Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 2, 2024, pp. 260 – 270

Analysis of K-Nearest Neighbor (KNN), Naive Bayes ands Decision Tree C4.5 Algorithm With Classification Method In Breast Cancer Using RapidMiner

Iqbal Muhammad*[1], Maulana Donny[2], Hadikristanto Wahyu[3], Kurniadi Nanang Tedi[4], Amali[5], Nawangsih Ismasari[6]

Department of Informatic Engineering, University Pelita Bangsa,
Jl. Inpeksi Kalimalang, Cikarang Pusat, Bekasi, Jawa Barat, 17530, Indonesia
Corresponding author: mhmmdiqbal@mhs.pelitabangsa.ac.id

Abstract - Breast cancer is cancer that forms in the cells of the breast. It is the most common cancer in women and the leading cause of cancer deaths in women worldwide. Breast cancer is usually divided into two types: benign, or usually called benign and malignant, or usually called malignant. Benign cancers are usually characterized by small, round, tender lumps. In the fields of medicine, finance, marketing, and social science, data mining is a popular tool for performing proven analysis. This study will compare K-Nearest Neighbor (KNN), Naive Bayes, and Decision Tree C4.5 approaches for classifying breast cancer. The problem of this research is which algorithm has a high level of accuracy that can be used with breast cancer datasets and can provide information about patterns or models for early detection of breast cancer. The results of the research conducted using CRISP-DM show that K-Nearest Neighbor (KNN) has the highest accuracy value with 97.14% and its AUC value is 0.976. The AUC value also showed excellent classification, with an AUC value between 0.90 and 1.00.

Keywords - Breast cancer, K-Nearest Neighbor, Decision Tree C4.5, Naive Bayes, ROC Curve, Matrix Confusion

1. INTRODUCTION

Breast cancer is cancer that forms in the cells of the breast. Breast cancer is the most common cancer in women and the leading cause of cancer deaths among women worldwide [1]. Globocan data in 2020 the number of world cancer cases reached a total of 19,292,789 cases, for breast cancer cases amounted to 2,261,419 (11.7%) Breast cancer is a tumor with a high incidence that threatens women's health seriously. For cases of death caused by cancer disease totaled 9,958,133, while cases of death caused by breast cancer amounted to 684,996 accounted for (6.9%) of cases of death caused by cancer disease [2].

Breast cancer is generally divided into 2, namely benign or commonly called benign and malignant or commonly called malignant, usually benign breast cancer is characterized by a small round, soft lump. Breast cancer in benign levels will usually have conditions and growths that are not cancerous. This cancer can be detected but will not spread and damage nearby tissues. Malignant breast cancer is characterized by a shape that is asymmetrical, rough, painful, and others [2].

Data mining methods are applied but need to be adjusted to the purpose of their use. Examples of data mining methods are K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree C4.5 and others. Decision Tree C45, K-Nearest Neighbor (KNN) and Naïve Bayes Classifier algorithms are algorithms used for comparison of accuracy and error values. Accuracy is the accuracy of the correct measurement value over the total number of samples considered. Error is the scale of incorrect trial values over the total number of samples considered [3].

In this paper, three algorithms including K-Nearest Neighbor, Decision Tree C4.5 and Naive Bayes algorithm are combined to present an efficient predictor model for diagnosing the type of breast cancer. In the proposed model, raw data is loaded first and then they are pre-processed. Next, all data is divided into two training set and test set. Training data set is given to all the three algorithms in parallel so that three independent predictor models are created. Then, the test dataset is given to each model and the results are combined using the voting approach to obtain the final result. The rest of this paper is organized as follows. Section II presents related work, breast cancer dataset, and the proposed model. Section III presents the simulation results. Finally, the paper is concluded in Section IV.

2. RESEARCH METHOD

In this section, some existing works are studied first. Then, the breast cancer dataset used in this study is introduced. Finally, the proposed model is presented.

2.1 Related Work

In [1] Comparison of K-Nearest Neighbor (KNN) and Gaussian Naive Bayes (GNB) Algorithms in Breast Cancer Coimbra Classification. the KNN algorithm obtained the greatest results with an accuracy of 86.9%, precision of 87.3%, and recall of 86.7% on the 5th trial. The GNB algorithm gets the biggest results with an accuracy of 78.2%, precision of 80.4%, and recall of 77.6% on the 5th try.

In [4] Comparison of Neural Network, Support Vector Machine, and Naive Bayes Classification Techniques in Detecting Breast Cancer. Breast cancer is a type of cancer that is often found by most women. In Indonesia, breast cancer ranks first in hospitalized patients in all hospitals. The purpose of this study is to conduct a computation-based diagnosis of breast cancer that can produce how a person's cancer condition is based on the accuracy of the algorithm. This research uses orange python programming and Wisconsin Breast Cancer dataset for modeling breast cancer classification. The data mining methods applied are Neural Network, Support Vector Machine, and Naive Bayes. In this study, the best classification algorithm was obtained, namely the Kernel SVM algorithm with an accuracy rate of 98.9% and the lowest algorithm was Naive Bayes worth 96.1% [4].

In [2] Early Detection of Breast Cancer Using K-Nearest Neighbor (KNN) Algorithm and Decision Tree C4.5. Breast cancer is a type of cancer that generally forms in breast cells and the cancer cells grow out of control. Breast cancer can occur in all genders. In Indonesia, the number of breast cancer cases ranks first compared to other types of cancer and is one of the first contributors to death. Based on the number of deaths and considering that breast cancer does not look at gender, both men and women should be aware of their health by taking actions such as early detection and avoiding risks that cause cancer. The data used in this research comes from https://www.kaggle.com/datasets/. The purpose of this research is to utilize several existing data mining algorithms and compare two data mining algorithms in classifying breast cancer. In this study, the algorithms used in making comparisons are the Decision Tree C4.5 algorithm, and K-Nearest Neighbors. The author combines the Decision Tree C4.5 Classifier algorithm which has good ability to process large databases as feature selection then with the K-Nearest Neighbors (KNN) algorithm which is feasible and relevant to use in analyzing and diagnosing Cancer. The results of the test show that the K-Nearest Neighbors algorithm produces the best result.

In [6] COMPARISON OF THE ACCURACY OF RANDOM FOREST AND KNN FOR DIAGNOSING BREAST CANCER. Breast cancer is cancer that forms in the cells of the breast. According to data from the Observatory breast cancer contributes as much as 30.8% to cancer deaths in women of all ages in 2020. This research uses breast cancer data sets to increase awareness, because, awareness of breast cancer is important and should be a common science. KNN algorithm is often used for classification cases and Random Forest has versatile properties and without tuning can provide good accuracy in classification. From previous research, SVM has 96.47% accuracy, Neural Network as much as 97.06%, and Naive Bayes 91.18% accuracy. In this study, researchers have an interest in comparing the two algorithms with the ROC Curve. The data source comes from Kaggle. Diagnosis 'M' (malignant) and 'B' (benign). Consists of 569 data and 33 columns. Training data is 75% and uses 31 columns. From this study it can be concluded that the ROC value owned by KNN is 0.9959 and Random Forest is 0.9951.

In [5] Comparison of Data Mining Methods for Rainfall Classification with C4.5, Naïve Bayes, and KNN Algorithms. Rain is one of the things that must be observed because it is classified as rainfall. The Meteorology Climatology and Geophysics Agency (BMKG) is one of the government agencies in charge of conveying weather information. For rainfall, there are standards that will be achieved by BMKG, namely temperature, humidity and wind speed. This rainfall dataset is taken from the database of BMKG Jatiwangi, Majalengka from 01/2/2008 to 26/12/2018 taken from www.bmkg.go.id. To estimate the rainfall, data mining method with classification function is used. The Knowledge Discovery of Databases (KDD) process usually begins with the steps of data selection, pre-processing (data cleaning), transforming data, data mining and evaluation. In this research, 3 (three) algorithm methods are used, namely C4.5 or Decision Tree C4.5, K-Nearest Neigbor (KNN,) and Naïve Bayes. The software used to process data is RapidMiner. The conclusion of the three algorithms used is that the C4.5 algorithm is the best algorithm for estimating rainfall with an accuracy value (88.03%) and error (11.97%).

2.2 Breast Cancer Dataset

The studied dataset is called Breast Cancer Wisconsin [6] taken from UCI data repositories. This dataset is collected by Dr. William at the University of Wisconsin. This dataset includes 699 samples and 11 attributes as presented in Table 1, values of all features are an integer. Output field of this dataset is class. All samples of this dataset are classified as benign and malignant. 458 cases are benign and 241 cases are malignant.

Table 1. DATASET OVERVIEW

No.	Atribut	Nilai
1	*ID*	*Id Number*
2	*Clump Thickness*	1-10
3	*Uniformity of Cell Size*	1-10
4	*Uniformity of Cell Shape*	1-10
5	*Marginal Adhesion*	1-10
6	*Single Epithelial Cell Size*	1-10
7	*Bare Nuclei*	1-10
8	*Bland Chromatin*	1-10
9	*Normar Nucleoli*	1-10
10	*Mitoses*	1-10
11	*Class*	*Benign* *Malignant*

2.3 The Proposed Method

The research method used in this experiment uses the Cross-Standard Industry for Data Mining (CRISP-DM) model which consists of 6 phases, namely:

1. *Research Understanding Phase*

   The data used in this study is a breast cancer dataset taken from the UCI Repository, which was accessed on December 17, 2023 on the page https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original. The amount of data obtained is 699 data and there are 11 attributes. There are 241 of them affected by malignant breast cancer. This is a problem that occurs because there is no accurate analysis for that, so in this study the classification of the K-Nearest Neighbor (KNN), Naive Bayes, and Decision Tree C4.5 algorithms will be carried out.

   To determine the classification of breast cancer, there are 10 predictor attributes and 1 class attribute. The attributes that become parameters are shown in table

2. *Data Understanding*

   To determine the classification of breast cancer, there are 10 predictor attributes and 1 class attribute. The attributes that become parameters are shown in Table I.

3. *Data Preparation*

   The data obtained for this study were 699 breast cancer records, both malignant and benign. To obtain quality data, several preprocessing techniques were used [7], namely:

   a. Data Validation, to identify and remove outliers/noise, inconsistent data, and missing data.

   b. Data Integration and Transformation, to improve the accuracy and efficiency of the algorithm. The data used in this paper is categorical. For the neural network model, the data is transformed into numbers using RapidMiner software.

   c. Data Size reduction and dicretization, to obtain a dataset with fewer attributes and records but informative. In the training data used in this study, attribute selection and deletion of duplicate data were carried out using RapidMiner software.

   After preprocessing the data obtained from UCI Repostory as many as 699 records and 11 Attributes. After that, it is transformed from Binominal data type to numeric by changing the class attribute from Benign to number 2 and Malignant to number As table 2 which is a sample of training data.

Table 2. SAMPLE OF 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	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2

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
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2

4. *Modeling*
   This stage can also be called the learning stage because at this stage the training data is classified by the model and then produces a number of rules. In this research, modeling uses three algorithms, namely the K-Nearest Neighbor (KNN) algorithm, Naive Bayes, and Decision Tree C4.5.

5. *Evaluation*
   At this stage, testing of the classified models is carried out to obtain the most accurate model information. Evaluation and validation using Split Validation, Confusion matrix, and ROC Curve methods.

6. *Deployment*
   After model formation and analysis and measurement in the previous stage, the most accurate model is applied to breast cancer to determine the most accurate algorithm.

3. RESULTS AND DISCUSSION

3.1 *Model Testing*

The next process to classify breast cancer in this test uses RapidMiner Tools.

1. *Decision Tree C4.5*
   Testing at this stage is carried out to implement the Decision Tree C4.5 procedure using RapidMiner Tools. The structured division process regarding in RapidMiner Displays the C4.5 Decision Tree Algorithm Model with the process of entering the read excel input operator, then enter the Decision Tree C4.5 operator connect all operators then click the run button. Can be seen in Figure 1.

Figure 1. Initial testing stage of the C4.5 Decision Tree algorithm

So as to produce a C4.5 Decision Tree Algorithm model. Can be seen in Figure 2.

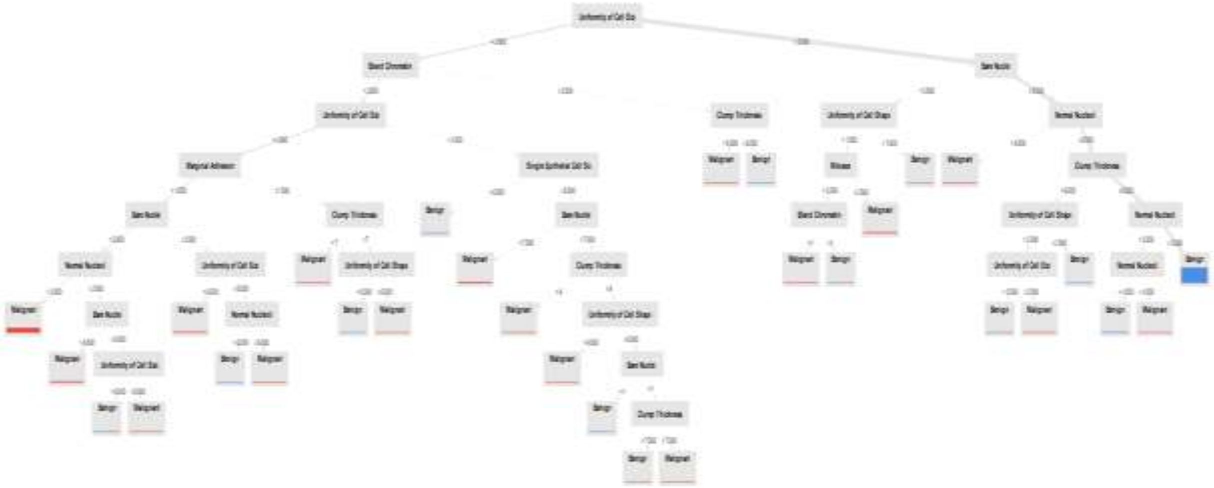


Figure 2. Decision Tree C4.5 Algorithm Model

Furthermore, to get the value of accuracy, precision, and recall by adding the apply model and performance operator can be seen in Figure 3.

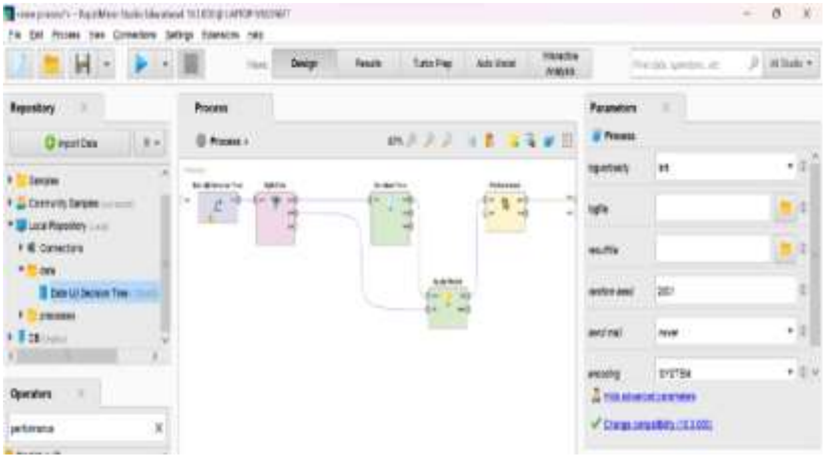
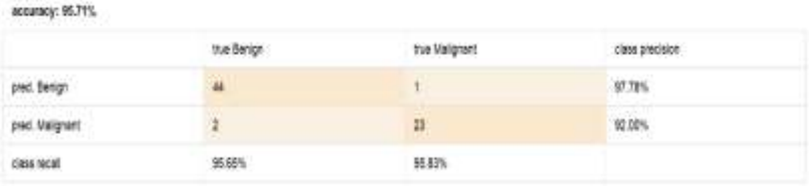


Figure 3. Final testing stage of Decision Tree C4.5 algorithm

At this stage, the results are assessed using performance tools aimed at displaying confusion tables, which are used to display the results of accuracy, precision and recall.

This discussion is carried out to obtain accuracy and precision values, respectively. Accuracy is the level of accuracy between the information requested by the user and the answer provided by the system, and precision is the level of closeness between the prediction results and the factual results. Recall is the success rate of the system in retrieving 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. Accuracy result of Decision Tree C4.5 algorithm

In Figure 4. is a calculation based on a dataset divided by split validation resulting in 90% training data and 10% testing data, it is known from 70 testing data, 44 are classified Benign in accordance with predictions made by the C4.5 Decision Tree algorithm, then 1 data is predicted benign but turns out to be malignant, 23 malignant data is predicted accordingly, and 2 data predicted malignant turns out to be benign.

Then the calculation results are visualized with the ROC curve. Can be seen in Figure 5.

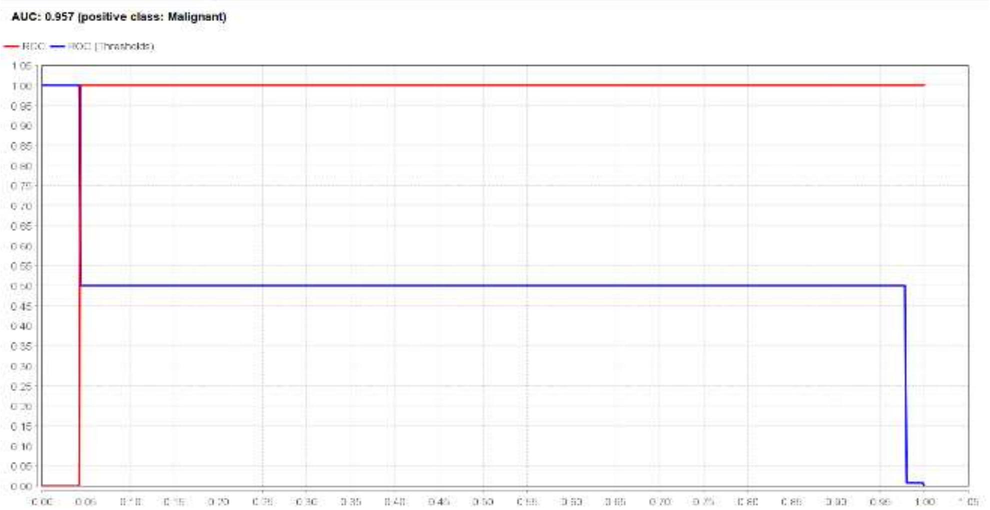
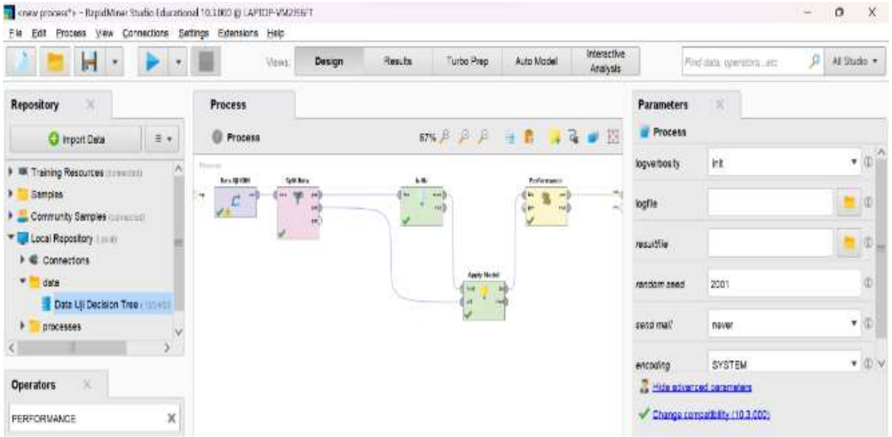


Figure 5. ROC curve with Decision Tree C4.5 algorithm

The ROC curve in Figure 5 expresses the confusion matrix from Figure 4. Horizontal lines are False Positives and Vertical lines are True Positives.

2. *K-Nearest Neighbor (KNN)*

Testing at this stage is carried out to implement the K-Nearest Neighbor (KNN) procedure using RapidMiner Tools. Structured sharing process regarding in RapidMiner Displaying the K-Nearest Neighbor (KNN) Algorithm Model by entering the read excel input operator, then entering the split data operator, the K-Nearest Neighbor (KNN) operator and to get the accuracy, precision, and recall values by adding the apply model and performance operators. Then connect all operators then click the run button. Can be seen in Figure 6.

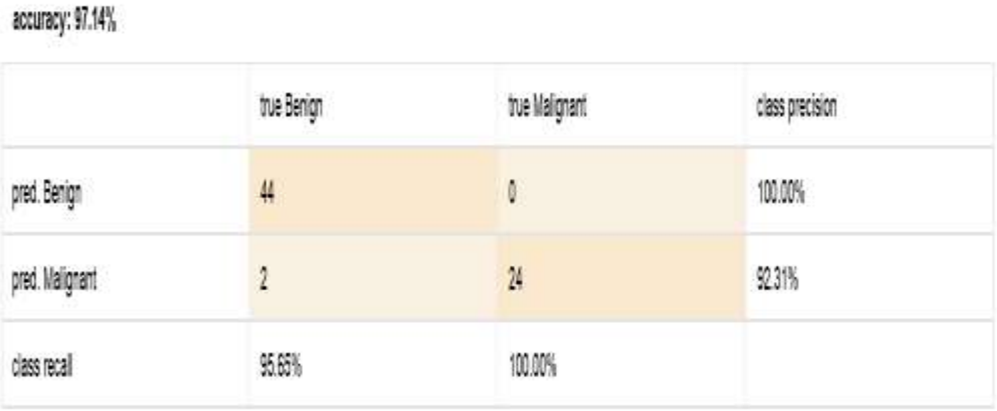


logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

Figure 6. K-Nearest Neighbor (KNN) algorithm testing phase

At this stage, the results are assessed using performance tools aimed at displaying confusion tables, which are used to display the results of accuracy, precision and recall.

This discussion is carried out to obtain the accuracy, precision and recall values, respectively. Accuracy is the level of accuracy between the information requested by the user and the answer given by the system, and precision is the level of closeness between the prediction results and the factual results. Recall is the success rate of the system in retrieving 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 7. Decision Tree C4.5 Algorithm Model

In Figure 7 is a calculation based on a dataset that is divided by split validation resulting in 90% training data and 10% testing data, known from 70 testing data, 44 are classified as Benign in accordance with predictions made by the K-Nearest Neighbor (KNN) algorithm, then 24 malignant data is predicted accordingly, and 2 data predicted malignant turns out to be benign.

Then the calculation results are visualized with a ROC curve. Can be seen in Figure 8.

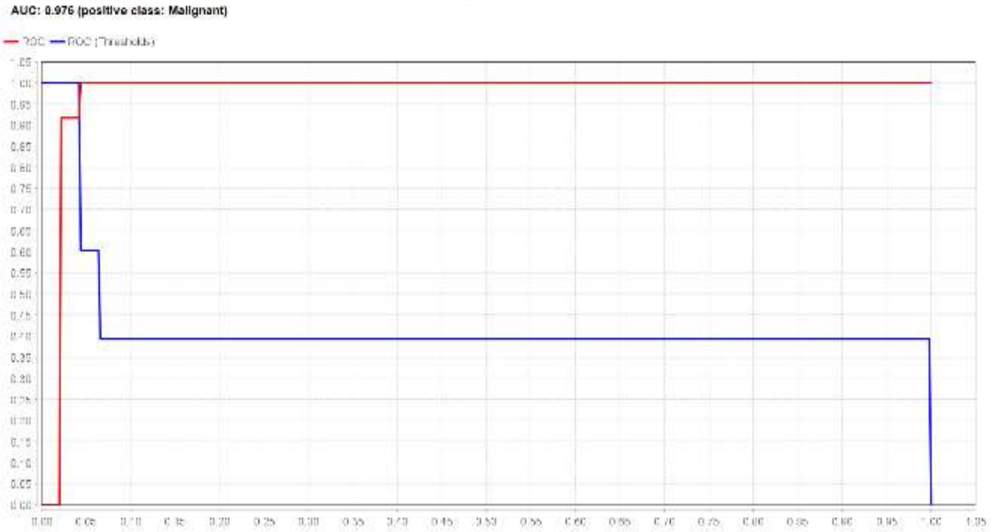
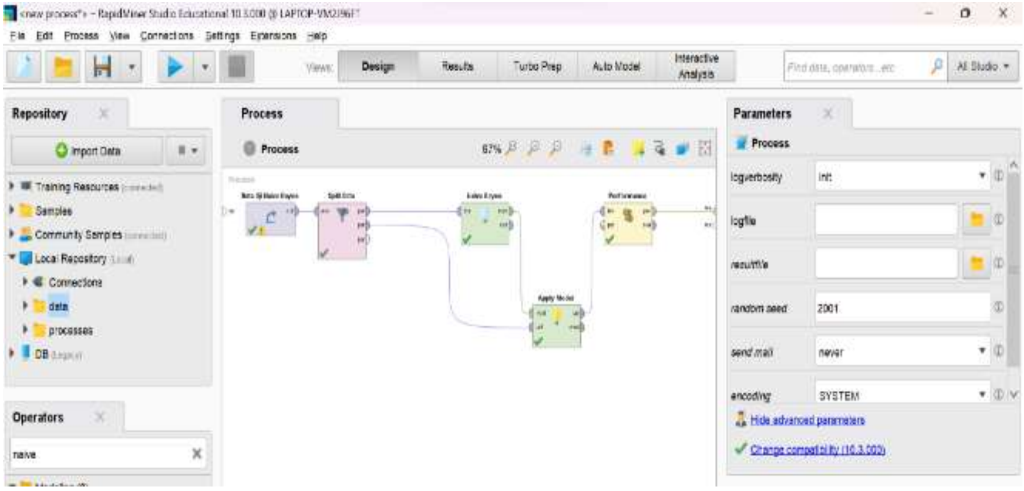


Figure 8. ROC curve with K-Nearest Neighbor (KNN) algorithm

The ROC curve in Figure 8 expresses the confusion matrix from Figure 7. Horizontal lines are False Positives and Vertical lines are True Positives.

3. *Naive Bayes*

Testing at this stage is carried out to implement the Naive Bayes procedure using RapidMiner Tools. The structured division process regarding in RapidMiner Displays the Naive Bayes Algorithm Model by entering the read excel input operator, then entering the split data operator, the Naive Bayes operator and to get the accuracy, precision, and recall values by adding the apply model and performance operators. Then connect all operators then click the run button. Can be seen in Figure 9.

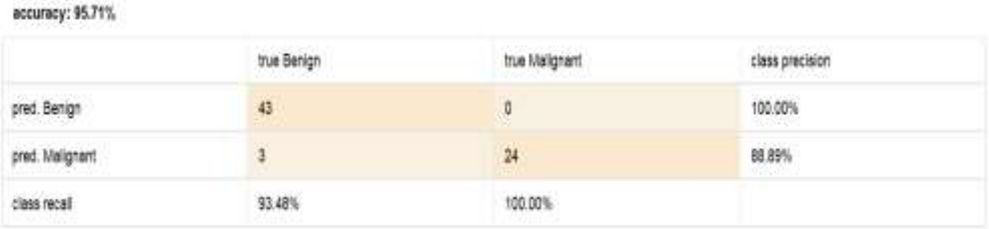


logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Figure 9. Naive Bayes algorithm testing phase

At this stage, the results are assessed using performance tools aimed at displaying confusion tables, which are used to display the results of accuracy, precision and recall.

This discussion is carried out to obtain the accuracy, precision and recall values, respectively. Accuracy is the level of accuracy between the information requested by the user and the answer given by the system, and precision is the level of closeness between the prediction results and the factual results. Recall is the success rate of the system in retrieving information.



accuracy: 95.71%

	true Benign	true Malignant	class precision
pred. Benign	43	0	100.00%
pred. Malignant	3	24	88.89%
class recall	93.48%	100.00%	

Figure 10. Accuracy result of Naive Bayes algorithm

In Figure 10 is a calculation based on a dataset that is divided by split validation resulting in 90% training data and 10% testing data, known from 70 testing data, 43 are classified Benign in accordance with predictions made by the K-Nearest Neighbor (KNN) algorithm, then 24 malignant data are predicted accordingly, and 3 data predicted malignant turned out to be benign.

Then the calculation results are visualized with the ROC curve. Can be seen in Figure 11.

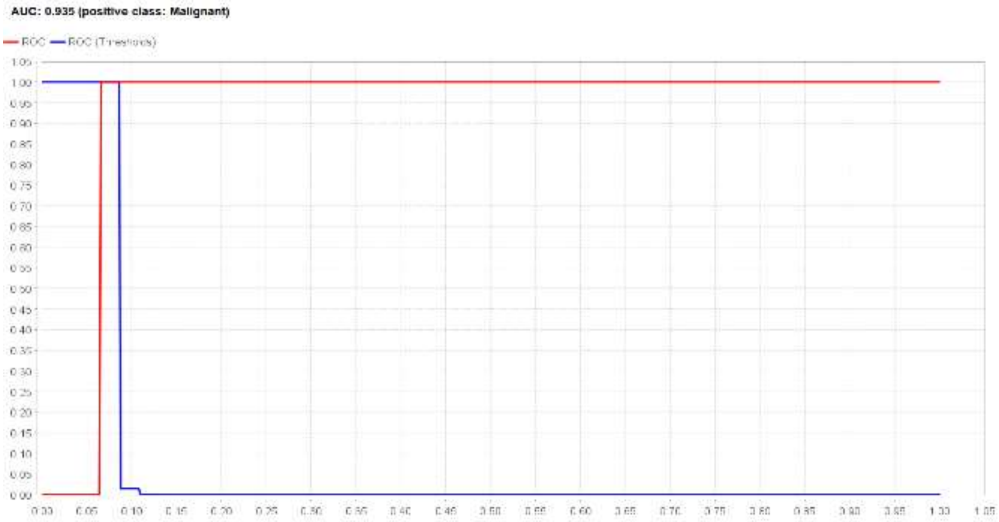


Figure 11. ROC curve with Naive Bayes algorithm

The ROC curve in Figure 11. expresses the confusion matrix from Figure 10. Horizontal lines are False Positives and Vertical lines are True Positives.

3.2 Analysis of Results

The model generated by the algorithm is tested using the split validation method, showing the highest comparison of accuracy, precision, recall, ROC curve values.

Table 3. ACCURACY AND AUC VALUE ON DECISION TREE C4.5, K-NEAREST NEIGHBOR (KNN), NAIVE BAYES

	Decision Tree C4.5	K-Nearest Neighbor (KNN)	Naive bayes
Accuracy	95,71%	97,14%	95,71%
AUC	0,957	0,976	0,935

Table 3. compares the Accuracy and AUC of each algorithm. It can be seen that the K-Nearest Neighbor Accuracy value is the highest as well as the AUC value. For Decision Tree Algorithm C4.5 and Naive Bayes also show the corresponding values.

Based on the grouping above and Table 3. it can be concluded that the Decision Tree C4.5, K-Nearest Neighbor (KNN), Naive Bayes algorithms are classified as very good because they have an AUC value between 0.90-1.00.

3.3 Research Implications

From the evaluation results, the K-Nearest Neighbor algorithm proved to be the most accurate compared to Decision Tree C4.5 and Naive Bayes. The three classification methods were applied to breast cancer data. With these results, it shows that the K-Nearest Neighbor algorithm is a good enough method in classifying data, thus the K-Nearest Neighbor algorithm can provide a solution to the problem of determining whether people can be diagnosed with benign or malignant breast cancer.

4. CONCLUSION

In this study, a model was created using the Decision Tree C4.5, K-Nearest Neighbor and Naive Bayes algorithms using breast cancer data. The resulting model is compared to find out the best algorithm in determining whether people have benign or malignant breast cancer. To measure the performance of the three algorithms, Split Validation, Confusion Matrix and ROC Curve testing methods are used, it is known that the K-Nearest Neighbor algorithm has the highest accuracy and AUC values, followed by the C4.5 Decision Tree algorithm and the lowest Naive Bayes algorithm.

Thus, the K-Nearest Neighbor algorithm is a fairly good algorithm in classifying data, thus the K-Nearest Neighbor algorithm can provide a solution to the problem of whether someone is diagnosed with benign or malignant breast cancer.

REFERENCES

[1] J. T. Wijaya, H. Oktavianto, and H. A. Al Faruq, “Perbandingan Algoritma K-Nearest Neighbor (Knn) Dan Gaussian Naive Bayes (Gnb) Dalam Klasifikasi Breast Cancer Coimbra,” *J. Smart Teknol.*, vol. 3, no. 3, pp. 233–237, Mar. 2022.

[2] Fahrurrozi and Wasilah, “Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor (KNN) Dan Decision Tree C-45,” *J. Tek.*, vol. 17, no. 2, pp. 427–434.

[3] A. A. Arif, M. Firdaus, Rahmaddeni, and Y. Maruhawa, “Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN,” *Inst. Ris. Dan Publ. Indones. IRPI*, pp. 187–197, Agustus 2022.

[4] D. Derisma and F. Febrian, “Perbandingan Teknik Klasifikasi Neural Network, Support Vector Machine, dan Naive Bayes dalam Mendeteksi Kanker Payudara,” *BINA INSANI ICT J.*, vol. 7, no. 1, p. 53, Jun. 2020, doi: 10.51211/biict.v7i1.1343.

[5] V. Angkasa and J. J. Pangaribuan, “KOMPARASI TINGKAT AKURASI RANDOM FOREST DAN KNN UNTUK MENDIAGNOSIS PENYAKIT KANKER PAYUDARA,” no. 1, 2022.

[6] I. Nawangsih, I. Melani, and S. Fauziah, “PREDIKSI PENGANGKATAN KARYAWAN DENGAN METODE ALGORITMA C5.0 (STUDI KASUS PT. MATARAM CAKRA BUANA AGUNG”.

[7] I. Sutoyo, “IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI DATA PESERTA DIDIK,” *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, Sep. 2018, doi: 10.33480/pilar. V 14i2.926.