A COMPARATIVE STUDY ON DATA MINING TECHNIQUES FOR BREAST CANCER SURVIVABILITY PREDICTION

Christian Zamiela, Haifeng Wang, Wenmeng Tian, and Linkan Bian

Industrial and Systems Engineering Department, Mississippi State University, Starkville, MS 39762

Corresponding Author: Christian Zamiela <Cez39@msstate.edu>

ABSTRACT

This research presents a comparative study of using common data mining techniques for breast cancer survivability prediction. Breast cancer is one of the leading cancer instances in the United States. The prediction of the survival rate can certainly help healthcare providers to understand the disease and generate prevention strategies to reduce breast cancer mortality. However, a clinical test is quite costly and time-consuming for breast cancer prognosis. It is also a tedious task to have a long-term monitoring process after breast cancer treatment. Data mining models are potential approaches that can provide early breast cancer detection and improve the efficiency of the whole diagnosis process. This research aims to apply data mining techniques to predict the survivability of patients after diagnosed as breast cancer. A comparative study is conducted for six classification models, i.e., linear discriminant analysis, logistic regression, naïve Bayes classifier, support vector machine, random forest, and gradient boosting on a practical large-scale dataset. The data is based on the breast cancer dataset from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program, which includes more than 80,000 instances over 40 years' collection. This research shows the potential of using advanced data mining techniques to solve practical breast cancer prediction problems. The outcome of this research could also be used for other disease diagnoses and benefits decision making in the healthcare process.

Keywords – SEER, Data mining, Breast cancer survivability, Classification methods

INTRODUCTION

Mathematical modeling in the health fields is a growing need to make a less costly diagnosis and provide a better measurement of a patient's health. Cancer research and treatment was funded with 3.52 billion dollars by the National Cancer Institute [7]. Breast cancer is the second leading cause of death for women in the United States. Nearly one-fifth of women in the United States die from cancer, and the need for treatments is growing. Additionally, cancer is more common in the elderly than any other population because our cells deteriorate over time. The understanding of cancer survivability is critical to prevent cancer death and reduce the mortality rate.

Accurate cancer survivability predictions are useful to insurance companies that are setting insurance rates. For example, actuaries are predicting whether a person will survive to their next birthday. Using math models to predict survivability within the next five years would help actuaries create more accurate insurance rates. Prediction models are useful to doctors that are following the progress of patients and