International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 4 ǁ April 2023 ǁ PP. 519-522

Cancer Prediction Using Machine Learning

Tapan Pandey, Tanishq Beri, Umang Garg, Vivek Verma
Dept. of Mechanical Engineering
ABES Engineering College, Ghaziabad, UP, India

Abstract
Today where cancer has been critical disease subjected various research and studies to find out the most effective treatment focus on early prognosis and diagnosis is also vital for an effective treatment and researches are being carried out, with breast cancer being center of focus as its cases have significantly increased and are continually on rise, in women breast cancer is most prevalent cancer for death. India is having 10 lakh new cases every year and breast cancer has also overtaken cervical cancer and become the most common form of cancer in women living in India. we know that early detection is important which will result in early diagnosis of breast cancer thus resulting in early treatment which can be done. In this study, machine learning algorithm are used, for the early prognosis of breast cancer in the past various algorithms are used for cancer prediction the five most common machine learning algorithm which can be used for prediction model i.e., support vector machine, random forest, logistic regression, decision tree and k-nearest neighbors. Are widely used for cancer prediction model in this paper we will utilize logistic regression for cancer prediction model for breast cancer the methodology and result obtain for the model.

Date of Submission: 08-04-2023 Date of acceptance: 22-04-2023

I. Introduction

Breast cancer is the cancer in which the cells are formed and spreads in different regions of the breast. the breast cancer is most diagnosed in women the highest percentage of deaths occurring because of it.

According to WHO, Breast cancer arises in the ducts or lobules of the breast in a tissue known as glandular tissue. The initial cancerous growth begins in the duct or lobule, Over the period of time these as cancer progress and invades the neighbouring breast tissue then spread to the nearby lymph node and then to other organs in the body the spreading of the cancerous cell in different regions or parts of the body is known as metastasis. The cause of death is due to wide spread metastasis. Early diagnosis and treatment of breast cancer is vital for the survival of the patient, there is a high rate of survival for patient if breast cancer is identified early. Treatment of breast cancer consists of a combination of surgical removal of lumped part from the breast, radiation therapy, medication and, chemotherapy as well, which can prevent growth and spread, in 2020, according to World Health Organisation there were more than 2 million women having breast cancer with 600 thousand deaths globally. making it the most prevalent form of cancer.

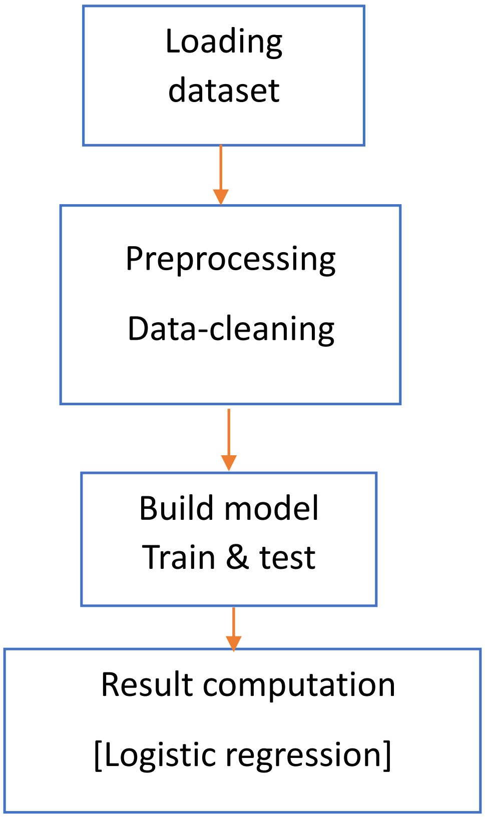
Our focus has been to integrate machine learning to address breast cancer and to provide key insights in prognosis of breast cancer and classify whether it is a benign tumour or malignant tumour.

Benign tumour is basically those tumours where the growth of cancerous cells is slow and that do not spread to other parts. Malignant tumour is those which are fast growing and spread to other parts of the body.
logistic regression is used here to classify the dataset into cateogaries of benign and malignant tumour.

II. Methodology

The methodology works out with logistic regression which is a supervised learning algorithm which will be effectively implemented for prediction, this algorithm is highly effective for classification of tumors it uses the sigmoid function also known as cost function where the probability always lies between 0 and 1.

Our methodology starts from data acquisition we have taken Breast Cancer Winscoin dataset, with data pre-processing which will contain steps like i.e., data cleaning, selecting attribute, setting target role, feature extraction. Followed by the build model is used for testing and training done, concluding with the result computing the accuracy the code is written in python for the best implementation given python's excellent integration for data analytics and machine learning algorithms libraries like scikit-learn, SciPy etc. were used.



Logistic regression (MACHINE LEARNING ALGORITHM)

Logistic regression is a widely used Machine Learning algorithm, which comes under the Supervised Learning techniques. It is used for predicting the categorical variable used in each set of independent variables. Logistic regression predicts the output of a categorical variable. Therefore, the outcome must be in a form of a binary category i.e., in the form 0 or 1, yes or no, etc. The values obtained here are probabilistic values ranges between 0 or 1. Logistic Regression is much like the Linear Regression except that how they are used. Linear Regression is used for problems related with Regression line for fitting, whereas Logistic regression is used for classification problems solving the. In Logistic regression, a regression line curve is not utilized, instead an "S" shaped logistic function, which predicts binary values (0 or 1). The sigmoid function curve of the logistic function describe about the occurrence of an event such as whether the tumour is benign or malignant, or will the batsman be able to score 100 or not etc. the logistic regression uses cost function to calculate the probability if the value is more than the threshold value it returns 1 if not it returns 0 this can be used to classify whether a patient has a malignant tumour or a benign tumour patients with malignant tumour can undergo diagnosis for breast cancer which is crucial for the treatment as early diagnosis is vital for the survival of the patients.

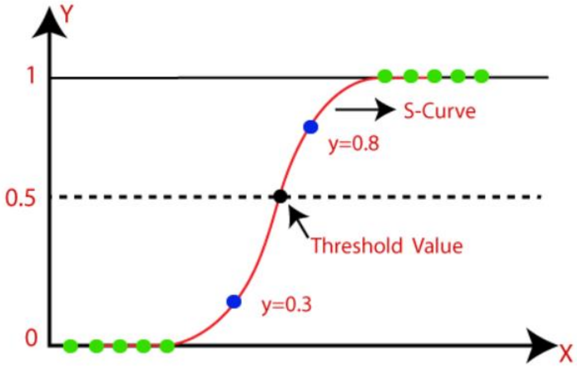


Image taken from Javatpoint

Related work or literature review

[1]UG This study provides a primary evaluation of the application of ML to predict breast cancer prognosis. They analysed 1021 patients who underwent surgery for breast cancer in our Institute and they included 610 of them. They developed two types of models i.e. (SVM)support vector machine and artificial neural network. But there are some merits also, Additional costs were the software license and the amount of time required to build the models. The purpose of applying ML models in the same institute in which the patients forming the training set were treated was to reduce hidden variables.

[2] various advancement that can be done in generic techniques that are implemented nowadays for detecting breast cancer and focused on the various imaging techniques that can be used without harming the subject.

TECNOLOGIES

- Mammography
- Ultrasonography
- MRI
- Radionuclide imaging
- Digital breast tomosynthesis
- Computed tomography

[3] various technologies used for detection of prostate cancer and how we can use machine learning for fast and accurate prediction as primitive methods use complex data for which the specialist in that particular field is also required which increases the time and cost.

TECNOLOGIES DISCUSSED

- MRSI

In this technology spectroscopy is used in which the levels of certain elements like Cl, Cr, Ni were measured from non-cancerous and cancerous patients and with the help of ANN algorithms they get results in binary terms 0,1.

- MRSI with Anatomical samples

The only difference in above model and this is of anatomical samples which increases the no. Of variables which results to higher accuracy than above model

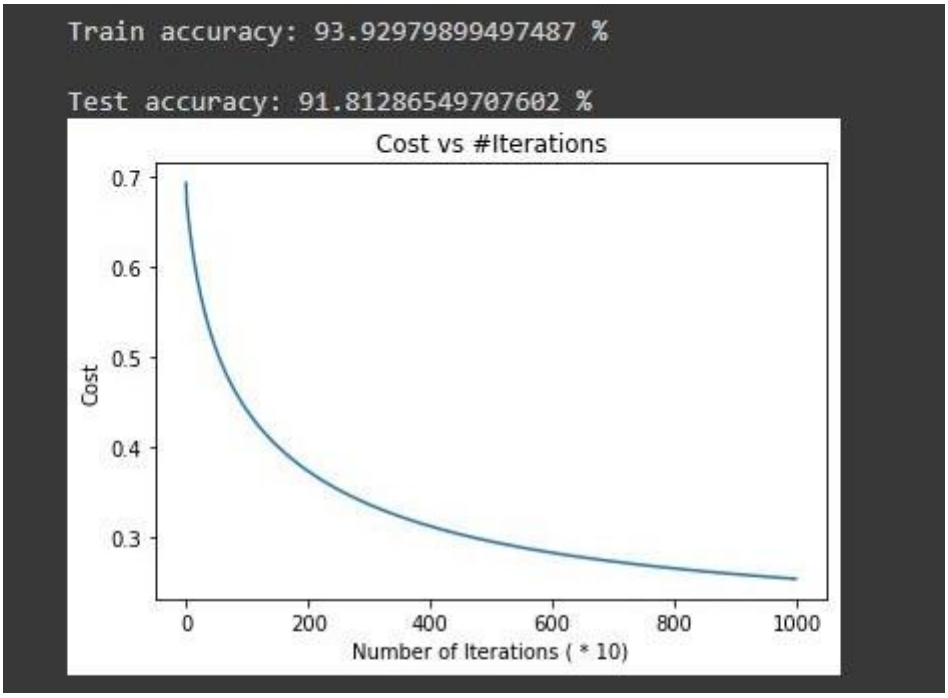
[4] UG This study tells us about the Random Forest. It is an approach which is used to analyse large data set by creating random sample sub sets from sample data heap, constraining them in multiple decision trees which is referred as the forest and the final decision is taken as the one which will obtained as an output from majority of trees. In this paper discussed about reducing the size of decision trees and approached towards the particular branches of trees which plays the significant role in the decision making for getting the desired output. One of the major advantages of random forest is that it allows us to process large amount of data.

III. Result and conclusion

The accuracy obtained from the model validates its functioning the idea to utilize two different algorithms for ensemble modelling was difficult to implement the most of the models developed have average accuracy of 80% or more in case of our model we have been able to reach our intended target of 90% accuracy with the respective accuracies as follows

Train Accuracy: 93.9297989949786 %

Test Accuracy: 91.81286549707602 %



```
Train accuracy: 93.92979899497487 %

Test accuracy: 91.81286549707602 %
```

References

[1]. Boeri, C., Chiappa, C., Galli, F., de Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Medicine*, *9*(9), 3234–3243. https://doi.org/10.1002/cam4.2811

[2]. Karellas, A., & Vedantham, S. (2008). Breast cancer imaging: A perspective for the next decade. In *Medical Physics* (Vol. 35, Issue 11, pp. 4878–4897). John Wiley and Sons Ltd. https://doi.org/10.1118/1.2986144

[3]. Matulewicz, L., Jansen, J. F. A., Bokacheva, L., Vargas, H. A., Akin, O., Fine, S. W., Shukla-Dave, A., Eastham, J. A., Hricak, H., Koutcher, J. A., & Zakian, K. L. (2014). Anatomic segmentation improves prostate cancer detection with artificial neural networks analysis of 1H magnetic resonance spectroscopic imaging. *Journal of Magnetic Resonance Imaging*, *40*(6), 1414–1421. https://doi.org/10.1002/jmri.24487

[4]. Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3d deep learning on medical images: A review. In *Sensors (Switzerland)* (Vol. 20, Issue 18, pp. 1–24). MDPI AG. https://doi.org/10.3390/s20185097

[5]. Zhang, H., & Wang, M. (n.d.). *Search for the smallest random forest*.

[6]. Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. In *Cancer Informatics* (Vol. 2).

[7]. Octaviani, T. L., & Rustam, Z. (2019). Random forest for breast cancer prediction. *AIP Conference Proceedings*, *2168*. https://doi.org/10.1063/1.5132477

[8]. Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The application of deep learning in cancer prognosis prediction. In *Cancers* (Vol. 12, Issue 3). MDPI AG. https://doi.org/10.3390/cancers12030603

[9]. Rafique, R., Islam, S. M. R., & Kazi, J. U. (2021). Machine learning in the prediction of cancer therapy. In *Computational and Structural Biotechnology Journal* (Vol. 19, pp. 4003–4017). Elsevier B.V. https://doi.org/10.1016/j.csbj.2021.07.003

[10]. Naji, M. A., Filali, S. el, Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*, *191*, 487–492. https://doi.org/10.1016/j.procs.2021.07.062

[11]. Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. v., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. In *Genome Medicine* (Vol. 13, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-021-00968-x

[12]. jahanvi joshi, r. D. (N.D.). *Diagnosis of breast cancer using clustering data mining approach* . retrieved from https://www.Hindawi.Com/journals/jhe/2019/7294582

[13]. Konstantina kourou, D. I. (n.d.). *Machine learning applications in cancer prognosis and prediction*. Retrieved from https://www.sciencedirect.com/science/article/pii/S2001037014000464

[14]. Puja Gupta, s. g. (n.d.). *breast cancer prediction using varying parameters.* https://scholar.google.co.in/scholar_url?url=https://www.sciencedirect.com/science/article/pii/S1877050920310310/pdf%3Fmd5%3D050bc4e8c2c797b9f08a47cc75fd58d7%26pid%3D1-s2.0-S1877050920310310-main.pdf&hl=en&sa=X&ei=wGWjY4CkJtKxywSqmKWoBQ&scisig=AAGBfm2aY6B

[15]. thomas noel, h. a. (n.d.). *Using Machine Learning Algorithms for Breast Cancer Risk*. Retrieved from https://reader.elsevier.com/reader/sd/pii/S1877050916302575?token=001F66C9E2132E44B6C5DF22CD9F8A3875BDC2527D6D8B412477C1F81EDF9FA