Machine Learning in Tumor Classification in Breast Cancer

Ana Sofia Lima[1], Carolina Coutinho[1], Raquel Machado[1*], Alexandra Alves Oliveira[1,2], Brígida Mónica Faria[1,2]

1. ESS, Polytechnic Institute of Porto, 4200-072 Porto, Portugal; 2. Artificial Intelligence and Computer Science Laboratory, 4150-181 Porto, Portugal

* Corresponding author email: 10231153@ess.ipp.pt

Introduction: Breast cancer is the primary cause of mortality among women worldwide (1). Discernible patterns can be found within the disease, presenting an opportunity for the application of machine learning (ML), garnering effective results in screening and diagnosis. **Objectives:** Different ML algorithms were tested – Decision Tree, Deep Learning (DL), k-Nearest Neighbors (k-NN) and Naïve Bayes – to construct a predictive model allowing the early classification of a breast tumor as benign or malignant, avoiding the need to proceed to a more invasive technique. **Methods:** The ML models were constructed and applied to a database of 201 individuals with breast cancer and descriptive attributes (e.g. age, tumor size, presence of invasive nodes) (2) by using RapidMiner Studio. The evaluation of the models was done by analyzing their accuracy, true negative (TNR) and true positive rates (TPR), their ROC (Receiver Operating Characteristic) curves and AUC (Area Under Curve). **Results:** During a first exploratory phase, fours clusters were detected: smaller tumor sizes, younger patients, and a benign diagnosis; older age, bigger tumor sizes and a malignant diagnosis; and two more with the opposite characteristics. These characteristics were later found to be important factors in the construction of the Decision Tree. When comparing the models accuracy, the best model was Naïve Bayes (91.04%), followed by the Decision Tree (90.55%), DL (90.02%) and k-NN (86.32%). There is a statistically significant difference between the performances of every model ($p<0.05$) except between the DL and the Decision Tree models. Naïve Bayes presented the highest TPR (98.21%) while DL presented the highest TNR (83.15%). The Decision Tree model presented the highest AUC (0.976), followed by Naïve Bayes (0.961). **Conclusions:** The Decision Tree model best achieved our goal by having the highest AUC which denotes an exceptional sensitivity rate, surpassing Naïve Bayes while maintaining a similar accuracy and TNR.

Keywords: machine learning, predictive models, breast cancer

References:

1. World Health Organization. Global breast cancer initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer: executive summary. Geneva; 2023. Available from: https://www.who.int/publications/i/item/9789240067134
2. Eteng I, Bisong E, Fagbola T, Ibrahim M, Udosen J, Akpotuzor S. UCTH Breast Cancer Dataset. Mendeley Data: V2; 2023. Available from: https://data.mendeley.com/datasets/63fpbc9cm4/2 doi: 10.17632/63fpbc9cm4.2