Available *online* at: http://jarpet.ft.unand.ac.id/

Jurnal Andalas: Rekayasa dan Penerapan Teknologi

ISSN (Online) 2797-9024

Leveraging Naive Bayes Classification for Early Detection of Breast Cancer: A Data-Centric Diagnostic Approach

Baik Budi [1], *Refki Budiman* [2], *Queen Hesti Ramadhamy* [3]

[1,2,3] *Department of Electrical Engineering, Andalas University, Padang, Indonesia*

ARTICLE INFORMATION

Received: April, 19, 2025
Revised: April, 30, 2025
Available online: June, 01, 2025

KEYWORDS

Naïve Bayes, K-Means, Cancer, Machine, Learning

CORRESPONDENCE

E-mail: baikbudi@eng.unand.ac.id

ABSTRACT

This study aims to develop a breast cancer detection model using two distinct approaches: the Naive Bayes algorithm for classification and the K-Means algorithm for clustering. The methodology involves the collection of diagnostic clinical feature data, data preprocessing for normalization, and the separate training and evaluation of each model. Naive Bayes is employed to classify breast cancer as malignant or benign based on training and testing datasets, while K-Means is applied to unlabeled data as an additional analytical method. The performance of the Naive Bayes classifier is assessed using a confusion matrix, whereas the clustering results from K-Means are evaluated based on cluster validity metrics. The results indicate that Naive Bayes achieves a high level of accuracy (93%) in breast cancer classification, while K-Means offers additional insights through data pattern clustering. Together, these approaches demonstrate potential to effectively support the medical diagnostic process.

INTRODUCTION

According data from the World Health Organization (WHO), breast cancer accounts for approximately 25% of all cancer cases among women [1]. Despite advancements in the medical field that have introduced various early detection methods—such as mammography and biopsy—achieving fast and accurate diagnosis remains a significant challenge, particularly in regions with limited medical infrastructure. One of promising approach to improving detection accuracy involves the use of artificial intelligence techniques, particularly machine learning algorithms. The Naive Bayes algorithm, a simple yet effective classification method, has been widely utilized across various domains to categorize data into specific classes. In the context of breast cancer detection, this algorithm can be applied to distinguish between benign and malignant breast cancer cases based on available clinical diagnostic data [2].

In addition to Naive Bayes, the K-Means algorithm—an unsupervised learning technique for clustering—also holds significant potential in breast cancer data analysis. K-Means can group diagnostic data into clusters based on feature similarity, which is can help to uncover patterns that are difficult to detect through manual analysis. By clustering the data into benign and malignant groups, K-Means also can provide an initial overview of data distribution, which can be used as a foundation for further analysis using classification algorithms such as Naive Bayes.

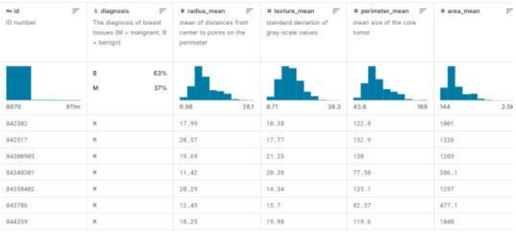
Previous studies have highlighted the strengths of both algorithms in medical data analysis. For instance, research by Smith et al. reported that Naive Bayes achieved a classification accuracy of up to 85% in breast cancer detection [3]. Meanwhile, a study C.K. Putra (2022) demonstrated that K-Means can effectively identify data patterns with high clustering accuracy, especially when preceded by proper data preprocessing steps such as feature normalization and dimensionality reduction [4]. The primary challenge that remains is how to integrate K-Means clustering results with classification models to further enhance overall prediction accuracy.

METHOD

This study aims to develop a breast cancer detection system using the Naive Bayes algorithm, applied to a dataset obtained from Kaggle. The research process is carried out through five main stages: data collection, data exploration, data preprocessing, model training, and model evaluation. Each stage is described in detail in the methodology section.

https://doi.org/10.25077/jarpet.v5i1.114

1. **Data Collection**
   The dataset used in this study was sourced from Kaggle [5], a platform that provides various datasets for research purposes. This dataset contains medical data related to breast cancer, including various features such as tumor size, tumor shape, texture, and cancer classification (benign or malignant).



id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
ID number	The diagnosis of breast tissues (M = malignant, B = benign)	mean of distances from center to points on the perimeter	standard deviation of gray-scale values	mean size of the core tumor	
8670 911m	B 63% M 37%	6.98 28.1	9.71 39.3	43.8 189	144 2.5k
842302	M	17.99	10.38	122.8	1001
842517	M	20.57	17.77	132.9	1326
84300903	M	19.69	21.25	130	1203
84348301	M	11.42	20.38	77.58	386.1
84358402	M	20.29	14.34	135.1	1297
843786	M	12.45	15.7	82.57	477.1
844359	M	18.25	19.98	119.6	1040

Figure 1. Breast Cancer Dataset from Kaggle

   The dataset is labeled, with each instance indicating whether the tumor is benign or malignant. Benign tumors are denoted by the letter "B," while malignant tumors are labeled with the letter "M." The data collection process involves downloading the relevant dataset and verifying it to ensure accuracy and completeness before proceeding with further processing.

2. **Exploratory Data Analysis (EDA)**
   Once the data is successfully collected, the next step is data exploration to understand the structure and characteristics of the dataset. During this phase, the data is analyzed to identify missing values, outliers, and the distribution of the features. Data visualization is performed using various graphs such as histograms, box plots, and scatter plots to illustrate feature distributions and relationships between them. This step aims to gain initial insights into the data and detect issues that need to be addressed, such as class imbalance or incomplete data.

3. **Data Preprocessing**
   After the exploration phase, the analyzed data is processed to prepare it for use in the model. The data preprocessing involves several steps, including:
   a. **Handling Missing Data:** If missing values are identified in the dataset, the first step is to perform imputation to replace the missing values. Imputation techniques may include mean or mode imputation for numerical data, or other methods such as interpolation for more complex values.
   b. **Normalization and Standardization:** To ensure better performance of machine learning algorithms, features with different scales need to be normalized or standardized. Normalization is used to scale the feature values to a specific range, while standardization adjusts the data to have a mean of 0 and a standard deviation of 1.
   c. **Encoding Categorical Data:** For categorical features, such as diagnosis type, encoding is applied to convert them into numerical formats that can be processed by machine learning algorithms, for example, using label encoding techniques.
   d. **Data Splitting:** The dataset is divided into two parts: a training set (80%) and a testing set (20%) using the train-test split method. This division ensures that the model is trained on one subset of the data and tested on a different subset, allowing for the assessment of the model's generalization ability *[6]*.

4. **Model Training**
   Once the data is prepared, the next step is to train the model using the Naive Bayes algorithm. Naive Bayes is a classification algorithm that assumes the features in the dataset are independent and applies Bayes' theorem to calculate the class probabilities based on the given features. The **GaussianNB** function is chosen because the features in the dataset exhibit a normal distribution.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure 2; Gaussian Naïve Bayes Formula

   In this case, we assume that the data used (features related to breast tumors) follows a normal or Gaussian distribution. **GaussianNB** is the specific implementation of Naive Bayes used when the features in the data adhere to a normal distribution.

Meanwhile, for the use of K-Means Clustering, the steps involved in K-Means are as follows:

1. **Data Collection:**
   Breast cancer diagnostic data is obtained from trusted sources, containing relevant clinical features for analysis. This data includes attributes such as tumor size, texture, and perimeter.
2. **Data Preprocessing:**
   The data is imported into RapidMiner [7] for processing.
3. **Application of the K-Means Algorithm:**

   a. The K-Means algorithm is applied using RapidMiner, with varying values of K for the number of clusters, ranging from 2 to 7.
   b. For each value of K, RapidMiner calculates the average distance between the data points and the centroids of each cluster.

4. **Result Collection:**
   The average distance results for each value of K are recorded in a table 1, as shown below:

Figure 1: Average distance based on K Value

K	Avg Distance
2	136.983
3	83.067
4	51.365
5	36.098
6	29.134
7	23.317

5. **Result Visualization:**
   a. The average distance data obtained from RapidMiner is imported into Microsoft Excel.
   b. A scatter plot is created, with the X-axis representing the K values and the Y-axis representing the corresponding average distances.
   c. The points on the graph are connected with lines to facilitate interpretation.
6. **Elbow Method Analysis:**
   a. The elbow method is used to determine the optimal value of K.
   b. The "elbow point" is identified as the point where the decrease in average distance begins to level off significantly.
   c. Based on the graph, the optimal value of K is selected for further analysis.

RESULTS AND DISCUSSION

Confusion Matrix

In this study, analysis was conducted using the Naive Bayes method with the GaussianNB model to classify the breast cancer dataset into two categories: malignant (cancerous) and benign (non-cancerous). After splitting the dataset using the train-test split technique—allocating 80% of the data for training and 20% for testing—the model was evaluated using a confusion matrix to measure its performance in terms of accuracy, precision, recall, and F1-score. Table 2 was presents the confusion matrix results of the GaussianNB model on the test dataset.

Table 2. Confusion Matrix of GaussianNB Model on Test Data

	Benign Prediction	Malignan Prediction
Benign Actual Label	1.00	0.00
Malignan Actual Label	0.07	0.93

In this study, the confusion matrix displays four key components: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). These components are essential for calculating various important evaluation metrics, such as accuracy, precision, recall, and F1-score. These metrics offer a comprehensive assessment of the model's effectiveness in correctly identifying and distinguishing between different classes. The detailed components of the confusion matrix are presented in Table 3.

	Benign Prediction	Malignan Prediction
Benign Actual Label	TP	FN
Malignan Actual Label	FP	TN

Based on the resulting confusion matrix, the model demonstrates excellent performance in classifying data labeled as *Benign*. This is evidenced by the perfect accuracy score of 1.00 in the *Benign* prediction column for the actual *Benign* label, indicating that all instances with this label were correctly classified by the model (*True Positive*, TP). There were no cases where *Benign* data was misclassified as *Malignant* (*False Negative*, FN).

https://doi.org/10.25077/jarpet.v5i1.114

For the *Malignant* label, the model also shows reasonably good performance, although with a slight misclassification. Approximately 7% of the data labeled as *Malignant* were incorrectly predicted as *Benign* (*False Positive*, FP), while the remaining 93% were correctly identified as *Malignant* (*True Negative*, TN).

Overall, this confusion matrix illustrates that the classification model is highly effective, particularly in identifying the *Benign* class with a very low error rate. This is especially important in the context of medical diagnosis, where accurate differentiation between *Benign* and *Malignant* conditions is critical. However, the 7% misclassification rate in *Malignant* cases suggests that there is still room for improvement, especially in detecting more critical conditions.

K-Means Clustering

In this study, an analysis was conducted using the K-Means clustering method to group a breast cancer dataset into several clusters based on similarities in diagnostic features. The dataset was processed using the K-Means algorithm, where the number of clusters (K) was varied from 2 to 7 in order to determine the optimal cluster count. The analysis began by utilizing RapidMiner software to compute the average distance of data points to their respective centroids. The average distance values for each K were then input into Microsoft Excel, and a graph was plotted to identify the optimal K value using the Elbow Method. This graph is presented in Figure 3.

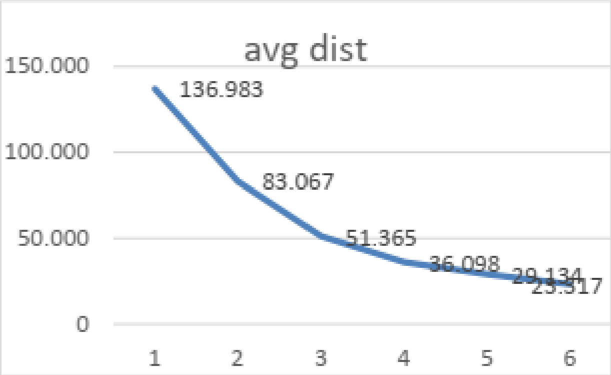


Figure 3. Average Distance

The K-Means method was employed to cluster the diagnostic data, aiming to gain deeper insight into the structure of the dataset. To determine the optimal number of clusters, the *elbow method* was applied. Based on the elbow plot, the optimal number of clusters was found to be K = 4. This point represents where the reduction in the average distance to the cluster centroid begins to level off, indicating an optimal trade-off between model complexity and clustering efficiency. The following visualization shows the decreasing average distance up to K = 4:

CONCLUSIONS

Based on the analysis of the confusion matrix, it can be concluded that this classification model demonstrates excellent performance in detecting the *Benign* class, with an accuracy rate of 100%. All instances labeled as *Benign* were correctly classified as *Benign* (*True Positive*, TP), with no misclassifications into the *Malignant* class (*False Negative*, FN). This indicates that the model is highly efficient in identifying *Benign* conditions.

However, for the *Malignant* class, the model performs reasonably well, with 93% of the data correctly identified as *Malignant* (*True Negative*, TN). Despite this, a 7% misclassification rate exists, where *Malignant* data were mistakenly classified as *Benign* (*False Positive*, FP). While this error rate is relatively low, it still highlights the potential for improvement.

The K-Means method was applied to cluster the data based on the available diagnostic features. According to the Elbow Method analysis, K = 4 was chosen as the optimal number of clusters, indicating a fairly effective partitioning of the data into several groups. However, although the resulting clusters provide a reasonable overview, some data points did not fully align with the desired clusters, suggesting some imperfection in the data separation process. Additionally, the K-Means model does not provide explicit classification results like Naive Bayes but can be used to uncover patterns in the data and group similar instances together.

To improve the model's performance, it is recommended to increase sensitivity to the *Malignant* class. One potential approach is to adjust the decision threshold or employ oversampling techniques for the *Malignant* data, allowing the model to focus more on detecting *Malignant* instances that might be overlooked. Furthermore, applying more advanced regularization techniques could help reduce bias towards the *Benign* class.

In conclusion, while this model has demonstrated strong performance, particularly in detecting *Benign* cases, efforts are needed to reduce misclassifications within the *Malignant* class in order to achieve more optimal results. It is also advisable to use more comprehensive evaluation metrics, such as precision, recall, and F1-score, to better understand the trade-offs between detecting more *Malignant* cases and minimizing the misclassification of *Benign* instances.

.

REFERENCES

[1] Organization, World Health, “https://www.who.int/news-room/fact-sheets/detail/breast-cancer,” 2 11 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] J. Smith, “Application of Naive Bayes for breast cancer classification,” *Journal of Medical Informatics,* vol. 35, no. No, 2, pp. 123-135, 2020.

[3] S. White, “Evaluation of machine learning techniques in early cancer diagnosis,” *Medical Journal of Artificial Intelligence,* vol. 29, pp. 32-45, 2022.

[4] A. Putra, “Increase Accuracy of Naïve Bayes Classifier Algorithm with K-Means Clustering for Prediction of Potential Blood Donors,” *Journal of Advances in Information Systems and Technology,* vol. 4, pp. 42-49, 2022.

[5] Kaggle, “Breast Cancer Wisconsin (Diagnostic) Data Set,” [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. [Diakses 2 11 2024].

[6] G. S. Gupta P, “Breast Cancer Prediction Using Varying Parameters of Machine Learning Models,” *Procedia Computer Science,* vol. 171, pp. 593-601, 2020.

[7] “https://altair.com/altair-rapidminer,” Altair, [Online]. Available: https://altair.com/altair-rapidminer. [Diakses 02 11 2024].

https://doi.org/10.25077/jarpet.v5i1.114