Prediction of Breast Cancer Recurrence using Classification Methods

Ade Adeoye[1]

[1] *Email: adewale.adeoye@ryerson.ca*

Abstract

Breast cancer is one of the leading causes of cancer-related deaths in the world. Its incidence is however quite biased. According to the Centers for Disease Control and Prevention (CDC), most cases of breast cancer are found in women, with estimates putting the likelihood at about a 100 times more likely than in males. For cancer-related diseases, early detection is crucial. If discovered early, a patient is highly likely to survive the disease, even if there may be a few cases of recurrence. This has given rise to a suite of data mining and machine learning efforts attempting to gain an advantage against this ailment by predicting the chances of a patient developing the disease, or if diagnosed already, the chances of the cancer remaining benign or turning malignant.

In this study, we examine a breast cancer data set obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia (https://archive.ics.uci.edu/ml/datasets/breast+cancer); and, applying a range of machine learning techniques, we predict the chances of recurrence of breast cancer in 286 patients. The essential part of the data mining procedure conducted is classification, where we categorize the sample according to its original binary split using logistic regression, mainly on the test data. As conclusive analysis, we also apply two additional classification methods, decision tree and naive-bayes, and conduct a comparison of the metrics of these classification models, examining how the different models perform on the sample data compared to each other.

Keywords: breast cancer, classification, logistic regression, decision tree, naive-bayes.