Journal of Applied Probability and Statistics
2021, Vol. 16, No. 3, pp. 01-17
Copyright ISOSS Publications

BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION WITH NAÏVE BAYES CLASSIFIER: A CASE STUDY IN INDONESIAN BREAST CANCER PATIENTS

Sarini Abdullah[1], Nabilla Fauziyyah[1], Siti Nurrohmah[1] and Andhika Rachman[2]
[1] *Mathematics Department, University of Indonesia, Kampus UI Depok, 16424, Indonesia*
[2] *Department of Internal Medicine, Faculty of Medicine, University of Indonesia, Indonesia*
Email: sarini@sci.ui.ac.id

SUMMARY

Naïve Bayes Classifier (NBC) is one of the most popular machine learning methods due to its simplicity yet great overall accuracy. Previous studies stated a number of claims and reasons as to why this method is suitable to apply in analysing medical data. While there is abundant literature which cover the application of NBC on medical data, there is little to none which covers its application on medical data in Indonesia, especially Indonesian breast cancer data. The main goal of this study is to apply NBC in classifying 101 breast cancer patients in private hospital in Indonesia into five classes of molecular subtypes, assess its accuracy, and compare its performance with another popular classification model, Decision Tree (DT). Results showed that NBC outperformed DT, predicting the classes of most patients in the test set correctly by having an overall accuracy of 85.7%. NBC also could identify all patients in the human epidermal growth factor receptor 2 (HER2)-Enriched and Triple Negative subtypes faultlessly. Another important result to be noted is that oestrogen (ER), progesterone (PR), and HER2 statuses are dominant factors in differentiating subtypes between each patient, matching the official guide made by the breast cancer expert panel. Several empirical results were also found regarding NBC: it does not need a large data to produce high overall accuracy and it could still predict the classes of most patients correctly even with the presence of missing and noisy data.

Keywords and phrases: Bayes rule; breast cancer subtypes; classification; decision tree; and posterior class probability.

2020 Mathematics Subject Classification: Primary 62P10, secondary 62H30.