International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/IJSRCSEIT

Breast Cancer Predication Using Machine Learning and Data Mining

Swetha K, Ranjana R
Sri Krishna College of Technology, Kovaipudur, Tamil Nadu, India

ABSTRACT

Breast cancer is a type of cancer that emerge in breast of a women or a men. mostly women are affected by breast cancer than men. it is important to know that most breast lumps are benign and not a cancer. There are different types of breast cancer and most common type of breast cancer includes ductal carcinoma in situ (DCIS) and invasive carcinoma. classification and data mining methods are an effective way to classify data. Most commonly in medical field, where those are widely used in diagnosis of breast cancer. Women with 40 to 50 or older are average risk of breast cancer. women nearly age of 30 are mostly affected by the risk of breast cancer. Closely in 2012 1. 7 million new breast cancer cases were diagnosed. Breast cancer is mosttly diagnosed among women for breast cancer for 140 of 180 countries. After skin cancer Breast cancer is the most common cancer among American women. Nearly 500 men will die in breast cancer. 62 percent of breast cancer cases are diagnosed at a sectarian stage, for which 5 years of survival rate is 99%. Machine learning technique is used for prediction of breast cancer using data set. we have used WEKA tool for predicting the accuracy rate using various algorithms. Here we have used algorithms of IBK, Simple logistic, Naive bayes, Decision table, Multilayer perception.

Keywords : WEKA, IBK, Simple Logistic, Naive Bayes, Decision Table, Multilayer

I. INTRODUCTION

Breast cancer is a type of cancer that emerge in breast of a women or a men. mostly women are affected by breast cancer than men. it is important to know that most breast lumps are benign and not a cancer. There are different types of breast cancer and most common type of breast cancer includes ductal carcinoma in situ (DCIS) and invasive carcinoma. classification and data mining methods are an effective way to classify data. Most commonly in medical field, where those are widely used in diagnosis of breast cancer. Women with 40 to 50 or older are average risk of breast cancer. women nearly age of 30 are mostly affected by the risk of breast cancer. Closely in 2012 1. 7 million new breast cancer cases were diagnosed. Breast cancer is mosttly diagnosed among women for breast cancer for 140 of 180 countries. After skin cancer Breast cancer is the most common cancer among American women. Nearly 500 men will die in breast cancer. 62 percent of breast cancer cases are diagnosed at a sectarian stage, for which 5 years of survival rate is 99%. Machine learning technique is used for prediction of breast cancer using data set. we have used WEKA tool for predicting the accuracy rate using various algorithms. Here we have used algorithms of IBK, Simple logistic, Naive bayes, Decision table, Multilayer perception

II. LITERATURE SURVEY

There are several studies based on the diagnosis of breast cancer using statistical approaches and artificial neural networks. Prediction of breast cancer

or any others disease using data set has created a affective impact on past two decades. Here the various author had studies on outcomes of breast cancer. Lerman et al at 1995 had a study design on cross section on breast cancer. Ganz et al at 1995 had a study design on RCT the outcome is QOL. Stefanek et al at 1995 had a study design on cross section and the outcomes is cancer related worry. Among the all gone through article the most common outcome was HRQOL.

[1]. M. Navya sri et al at 2019 has used WELKA tool to has made the comparative analysis between Decision Tree J48 algorithm and Bayesian classification to determine the breast cancer among the women and the result of this experiment is J48 have given 75. 875% accuracy and 75. 17% of Bayesian

[2]. Chintan et al at 2013 had using algorithms namely Naive Bayes, DT, KNN to predict the cancer and the result shown that naive bayes works effeciently than the other two algorithms.

[3]. K. Goyal et al in 2019 used Adaboost, SVM, Random Forest, Naive Bayes, Decision tree algorithms in WEKA tool, J48, Logistic Regression.

[4]. Askhya yadav et al at 2019 used anaconda a, python tool and SVM, ANN, KNN, DT, RF algorithms they made the performance comparison and concluded that SVM & Random Forest have highest accuracy whereas Naive bayes classifier have a highest precision rate.

Table 1

ALGORITHM	ACCURACY	FP Rate	PRECISION	RECALL	F-MEASURE	MCC	ROC AREA	PRC AREA
NAIVEBAYES	97. 6731	0. 104	0. 939	0. 944	0. 941	0. 842	0. 976	0. 983
IBK	95. 6892	0. 097	0. 967	0. 972	0. 969	0. 917	0. 956	0. 955
DECISION TABLE	98. 2331	0. 104	0. 941	0. 966	0. 953	0. 872	0. 979	0. 980
RANDOM FOREST	99. 1678	0. 75	0. 966	0. 980	0. 968	0. 913	0. 991	0. 991
SIMPLE LOGISTIC	99. 7612	0. 052	0. 976	0. 986	0. 976	0. 940	0. 996	0. 997

DATASET DISCRIBTION

I

n the dataset used for predication of breast cancer there many attributes were used. In that dataset patients id, diagnosis list, radius_ mean, texture_ mean perimeter _mean, area_ mean, smoothness_ mean, compactness _mean concavity _mean, concave points _ mean symmetry_ mean, fractal _dimension _mean, radius _se, texture_ se, perimeter_ se , area _se, smoothness_ se compactness _ se, concavity _se, concave points _se the above listed columns were used.

III. PROPOSED SYSTEM

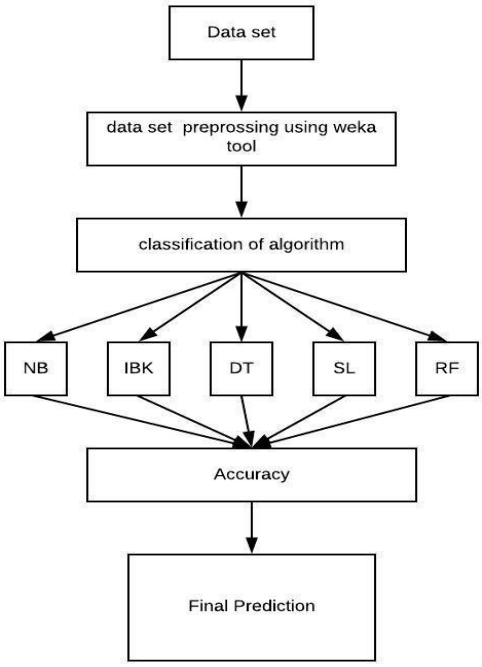

Fig 2

The above diagram fig 1. 1 shows the mechanism of predicting the accuracy for breast cancer. The data set is put into the weka tool and the accuracy is predicted by selecting the different kind of algorithms. we have used five algorithms in our prediction namely naive bayes, ibk, decision table, simple logistic, random forest. The classifier output have the FP rate, TP rate, Precision, Recall, F-measure, MCC, ROC area, PRC Area and class. we have calculate the accuracy using the precision. accuracy of all algorithms are calculated. And the final predication is made according to the algorithm having highest accuracy rate.

IV. DATA PREPROCESSING

Data preprocessing is used for removing the unwanted columns or features in the dataset, and keeping only the required attributes. The essential set in the data mining process is data preprocessing In this dataset we take two tasks for carried out for preprocessing the data. We performed the missing values value analysis and it perform that there is no missing values in the dataset. And we expect that abstraction that the classifier to learn during training which might not achieve with the simple dataset. It leads to good classification results and improve the mining process efficiency. The data processing tasks such as data cleaning, data reduction and the data integration. Data integration is important and useful when the data merged from different and multiple data sources. ng. Data preprocessing are used in various stages to remove errors from the input data so that it cannot effect the experimental results. In classification we remove the unwanted columns like perimeter _mean, area _mean and use the desired columns like precision, Recall etc…. And we calculate the result using the accuracy. During preprocessing stage, the data is partitioned into the training set and the validation set.

V. TYPES OF ALGORITHM

NAIVE BAYES

It is a classification algorithm. it makes the algorithm with highest probability. it is a classifier. There is no any other difference than olden machine learning algorithms. Naive bias is an powerful algorithm for the predictive modeling. It is a supervised machine learning algorithm that uses Bayes theorem. This theorem explains that naive assumption that the input variables are independent of each other. It allows the hypothesis to update each time and then new evidence can be introduced. And it examines that probability of event is based on the prior knowledge of any event related to former event.

The mathematical formula for evaluating accuracy

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

DECISION TREE

Decision tree consists of two main parts they are nodes and rules(tests). The concept of this algorithm is to draw the flowchart for root at top. All other non

leaf nodes represent a test to single or of multiple attributes until it reach a leaf node(final result). Decision trees are used in data mining and classification. Decision tree provides a clear indication to important attributes major part of establishing rules between attributes is indicating the importance level of each one. It provides less computations compared to other classification algorithm such as mathematical formulae.

The mathematical formula for evaluating accuracy:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} . Entropy(S_v)$$

$$IG(S, A) = H(S) - H(S, A)$$

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

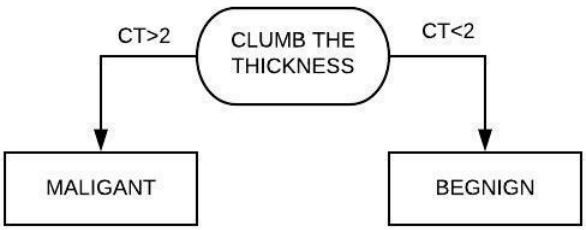

Fig 3

The above diagram shows the decision tree working in flow chart approach.

SIMPLE LOGISTIC REGRESSION

Simple logistic regression(SLR)[14] is a linear logistic model using logic boot algorithm. As logic boot algorithm uses a symmetric model, a sufficient number of iteration is performed in simple logistic regression to train a model. Built in attribute selection is performed in SLR as an additional advantage If we use simple logistic regression and it also gives the best performance of the classifier.

The mathematical formula for evaluating accuracy:

$$g(z) = \frac{1}{1+e^{-z}}$$

$$h(x_i) = g(/beta^T x_i) = /frac11 + e^{-/beta^T x_i}$$

$$J(/beta) = /sum_{i=1}^{n} - y_i log(h(x_i)) - (1 - y_i) log(1 - h(x_i))$$

IBK

The IBK algorithm is inspired by nearest algorithm(knn). IBK algorithm does not generate a model while classification but generates a prediction for a test sample just in time. The IBK algorithm uses a distances measure to find ' k' close instances in the training data of each test sample and uses a selected distance to estimate. The aim is to elaborate on simple IBL algorithm(IBI) and provide an analysis of which class of concepts can be learned. It is in similarity of numerical values. The classification function takes the results and classification of the similarity function and it performs the recording of the instances in the concept of description.

The mathematical formula for evaluating accuracy:

```
AD(xi) = k f=1 SD xi, xf I
fu(AD) = 1 nu n i=1 K
AD     ADi u
```

RANDOM FOREST

Random forest is a set of decisions which is used to classify the data set. Random forest is also known as classification and regression tree(CART). collection of decision trees. is random forest. Random forest iteratively takes a series of questions based on that answer it will ask another set of questions to classify the data. Take the test data sets and randomly creates the decision trees. Find the decision of each decision tree according to the majority. Choose the high value as the final decisions.

The mathematical formula for evaluating accuracy:

$$mg(X,Y) = avk\ I(hk\ (X)=Y) - max\ j \neq Y\ avk\ I(hk\ (X)= j)$$

$$PE* = PX,Y\ (mg(X,Y) < 0)$$

$$PX,Y\ (P\ominus(h(X,\ominus)=Y) - max\ j \neq Y\ P\ominus(h(X,\ominus)= j) < 0)$$

VI. RESULT

NAIVE BAYES

```
== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   MCC
               0.836     0.565     0.778       0.836    0.806       0.288
               0.435     0.164     0.529       0.435    0.477       0.288
eighted Avg.   0.717     0.446     0.704       0.717    0.708       0.288

== Confusion Matrix ===

  a   b   <-- classified as
168  33 |   a = no-recurrence-events
 48  37 |   b = recurrence-events
```

fig 4

RANDOM FOREST

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure
               0.871     0.718     0.742       0.871    0.801
               0.282     0.129     0.480       0.282    0.356
Weighted Avg.  0.696     0.543     0.664       0.696    0.669

=== Confusion Matrix ===

   a   b   <-- classified as
 175  26 |   a = no-recurrence-events
  61  24 |   b = recurrence-events
```

fig 5

IBK

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   MCC
               0.896     0.682     0.756       0.896    0.820       0.2
               0.318     0.104     0.563       0.318    0.406       0.2
Weighted Avg.  0.724     0.511     0.699       0.724    0.697       0.2

=== Confusion Matrix ===

   a   b   <-- classified as
 180  21 |   a = no-recurrence-events
  58  27 |   b = recurrence-events
```

fig 6

SIMPLE LOGISTIC

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
               0.930    0.671    0.766      0.930   0.840      0.335   0.675     0.797     no-recurrence-events
               0.329    0.070    0.667      0.329   0.441      0.335   0.675     0.490     recurrence-events
Weighted Avg.  0.752    0.492    0.737      0.752   0.722      0.335   0.675     0.706

=== Confusion Matrix ===

   a   b   <-- classified as
 187  14 |   a = no-recurrence-events
  57  28 |   b = recurrence-events
```

fig 7

DECISION TREE

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
               0.925    0.718    0.753      0.925   0.830      0.277   0.658     0.794     no-recurrence-events
               0.282    0.075    0.615      0.282   0.387      0.277   0.658     0.469     recurrence-events
Weighted Avg.  0.734    0.527    0.712      0.734   0.699      0.277   0.658     0.697

=== Confusion Matrix ===

   a   b   <-- classified as
 186  15 |   a = no-recurrence-events
  61  24 |   b = recurrence-events
```

fig 8

We have used different algorithms for predication of breast cancer using weka tool. There we have used five types of algorithms for the predication. naive bayes, simple logistic, random forest, decision table, IBK. we have calculated the accuracy for each of the algorithms. Naive bayes have produced the accuracy of 97. 6731, IBK have given the accuracy of 95. 6892, Decision table have produce the accuracy of 98. 2331, Random forest have stated accuracy of 99. 1678. simple logistic produced the accuracy of 99. 7612.

VII. CONCLUSION

To predict the outcomes in medical field, machine learning tools and data mining tools are mainly used. we have used five algorithms for predication of breast cancer, the accuracy are analyzed for five algorithms like Naive bayes, IBK, Simple logistic, Decision table, Random forest. in those algorithms Simple logistics has produced the accuracy of 99. 7612 % in comparison with other algorithms. finally concluded that the simple logistic have given the effective algorithm for the predication of breast cancer and diagnosis of ailment with lowest error rate.

VIII. REFERENCES

[1]. Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.

[2]. Gayathri, B. M., C. P. Sumathi, and T. Santhanam. "Breast cancer diagnosis using machine learning algorithms-a survey." International Journal of Distributed and Parallel Systems 4, no. 3 (2013): 105.

[3]. Jerez, José M., et al. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine 50.2 (2010): 105-115.

[4]. Wang, Deling, Jia-Rui Li, Yu-Hang Zhang, Lei Chen, Tao Huang, and Yu-Dong Cai. "Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms." Genes 9, no. 3 (2018): 155.

[5]. bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520-523. Springer, Berlin, Heidelberg, 2007.

[6]. Ahmad, L. Gh, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi. "Using three machine learning techniques for predicting breast cancer recurrence." J Health Med Inform 4, no. 124 (2013): 3

[7]. Wolberg, William H., W. Nick Street, and Olvi L. Mangasarian. "Image analysis and machine learning applied to breast cancer diagnosis and prognosis." Analytical and Quantitative cytology and histology 17, no. 2 (1995): 77-87.

[8]. Aswathy, M. A., and Jagannath Mohan. "Analysis of Machine Learning Algorithms for Breast Cancer Detection." In Handbook of Research on Applications and Implementations of Machine Learning Techniques, pp. 1-20. IGI Global, 2020.

[9]. Goyal, Kashish, Preeti Aggarwal, and Mukesh Kumar. "Prediction of Breast Cancer Recurrence: A Machine Learning Approach." In Computational Intelligence in Data Mining, pp. 101-113. Springer, Singapore, 2020.

[10]. Kashif, Muhammad, Kaleem Razzaq Malik, Sohail Jabbar, and Junaid Chaudhry. "Application of machine learning and image processing for detection of breast cancer." In Innovation in Health Informatics, pp. 145-162. Academic Press, 2020.

[11]. Shrivastava, Deepshikha, Sugata Sanyal, Arnab Kumar Maji, and Debdatta Kandar. "Bone cancer detection using machine learning techniques." In Smart Healthcare for Disease Diagnosis and Prevention, pp. 175-183. Academic Press, 2020.

[12]. Ganggayah, Mogana Darshini, Nur Aishah Taib, YipCheng Har, Pietro Lio, and Sarinder Kaur Dhillon. "Predicting factors for survival of breast cancer patients using machine learning techniques." BMC medical informatics and decision making 19, no. 1 (2019): 48.

[13]. Li, C. (2019, September). Classification of Breast Cancer Malignancy Using Machine Learning Mechanisms in TensorFlow and Keras. In Future Trends in Biomedical and Health Informatics and Cybersecurity in Medical Devices: Proceedings of the International Conference on Biomedical and Health Informatics, ICBHI 2019, 17-20 April 2019, Taipei, Taiwan (Vol. 74, p. 42). Springer Nature.

[14]. Ajay, Kumar, Rama Sushil, and Arvind Tiwari. "Cancer Survival Analysis Using Machine Learning." Available at SSRN 3354469 (2019)

[15]. Nilashi, Mehrbakhsh, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. "An analytical method for diseases prediction using machine learning techniques." Computers & Chemical Engineering 106 (2017): 212-223.

[16]. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health. 2019 Sep 1;7(3):293-9.

[17]. Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018, April). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.

[18]. Sharma, Shubham, Archit Aggarwal, and Tanupriya Choudhury. "Breast Cancer Detection Using Machine Learning Algorithms." In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 114-118. IEEE, 2018.

[19]. Turgut, Siyabend, Mustafa Dağtekin, and Tolga Ensari. "Microarray breast cancer data classification using machine learning methods." In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-3. IEEE, 2018.

[20]. Abreu, Pedro Henriques, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, and Daniel Castro Silva. "Predicting breast cancer recurrence using machine learning techniques: a systematic review." ACM Computing Surveys (CSUR) 49, no. 3 (2016): 1-40.

[21]. Kim, I., H. J. Choi, J. M. Ryu, S. K. Lee, J. H. Yu, S. W. Kim, S. J. Nam, S. W. Seo, and J. E. Lee. "Abstract P2-08-52: A predictive model for distant metastasis in breast cancer patients using machine learning." (2019): P2-08.

[22]. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital signal processing. 2007 Jul 1;17(4):694-701.

Cite this article as :

Swetha K, Ranjana R, "Breast Cancer Predication Using Machine Learning and Data Mining", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.610-615, May-June-2020.
Journal URL : http://ijsrcseit.com/CSEIT206219