Research Exhibition in Mathematics and Computer Sciences (REMACS 6.0)/
ISBN: 978-629-97440-5-4

FEASIBLE FEATURES OF BREAST CANCER RECURRENCE (BCR) PATIENTS USING MACHINE LEARNING ALGORITHMS

Rusyada Ardina Rushdi and Balkiah Moktar
College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Perlis Branch
rusyada.ardina@gmail.com and balkiah@uitm.edu.my

ABSTRACT – Cancer is a disease in which cells in the body grow out of control and is consistently named for the parts of the body where it starts from the breast. BCR is breast cancer that returns after initial treatment and may occur within months or years. This study aims to identify the feasible feature in predicting BCR using four machine learning algorithms. The study utilized 10377 secondary data from the official statistic of the Ministry of Health and Medical Education and the Iran Cancer Research Center. Naïve Bayes (NB), Random Forest (RF), Gradient Boosted Tree (GBT), and Logistic Regression (LR) were utilized by using RapidMiner to obtain a good classifier's performance to be evaluated to determine the best model that can accurately predict the BCR, and the significant risk factors of BCR using the best model. The results show that the best model with the highest accuracy is GBT, with a ratio of 51%, and the most essential feature of this algorithm is radiotherapy.

Keywords: Breast cancer recurrence, machine learning algorithm, accuracy, RapidMiner

1. INTRODUCTION

According to World Health Organization, in 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths worldwide. This study aims to identify the feasible features that predict breast cancer recurrence (BCR) in Iran patients. Secondly, to predict the BCR and to identify the best model with the highest accuracy by incorporating the Naïve Bayes, Random Forest, Gradient Boosted Tree, and Logistic Regression using RapidMiner. This study can instill consciousness in the masses on the dire risk of BCR so women can take preventative measures to avoid spreading cancer.

2. METHODOLOGY

Data was retrieved from PLOS ONE and was acquired from official statistics of the Ministry of Health and Medical Education and the Iran Cancer Research Center, which consists of 10377 records of patients with BCR. The data consist of 15 features relating to BCR. The data mining approach was utilized in this study by undergoing the training and testing phase, the percentage of training is 70%, and testing is 30%. We evaluated the performance of Naïve Bayes, Random Forest, Gradient Boosted Tree, and Logistic Regression using RapidMiner. These will be considered on the accuracy, precision, recall, sensitivity, specificity, and F – measure.

3. RESULTS AND DISCUSSION

The best model with the highest accuracy is the Gradient Boosted Tree with a ratio of 51%; regardless of the attribute weights, *radiotherapy* has been selected as the feasible feature by all algorithms. Whereas *Chemotherapy, Estrogen Receptor Value, LN involvement rate,* and *result of biopsy of pathology* have been selected by three algorithms. Any algorithms do not select *Progesterone Receptor Value* and *tumor size*. However, this observation differs from the original dataset curators Mosayebi et al. (2020), whereby the best model is C5.0 with an accuracy of 81.90%, and the most crucial feature is the *Her2 value*.

4. NOVELTY OF RESEARCH / PRODUCT

Several studies and research projects have investigated and studied the best algorithm for predicting the probability of breast cancer recurrence. Lu et al. (2018) utilized machine learning techniques to examine predictive indicators for breast cancer recurrence using Wisconsin Prognostic Breast Cancer data, similar to Gracia – Murillas et al. (2012) and Rana et al. (2015), but the only difference is the machine learning algorithm which is Gradient Boosted Tree, Naïve Bayes, Random Forest, and Logistic Regression. However, this researcher does not include the most significant

features from the best algorithm model. This study used breast cancer recurrence data from official statistics of the Ministry of Health and Medical Education and the Iran Cancer Research Center. It used Naïve Bayes, Random Forest, Gradient Boosted Tree, and Logistic Regression to determine the best model with the highest accuracy and the most significant feature from the best algorithm.

5. CONCLUSION

In a nutshell, all the objectives are achieved in this study. The study benefits women who have completed breast cancer treatment, providing valuable information about the risk factors associated with BCR and serving as an essential resource for raising awareness among women.

REFERENCES

Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R. J., Cheang, M., Osin, P., Nerurkar, A., Kozarewa, I., Garrido, J. A., Dowsett, M., Reis-Filho, J. S., Smith, I. E., & Turner, N. C. (2012.). *Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer*. www.ScienceTranslationalMedicine.org

Lu, H., Wang, H., & Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Expert Systems with Applications*, *116*, 340–350. https://doi.org/10.1016/j.eswa.2018.08.040

Mosayebi, A., Mojaradi, B., Naeini, A. B., & Hosseini, S. H. K. (2020). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLoS ONE*, *15*(10 October). https://doi.org/10.1371/journal.pone.0237658

Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques. In *IJRET: International Journal of Research in Engineering and Technology*. http://www.ijret.org