Predicting breast cancer using combined K-fold cross-validated decision tree models

Mohamed Salem *(msalem2@my.shu.ac.uk)* ***and Dr Kassim Mwitondi*** *(k.mwitondi@shu.ac.uk)*

Sheffield Hallam University, Faculty of Arts, Computing, Engineering and Sciences

Abstract

Different forms of cancer have been widely studied and documented in various studies across the world. However, there have not been many similar studies in the developing countries - particularly on the African continent (Parkin, et al., 2005). This paper seeks to uncover the geo-demographic occurrence patterns of the disease by applying decision tree models to learn the underlying rules in the overall behaviour of breast cancer. The data, 3,057 observations on 29 variables, obtained from four cancer treatment centres in Libya (2004-2008) were interrogated using multiple K-fold cross-validated decision tree models. The results from the selected optimal models exhibit greater accuracy and reliability as compared to using conventional decision models. The proposed strategy is therefore strongly recommended for use as a predictive tool in health and clinical centres to help minimise high costs of pathological tests. It is expected that the findings from this paper will provide an input into comparative geo-ethnic studies of cancer and provide informed intervention guidelines in the prevention and cure of the disease not only in Libya but also in other parts of the world.

Keywords

Breast cancer, cross validation, data mining, decision trees, over-fitting and risk factors