Data Analysis of Early Detection and Clinical Stages of Breast Cancer in Libya

H Saad and N Nagarur

Department of Systems Science and Industrial Engineering,
Binghamton University
Binghamton, NY

Corresponding Author's email: hsaad1@binghamton.edu

Author Note: Hamza Saad is at present pursuing his doctoral degree in the department of Systems Science and Industrial Engineering, Binghamton University. Nagen Nagarur is a professor of Systems Science and Industrial Engineering in the Department of Systems Science and Industrial Engineering, Binghamton University. His research interests include supply chains, operations management, and healthcare delivery.

Abstract: Going for testing can help detect the cancer at an early stage, increasing the chance of surviving the disease. However, such instances of going for early testing are relatively in developing countries. This study deals with 130 cases of women suffering from breast cancer in a Libyan hospital. The analysis is carried out two steps. In the first step, socio-behavioral attributes such as age, education, marital status and healthy living are used to predict if a patient would go for early testing. In the next step, physical attributes like age, tumor size, and lymph node LN status are used as input variables, with clinical grade or stage as the output variable. The data of the 130 cases is used for the analysis. Five different algorithms, namely, J48, Bayes Net, K Star, Logistic Regression, and Support Vector Machine are used for both the stages. The response variable for the first step of early testing is the outcome of the results– positive or negative. The second response variable is categorized into four groups, based on the clinical grade of the disease. For the early testing results, there is not too much variation in the accuracies, with J48 scoring the highest level of accuracy at 68%. The prediction accuracies are not high, with the average of the five algorithms just above 65%. For the next step, the prediction accuracies are higher across the algorithms, with J48 again leading at 87%. The accuracy for clinical stage 4 is the highest among the four clinical stages.

Keywords, Early Breast Cancer Detection, Clinical Stages, Data Mining algorithms, Classification.