display the results of accuracy, precision and recall.

The purpose of this conversation was to determine the accuracy or precision values. Precision is the degree of similarity between the predicted and actual results, while accuracy is the degree of accuracy between the user's requested information and the system's response. The system's success rate in retrieving data is referred to as recall.

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 *Decision Tree* Algorithm

A calculation based on a dataset split by split validation, which yields 90% training data and 10% testing data, is shown in Figure 4.3. Of the 70 testing data, 44 are classified as benign based on the Decision Tree method's predictions, 1 is predicted to be benign but turns out to be malignant, 23 malignant data is predicted to be appropriate, and 2 is predicted to be malignant.

ROC Curve Decision Tree

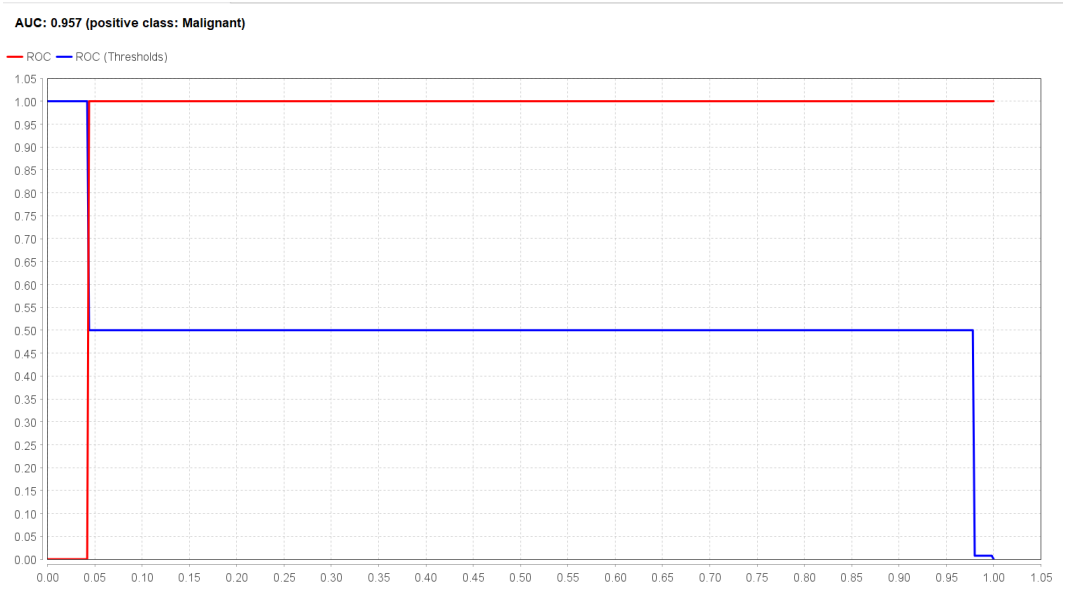


Figure 4. 1 ROC Decision Tree Algorithm

Figure 4.4's confusion matrix is expressed by the ROC curve in the above figure. Vertical lines indicate true positives, while horizontal lines indicate false positives.

To make it easier to read breast cancer data, it is necessary to enter performance tools to find Root Mean Squared Error and Squared Error. Here are the results in Figure 4.6

PerformanceVector

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

Figure 4. 6 Performance Vector Results

Figure 4. 6 Root Mean Square Error and Square Error Test Results. The second step is to implement a linear regression algorithm using rapidminer tools. Here are the stages in the application of the Decision Tree algorithm:

1. Determines the classification of test data performed by rapidminer and produces a classified confidence value.
2. Determine performance with output to find Root Mean Squared Error and Squared Error.

The two components of split validation modeling are the training phase, which is used for classification algorithms, and the testing phase, which uses the Performance feature to show the Root Mean Squared Error and Squared Error and the Apply Model feature to apply the model to the testing data.

4.3.2. *K-Nearest Neighbor* (KNN)

At this point, testing is done by using the RapidMiner tool to implement the K-Nearest Neighbor (KNN) procedure. The K-Nearest Neighbor (KNN) Algorithm Model is displayed by RapidMiner's structured sharing process, which involves reading Excel input, entering the KNN operator, connecting all operators, and then clicking the run button.

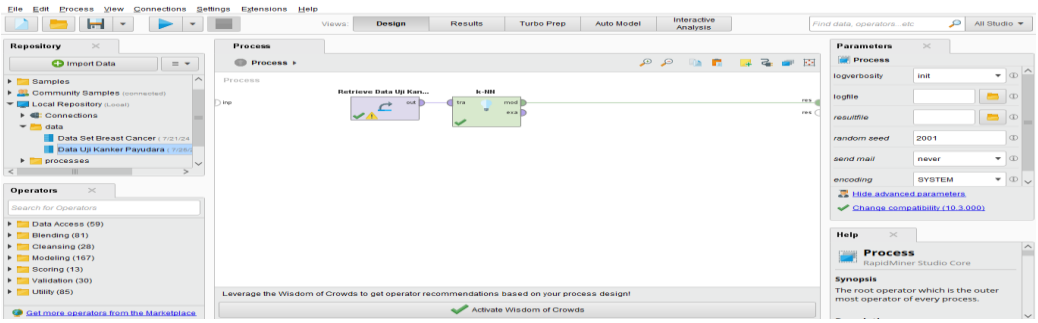


Figure 4.7 Initial Testing Stage of the K-Nearest Neighbor (KNN) Algorithm

Thus producing the K-Nearest Neighbor (KNN) Algorithm model. It can be seen in Figure 4.8

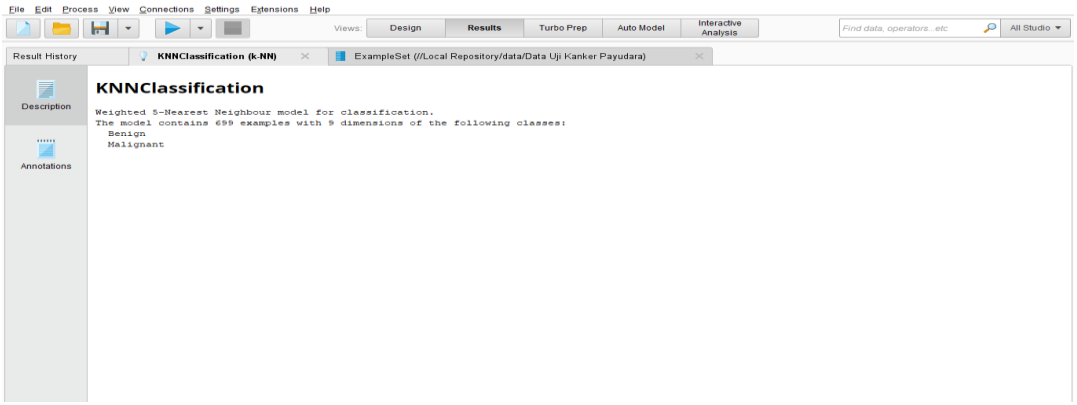


Figure 4.8 K-Nearest Neighbor Algorithm Model

Furthermore, to get the values of accuracy, precision, and recall by adding an operator, apply the model, and performance can be seen in the image.

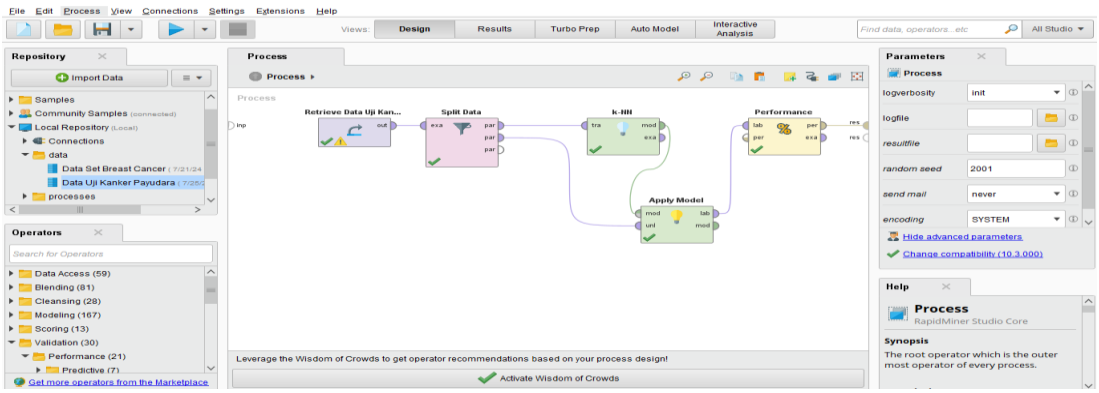


Figure 4.9 Final Testing Stage of the K-Nearest Neighbor (KNN) Algorithm

At this point, performance tools are used to evaluate the results in order to display the confusion table, which shows the accuracy, precision, and recall results.

This discussion was conducted in order to determine the accuracy and precision values. Precision refers to how close the prediction result is to the actual result, while accuracy is the degree of correspondence between the user's requested information and the system's response. The system's recall rate is its ability to retrieve information.

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 K-Nearest Neighbor (KNN) Algorithm

Figure 4.9 displays a computation based on a dataset split by split validation, which produces 90% of the training data and 10% of the testing data. Based on the C4.5 method's predictions, 44 of the 70 testing data are categorized

as benign. One data set is expected to be benign but turns out to be malignant, 23 malignant data sets are predicted to be appropriate, and two data sets are predicted to be malignant.

Kurva ROC *K-Nearest Neighbor* (KNN)

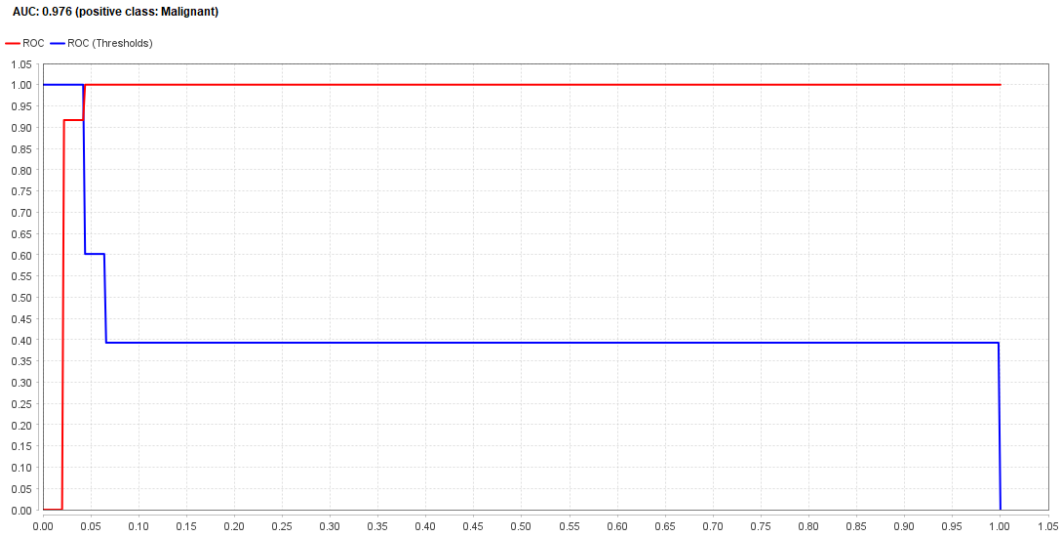


Figure 4.11 K-Nearest Neighbor (KNN) ROC Curve

The confusion matrix from graphic 4.9 is expressed by the ROC curve in the preceding graphic. True positives are shown by vertical lines, whereas false positives are represented by horizontal lines.

To make it easier to read breast cancer data, it is necessary to input performance tools to look for Root Mean Squared Error and Squared Error. Here are the results in the picture.

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 Root Mean Square Error and Square Error Test Results

The second step is to implement the linear regression algorithm using the Rapidminer tool. The following are the steps in implementing the linear regression algorithm:

1. Determine the classification of test data carried out by rapidminer and produce a confidence value that has been classified.
2. Determine performance with output to find Root Mean Squared Error and Squared Error.

In split validation modeling, there are two parts, namely the training part (used for classification algorithms) and the testing part (using the Apply Model feature to apply the model to the testing data and the Performance feature to display the Root Mean Squared Error and Squared Error)..

4.5 Result Analysis

The model produced by the algorithm was tested using the split validation method, and the comparison of accuracy, precision, recall, and ROC curves was seen with the highest.

Table 4.10 Accuracy and AUC Values in Decision Tree and KNN

	Decision Tree	*K-Neasrest Neighbor* (KNN)
Accuracy	95,71%	97,14%
AUC	0.957	0.976

Table 4.10 compares the Accuracy and AUC of each algorithm. It can be seen that the Accuracy value of K-Neasrest Neighbor (KNN) is the highest as well as the AUC value. For the Decision Tree algorithm also shows the corresponding value. Based on the above grouping and Table 4.10, it can be concluded that the Decision Tree and K-Nearest Neighbor (KNN) algorithms are classified very well because they have an AUC value between 0.90-1.00.

DISCUSSION

The findings underscore the importance of algorithm selection in medical diagnostics. KNN's higher accuracy and reliability can be attributed to its ability to capture local patterns in the dataset, making it well-suited for classification tasks involving complex, non-linear relationships. However, KNN's performance is highly dependent on the choice of k and computational efficiency, which could be limitations in larger datasets.

In contrast, the Decision Tree algorithm, while interpretable and easy to implement, showed slightly lower accuracy and reliability. This might be due to overfitting tendencies or its reliance on discrete feature splits, which may not capture subtle patterns in the data.

Future research could explore ensemble methods, such as Random Forest or Gradient Boosting, to enhance classification accuracy further. Additionally, incorporating feature selection techniques may improve model efficiency and reduce computational demands.

CONCLUSION

This study compared the performance of K-Nearest Neighbor and Decision Tree algorithms in classifying breast cancer cases using the Wisconsin Breast Cancer dataset. The results demonstrate that KNN achieves superior accuracy (96.78%) compared to the Decision Tree (93.48%), making it a more effective choice for breast cancer diagnosis. These findings highlight the role of machine learning in advancing medical diagnostics and emphasize the need for careful algorithm selection to ensure optimal results.

RECOMMENDATION

Based on the research conducted, the author can provide the following suggestions:

1. Maximize or add more specific and more attributes in determining the classification of breast cancer.
2. Further research is needed in testing with other methods or algorithms in order to obtain a comparison with the best level of accuracy in identifying breast cancer data.
3. Adding the number of datasets or samples in the study in order to get a better accuracy value. 3.

ACKNOWLEDGMENTS

The authors thank the UCI Machine Learning Repository for providing the Wisconsin Breast Cancer dataset and RapidMiner for its robust data analysis platform.

REFERENCE

[1] N. A. Madyaningrum and Sulastri, "Analisa Prediksi Kekambuhan Kanker Payudara Dengan Menggunakan K-Nearest Neighbor," *Proceeding SINTAK 2019*, pp. 180–185, 2019.

[2] Fahrurrozi and Wasilah, "Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor (KNN) Dan Decision Tree C-45," *Teknika*, vol. 17, no. 2, pp. 427–434, 2023, [Online]. Available: https://jurnal.polsri.ac.id/index.php/teknika/article/view/7565

[3] B. A. Farahdiba, D. Yusuf, and S. Nugroho, "Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio."

[4] V. Angkasa and J. J. Pangaribuan, "Information System Development Komparasi Tingkat Akurasi Random Forest Dan Knn Untuk Mendiagnosis Penyakit Kanker Payudara," *J. Inf. Syst. Dev.*, vol. 7, no. 1, pp. 37–38, 2022, [Online]. Available: http://dx.doi.org/10.19166/xxxx

[5] S. A. 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.

[6] Y. 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.

[7] D. Derisma and F. Febrian, "Perbandingan Teknik Klasifikasi Neural Network, Support Vector Machine, dan Naive Bayes dalam Mendeteksi Kanker Payudara," *Bina Insa. Ict J.*, vol. 7, no. 1, p. 53, 2020, doi: 10.51211/biict.v7i1.1343.

[8] M. Abdul Jabbar, E. Hasmin, C. Susanto, W. Musu, and I. Artikel, "Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification," *Oktober*, vol. 14, no. 3, pp. 258–270, 2022, [Online]. Available: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

[9] Hidayati, F. S. Rahmat Suwandi, D. Ediana, and F. Keperawatan dan Kesehatan Masyarakat Universitas Prima Nusantara Bukittinggi Sumatera Barat, "Pengalaman Pasien Pertama Kali Terdiagnosis Kanker Paru Ditinjau Dari Teori 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/

[10] R. Wulandari, W. Wijayanti, E. Hapsari, D. Widyastutik, and S. Putri H, "Upaya Peningkatan Ketrampilan Kader dalam Deteksi Dini kanker

Payudara dengan Pemeriksaan Payudara Sendiri (SADARI) di Posyandu Tanggul Asri RW 10 Kelurahan Kadipiro Kecamatan Banjarsari Surakarta," *J. Salam Sehat Masy.*, vol. 3, no. 2, pp. 47–52, 2022, doi: 10.22437/jssm.v3i2.18171.

[11] D. R. Aini Silvi Astuti, Yunia Renny Andhikatias, "Efektivitas Pendidikan Kesehatan Sadari Terhadap Tingkat Pengetahuan Remaja Putri Tentang Deteksi Dini Kanker Payudara Di Tegalsari Bendungan," *Angew. Chemie Int. Ed. 6(11), 951–952.*, vol. 2, 2019.

[12] N. Destria, "Sistem Pendukung Keputusan Perusahaan yang Berprestasi dalam Sektor Indutri dengan Metode Weighted Product," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 3, no. 2, pp. 1–11, 2021, doi: 10.52005/jursistekni.v3i2.88.

[13] I. Nawangsih, I. Melani, S. Fauziah, and A. I. Artikel, "Pelita Teknologi Prediksi Pengangkatan Karyawan Dengan Metode Algoritma C5.0 (Studi Kasus Pt. Mataram Cakra Buana Agung," *J. Pelita Teknol.*, vol. 16, no. 2, pp. 24–33, 2021.

[14] F. Kirsten, *Prediction (of metaphor).* 2021.

[15] T. C. F. Polo and H. A. Miot, "Aplicações da curva ROC em estudos clínicos e experimentais," *J. Vasc. Bras.*, vol. 19, pp. 13–16, 2020, doi: 10.1590/1677-5449.200186.

[16] K. Erwansyah, "Implementasi Data Mining Untuk Menganalisa Hubungan Data Penjualan Produk Bahan Kimia Terhadap Persedian Stok Barang Menggunakan Algoritma FP (Frequent Pattern) Growth Pada PT . Grand Multi Chemicals," *J. Teknol. Sist. Inf. dan Sist. Komput. TGD (J-SISKO TECH)*, vol. 2, no. 2, pp. 30–40, 2019.

[17] E. Manurung[1] and P. S. Hasugian[2], "DATA MINING TINGKAT PESANAN INVENTARIS KANTOR MENGGUNAKAN ALGORITMA APRIORI PADA KEPOLISIAN DAERAH SUMATERA UTARA," 2019.

[18] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[19] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[20] Y. E. Fadrial, "Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa Naive Bayes Algorithm for Finding Student Estimated Time Students," *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 20–29, 2021.

[21] B. Hermanto and A. Jaelani, "PENERAPAN DATA MINING UNTUK PREDIKSI PENERIMA BANTUAN PANGAN NON TUNAI (BPNT) DI

DESA WANACALA MENGGUNAKAN METODE NAÏVE BAYES."

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

[23] Mustakim, R. Hastarimasuci, P. Papilo, Zarkasih, Zaitun, and A. Nazir, "Variable Selection to Determine Majors of Student 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.

[24] M. Reza Noviansyah, T. Rismawan, D. Marisa Midyanti, J. Sistem Komputer, and F. H. MIPA Universitas Tanjungpura Jl Hadari Nawawi, "Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya)," *J. Coding, Sist. Komput. Untan*, vol. 06, no. 2, pp. 48–56, 2018.

[25] D. Setiawati, I. Taufik, J. Jumadi, and W. B. Zulfikar, "Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile," *J. Online Inform.*, vol. 1, no. 1, p. 24, 2016, doi: 10.15575/join.v1i1.7.

[26] K. P. Keputusan, "Rujukan Decision Tree 3," vol. 11, no. November, pp. 243–257, 2020.

[27] R. Rustam, S. Rahmatullah, S. Supriyato, and S. Wahyuni, "Penerapan Data Mining Untuk Prediksi Penjualan Produk Triplek Pada Pt Puncak Menara Hijau Mas," *J. Inf. dan Komput.*, vol. 8, no. 2, pp. 75–86, 2020, doi: 10.35959/jik.v8i2.186.

[28] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021, doi: 10.30813/jbase.v4i1.2729.

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

[30] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.