DATA MINING USING RANDOM FOREST, NAÏVE BAYES, AND ADABOOST MODELS FOR PREDICTION AND CLASSIFICATION OF BENIGN AND MALIGNANT BREAST CANCER

Bahtiar Imran[1*]; Hambali[2]; Ahmad Subki[3]; Zaeniah[4]; Ahmad Yani[5]; Muhammad Rijal Alfian[6]

[1]Computer System Engineering, [2]Informatics Engineering, [3,4]Software Engineering, [5]Information Technology, [6]Computer Engineering
Universitas Teknologi Mataram

[1]bahtiarimranlombok@gmail.com, [2]08mi071@gmail.com, [3]lp2mutm@gmail.com, [4]hambaaloh847@gmail.com, [5]m4dy45@gmail.com, kodeflek@gmail.com

(*) Corresponding Author

Abstract— This study predicts and classifies benign and malignant breast cancer using 3 classification models. The method used in this research is Random Forest, Naïve Bayes and AdaBoost. The prediction results get Random Forest = 100%, Naïve Bayes = 80% and AdaBoost = 80%. Results using Test and Score with Number of Folds 2, 5 and 10. Number of Folds 2 Random Forest model Accuracy = 95%, Precision = 95% and Recall = 95%, Naïve Bayes Accuracy = 93%, Precision = 93% and Recall 93%, AdaBoost Accuracy = 90%, Precision = 90% and Recall = 90%. With Number of Folds 5 with Random Forest = 96%, Precision = 96% and Recall 96%. Naïve Bayes Accuracy value = 94%, Precision = 94% and Recall = 94%, AdaBoost Accuracy value = 93%, Precision = 93% and Recall = 93%. With Number of Folds 10 Random Forest model = 96%, Precision = 96% and Recall 96%. Naïve Bayes Accuracy value = 94%, Precision = 94% and Recall = 94%, AdaBoost Accuracy value = 92%, Precision = 92% and Recall = 92%. Of the 3 models used, Random Forest got the best classification results compared to the others.

Keywords: prediction; data mining; classification; breast cancer; model classification

Abstrak—*Penelitian ini memprediksi dan klasifikasi kanker payudara jinak dan ganas degan menggunakan 3 model klasifikasi. Metode yang digunakan pada penelitian ini menggunakan Random Forest, Naïve Bayes dan AdaBoost. Hasil prediksi mendapatkan hasil Random Forest = 100%, Naïve Bayes = 80% dan AdaBoost = 80%. Hasil menggunakan Test and Score dengan Number of Folds 2, 5 dan 10. Number of Folds 2 model Random Forest nilai Akurasi = 95%, Precision = 95% dan Recall = 95%, Naïve Bayes nilai Akurasi = 93%, Precision = 93% dan Recall 93%, AdaBoost nilai Akurasi = 90%, Precision = 90% dan Recall = 90%. Dengan Number of Folds 5 dengan hasil Random Forest = 96%, Precision = 96% dan Recall 96%. Naïve Bayes nilai Akurasi = 94%, Precision = 94% dan Recall = 94%, AdaBoost nilai Akurasi = 93%, Precision = 93% dan Recall = 93%. Dengan Number of Folds 10 model Random Forest = 96%, Precision = 96% dan Recall 96%. Naïve Bayes nilai Akurasi = 94%, Precission = 94% dan Recall = 94%, AdaBoost nilai Akurasi = 92%, Precission = 92% dan Recall = 92%. Dari ke 3 model yang digunakan, Random Forest mendapatkan hasil klasifikasi yang paling baik dibanding dengan yang lainnya..*

***Kata Kunci**: prediksi, data mining, klasifikasi, kanker payudara, model klasifikasi.*

INTRODUCTION

Breast cancer is one of the primary diseases threatening humans, especially women worldwide (BV, 2019). In women aged between 45 years to 60 years (Gupta & Kaushik, 2018), breast cancer is the leading cause of death for women after lung cancer. Based on the data, about 1.7 million new cases worldwide were diagnosed, and 521,900 deaths in 2012 (Octaviani & Rustam, 2019). Breast cancer is one of the most widespread cancers among women and is one of the leading causes of death. Not all tumors in the breast are breast cancer. However, further examination is needed from a doctor to determine breast cancer diagnosis (Abd-Elrazek et al., 2018). The early stages of breast cancer may not cause any symptoms, depending on the type of cancer to cause various symptoms (Krishna & Rao, 2018).

Calculation of breast cancer risk factors can be determined using an algorithm or model for early detection of breast cancer risk through determinant factors and requires preventive action, using

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2912

machine learning by classifying breast cancer risk from predictor variables, making it easier to classify (Nindrea et al., 2018).

From several previous studies, there has been no research that uses Data Mining with Random Forest, Naïve Bayes, and AdaBoost models to predict benign and malignant breast cancer, as in the study (BV, 2019), researchers used data mining techniques to predict breast cancer and got an accuracy result of 98%. Octaviani (Octaviani & Rustam, 2019) uses the RF method for breast cancer prediction and produces an accuracy of 100%, Krishna (Krishna & Rao, 2018) proposes Machine Learning for breast cancer prediction. The proposed Machine Learning models include SVM (Support Vector Machine) classifier, Random Forest, Gradient boosting, Naive Bayes, Cart Model, Neural Network, and Linear Regression. The results obtained that using the SVM model gives the best results. Vikas Chaurasia (Chaurasia et al., 2018) proposed a Data Mining technique to predict benign and malignant breast cancer. The results obtained that the naive Bayes model gave the highest accuracy of 97.36%. Nikita Rane (Jean Sunny et al., 2020) proposed a Machine Learning method for the classification and precision of breast cancer. The proposed method can classify benign and malignant breast cancer. Swetha K (Swetha & Ranjana, 2020) suggested Machine Learning and Data Mining for breast cancer prediction and got the results that the Simple Logistics algorithm got better results than other methods with an accuracy rate of 99.7612%. Angela More's (More et al., 2022) this study uses Machine Learning Techniques to classify and predict breast cancer. The results obtained are that the SVM model gives good accuracy results of 97.87%. Morgana Darshini Ganggayah (Ganggayah et al., 2019) uses Machine Learning to predict the survival of breast cancer patients. The proposed model results that the model from Random Forest gets the best results with 82.7% accuracy. Nitasha (Nitasha, 2019) uses Data Mining to predict breast cancer by bringing the results that the proposed method can give good results. Aqua Anjum (Anjum, 2019) uses Machine Learning to diagnose breast cancer with the result that the proposed method can help in the diagnosis process. Iffat Khan's study (Khan et al., 2020) suggested the prediction of breast cancer using Data Mining. The focus of this study was on female breast cancer. Hiba Masood (Masood, 2021) uses a Machine Learning algorithm to detect breast cancer. This study proposes a model to provide the best accurate results. Ramik Rawal (Rawal, 2020) suggested using Machine Learning to predict breast cancer. The result was that the accuracy obtained by SVM (97.13%) was better than the accuracy obtained by C4.5, Naïve Bayes, and k-NN, which had varying accuracy between 95.12 % and 95.28%. Ayyoubzadeh S.M (Ayyoubzadeh et al., 2021) uses Data Mining to predict early breast cancer; This study at a hospital in Iran. The results obtained that the Random Forest model provides high accuracy results. Harjasdeep Singh (Singh, 2021) analyzed and predicted breast cancer using the Machine Learning method. The results obtained that the Random Forest algorithm shows promising results with an accuracy of 99.76%. Jiaxin Li (Li et al., 2021) this study predicts the survival of breast cancer for five years using Machine Learning, with the results of decision trees (19 studies, 61.3%, artificial neural networks (18 studies, 58.1%), support vector machines (16 studies, 51.6%, and ensemble learning (10 studies, 32.3%). Keerthana Rajendran (Rajendran et al., 2020) This research applies the Supervised Machine Learning method for breast cancer prediction. This study finds that the Bayesian Network algorithm achieves accurate results. In predicting breast cancer based on its risk factors.

There have been many studies that have been done by predicting breast cancer objects and using different methods. However, research using Data Mining with Random Forest, Naïve Bayes, and The AdaBoost model has not performed predicting and classifying benign and malignant breast cancer. So the purpose of this study is to predict benign and malignant breast cancer using Data Mining using Random Forest, Naïve Bayes, and AdaBoost models. The determination of the three models has reasons:. The Random Forest model is the model used because this model has good accuracy results (Ayyoubzadeh et al., 2021; Ganggayah et al., 2019; Octaviani & Rustam, 2019; Singh, 2021). Naïve Bayes is a model used for classification. This model also provides a good classification with a high level of accuracy (Kharya & Soni, 2016). AdaBoost is a method with good performance and performance in delivering classification (Kaur & Chopra, 2015; Kumari & Rani, 2017; Perveen et al., 2016). Before the classification stage uses the proposed method, the data is pre-processed using Normalize to Interval. It is done because there are still a lot of data that are still Null and redundant (Krishna & Rao, 2018).

MATERIALS AND METHODS

The stages of the research process carried out in this study are documented in the research methodology flow chart, which can be seen in Figure 1.

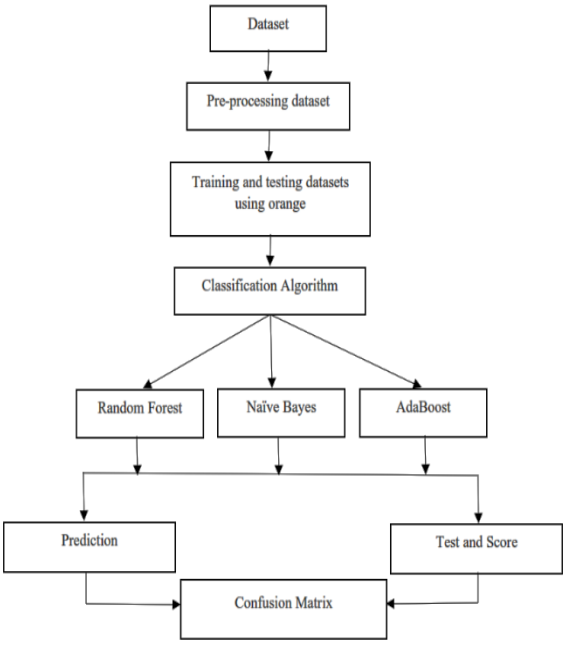


Figure 1. Research Methodology

a. Dataset

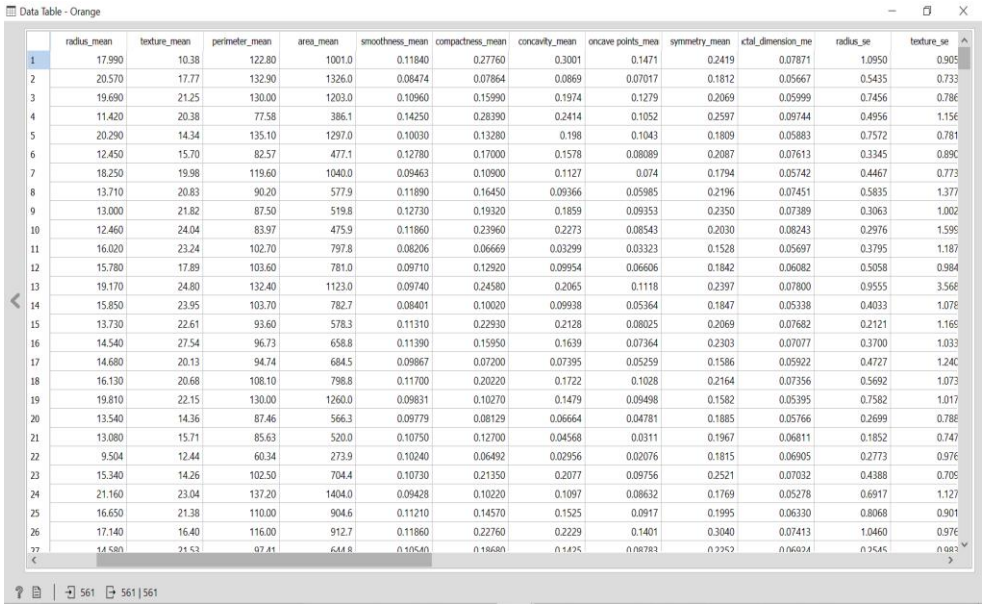
This data set takes from the UCI Machine Learning Repository website. This dataset was obtained from Dr. W.H. Wolberg at the University of Wisconsin-Madison, consisting of 569 data (Abd-Elrazek et al., 2018; BV, 2019; Chaurasia et al., 2018; Krishna & Rao, 2018; Octaviani & Rustam, 2019), cases classified into benign and malignant breast cancer? (BV, 2019). *This dataset is grouped into two classes, namely benign and malignant classes* (Chaurasia et al., 2018), while the tools used in this study are Orange. Orange is one data mining tool that produces better and more effective results than others (Kodati & Vivekanandam, 2018; Kukasvadiya et al., 2017; Manimannan et al., 2019; Sasikala, 2017). The attributes used are 29 attributes, and Table 1 is the attribute used in this study.

Table 1. Attributes Used

No	Nama Atribut	No	Nama Atribut
1	radius_mean	15	compactness_se
2	texture_mean	16	concavity_se
3	perimeter_mean	17	concave points_se
4	area_mean	18	symmetry_se
5	smoothness_mean	19	fractal_dimension_se
6	compactness_mean	20	radius_worst
7	concavity_mean	21	texture_worst
8	concave points_mean	22	perimeter_worst
9	symmetry_mean	23	area_worst
10	fractal_dimension_mean	24	smoothness_worst
11	radius_se	25	compactness_worst
12	texture_se	26	concavity_worst
13	perimeter_se	27	concave points_worst
14	area_se	28	symmetry_worst

b. Pre-processing Dataset

The data obtained on the UCI Machine Learning Repository site, the data is processed first. It is because there is still a lot of inaccurate and redundant data (Krishna & Rao, 2018). The method used in this pre-processing is Normalize Features with Normalize to Interval [-1,1] feature. An example of data that has not been pre-processed can be seen in Figure 2, while the size of the pre-processing results with Normalize Features is in Figure 2.


Data Table - Orange

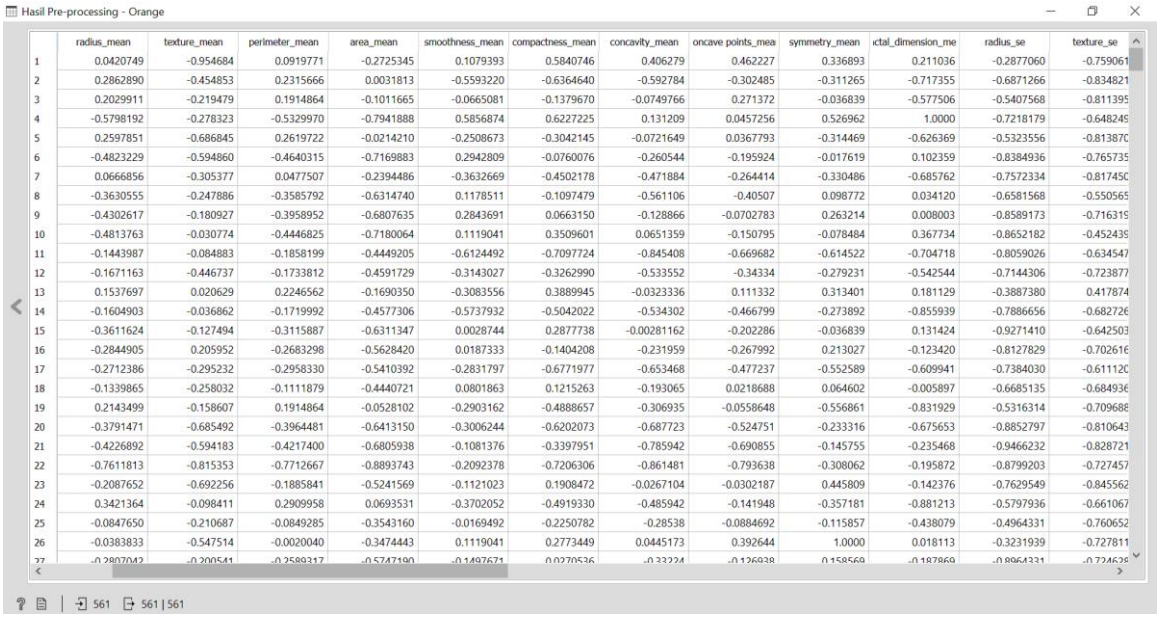
	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	oncave points_mea	symmetry_mean	ictal_dimension_me	radius_se	texture_se
1	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.1471	0.2419	0.07871	1.0950	0.905
2	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.733
3	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.1279	0.2069	0.05999	0.7456	0.786
4	11.420	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.1052	0.2597	0.09744	0.4956	1.156
5	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198	0.1043	0.1809	0.05883	0.7572	0.781
6	12.450	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	0.2087	0.07613	0.3345	0.890
7	18.250	19.98	119.60	1040.0	0.09463	0.10900	0.1127	0.074	0.1794	0.05742	0.4467	0.773
8	13.710	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	0.2196	0.07451	0.5835	1.377
9	13.000	21.82	87.50	519.8	0.12730	0.19320	0.1859	0.09353	0.2350	0.07389	0.3063	1.002
10	12.460	24.04	83.97	475.9	0.11860	0.23960	0.2273	0.08543	0.2030	0.08243	0.2976	1.599
11	16.020	23.24	102.70	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187
12	15.780	17.89	103.60	781.0	0.09710	0.12920	0.09954	0.06606	0.1842	0.06082	0.5058	0.984
13	19.170	24.80	132.40	1123.0	0.09740	0.24580	0.2065	0.1118	0.2397	0.07800	0.9555	3.568
14	15.850	23.95	103.70	782.7	0.08401	0.10020	0.09938	0.05364	0.1847	0.05338	0.4033	1.078
15	13.730	22.61	93.60	578.3	0.11310	0.22930	0.2128	0.08025	0.2069	0.07682	0.2121	1.169
16	14.540	27.54	96.73	658.8	0.11390	0.15950	0.1639	0.07364	0.2303	0.07077	0.3700	1.033
17	14.680	20.13	94.74	684.5	0.09867	0.07200	0.07395	0.05259	0.1586	0.05922	0.4727	1.240
18	16.130	20.68	108.10	798.8	0.11700	0.20220	0.1722	0.1028	0.2164	0.07356	0.5692	1.073
19	19.810	22.15	130.00	1260.0	0.09831	0.10270	0.1479	0.09498	0.1582	0.05395	0.7582	1.017
20	13.540	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.788
21	13.080	15.71	85.63	520.0	0.10750	0.12700	0.04568	0.0311	0.1967	0.06811	0.1852	0.747
22	9.504	12.44	60.34	273.9	0.10240	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.976
23	15.340	14.26	102.50	704.4	0.10730	0.21350	0.2077	0.09756	0.2521	0.07032	0.4388	0.709
24	21.160	23.04	137.20	1404.0	0.09428	0.10220	0.1097	0.08632	0.1769	0.05278	0.6917	1.127
25	16.650	21.38	110.00	904.6	0.11210	0.14570	0.1525	0.0917	0.1995	0.06330	0.8068	0.901
26	17.140	16.40	116.00	912.7	0.11860	0.22760	0.2229	0.1401	0.3040	0.07413	1.0460	0.976
27	14.580	21.53	97.41	644.8	0.10540	0.18680	0.1425	0.08783	0.2252	0.06924	0.2545	0.983

561 | 561 | 561

Figure 2. Data Before Pre-Processing

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2912

Figure 3 is an example of breast cancer data in pre-processing. Previously processed data still contains Null and redundant data.


Hasil Pre-processing - Orange

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	oncave points_mea	symmetry_mean	ıctal_dimension_me	radius_se	texture_se
1	0.0420749	-0.954684	0.0919771	-0.2725345	0.1079393	0.5840746	0.406279	0.462227	0.336893	0.211036	-0.2877060	-0.759061
2	0.2862890	-0.454853	0.2315666	0.0031813	-0.5593220	-0.6364640	-0.592784	-0.302485	-0.311265	-0.717355	-0.6871266	-0.834821
3	0.2029911	-0.219479	0.1914864	-0.1011665	-0.0665081	-0.1379670	-0.0749766	0.271372	-0.036839	-0.577506	-0.5407568	-0.811395
4	-0.5798192	-0.278323	-0.5329970	-0.7941888	0.5856874	0.6227225	0.131209	0.0457256	0.526962	1.0000	-0.7218179	-0.648245
5	0.2597851	-0.686845	0.2619722	-0.0214210	-0.2508673	-0.3042145	-0.0721649	0.0367793	-0.314469	-0.626369	-0.5323556	-0.813870
6	-0.4823229	-0.594860	-0.4640315	-0.7169883	0.2942809	-0.0760076	-0.260544	-0.195924	-0.017619	0.102359	-0.8384936	-0.765735
7	0.0666856	-0.305377	0.0477507	-0.2394486	-0.3632669	-0.4502178	-0.471884	-0.264414	-0.330486	-0.685762	-0.7572334	-0.817450
8	-0.3630555	-0.247886	-0.3585792	-0.6314740	0.1178511	-0.1097479	-0.561106	-0.40507	0.098772	0.034120	-0.6581568	-0.550565
9	-0.4302617	-0.180927	-0.3958952	-0.6807635	0.2843691	0.0663150	-0.128866	-0.0702783	0.263214	0.008003	-0.8589173	-0.716319
10	-0.4813763	-0.030774	-0.4446825	-0.7180064	0.1119041	0.3509601	0.0651359	-0.150795	-0.078484	0.367734	-0.8652182	-0.452439
11	-0.1443987	-0.084883	-0.1858199	-0.4449205	-0.6124492	-0.7097724	-0.845408	-0.669682	-0.614522	-0.704718	-0.8059026	-0.634547
12	-0.1671163	-0.446737	-0.1733812	-0.4591729	-0.3143027	-0.3262990	-0.533552	-0.34334	-0.279231	-0.542544	-0.7144306	-0.723877
13	0.1537697	0.020629	0.2246562	-0.1690350	-0.3083556	0.3889945	-0.0323336	0.111332	0.313401	0.181129	-0.3887380	0.417874
14	-0.1604903	-0.036862	-0.1719992	-0.4577306	-0.5737932	-0.5042022	-0.534302	-0.466799	-0.273892	-0.855939	-0.7886656	-0.682726
15	-0.3611624	-0.127494	-0.3115887	-0.6311347	0.0028744	0.2877738	-0.00281162	-0.202286	-0.036839	0.131424	-0.9271410	-0.642503
16	-0.2844905	0.205952	-0.2683298	-0.5628420	0.0187333	-0.1404208	-0.231959	-0.267992	0.213027	-0.123420	-0.8127829	-0.702616
17	-0.2712386	-0.295232	-0.2958330	-0.5410392	-0.2831797	-0.6771977	-0.653468	-0.477237	-0.552589	-0.609941	-0.7384030	-0.611120
18	-0.1339865	-0.258032	-0.1111879	-0.4440721	0.0801863	0.1215263	-0.193065	0.0218688	0.064602	-0.005897	-0.6685135	-0.684936
19	0.2143499	-0.158607	0.1914864	-0.0528102	-0.2903162	-0.4888657	-0.306935	-0.0558648	-0.556861	-0.831929	-0.5316314	-0.709688
20	-0.3791471	-0.685492	-0.3964481	-0.6413150	-0.3006244	-0.6202073	-0.687723	-0.524751	-0.233316	-0.675653	-0.8852797	-0.810643
21	-0.4226892	-0.594183	-0.4217400	-0.6805938	-0.1081376	-0.3397951	-0.785942	-0.690855	-0.145755	-0.235468	-0.9466232	-0.828721
22	-0.7611813	-0.815353	-0.7712667	-0.8893743	-0.2092378	-0.7206306	-0.861481	-0.793638	-0.308062	-0.195872	-0.8799203	-0.727457
23	-0.2087652	-0.692256	-0.1885841	-0.5241569	-0.1121023	0.1908472	-0.0267104	-0.0302187	0.445809	-0.142376	-0.7629549	-0.845562
24	0.3421364	-0.098411	0.2909958	0.0693531	-0.3702052	-0.4919330	-0.485942	-0.141948	-0.357181	-0.881213	-0.5797936	-0.661067
25	-0.0847650	-0.210687	-0.0849285	-0.3543160	-0.0169492	-0.2250782	-0.28538	-0.0884692	-0.115857	-0.438079	-0.4964331	-0.760652
26	-0.0383833	-0.547514	-0.0020040	-0.3474443	0.1119041	0.2773449	0.0445173	0.392644	1.0000	0.018113	-0.3231939	-0.727811
27	-0.2807042	-0.200541	-0.2589317	-0.5747190	-0.1497671	0.0270536	-0.33224	-0.126938	0.158569	-0.187869	-0.8964331	-0.724628

561 | 561 | 561

Figure 3. Data After Pre-Processing

c. Training and Testing Data

After the dataset is collected and pre-processing is carried out, it is necessary to conduct training and testing of the data used as a classification test. There are two groups of data files, namely training files and testing files. In the training data, the attributes used as targets are first determined. In this case, the target attribute is the diagnosis attribute. In the diagnosis attribute, two classes will be targeted in this case, coded M and B. In this case, M = Malignant and B = Benign (Krishna & Rao, 2018; Rawal, 2020).

d. Data Mining Stage Process

To evaluate the proposed categorization model's performance. It is necessary to compare the performance and performance of the model used. Of the three proposed classification models, several previous studies concluded that Random Forest is a model that provides a good classification (Ayyoubzadeh et al., 2021; Ganggayah et al., 2019; Octaviani & Rustam, 2019; Singh, 2021). Naïve Bayes is also a classification model that provides high accuracy results (Kharya & Soni, 2016). The AdaBoost classification model is a model that offers good classification performance (Kaur & Chopra, 2015; Kumari & Rani, 2017; Perveen et al., 2016). The previous stage uses the data as training data. The next step is to process test data to reduce invalid and redundant data.

e. Process Results in Evaluation Stage

In this process, the process for calculating the success rate of the Random Forest, Naïve Bayes, and AdaBoost methods for calculating the success rate using Test and Score. Meanwhile, to evaluate the success and performance of the proposed model, this study uses the Confusion Matrix.

The Confusion Matrix includes a table that is applied to represent the results of the classification process on a test data set whose actual level is known (BV, 2019), using the Confusion matrix to evaluate and calculate the performance of the classification model (More et al., 2022). Table 2 is an example of a Confusion Matrix.

Table 2. Confusion Matrix

		Predicted	
Actual		B	M
	B	True Positives (TP)	False Negatives (FN)
	M	False Positives (FP)	True Negatives (TN)

Accuracy is measured using (Ayyoubzadeh et al., 2021; Bissanum et al., 2021; Chaurasia et al., 2018):

$$\text{Accuracy } \frac{TP+TN}{TP+FP+FN+TN} \quad \text{.....(1)}$$

Precision is measured using (Abd-Elrazek et al., 2018; Bissanum et al., 2021; BV, 2019):

Precision $\frac{TP}{TP+FP}$ ..... (2)

The recall is measured using (Abd-Elrazek et al., 2018; Bissanum et al., 2021; BV, 2019; Chaurasia et al., 2018):

Recall $\frac{TP}{TP+FN}$ .....(3)

RESULTS AND DISCUSSION

a. Prediction Results with Prediction Widget

At this stage, predictions to determine whether the proposed model can provide the correct predictions following the input testing data. In the test data carried out, the model using Random Forest gives good predictive results, as evidenced by the prediction results with values for the Random Forest model = 100%, the Naïve Bayes model = 80%, and the AdaBoost model = 80%. Figure 4 is the result of predictions made using the prediction widget.

Predictions - Orange

	Random Forest	Naive Bayes	AdaBoost	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mear	ompactness_mea	concavity_mean	ncave points_mea	symmetr
1	B	B	B	923465	10.820	24.21	68.89	361.6	0.08192	0.06602	0.01548	0.00816	0.1976
2	B	B	B	923748	10.860	21.48	68.51	360.5	0.07431	0.04227	0	0.00000	0.1661
3	B	B	B	923780	11.130	22.44	71.49	378.4	0.09566	0.08194	0.04824	0.02257	0.2030
4	B	B	B	924084	12.770	29.43	81.35	507.9	0.08276	0.04234	0.01997	0.01499	0.1539
5	B	B	B	924342	9.333	21.94	59.01	264.0	0.09240	0.05605	0.03996	0.01282	0.1692
6	B	B	B	924632	12.880	28.92	82.50	514.3	0.08123	0.05824	0.06195	0.02343	0.1566
7	B	B	B	924934	10.290	27.61	65.67	321.4	0.09030	0.07658	0.05999	0.02738	0.1593
8	B	B	B	924964	10.160	19.59	64.73	311.7	0.10030	0.07504	0.005025	0.01116	0.1791
9	B	B	B	925236	9.423	27.88	59.26	271.3	0.08123	0.04971	0	0.00000	0.1742
10	B	M	B	925277	14.590	22.68	96.39	657.1	0.08473	0.13300	0.1029	0.03736	0.1454
11	B	B	B	925291	11.510	23.93	74.52	403.5	0.09261	0.10210	0.1112	0.04105	0.1388
12	B	M	M	925292	14.050	27.15	91.38	600.4	0.09929	0.11260	0.04462	0.04304	0.1537
13	B	B	M	925311	11.200	29.37	70.67	386.0	0.07449	0.03558	0	0.00000	0.1060
14	M	M	M	925622	15.220	30.62	103.40	716.9	0.10480	0.20870	0.255	0.09429	0.2128
15	M	M	M	926125	20.920	25.09	143.00	1347.0	0.10990	0.22360	0.3174	0.14740	0.2149
16	M	M	M	926424	21.560	22.39	142.00	1479.0	0.11100	0.11590	0.2439	0.13890	0.1726
17	M	M	M	926682	20.130	28.25	131.20	1261.0	0.09780	0.10340	0.144	0.09791	0.1752
18	M	M	M	926954	16.600	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590
19	M	M	M	927241	20.600	29.33	140.10	1265.0	0.11780	0.27700	0.3514	0.15200	0.2397

Figure 4. Prediction Results

b. Evaluation Results with Test and Score

1) Testing with Number of Folds 2

The results of the tests carried out on a set of test data with 1 attribute that is the target in this case the diagnosis. In the tests that have been carried out the sampling used is Cross Sapling and with a Number of Folds of 2, 5 and 10. For testing using Number of Folds 2, the evaluation results from the Random Forest model with AUC = 0.985, accuracy = 0.946, F1 = 0.945 , precision = 0.945 and recall = 0.946. And for the Naïve Bayes model with AUC = 0.983, accuracy = 0.944, F1 = 0.944, precision = 0.944 and recall = 0.944. Meanwhile for AdaBoost with AUC = 0.906, accuracy = 0.910, F1 = 0.910, precision = 0.911 and recall = 0.910. Figure 5 is the result of testing using Number of Folds 2.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.985	0.946	0.945	0.945	0.946
Naive Bayes	0.983	0.944	0.944	0.944	0.944
AdaBoost	0.906	0.910	0.910	0.911	0.910

Figure 5. Number of Folds 2

Figure 6 is a Confusion Matrix from the test results using the Random Forest model for the actual data successfully classified were B = 344 and M = 194, while those that failed to be classified were B = 13 and M = 18.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2912

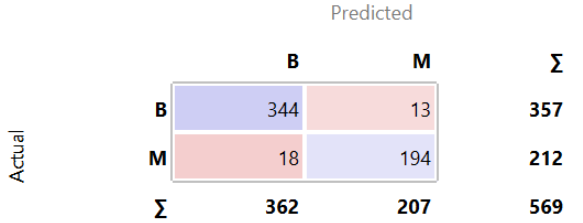


Figure 6. Confusion Matrix with Random Forest Model.

Figure 7 is a Confusion Matrix from the test results using the Naïve Bayes model for the actual data successfully classified were B = 340 and M = 197, while those that failed to be classified were B = 17 and M = 15.

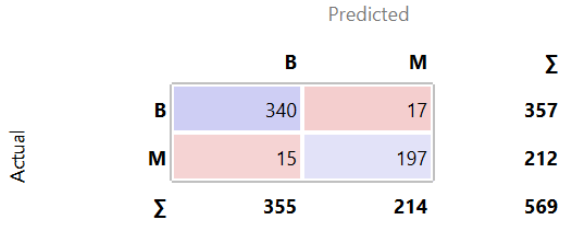


Figure 7. Confusion Matrix with Naïve Bayes Model

Figure 8 is a Confusion Matrix from the test results using the Naïve Bayes model, for the actual data successfully classified were B = 330 and M = 188, while those that failed to be classified were B = 27 and M = 24.

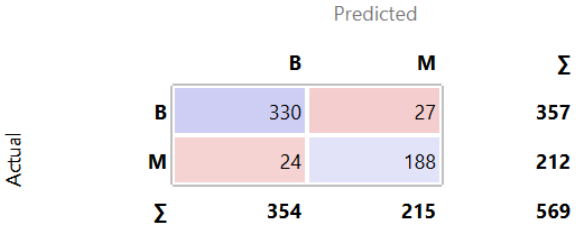


Figure 8. Confusion Matrix with Adaboost. Model

2) Testing with Number of Folds 5

For testing using Number of Folds 5, the evaluation results from the Random Forest model with AUC = 0.987, accuracy = 0.951, F1 = 0.951, precision = 0.951 and recall = 0.951. And for the Naïve Bayes model with AUC = 0.983, accuracy = 0.937, F1 = 0.937, precision = 0.937 and recall = 0.937. Meanwhile for AdaBoost with AUC = 0.911, accuracy = 0.916, F1 = 0.916, precision = 0.916 and recall = 0.916. Figure 9 is the result of testing using Number of Folds 5.

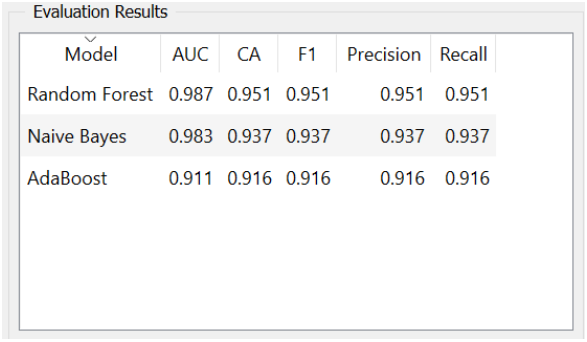


Figure 9. Number of Folds 5

Figure 10 is a Confusion Matrix from the test results using the Random Forest model for the actual data successfully classified were B = 343 and M = 198, while those that failed to be classified were B = 14 and M = 14.

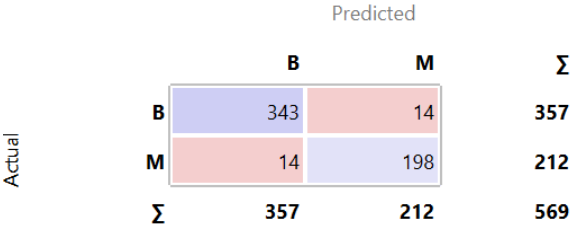


Figure 12. Confusion Matrix with Random Forest Model

Figure 11 is the Confusion Matrix from the test results using the Random Forest model, and the actual data successfully classified were B = 338 and M = 195, while those that failed to be classified were B = 19 and M = 17.

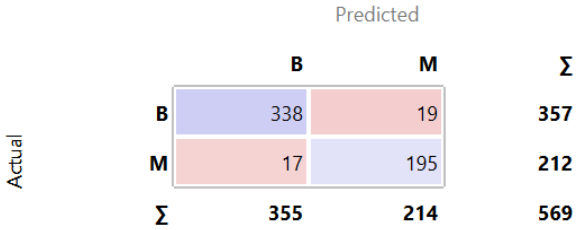


Figure 11. Confusion Matrix with Naïve Bayes Model

Figure 12 is the Confusion Matrix from the test results using the Random Forest model, and the actual data successfully classified were B = 332 and M = 189, while those that failed to be classified were B = 25 and M = 23.

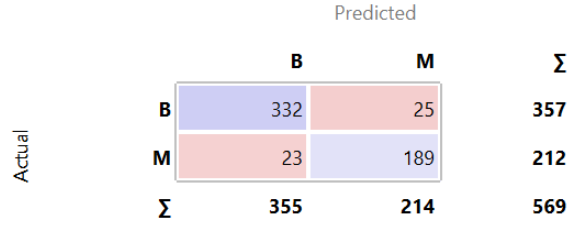


Figure 12. Confusion Matrix with Adaboost Model

3) Testing with Number of Folds 10

For testing using Number of Folds 10, the evaluation results from the Random Forest model with AUC = 0.984, accuracy = 0.947, F1 = 0.947, precision = 0.947 and recall = 0.947. And for the Naïve Bayes model with AUC = 0.983, accuracy = 0.940, F1 = 0.940, precision = 0.940 and recall = 0.940. Meanwhile for AdaBoost with AUC = 0.906, accuracy = 0.912, F1 = 0.912, precision = 0.912 and recall = 0.912. Figure 13 is the result of testing using Number of Folds 10.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.984	0.947	0.947	0.947	0.947
Naive Bayes	0.983	0.940	0.940	0.940	0.940
AdaBoost	0.906	0.912	0.912	0.912	0.912

Figure 13. Number of Folds 10

Figure 14 is the Confusion Matrix from the test results using the Random Forest model, for the actual data successfully classified were B = 347 and M = 192, while failed to be classified were B = 10 and M = 20.

Predicted

Actual	B	M	Σ
B	347	10	357
M	20	192	212
Σ	367	202	569

Figure 14. Confusion Matrix with Random Forest Model.

Figure 15 is a confusion matrix from the test results using the Naïve Bayes model for the actual data successfully classified was B = 339 and M = 196, while those that failed to be classified were B = 18 and M = 16.

Predicted

Actual	B	M	Σ
B	339	18	357
M	16	196	212
Σ	355	214	569

Figure 15. Confusion Matrix with Naïve Bayes Model.

Figure 16 is a Confusion Matrix from the test results using the AdaBoost model for the actual data successfully classified was B = 332 and M = 187, while those that failed to be classified were B = 25 and M = 15.

Predicted

Actual	B	M	Σ
B	332	25	357
M	25	187	212
Σ	357	212	569

Figure 16. Confusion Matrix with Adaboost Model.

CONCLUSION

From the results of the classification tests carried out, using three models from Data Mining, namely Random Forest, Naïve Bayes, and AdaBoost, the results obtained that the Random Forest model produces better and consistent results in classifying benign and malignant breast cancer using either Prediction or Test and Score. Prediction results using the Prediction widget get results with Random Forest model = 100%, Naïve Bayes model = 80% and AdaBoost model = 80%. The Number of Folds used in this study are 2, 5 and 10, using Number of Folds 2 to get Random Forest results for Accuracy = 95%, Precision = 95% and Recall = 95%, for the Naïve Bayes model with Accuracy value = 93%, Precision = 93% and Recall 93%, while for AdaBoost the Accuracy = 90%, Precision = 90% and Recall = 90%. Using Number of Folds 5, the results are Random Forest = 96%, precision = 96%, and Recall 96%. For the Naïve Bayes model, Accuracy = 94%, Precision = 94% and Recall = 94%, while for the AdaBoost model, Accuracy = 93%, Precision = 93% and Recall = 93%. Using the Number of Folds 10 to get the results, the Random Forest model = 96%, precision = 96%, and Recall 96%. For the Naïve Bayes model, Accuracy = 94%, Precision = 94% and Recall = 94%, while for the AdaBoost model, Accuracy = 92%, Precision = 92% and Recall = 92%. Results Based on the accuracy obtained, the use of 29 attribute data for benign and malignant cancers gave good results. As a suggestion for further development, the prediction and classification models used are more diverse by adding a more varied Number of Folds and using different data pre-processing methods to get maximum results.

REFERENCE

Abd-Elrazek, M. A., Othman, A. A., Abd Elaziz, M. H., & Abd-Elwhab, M. N. (2018). Intelligent

Prediction of Breast Cancer: A Comparative Study. *Egyptian Computer Science Journal*, *42*(3), 29–43.

Anjum, A. (2019). Role of Machine Learning in Diagnosis of Breast Cancer. *International Journal of Innovative Science and Research Technology*, *4*(5), 280–284.

Ayyoubzadeh, S. M., Almasizand, A., & ... (2021). Early Breast Cancer Prediction Using Dermatoglyphics: Data Mining Pilot Study in a General Hospital in Iran. *Health Education and Health Promotion*, *9*(3), 279–285.

Bissanum, R., Chaichulee, S., Kamolphiwong, R., Navakanitworakul, R., & Kanokwiroon, K. (2021). Molecular classification models for triple negative breast cancer subtype using machine learning. *Journal of Personalized Medicine*, *11*(9). https://doi.org/10.3390/jpm11090881

BV, E. (2019). Application of Data Mining Techniques to Predict Breast Cancer. *Procedia Computer Science*.

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, *12*(2), 119–126. https://doi.org/10.1177/1748301818756225

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*, *19*(1), 1–17. https://doi.org/10.1186/s12911-019-0801-4

Gupta, A., & Kaushik, B. N. (2018). Feature selection from biological database for breast cancer prediction and detection using machine learning classifier. *Journal of Artificial Intelligence*, *11*(2), 55–64. https://doi.org/10.3923/jai.2018.55.64

Jean Sunny, Nikita Rane, Rucha Kanade, & Sulochana Devi. (2020). Breast Cancer Classification and Prediction using Machine Learning. *International Journal of Engineering Research And*, *V9*(02), 576–580. https://doi.org/10.17577/ijertv9is020280

Kaur, E. R., & Chopra, V. (2015). Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining. *International Journal of Adanced Research in Computer and Communication Engineering*, *4*(7), 306–311. https://doi.org/10.17148/IJARCCE.2015.4771

Khan, I., Gandhi, A., Parmar, N., & Garg, B. (2020). Breast cancer prediction using data mining. *International Journal of Scientific Research and Engineering Development*, *3*(2), 978–980.

Kharya, S., & Soni, S. (2016). Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection. *International Journal of Computer Applications*, *133*(9), 32–37. https://doi.org/10.5120/ijca2016908023

Kodati, S., & Vivekanandam, R. (2018). Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Global Journal of Computer Science and Technology: C Software & Data Engineering*, *18*(1), 16–22.

Krishna, M. H., & Rao, D. K. N. (2018). PREDICTION OF BREAST CANCER USING MACHINE LEARNING TECHNIQUES. *International Journal of Management, Technology And Engineering*, *8*(12), 150–153. https://doi.org/10.2174/2213275912666190617160834

Kukasvadiya, M. S., Nidhi, D., & Divecha, H. (2017). Analysis of Data Using Data Mining tool Orange. *International Journal of Engineering Development and Research*, *5*(2), 1836–1840.

Kumari, G. T. P., & Rani, M. U. (2017). A Study of AdaBoost and Bagging Approaches on Student Dataset. *Journal of Advanced Engineering and Science*, *2*(2), 375–380.

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS ONE*, *16*(4 April), 1–23. https://doi.org/10.1371/journal.pone.0250370

Manimannan, G., Priya, R. L., & Kumar, C. A. (2019). Application of Orange Data Mining Approach of Argiculture Productivity Index Performance in Tamilnadu. *International Journal of Scientific and Innovative Mathematical Research*, *7*(8), 8–16. https://doi.org/10.20431/2347-3142.0708003

Masood, H. (2021). Breast Cancer Detection Using Machine Learning Algorithms. *International Research Journal of Engineering and Technology (IRJET)*, *8*(2), 738–747. https://doi.org/10.1007/978-981-16-6309-3_34

More, A., Mhatre, S., Kamble, V., Patil, V., & Bhairnallykar, S. (2022). Breast Cancer Prediction Using Classification Techniques of Machine Learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *10*(1). https://doi.org/10.26483/ijarcs.v10i5.6464

Nindrea, R. D., Aryandono, T., Lazuardi, L., & Dwiprahasto, I. (2018). Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: A meta-

analysis. *Asian Pacific Journal of Cancer Prevention*, *19*(7), 1747–1752. https://doi.org/10.22034/APJCP.2018.19.7.1747

Nitasha. (2019). Review on Breast Cancer Prediction Using Data Mining Algorithms. *International Journal of Computer Science Trends and Technology (IJCST)*, *7*(4), 42–45.

Octaviani, T. L., & Rustam, Z. (2019). Random forest for breast cancer prediction. *AIP Conference Proceedings*, *2168*(November). https://doi.org/10.1063/1.5132477

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*, *82*(March), 115–121. https://doi.org/10.1016/j.procs.2016.04.016

Rajendran, K., Jayabalan, M., & Thiruchelvam, V. (2020). Predicting breast cancer via supervised machine learning methods on class imbalanced data. *International Journal of Advanced Computer Science and Applications*, *11*(8), 54–63. https://doi.org/10.14569/IJACSA.2020.0110808

Rawal, R. (2020). BREAST CANCER PREDICTION USING MACHINE LEARNING. *Journal of Emerging Technologies and Innovative Research (JETIR)*, *7*(5), 13–24. https://doi.org/10.2478/acss-2020-0018

Sasikala, R. (2017). a Comparative Analysis for Smart Water Resource Using Data Mining Tools. *International Journal of Research - GRANTHAALAYAH*, *5*(7(SE)), 24–30. https://doi.org/10.29121/granthaalayah.v5.i7(se).2017.2039

Singh, H. (2021). Breast Cancer Analysis and Prediction by Using Machine Learning. *International Journal of Research in Engineering and Science (IJRES)*, *9*(6), 69–73.

Swetha, K., & Ranjana, R. (2020). Breast Cancer Predication Using Machine Learning and Data Mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*(3), 610–615.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2912