Machine Learning Model for Risk of Breast Cancer Relapse

Aim:

The aim of this project is to analyse various factors and create a model that can determine the possibility of breast cancer relapse for a patient.

Objectives:

- One key objective is to determine key features related to recurrence of breast cancer. By determining these features, treatments can be tailored to focus on particular features.
- Another key step is knowing if the cancer might return. For this a predictive model is to be built to determine if breast cancer will reoccur in a patient.
- Related to predicting the reoccurrence is knowing how likely it is to reoccur, in the medical field nothing is without risk so knowing what the chances are pre-emptively will allow for tailored approach on what kind of treatment plan is to be followed.

Literature Review:

Breast cancer remains a deadly disease, even with all the recent technological advancements. Early intervention has made an impact, but an overwhelmingly large number of breast cancer patients still live under the fear of "recurrent" disease. Breast cancer recurrence is clinically a huge problem and one that is largely not well understood (Ahmad, 2013).

The challenges in managing breast cancer patients are very many. First, although many risk factors have been associated with the possible initiation and progression of disease, nothing concrete is established that can potentially prevent the primary disease or its progression and metastases. Additionally, there are well-studied disparities in breast cancer that include socioeconomic disparities as well as the racial disparities. All this information seems to suggest that no two women have equal chances of developing the disease. Even when comparing among breast cancer patients, there are not very reliable predictors of aggressiveness (Ahmad, 2013).

Recurrence of breast cancer leads to a high lifetime risk and a low 5-year survival rate. Researchers have tried predicting the risk of recurrence in patients with breast cancer, but the predictive performance remains controversial (Long, 2023).

Comprehensive breast cancer risk prediction models enable identifying and targeting women at high-risk, while reducing interventions in those at low risk. Breast cancer risk prediction models used in clinical practice have low discriminatory accuracy (0.53–0.64). Machine learning (ML) offers an alternative approach to standard prediction modelling that may address current limitations and improve accuracy of those tools. The purpose of this study was to compare the discriminatory accuracy of ML-based estimates against a pair of established methods—the Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models (Katapodi, 2019).

Breast cancer mortality is largely related to either resistance to therapies or metastases to distant organs, all of which contribute to recurrence. One factor that has greatly hampered our progress in the field is the absence of any acceptable model for the study of tumour recurrence (Ahmad, 2013).

Study demonstrates that patients with early-stage breast cancer who are disease free at 5 years after AST have a substantially increased residual risk of recurrence. Extended adjuvant endocrine therapy

is currently available only for postmenopausal patients with hormone receptor–positive disease, and these patients should be considered for treatment after careful evaluation of the risks and benefits. More research is needed to identify host and tumour characteristics that are associated with late breast cancer recurrences to individualize initial AST and extended adjuvant endocrine therapy (Abenaa M. Brewster, 2008).

Methodology:

Executive Summary -

- Collect data directly as dataset online due to time constraint.
- Wrangle data – by filtering the data, handling missing values and applying one hot encoding to prepare the data for analysis and modelling.
- Explore data via EDA data visualization techniques.
- Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters.

Data Collection:

- Dataset: (Anon., 2018)
- Breast cancer data was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

Data Processing:

- Missing values were handled by either dropping the row or replacing with mean.
- Data Encoding was done, One hot encoding with dummy trap method was performed on categorical variables and Label encoding was done for numerically significant variables like Tumor-Size, etc
- Data Imbalance was addressed by synthetically scaling the data using SMOTE library.

Model Building:

- Data was split into training and testing dataset.
- Grid Search was used with Cross Validation to find hyper parameters.
- Logistic Regression and Random Forest model were trained and Logistic Regression performed better by having lesser False Negatives.
- Logistic Regression had an accuracy of 68% and a recall rate of 78% for determining Recurrence.
- The Probability of recurrence was also calculated for each patient.

Conclusion:

In conclusion, this project addresses the crucial issue of breast cancer recurrence through machine learning. The study effectively developed a model to predict breast cancer relapse. A comprehensive literature review to emphasize the need for accurate predictors due to the challenges in breast cancer management.

Data collection from a reliable source provided a strong foundation, and data processing techniques prepared the dataset for modelling. Data Analysis helped determine key relevant features.

The model building phase involved training Logistic Regression and Random Forest models. Logistic Regression outperformed with a 68% accuracy and a 78% recall rate for recurrence prediction. It also provided probabilities for recurrence, allowing tailored treatment plans.

This project contributes to breast cancer research by identifying key recurrence factors and offering a predictive model. However, it's important to note that model performance depends on data quality and may require further validation. Such as the breast-quad data distribution imbalance is still a question. Thus continuous refinement and updates will be essential to refine and fine tune this project.

References

Abenaa M. Brewster, G. N. H. K. R. B. S.-W. K. C. A. S.-M. B. A. A. U. B. D. J. B. V. V. M. B. F. J. E., 2008. *Oxford Study.* [Online]
Available at: https://academic.oup.com/jnci/article/100/16/1179/913378
[Accessed 25 10 2023].

Ahmad, A., 2013. *Hindawi.* [Online]
Available at: https://www.hindawi.com/journals/isrn/2013/290568/
[Accessed 25 10 2023].

Anon., 2018. *OpenML.* [Online]
Available at: https://www.openml.org/search?type=data&sort=runs&id=13&status=active
[Accessed 24 10 2023].

Katapodi, C. M. &. M. C., 2019. *BMC.* [Online]
Available at: https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-019-1158-4
[Accessed 25 10 2023].

Long, D. L. a. X., 2023. *Springer Link.* [Online]
Available at: https://link.springer.com/article/10.1007/s00432-023-04967-w
[Accessed 25 10 2023].

Author

Simon Nadar, 25th Oct 2023

Contact:

+91 8779214328

simonnadar.work@gmail.com