

***Classification of Breast Cancer (C50) Using the Naïve Bayes Algorithm
at Level III Baladhika Husada Hospital Jember***

Mudafiq Riyan Pratama, S.Kom., M.Kom. (Supervisor)

Sefia Ayu Maharani

Health Information Management Study Program

Department of Health

ABSTRACT

Breast cancer (C50) is one of the most common cancers and a leading cause of cancer-related mortality in Indonesia. Most cases are diagnosed at an advanced stage, highlighting the need for methods to identify the stage of the disease quickly and accurately. This study aimed to analyze the classification of breast cancer (C50) stages using the Naïve Bayes algorithm at Baladhika Husada Level III Hospital, Jember. This study employed a quantitative approach using a data mining method. A total of 476 medical records of breast cancer patients from 2024 to 2025 were analyzed using Google Colab. The study variables included age, family history of breast cancer, history of previous breast disease, body mass index (BMI), breast pain, breast lump, axillary lump, changes in breast shape and size, nipple discharge, nipple changes, breast skin abnormalities, breast cancer localization status, and tumor size. Model performance was evaluated using a confusion matrix. A 90%:10% training-testing data split achieved an accuracy of 87.50%, a precision of 86.36%, and a recall of 86.36%. The likelihood analysis showed that advanced-stage breast cancer was more frequently characterized by a tumor size greater than 5 cm, axillary lump, changes in breast shape and size, nipple changes, and breast skin abnormalities, whereas early-stage breast cancer was more frequently characterized by a smaller tumor size with clinical findings that had not indicated more extensive disease spread. The Naïve Bayes algorithm demonstrated good performance in classifying early-stage and advanced-stage breast cancer. The resulting model has the potential to support data-driven identification of breast cancer stages and can be further developed into a clinical decision support system. Future studies are recommended to use larger datasets, incorporate additional relevant variables, and develop systems based on Optical Character Recognition (OCR) and Natural Language Processing (NLP).

Keywords: Breast Cancer, Naïve Bayes, Data Mining, Classification, Cancer Stage