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

Bahtiar Imran1; Hambali2; Ahmad Subki3; Zoeniah4; Ahmad Yani5; Muhammad Rijal Alfian6

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

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

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
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, Naïve 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 SM (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 dataset 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

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

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.
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; Ganggaya 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>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>True Positives (TP)</td>
</tr>
<tr>
<td>M</td>
<td>False Positives (FP)</td>
</tr>
</tbody>
</table>

Accuracy is measured using (Ayyoubzadeh et al., 2021; Bissam et al., 2021; Churasia et al., 2018):

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

Precision is measured using (Abd-Elrazek et al., 2018; Bissam 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.

![Figure 4. Prediction Results](image)

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.

![Figure 5. Number of Folds 2](image)

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.
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 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 $M = 24$.

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 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 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 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.

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.

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.

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.

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


