International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Breast Cancer Diagnosis and Prediction Using Machine Learning Algorithm

Shilpa M[1], C. Nandini[2]

[1]Assistant Professor, Department of CSE, DSATM, Bangalore, Karnataka, India

[2]Professor, Department of CSE, DSATM, Bangalore, Karnataka, India

Abstract: ***Breast cancer is now the most common cancer in most cities in India, and 2nd most common in the rural areas. Early predication of cancer allows for the proper treatment of the disease and increases the survival rate. Various traditional methods, based on physical and chemical tests, are available for diagnosing. The Machine Learning algorithms allows for the prediction of the disease based on the past data. The Machine Learning algorithms can be used for an effective mechanism for early and accurate method for breast cancer prediction. Various Machine Learning algorithms like Decision Tree, Support Vector Machine (SVM), Naïve Bayes, Nearest Neighbor (KNN), Linear Regression, Neural Networks are available. In this paper we use the Naïve Bayes algorithm for breast cancer prediction. The Wisconsin original breast cancer data set was used as a training set. The classifiers are fed with data set of fixed number of attributes. The algorithm was implemented in Python using Anaconda.***

Keywords: Breast Cancer, Machine Learning (ML), Navies Bayes Algorithm

1. Introduction

Breast cancer is the malignant tumor that starts in the cells of the breast. It occurs both in men and women. However male breast cancer is rare. Breast cancer is now the most common cancer in most cities in India, and 2nd most common in the rural areas. According National Cancer Registry Programme there is an increasing incidence of Breast cancer in younger age groups and rising numbers of cases of breast cancer in India. Early predication of cancer allows for the proper treatment of the disease and increases the survival rate.

Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. The hidden patterns and relationships in the data are mostly overlooked.

Diagnosing patients is a difficult task and doctors who can accurately predict such diseases are few in number. Data Mining refers to using a variety of techniques to identify information or decision-making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, predictions, forecasting and estimation. The improvement and exploitation of a number of prominent Data Mining techniques in numerous real-world application areas (e.g. Industry, Healthcare and Bioscience) has led to the utilization of such techniques in machine learning environments, in order to extract useful pieces of information of the specified data and support decision making

Machine learning is a branch of artificial intelligence that incorporate a variety of statistical, probabilistic and optimization techniques that allow computers to "learn" from past examples and to detect hardto- diagnosed patterns from massive, noisy or complex data sets. Thesefeatures are particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. Machine learning techniques like support vector machine, Bayesian belief network, artificial neural network is frequently used in cancer diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction. The survey has shown that there are lots of best performing algorithms for the analysis of features of data sets.

2. Related Work

The use of Artifical Intelligence and Machine Learning in the field of Medical Science is one of the important research trends which is prevalent from a long time. But the current changes and improvements in machine learning gives us more opportunity for research. A lot of research has been done in the prediction of cancer using machine learning techniques.

In [2], B.M.Gayathri, Dr.C.P.Sumathicompare study of RVM with various ML algorithms, to show that RVM classifies better than other ML algorithms even when the variables are reduced.As a future work they suggest RVM can be combined with other ML algorithms so that it can be fine-tuned to improve the accuracy.

In [9], Ayush Sharma, Sudhanshu Kulshrestha, Sibi Daniel use sophisticated classifiers such as Logistic Regression, Nearest Neighbor, Support Vector Machines for predicting breast cancer. A concrete relationship between precision, recall and the number of features in the data set is achieved, which is shown graphically. But there is a need to evaluate these techniques rigorously before using them commercially. In [1], the authors present a novel modality for the prediction of breast cancer and introduce with the Support Vector Machine and KNearest Neighbors which are the supervised machine learning techniques for breast cancer detection by training its attributes. The proposed system uses 10-fold cross validation to get an accurate outcome.

The papers propose the use of Machine Learning classifiers for Breast Cancer Prediction.

3. Methodology

a) Dataset description

The Breast Cancer Wisconsin (Diagnostic) DataSet, obtained from Kaggle, contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and describe characteristics of the cell nuclei present in the image. There are 569 instances of the dataset. Each instance has 32 attributes with ID, diagnosis and 30 real-valued input features. The attribute information is as follows

Attribute information

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. Ten real-valued features are computed for each cell nucleus:
   - radius (mean of distances from center to points on the perimeter)
   - texture (standard deviation of gray-scale values)
   - perimeter
   - area
   - smoothness (local variation in radius lengths)
   - compactness (perimeter^2 / area - 1.0)
   - concavity (severity of concave portions of the contour)
   - concave points (number of concave portions of the contour)
   - symmetry
   - fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

b) Flow

The proposed system consists of four stages namely pre-processing, features extraction, classification, training/ testing of data, and to classify images into normal and malignant images using SVM, Decision Tree and Bayesian classifier. The extracted individual and combination of features are then passed as input to the Machine learning classifiers.

c) Naïve Bayesion Classifier

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes performs consistently after reduction of number of attributes. According to Bayesian theorem

P (A|B) =P (A)*P (B/A)/P(B), Where *P (B|A)=P(A∩B)/P(A)*

Based on above formula, Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression it has the excellent interpretability same as decision tree and is able to use previous data to build analysis model for future prediction.

d) Evaluation Measures

The four terms used in computing evaluation measures are used for evaluating the model and are described here i.e. TP, FP, TN and FN. A confusion matrix for actual and predicted class is formed comprising of TP, FP, TN, and FN to evaluate the parameter. The significance of the terms is given below.

TP= True Positive (Correctly Identified)
TN= True Negative (Incorrectly Identified)
FP = False Positive (Correctly Rejected)
FN = False Negative (Incorrectly Rejected)

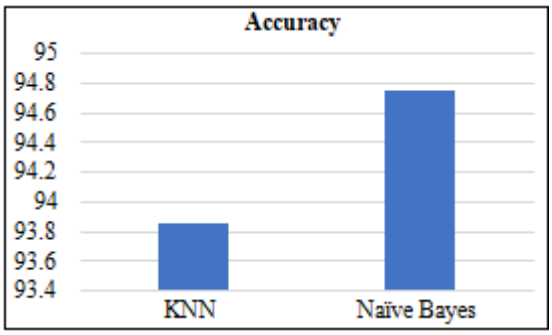
The performance of the proposed system is measured by the following formulas:

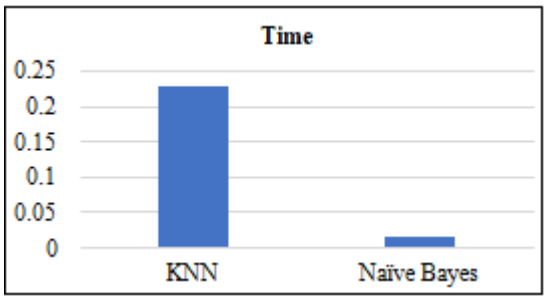
Accuracy (Acc)= (TP+TN)/(TP+TN+FP+FN)
Sensitivity (Sen) = TP/(TP+FN)
Specificity (Spec) = TN/(TN+FP)
Precision (PREC) – The proportion of the predicted relevant materials data sets that are correct:
Precision (%) = TP / (FP + TP)
Recall (REC) – The proportion of the relevant materials data sets that are correctly identified
Recall (%) = TP / (FN + TP)
F-Measure (FM) – Derives from precision and recall values:
F-Measure (%) = (2 x REC x PREC)/(REC + PREC)

4. Implementation and Result

The proposed model has been implemented in using Navies Bayes algorithm. We have used Scikit-learn which is an open-source software developed in Python for machine learning library. An Integrated development environment named as Jupyter is used to run the program. The results show that Naïve Bayes performs better than KNN in terms of accuracy and time taken in seconds.

Algorithm	Accuracy	Time
KNN	93.86	0.22
Naïve Bayes	94.74	0.01





5. Conclusion & Future Work

Breast cancer prediction is playing a crucial role in the current medical diagnosis. In this paper we propose the usage of Naive Bayes algorithm for the accurate and early predication of breast cancer. We have implemented the algorithm using Python and tested the same using data set. The test results show an accuracy of 94.74 and also reduces the time taken. The other evaluation parameters need to be calculated and compared with existing values. In future, we would like to extend the algorithm to test with larger dataset and check the scalability of the algorithm. More features like age, hereditary factors can be used for classification and training.

References

[1] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan, "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors ", 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) 21 - 23 Dec 2017, Dhaka, Bangladesh

[2] B.M.Gayathri, Dr.C.P.Sumathi,"Comparative study of Relevance Vector Machine with various machine learning techniques used for detecting breast cancer", 2016 IEEE International Conference on Computational Intelligence and Computing Research, 978-1-5090-0612-0/16/$31.00 ©2016 IEEE"

[3] Madhuri Gupta, Bharat Gupta," An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)", Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India

[4] Akshitha Shetty, Vrushika Shah "Survey of cervical cancer Prediction using Machine Learning: A comparative approach", 9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru, India

[5] Madhuri Gupta, Bharat Gupta "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques ",Proceedings of the Second International Conference on Computing Methodologies and Communication (ICCMC 2018) IEEE Conference Record # 42656; IEEE Xplore ISBN:978-1-5386-3452-3

[6] Naresh Khuriwal, Nidhi Mishra "Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm", 978-1-5386-1138-8/18/$31.00 ©2018 IEEE

[7] Pragya Chauhan, Amit Swami,"Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach",9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru Bengaluru, India

[8] Ayush Sharma, Sudhanshu Kulshrestha,Sibi Daniel," Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis" https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[9] https://www.kaggle.com/saramille/breast-cancer-prediction-knn-svc-and-logistic

[10] Mitchell T. 1997. Machine Learning. New York: McGraw Hill.