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: This study involves various factors that affect breast cancer in women, negatively or positively. Data has been collected about various biological and sociological factors and the clinical stage at which the disease was collected. Mining association rules is a useful technique to extract the critical relationship between variables by generating some rules used to build prediction. However, association rules did not give useful results or supported rules for this type of data. Hence a decision tree approach is applied in the study. A detailed tree graph including 19 nodes and 10 leaves was utilized to construct rules based on each decision node. Each leaf node in the tree has a value of accuracy; the highest accuracy is selected to confirm the strong rules. From the 10 rules, only five rules were selected to build prediction. The process of rules selection is supported by the accuracy of each leaf and impact of each variable in the output. Predictor screening was used to predict robust variables based on P-value and Chi-square value. Tumor size variable exists in all extracted rules with a strong impact.

Keywords. Classification, Decision tree, Predictor screening, Rules extraction. Breast cancer.

1. Introduction

Breast cancer is a major health problem among women worldwide, it is one of the most known malignancies with 23% of all cancers, with over one million new cases diagnosed annually (American Cancer Society: Cancer Facts and Figures 2010, Sabratha Cancer Registry, 2008, and Siegel et al., 2012). Roughly 4.4 million women are living with the breast cancer, and more than 400,000 die from the disease annually. This disease is responsible for 14% of all cancer deaths (Ermiah et al., 2012). It is the most common cause of female death in industrialized countries (Ermiah, 2013), the second most common cause in the world, and the third most common in developing countries (American Cancer Society, 2011). If breast cancer is discovered early, then there is a better chance of survival. Despite development in treatment strategies, advanced breast cancer remains incurable and the goals of therapy range from symptom palliation to extending survival.

In developing countries, the administration of breast cancer confronts significant medical, social and economic problems. Despite advances in the treatment, the mortality average is still high. Therefore, it is necessary to secure a good cancer control by applying different strategies such as improved early detection and find the prognostic variables, which applied with traditional factors can predict the output of the individual patient and allow selection of appropriate therapy (Ahmad, 2013).

Many studies such as Ermiah, 2013, focused on early detection which leads to getting early treatment for the disease, resulting sometimes in the cure, however, there is also a possibility that the patient would enter into more critical stages in spite of treatment. Early detection attribute does not give important correlation when applied as output in the classification because the data is collected based on the clinical stages for each patient. However, when early detection is applied as an input variable in the dataset, a strong relationship was detected between clinical stages (output) and early detection (input). Low accuracy does not help to predict disease or extract important variables which impact the disease.