Journal of Population Therapeutics & Clinical Pharmacology

Research Article
DOI: 10.47750/jptcp.2023.1002

Application of forward selection strategy using C4.5 algorithm to improve the accuracy of classification's data set

Etika Kartikadarma,[1] Pandu Adi Cakranegara,[2] Faisal Syafar,[3] Akbar Iskandar,[4] Arman Paramansyah,[5] Robbi Rahim[6]*

[1]Department of Informatics, Universitas Dian Nuswantoro, Semarang, Indonesia

[2]Department of Managgement, Universitas Presiden, Presiden, Indonesia

[3]Department of Electronics, Faculty of Engineering, Universitas Negeri Makassar, Makassar, Indonesia

[4]Department of Informatics, Universitas Teknologi AKBA Makassar, Makassar, Indonesia

[5]Department of Education, Institut Agama Islam Nasional Laa Roiba, Bogor, Indonesia

[6]Department of Management Technology, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia

***Corresponding author:** Robbi Rahim, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia. Email: usurobbi85@zoho.com

Submitted: 10 September 2022. Accepted: 28 October 2022. Published: 10 January 2023.

ABSTRACT

The purpose of this study is to improve the classification accuracy of the C4.5 Algorithm utilizing the forward selection technique. Breast Cancer from the UCI Machine Learning Repository is the dataset utilized. There are 286 records in the dataset with nine attributes and one class (label). The suggested model was evaluated with two existing classification models (C4.5 and Naïve Bayes) using the RapidMiner program. The procedure consists of multiple stages, the first of which consists of selecting the dominant trait using the feature selection technique (weight by information gain). The second step is forward selection based on the outcome of feature selection. Before processing, the dataset is separated into training and testing halves, where the ratios of comparison are 70:30, 80:20, and 90:10. The final step is examining the output. The experimental results demonstrate that the forward selection methodology employing the C4.5 (C4.5 + FS)

method outperforms the C4.5 and Naïve Bayes classification techniques. C4.5 + FS (Split Data 70:30) has an accuracy value of 76.74%, C4.5 + FS (Split Data 80:20) has an accuracy value of 78.95%, C4.5 + FS (Split Data 90:10) has an accuracy value of 78.57%, C4.5 (Split Data 70:30) has an accuracy value of 65.12%, and Naïve Bayes (Split Data is 70:30) has an accuracy value 85.55%. In comparison to typical classification algorithms (C4.5 and Naïve Bayes), the average accuracy values increased by 12.97% and 8.32%, respectively. In terms of precision, recall, and F-measure, the forward selection strategy utilizing the C4.5 method beat all other classification techniques, achieving 79.84%, 92.50%, and 85.55%, respectively. In addition, the results demonstrated an increase in the average Area Under Curve (AUC) from 0.628 to 0.732%. Therefore, it can be inferred that the forward selection strategy can be applied to the Breast Cancer Data Set in order to increase the accuracy value of classification method C4.5.

Keywords: *Forward selection, data mining, classification, method selection, data mining, classification, method C4.5, breast cancer*

INTRODUCTION

Breast cancer is a cancerous growth that targets breast tissue.[1] Women continue to have a high mortality rate due to breast cancer. According to data from the WHO (World Health Organization), breast cancer is a disease with a worldwide mortality rate of 42.5% and an annual average of 9.3 deaths per 100 woman.[2] Age, gender, race, family history, genetics, and personal behaviors such as smoking, consuming alcoholic drinks, and food can all contribute to the development of breast cancer.[3] Breast cancer continues to be a significant health issue.[4] Using Soft computing reasoning approaches,[4–6] breast cancer can now be detected due to the rapid growth of modern technology. One of the reasoning approaches of Soft computing is data mining, which is employed in health-related research.[7,8] Data mining is a method that identifies patterns with promise and utility for handling massive databases. In data mining, classification approaches, such as the C4.5 algorithm, Nave Bayes, Neural Network, and *K*-Nearest Neighbor, are frequently utilized by academics to solve difficulties.[9–12] The classification technique is also one of the most extensively researched algorithms.[13–16]

During the past decade, numerous researchers have employed classification approaches to solving breast cancer cases.[3] Research by Bahmani[17] suggested a hybrid model for breast cancer prediction, in which Naïve Bayes Network, Radial Basis Function (RBF) Network, and *K*-means clustering are utilized. Breast Cancer Wisconsin is the name of the dataset utilized, which was obtained from the UCI data repository. The results demonstrate that the hybrid model provided achieves an accuracy of 99% and an average absolute error of 0.019, which is superior to previous models. The subsequent work by Wu & Hicks[18] proposed a Machine Learning (ML) method for classifying breast cancer patients. The proposed models to be evaluated are Support Vector Machines (SVMs), *K*-NN, Naïve Bayes, and Decision Tree, which are trained to classify two forms of breast cancer using specified features at varying threshold levels (triple-negative and non-triple-negative). The dataset is derived from The Cancer Genome Atlas RNA-Sequence data from 110 triple negative and 992 non-triple negative breast cancer tumor samples to determine the characteristics (genes). The experimental results demonstrate that the SVM model classifies breast cancer

into triple-negative and non-triple-negative breast cancers more reliably and with fewer misclassifications than the other three models. In addition, (2021)[19] proposed a classification model for breast cancer detection by optimizing the Optimization-Based Feature Classification. The proposed model is a combination of the Whale Optimization Algorithm (WOA) model and the SVM model. Utilizing the Breast cancer dataset from the UCI repository, the dataset contains information about breast cancer. The results demonstrate that our system beats PSO-SVM and GA-SVM with a 98.82%higher accuracy (WOA-SVM). In addition, research was conducted[20] on the categorization of breast cancer by comparing three methods: Naïve Bayes, Neural Network, and SVM. The dataset consists of 2,000 digital mammography pictures, with 70%training data and 30%testing data. During the feature extraction process, the Gray Level Co-occurrence Matrix (GLCM) approach is utilized to represent two dimensions of gray level variation in the image. SVM offered the most consistent results in properly identifying the breast as 'Normal' or 'Cancer', with an accuracy of 99.4% on the training dataset and 98.76% on the test dataset.

Based on these considerations, this work proposes a classification model with a forward selection strategy in the classification algorithm to raise the value of classification accuracy using the breast cancer dataset. This model will be compared to the categorization model's standard version.

RESEARCH METHODOLOGY

In the research, the forward selection technique is applied to the C4.5 algorithm to improve classification accuracy, in this case with the aid of a dataset to evaluate the results of data analysis. Breast Cancer from the UCI Machine Learning Repository is the dataset utilized. This dataset is one of three domains made available by the Institute of Oncology that have been utilized regularly in classification research. There are 286 records in the dataset with 9 attributes and 1 class (label). The identity of the attribute is Class (no-recurrence-events, recurrence-events); age (10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99); menopause (lt40, ge40, premen); tumor-size (0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59); inv-nodes (0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39); node-caps (yes, no); deg-malig (1, 2, 3); breast (left, right); breast-quad (left-up, left-low, right-up, right-low, central); and irradiated (yes, no). The method proposed in this study is to apply the forward selection technique in increasing the accuracy of the C4.5 algorithm. This research uses the Rapid Miner Studio 9.10 software in the process of conducting the analysis. The experimental stages in this research are:

1. Prepare a dataset for use in experiments conducted by UCI Machine Learning Repository (Breast Cancer).
2. Apply the forward approach to the Breast Cancer dataset by means of split data (Training 70% and Testing 30%). Then, record the outcomes of the validation, which yields measurable statistics such as the area under curve Area Under Curve (AUC) and Accuracy.
3. Test the forward technique with the Breast Cancer dataset by splicing the data (Training 80% and Testing 20%). Then, record the outcomes of the validation, which yields measurable statistics such as AUC and Accuracy.
4. Test the forward approach with the Breast Cancer dataset by executing split data (Training 90% and Testing 10%). Then, record the outcomes of the validation, which yields measurable statistics such as AUC and Accuracy.
5. Test C4.5 without the forward approach on the Breast Cancer dataset through split data (Training 70% and Testing 30%). Then, record the outcomes of the validation, which yields measurable statistics such as AUC and Accuracy.
6. Test using Naïve Bayes with the Breast Cancer dataset by performing split data (Training 70%

and Testing 30%). Then, record the outcomes of the validation, which yields measurable statistics such as AUC and Accuracy.

7. Compare the top results for accuracy and retrieve the best findings.
8. Incorporate the most effective classification algorithm's results.

The proposed algorithm for this study is shown in Figure 1:

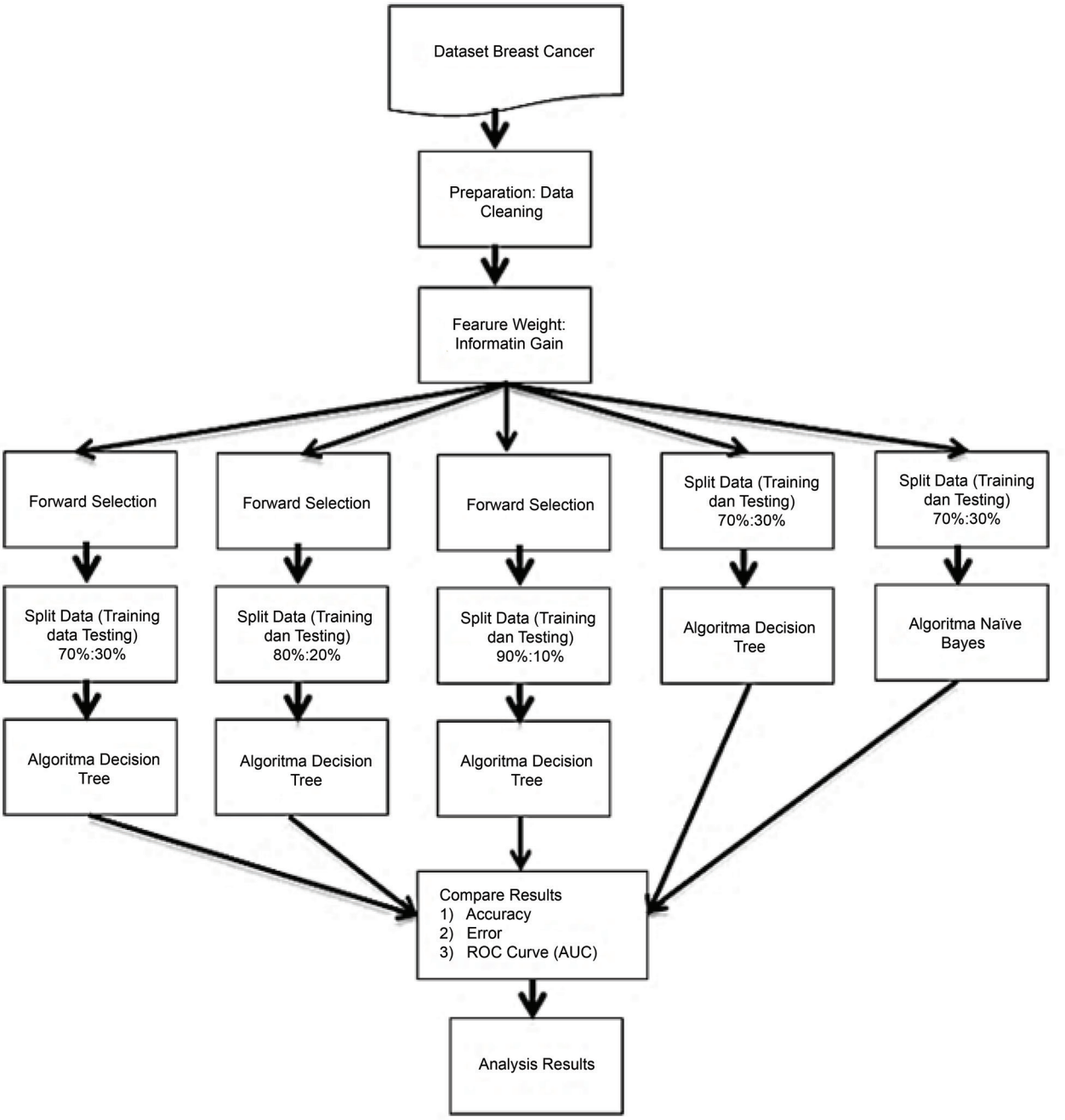

FIG 1. The proposed model.

J Popul Ther Clin Pharmacol Vol 30(1):e14–e23; 10 Janaury 2023.

RESULTS AND DISCUSSION

Figures 2 and 3 shows how the Breast Cancer dataset was used to test the C4.5 methods with a forward selection technique in Rapidminer 9.10. Figure 2 shows the details of the overall model, which combines the C4.5 algorithm with the forward selection technique (C4.5 + FS) and the standard classification model (C4.5 and Naïve Bayes). The C4.5 + FS model chooses the most important attribute by using feature selection (weighted by information gain). The results of the selection move on to the Forward Selection stage, where the dataset is split into two parts: training and testing. 70:30, 80:20, and 90:10 are the ratios that are used to compare. Then, Figure 3 shows how the C4.5 + FS model and the standard classification model were checked to make sure they were right. The accuracy, error, recall precision, and F-measure values will be used. These values will be found by evaluating the confusion matrix and the AUC curve.

For attribute selection using feature selection (weight by information gain), there are seven attributes that are selected or relevant to the classification results using Rapidminer 9.10, while three attributes are considered to have no effect or are irrelevant to the classification results, as shown in Table 1.

In Table 1, it is shown that optimal prediction can be made using nine attributes selected via feature selection (weight by information gain). These findings will be analyzed to determine the accuracy of comparison between the C4.5 + FS model and other categorization models listed in Table 2.

Figure 2 depicts the percentage of accuracy achieved by each model. The experimental results demonstrate that the forward selection methodology utilizing the C4.5 method (C4.5 + FS) beats the traditional classification technique (C4.5 and Naïve Bayes). The Naïve Bayes approach has a maximum accuracy of 69.77%and the C4.5 method has a maximum accuracy of 65.12%. In comparison to the C4.5 + FS model, the average accuracy values increased by 12.97%and 8.32%, respectively. As depicted in Figure 4, the following graph compares the accuracy rates of all models where the C4.5 + FS model has an accuracy value greater than 75% for all dataset comparison ratios.

Table 3 displays the precision, recall, Recall, and *F*-measure percentages for each model employed. The precision of all C4.5 + FS models yields percentages of 77.03%, 83.33%, and 79.17%, respectively (an average of 79.17%). This result outperforms the Naïve Bayes and C4.5 models by 73.44% and 75.76%, respectively (better 4.08% and 6.40%).

Meanwhile, the recall and *F*-Measure percentages for all C4.5 + FS models are much better than the Naïve Bayes and C4.5 models. The average recall and *f*-measure percentages were 93 and 86% or better 14 and 9% of the Naïve Bayes model and C4.5 (recall) and 10 and 6% better than the Naïve Bayes model and C4.5 (*f*-measure). For clarity, the following chart compares all models based on the percentage of precision, recall, Recall, and *F*-measure as shown in Figure 5.

The receiver operating characteristics (ROC) curve is a technique or method that helps you see, organize, and choose the best classification model based on how well it works. ROC has an area called the AUC that is very useful for comparing the performance of different classification models to find out which one is the best.[21] In Table 4, the levels of accuracy of the different classification models that were used to find the best classification method are compared. Compared to Naïve Bayes and C4.5, Model C4.5 + FS has the best AUC. In the fair classification category, the average AUC value is 0.763 when using the C4.5 + FS model. In the poor classification category, Naïve Bayes and C4.5 have AUCs of 0.653 and 0.628, respectively.

The following is a graph of the AUC of all models as shown in Figure 6.

Table 5 shows how the errors of all the models that were tested with the Breast Cancer dataset are compare with each other. The rate of errors is lower when you use the C4.5 method with the forward

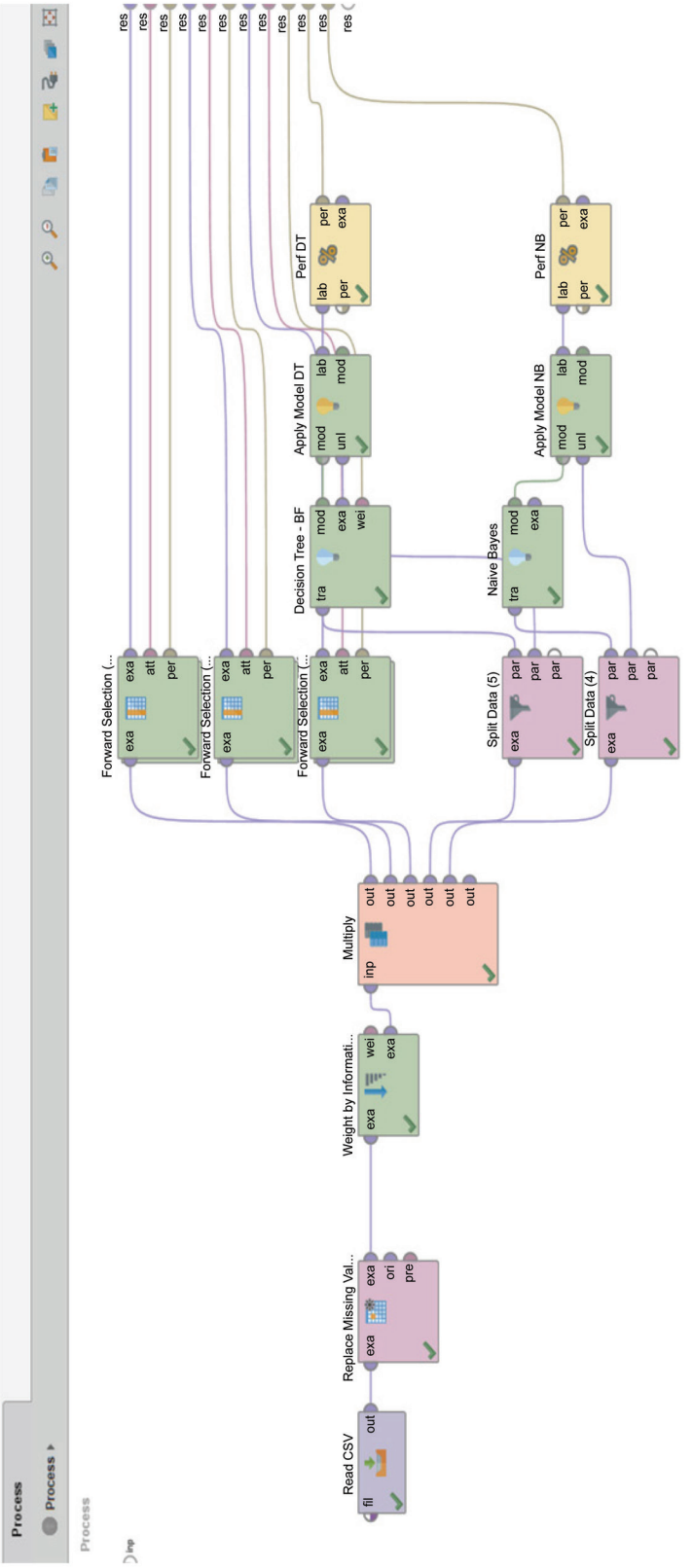

FIG 2. The proposed design utilizing RapidMiner.

J Popul Ther Clin Pharmacol Vol 30(1):e14–e23; 10 Janaury 2023.

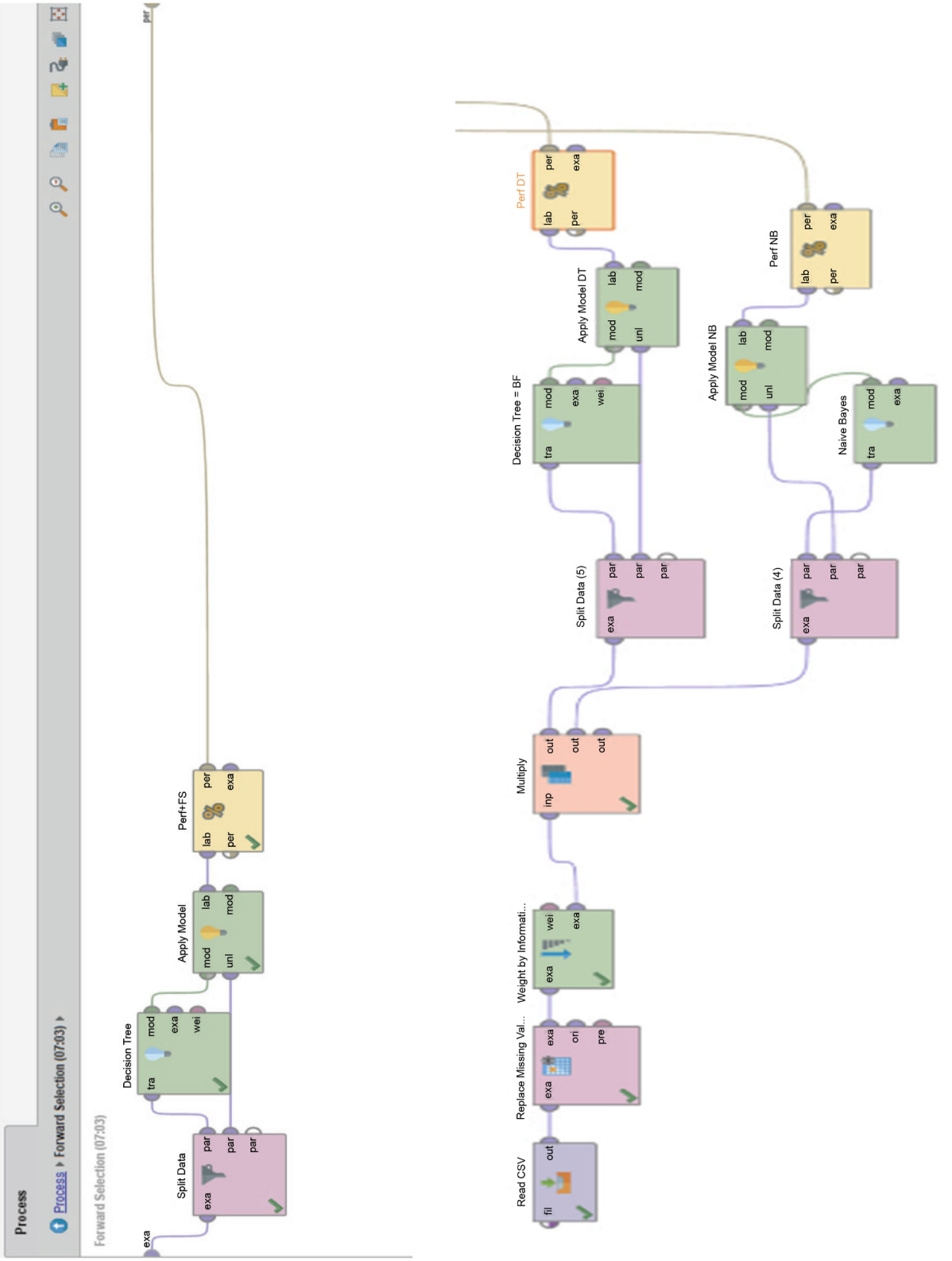

FIG 3. Validation procedure for all models tested using RapidMiner.

J Popul Ther Clin Pharmacol Vol 30(1):e14–e23; 10 Janaury 2023.

TABLE 1. Results of attribute selection.

Attribute	Weight
Breast-quad	1
Age	0
Menopause	1
Tumor-Size	1
Inv-Nodes	1
Node-Caps	0
Deg-Malig	1
Breast	1
Irradiat	0

TABLE 2. Comparison of accuracy.

Parameter	Accuracy (%)
C4.5 + FS (Split Data 70:30)	76.74
C4.5 + FS (Split Data 80:20)	78.95
C4.5 + FS (Split Data 90:10)	78.57
C4.5 (Split Data 70:30)	65.12
Naïve Bayes (Split Data 70:30)	69.77

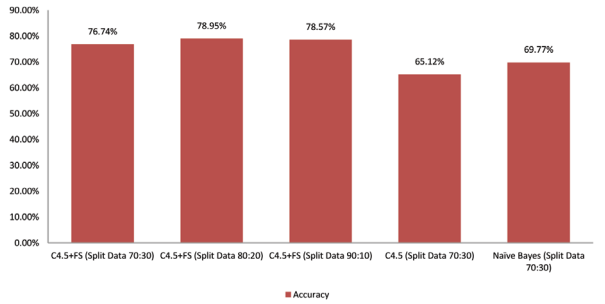

FIG 4. Comparative accuracy chart.

selection technique (C4.5 + FS) than when you use the standard classification model. The standard classification model and C4.5 + FS are not the same in a big way. In Table 5 and Figure 7, the error values for each model are shown in the form of tables and graphs.

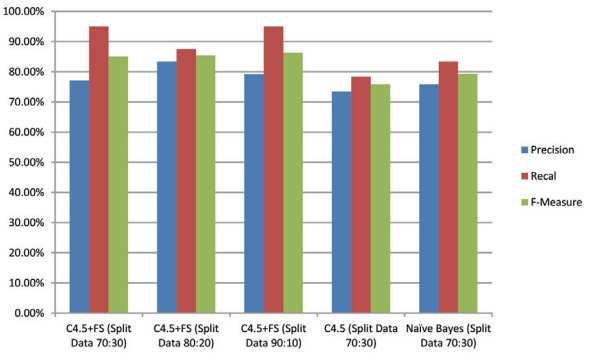

FIG 5. Comparison graph of all models' precision, recall, and *F*-measure levels.

TABLE 4. Comparative analysis of area under curve.

Parameter	AUC
C4.5 + FS (Split Data 70:30)	0.783
C4.5 + FS (Split Data 80:20)	0.732
C4.5 + FS (Split Data 90:10)	0.762
C4.5 (Split Data 70:30)	0.628
Naïve Bayes (Split Data 70:30)	0.653

AUC, Area Under Curve.

TABLE 3. Comparative analysis of precision, recall, and *F*-measure.

Parameter	Precision (%)	Recal (%)	*F*-Measure (%)
C4.5 + FS (Split Data 70:30)	77.03	95	85.07
C4.5 + FS (Split Data 80:20)	83.33	87.50	85.37
C4.5 + FS (Split Data 90:10)	79.17	95	86.36
C4.5 (Split Data 70:30)	73.44	78.33	75.81
Naïve Bayes (Split Data 70:30)	75.76	83.33	79.37

J Popul Ther Clin Pharmacol Vol 30(1):e14–e23; 10 Janaury 2023.

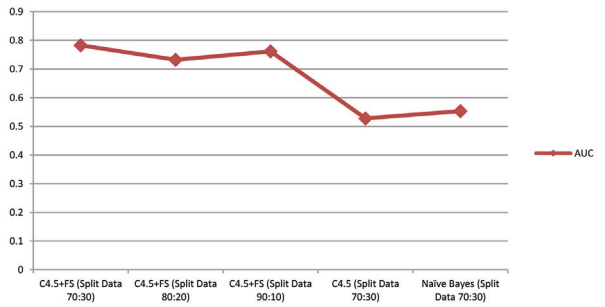

FIG 6. Comparison graph of all models' models' area under curve.

TABLE 5. Comparative analysis of error.

Parameter	Error (%)
C4.5 + FS (Split Data 70:30)	23.26
C4.5 + FS (Split Data 80:20)	21.05
C4.5 + FS (Split Data 90:10)	21.43
C4.5 (Split Data 70:30)	34.88
Naïve Bayes (Split Data 70:30)	30.23

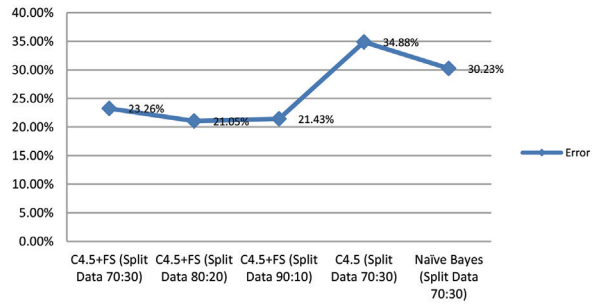

FIG 7. Comparison graph of all models' error.

CONCLUSION

Based on the results of this study's experiments and evaluations, it can be concluded that the integration of the C4.5 algorithm with the forward selection technique (C4.5 + FS) for the Breast Cancer dataset increased accuracy by 12.97% and 8.32%, respectively, when using training data samples with a ratio of 70:30, 80:20, and 90:10. Compared to other standard classification algorithms (C4.5 and Naïve Bayes), the increase in accuracy value is significant. In terms of precision, recall, and F-measure, the forward selection strategy using the C4.5 method beat all other classification techniques, achieving 79.84%, 92.50%, and 85.55%, respectively. Thus, it can be stated that there is a considerable difference in precision between the C4.5 + FS approach with C4.5 and the Naïve Bayes method.

REFERENCES

1. Tarawneh O, Otair M, Husni M, et al. Breast cancer classification using decision tree algorithms. Int J Adv Comput Sci Appl. 2022; 13(4): 676–680. https://doi.org/10.14569/IJACSA.2022.0130478
2. Oktavianto H, and Handri RP. Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes. INFORMAL Inform J. 2020; 4(3): 117. https://doi.org/10.19184/isj.v4i3.14170
3. Aslam MA, and Cui D. Breast cancer classification using deep convolutional neural network. J Phys Conf Ser. 2020; 1584(1): 1–10. https://doi.org/10.1088/1742-6596/1584/1/012005
4. Mathew TE, and Anil Kumar KS. A logistic regression based hybrid model for breast cancer classification. Ind J Comput Sci Eng. 2020; 11(6): 899–906. https://doi.org/10.21817/indjcse/2020/v11i6/201106201
5. He P, Zhang B, and Shen S. Effects of out-of-hospital continuous nursing on postoperative breast cancer patients by medical big data. J Healthc Eng. 2022; 2022. https://doi.org/10.1155/2022/9506915
6. Muhdi M, Buchori A, and Wibisono A. Whiteboard animation for android design using think talk write model to improve the post graduates students' concepts understanding. J Adv Res Dyn Contr Syst. 2019; 11(7): 535–543.
7. Guha Roy D, and Singh TN. Predicting deformational properties of Indian coal: soft computing and regression analysis approach. Meas J Int Meas Confed. 2020; 149: 106975. https://doi.org/10.1016/j.measurement.2019.106975
8. Charitopoulos A, Rangoussi M, and Koulouriotis D. On the use of soft computing methods in

educational data mining and learning analytics research: a review of years 2010–2018. Int J Artific Intellig Educ. 2020; 30(3): 371–430. https://doi.org/10.1007/s40593-020-00200-8

9. Novita R, Zakir S, Nur Khomarudin A, et al. Use of the C4.5 algorithm in determining scholarship recipients. J Phys Conf Ser. 2021; 1779(1). https://doi.org/10.1088/1742-6596/1779/1/012009
10. Windarto, A.P., Herawan, T. (2022). K-Means Algorithm with Rapidminer in Clustering School Participation Rate in Indonesia. In: Ab. Nasir, A.F., Ibrahim, A.N., Ishak, I., Mat Yahya, N., Zakaria, M.A., P. P. Abdul Majeed, A. (eds) Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering, vol 730. Springer, Singapore. https://doi.org/10.1007/978-981-33-4597-3_70
11. Sudarwanto AS, and Pujiyono. Responsibilities of banks to loss of customers using mobile banking. Int J Adv Sci Technol. 2020; 29(4): 1702–1706.
12. Buchori A, Setyosari P, Wayan Dasna I, et al. Developing character building learning model using mobile augmented reality on elementary school student in Central Java. Glob J Pure Appl Math. 2016; 12(4): 3433–3444.
13. Sonavane R, and Sonar P. Classification and segmentation of brain tumor using adaboost classifier. In Proceedings – International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016, 24–26 December 2016, India, pp. 396–403, 2017. https://doi.org/10.1109/ICGTSPICC.2016.7955334
14. Othman NA, Foozy CFM, Mustapha A, et al. A data mining approach for classification of traffic violations types. Int J Adv Intellig Inform. 2021; 7(3): 282–291. https://doi.org/10.26555/ijain.v7i3.708
15. Bardab SN, Ahmed TM, and Mohammed TAA. Data mining classification algorithms: an overview. Int J Adv Appl Sci. 2021; 8(2): 1–5. https://doi.org/10.21833/ijaas.2021.02.001
16. Al-Hawari A, Najadat H, and Shatnawi R. Classification of application reviews into software maintenance tasks using data mining techniques. Softw Qual J. 2021; 29(3): 667–703. https://doi.org/10.1007/s11219-020-09529-8
17. Bahmani E, Jamshidi M, and Shaltooki AA. Breast cancer prediction using a hybrid data mining model. Int J Inform Visual. 2019; 3(4): 327–331. https://doi.org/10.30630/joiv.3.4.240
18. Wu J, and Hicks C. Breast cancer type classification using machine learning. J Personal Med. 2021; 11(2): 1–12. https://doi.org/10.3390/jpm11020061
19. Raiesdana S. Breast cancer detection using optimization-based feature pruning and classification algorithms. Middle East J Cancer. 2021; 12(1): 48–68. https://doi.org/10.30476/mejc.2020.85601.1294
20. Samuri SM, Nova TV, Bahbibirahmatullah, et al. Classification model for breast cancer mammograms. IIUM Eng J. 2022; 23(1): 187–199. https://doi.org/10.31436/IIUMEJ.V23I1.1825
21. Fernanda JW. Boosting neural network dan Boosting Cart. J Mat. 2012; 2(2): 33–49.