Survival Outcome Prediction for Breast Cancer Patients

Dhivya S, Arulprabha R, Kowsalya M, Gaddam Hemanth Kumar, Mullagiri Bhavan Premchand Gandhi

Abstract*: The second most causative disease is breast cancer happening in women and a significant explanation behind expanding death rate among women. Observed rates of this cancer are increasing with industrialization and also with early detection facilities. As the finding of this ailment physically takes extended periods and the lesser accessibility of frameworks, there is a need to build up the programmed determination framework for early identification of malignant growth. We have used machine learning classification techniques to categorize benign and malignant tumors, in which the machine learns from past data and predicts the new input category. Models like logistic regression and Random Forest are Done on the UCI dataset. Our experiments have indicated that Random Forest has the best prescient examination with exactness of ~96%.

Keywords : Logistic Regression, Random Forest, Decision Tree

I. INTRODUCTION

Breast cancer is a common cause of death, and is the only type of cancer that is widespread among women all over the world. In the event of any sign or side effect, typically individuals visit a specialist quickly, Who can refer to an oncologist, whenever required. The oncologist can analyze breast cancer by undertaking an intensive clinical history, physical examination of the breasts as well as examining for the growth or solidification of any lymph hubs in the axis. Many imaging strategies have been created to detect and treat breast cancer early and to diminish the quantity of passing. Numerous researches are being led in this area by the use of different machine learning methods for various datasets on Breast cancer. Most of them indicate classification strategies give a decent exactness in expectation of the kind of tumour.

Revised Manuscript Received on May 15, 2020.

Correspondence Author:

Dhivya S, Assistant Professor, Department Of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.Email: dhivyasivanandan@siet.ac.in

Arulprabha R, Department Of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email: arulprabha.98@gmail.com

Kowsalya M, Department Of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email: kowsalyamarappan1999@gmail.com

Gaddam Hemanth Kumar, Department Of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India. Email: hemanth92.gaddam@gmail.com

Mullagiri Bhavan Premchand Gandhi, Department Of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India. Email: bhavanprem027@gmail.com

A. Existing system

The available technologies for detection of breast cancer were not sufficiently powerful to support definitive conclusions regarding its ultimate clinical value and use. In order to improve x-ray mammography, many of the companies tried to develop Digital mammography apparatus. Dissimilar to film mammography gadgets that creates an x-ray image of the breast legitimately on photographic film, computerized mammography gadgets Catch digital x-ray image. A whole slew of detectors makes a digitized picture that can be seen and controlled on a PC screen.In theory, this could empower the better location of tumors clouded by the thick breast tissue much of the time seen in young women. The capacity to expand or alter the differentiation of faulty zones without requiring a new x-ray introduction may encourage the discovery of sores that Film mammography has missed. The innovation could likewise improve mammography screening with permitting electronic capacity, recovery, And Mammogram Transmission. In any case, one significant confinement of advanced mammograms is that the pictures are not as precise as movie mammograms.

B. Proposed system

We'll use the UCI Machine Learning Repository for datasets on breast cancer. The dataset is created by Dr. William H. Wolberg, Physicist at Madison University of Wisconsin Hospital, Wisconsin, USA. To make the dataset Dr. Wolberg utilized liquid examples, Taken from patients with massive breast mass and a simple-to-utilize graphical PC program called Xcyt, which is equipped for playing out the examination of cytological highlights dependent on an advanced sweep. The program utilizes a curve-fitting algorithm, to figure ten highlights from each last of the cells in this example, at that point it ascertains the mean worth, extraordinary worth and standard error of each element for the image, returning 30 real-valued vectors. With the help of this dataset, along with the advanced machine learning algorithms, prediction has been done and the results have been displayed in a web page for easy understanding.

II. RELATED WORKS

K.Pravalika, C.Shravya and Dr.Shaik Subhani used three famous algorithms such as logistic regression, nearest neighbour algorithm and support vector machines for Building predictive models for predicting breast cancer and comparing its accuracy.

Vikas Chaurasia , Saurabh Pal and BB Tiwari used Naıve Bayes, RBF Network and J48 Decision Tree to find a best model for predicting breast cancer.

Youness Khourdif and Mohamed Bahaj experimented on breast cancer data using Support Vector Machines (SVM), Naive Bayes and K-Nearest Neighbour algorithms.

MandeepRana, PoojaChandorkar, Alishiba D'souza Worked on breast cancer diagnosis and prediction by applying the algorithms KNN, SVM, Naïve Bayes and Logistic Regression, programmed in MATLAB. These methods of classification are applied on two datasets taken from the UCI depository. A dataset of these is used for disease identification (WDBC) and the next for prediction of recurrence (WPBC).

NareshKhuriwal, Nidhi Mishra data took from Wisconsin Breast Cancer database and was functioning on the diagnosis of breast cancer. The results indicate that ANN and the Logistic Algorithm worked better and a good solution was provided.

III. METHODOLOGY

The dataset on breast cancer is collected from UCI repository and using Spyder to work on datasets and used Flask framework to display the results in a web application. This paper uses three popular classification algorithms like Logistic Regression and Random Forest on breast cancer dataset.

A.Data pre-processing and Data Exploration

Data pre-processing will be done for converting the dataset in a standardized way to be used for prediction by removing the variables which are not helpful for predicting the results. And also converting all the categorical variables into numerical variables so that will be suitable for prediction.

B.Feature Selection

Usually our dataset will contain features that vary greatly in magnitudes, units, and range. But, most machine learning algorithms use Eucledian distance in their computations between two data points. We need to bring all characteristics to the same magnitudes. We can achieve this by scaling. That means that data should be transformed to fit within a given scale, such as 0–100 or 0–1.

C.Model Selection

Identifying an algorithm is the most vigorous phase of applying Machine Learning to the dataset. Normally Data Scientists utilize various types of Machine Learning calculations to big data sets. But all those different algorithms can be ordered in two groups at a significant level: supervised and unsupervised learning.

Supervised learning: Supervised learning is a form of system that provides both input and output data as desired. Classification of Data to input and output Is labeled as a learning context for the future processing of data. Further, supervised learning is grouped into two types:

1. Regression
2. Classification

Regression algorithms will be used, when the variable we are going to predict is a continuous or real value, like “height” or “weight”.

Classifications algorithms will be used, when the predictor variable is like a category called “pass” or “fail”.

Unsupervised Learning: Unsupervised learning algorithms use information that is neither marked nor numbered, and requires the algorithm to operate without instruction on that information.

In the dataset, the dependent variable (outcome variable) has two sets of values like Benign (B) or Malignant (M). As the outcome variable is like a category, a classification algorithm is used.

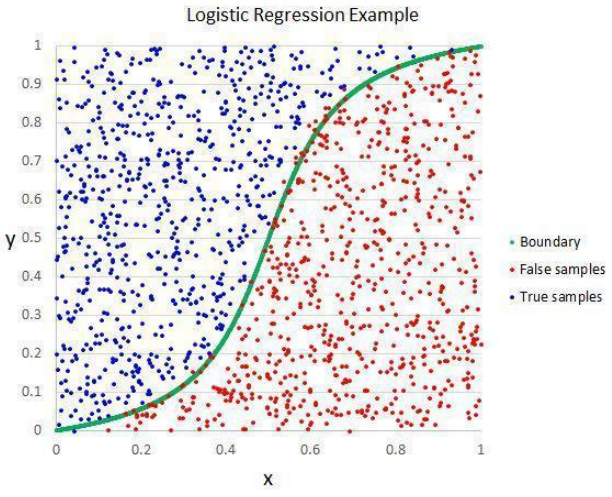
We have picked two different sorts of classification algorithms in Machine Learning.

1. Logistic regression
2. Random Forest
3. Decision tree

1. Logistic Regression

Logistic regression is the appropriate regression analysis algorithm to conduct when the dependent variable is dichotomous (binary) such as pass or fail, win or lose, men or women etc.. Logistic regression is like all regression analysis and also a predictive analysis. To describe data and to explain the relationship between one dependent binary variable and one or more independent variables, logistic regression is used. Logistic regression uses a logistic function to model a dependent variable(binary), in its basic form. In regression analysis, the parameters of a logistic model are estimated by logistic regression. A binary logistic model typically has a two-value dependent variable, such as pass / fail or win/lose are labeled "0" and "1" called as indicator variables. In the logistic model, the log-odds (the logarithm of the odds) is a combination of one or more independent variables (predictors); the independent variables can be a continuous variable (any real value) or a binary variable (two classes, coded by an indicator variable). [1]

This is a general statistical model was developed by Joseph Berkson, where he coined "logit";

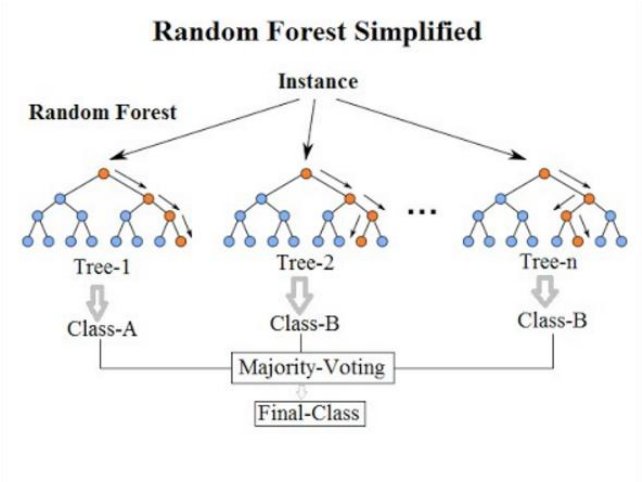


2. Random Forest

Random Forest is the most popular algorithm for machine learning in supervised learning technique. It can be used for both Regression and Classification problems in ML.

As per the name, "Random Forest is a classifier that contains a number of decision trees on different dataset subsets and takes the average to enhance that dataset's predictive accuracy." Rather than relying on one decision tree, the random forest Takes predictions out of every tree and predicts the final output based on the majority vote of predictions.

The greater number of trees in the forest leads to greater accuracy and avoids the problem of overfitting.

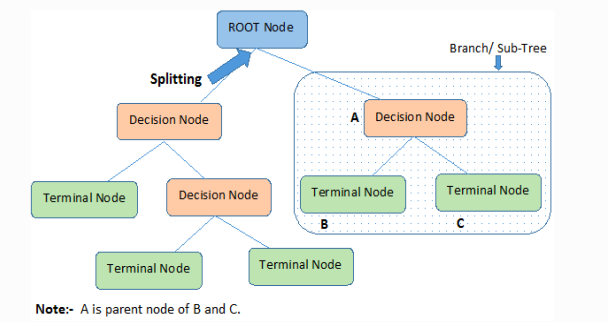


3. Decision Tree Algorithm

Decision Tree algorithm forms a part of the family of supervised learning algorithms. The decision tree algorithm can also be useful for resolving regression and classification problems, unlike other supervised learning algorithms.

The intent in using a Decision Tree is to develop a training model for predicting the target variable's class or value by learning simple rules of decision inferred from prior data (training).

In Trees of Decision we start from the tree root to predict the record class label. We compare the root attribute values with the attribute of that record. We follow the branch corresponding to that value on the basis of comparison, and jump to the next node.[2]



D. Webpage Development

Next step is developing the webpage for displaying the prediction results. The webpage includes text boxes so that predicting parameters can be entered and a method can be selected for prediction. This has been developed with the help of HTML, CSS and BOOTSTRAP.

E. Back-end connectivity

Connectivity to the webpage has been done with the FLASK framework, one of the most popular frameworks in python. Flask is a web-based framework. This means that flask provides tools, libraries, and technologies that enable you to build a web app.

IV. RESULTS AND DISCUSSION

As our dataset contains 32 characteristics, data pre-processing contributes a ton in diminishing the multidimensional data to a couple of measurements. Of all the two applied algorithms Random Forest and Logistic Regression, Random Forest gives the most noteworthy exactness of 98.6% when contrasted with logistic regression. In this way, we suggest that Random Forest with complex datasets is the most appropriate algorithm for predicting Breast Cancer occurrence. Well, this isn't always applicable to any dataset. We must always analyze our dataset and then use our machine-learning model to choose our model.

ALGORITHM	ACCURACY
Logistic Regression	95.8%
Random Forest	96.6%
Decision Tree	95.2%

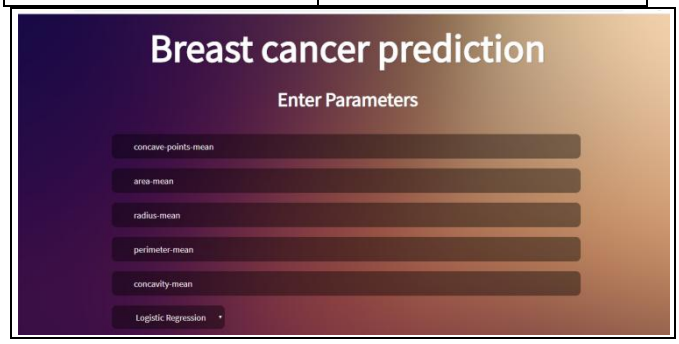


Fig 4.a Entering parameters and selecting model

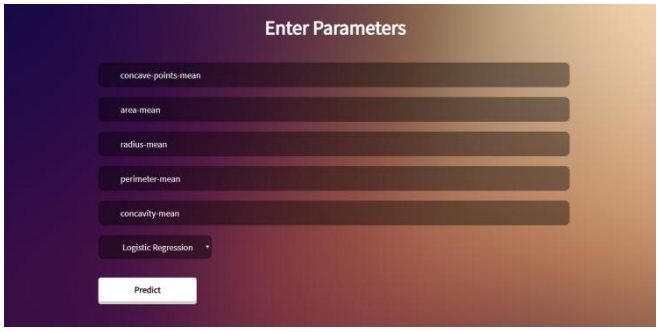


Fig 4.b Predicting after entering all the attributes

Fig 4.c Displaying results after prediction

Fig 4.d Content when entered incorrect value

V. CONCLUSION

Our work mainly focused on the easy understanding of breast cancer prediction by displaying the results in the web application using the python frameworks. For better performance of the classification techniques, more research should be carried out in that field, so that the accuracy can be increased even more.

REFERENCES

1. Ch. Shravya, K.Pravalika, Shaik Subhani "Prediction of Breast Cancer Using Supervised Machine Learning Techniques" Volume-8, Issue-6
2. Vikas Chaurasia , Saurabh Pal and BB Tiwari "Prediction of benign and malignant breast cancer using data mining techniques' Volume-12
3. Youness Khourdifi, Mohamed Bahaj "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification"
4. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.
5. Nidhi Mishra, Naresh Khuriwal- "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm", 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018
6. J. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, vol. 2, pp. 59–77, 2006.View at: Publisher Site | Google Scholar
7. Logistic Regression for Machine Learning - https://machinelearningmastery.com/logistic-regression-for-machine-learning/
8. Random Forest for Machine Learning -https://machinelearningmastery.com/implement-random-forest-scratch-python/
9. Random Forest for Machine Learning -https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/

AUTHORS PROFILE

Dhivya S was born in Erode, India, in 1990.She received the B.E degree in Computer Science and Engineering from the Avinashilingam University, India, in 2011, and the M.E. degree in Computer Science and Engineering from the Anna University, India, in 2014 .She has been with the Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology,Coimbatore, where she is an Assistant Professor. She is a Life Member of Institution of Engineers (India) [IEI] and [IEANG].

Arulprabha R was born in Tiruppur, India in 1998. She is currently doing her Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore. She is doing her internship in Bluebird India R&D centre, Bangalore, India. Her role in the company is Automation Engineer.

Kowsalya M was born in Coimbatore, India in 1998. She is currently doing her Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore. She is doing her internship in Techvolt software, Coimbatore, India. Her role in the company is PHP Developer.

Gaddam hemanth Kumar was born in Prakasam, India in 1998. He is currently doing his Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

Mullagiri Bhavan Premchaand Gandhi was born in Krishna, India in 1999. He is currently doing his Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.