International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206457

Breast Cancer Prediction Using Machine Learning

Gaurav Singh
Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Article Info
Volume 6, Issue 4
Page Number: 278-284
Publication Issue :
July-August-2020

Article History
Accepted : 25 July 2020
Published : 30 July 2020

ABSTRACT

Breast cancer may be a prevalent explanation for death, and it's the sole sort of cancer that's widespread among women worldwide. The prime objective of this paper creates the model for predicting breast cancer using various machine learning classification algorithms like k Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic Regression (LR), and Gaussian Naive Bayes (NB). And furthermore, assess and compare the performance of the varied classifiers as far as accuracy, precision, recall, f1-Score, and Jaccard index. The breast cancer dataset is publicly available on the UCI Machine Learning Repository and therefore the implementation phase dataset is going to be partitioned as 80% for the training phase and 20% for the testing phase then apply the machine learning algorithms. k Nearest Neighbors achieved a significant performance in respect of all parameters.

Keywords : Breast Cancer, Machine Learning, Classification, Accuracy, Precision, k Nearest Neighbors.

I. INTRODUCTION

Around the world, Breast cancer is the most widely recognized type of cancer alongside lung and bronchus cancer, prostate cancer, colon cancer, and pancreatic cancer among others. Breast cancer might be a prevalent reason for death, and it's the main kind of malignant growth that is boundless among ladies in the around the world. Breast Cancer causes are multifactorial and include family ancestry, weight hormones, radiation treatment, and even reproductive factors. As indicated by the report of the world health organization every year, 2.1 million ladies are recently affected by breast cancer, and furthermore cause the highest number of cancer-related deaths among ladies [1]. In 2018, it is assessed that 627,000 ladies died from breast cancer - that is roughly 15% of all cancer deaths among ladies [1]. While breast cancer growth rates are higher among ladies in extra developed areas, rates are expanding in about each locale internationally.

Many imaging techniques are developed for early identification and treatment of breast cancer and to scale back the amount of death and lots of aided breast cancer diagnosis methods are wont to increase the symptomatic precision.

Machine Learning algorithms are widely utilized in intelligent human services frameworks,

particularly for breast cancer diagnosis and guess. There are many many machine learning classification and algorithms for prediction of breast cancer outcomes but during this paper, we are comparing various sorts of classification algorithms like k Nearest Neighbors, Support Vector Machine, Logistic Regression, and Gaussian Naive Bayes. And furthermore, assess and compare the performance of the varied classifiers as far as accuracy, precision, recall, f1-Score, and Jaccard index. The outcomes obtained during this paper provide a summary of the condition of modern Machine Learning strategies for breast cancer detection.

II. MACHINE LEARNING ALGORITHMS

Figure 1 shows the bosom breast cancer classification model with machine learning calculations, where the breast cancer dataset is loaded, features need to be extracted and therefore the classification model is often trained and used for prediction of benign and malignant. Benign cases are considered noncancerous, which is non-perilous. Harmful cancer begins with irregular cell development and may quickly spread or attack close-by tissue all together that it is regularly hazardous.

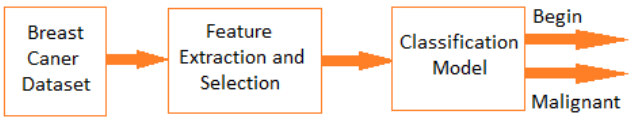

Figure 1. Breast Cancer Classification Model

A. k Nearest Neighbor (kNN)

k Nearest Neighbors algorithm utilizes 'feature similarity' to foresee the estimations of the most recent snippets of data which further methods the new information point will be assigned a value upheld how closely it matches the points inside the training set.

B. Support Vector Machine (SVM)

Support Vector Machine is of the Supervised Machine Learning characterization strategies that are broadly applied inside the field of cancer malignant growth determination and guess. Support Vector Machine works by choosing basic examples from all classes referred to as help vectors and isolating the classes by creating a linear function that partitions them as comprehensively as conceivable utilizing these help vectors. In this way, it is regularly said that planning between an input vector to a high dimensionality space is framed utilizing Support Vector Machine that intends to search out the preeminent reasonable hyperplane that separates the data set into classes. This linear classifier intends to expand the space between the decision hyperplane and along these lines the closest data, which is named the minimal distance, by finding the most appropriate hyperplane.

C. Logistic Regression (LR)

Logistic Regression is a key machine learning classification procedure. It has a place with the gathering of linear classifiers and is fairly practically like polynomial and statistical regression. Logistic regression is quick and similarly simple, and it's helpful for you to decipher the outcomes. In spite of the fact that it's basically a path for binary classification, it additionally can be applied to multi-class issues. This is frequently not the same as statistical regression, as statistical regression contemplates with the forecast of consistent qualities. Logistic regression models the likelihood that reaction falls into a specific classification. A logistic regression model helps us solve, via the Sigmoid function, for situations where the output can take but only two values, 0 or 1.

D. Naive Bayes (NB)

Naive Bayes is a classification method bolstered Bayes' Theorem with a presumption of independence among predictors. In straightforward terms, a Naive Bayes classifier considers that the nearness of specific features during a class is inconsequential to the nearness of the other element. despite the fact that these features rely on each other or upon the presence of the contrary features, those properties freely add to the likelihood of a class which is the reason it's referred to as 'Naive'. Naive Bayes (NB) is 'naive' in light of the fact that it makes that features of estimation are free of each other. this is frequently naive in light of the fact that it's (nearly) never evident. Naive Bayes model is easy to make and especially valuable for huge data sets. nearby straightforwardness, Naive Bayes is comprehended to beat even profoundly sophisticated classification methods.

III. ABOUTUT DATASET

This paper is predicated on a dataset that is openly accessible from the UCI Machine Learning Repository [2]. The dataset comprises of a few hundred human cell test records, every one of which contains the estimations of a gathering of cell qualities. The dataset having the resulting attributes:

i. ID Number Clump Thickness
ii. Uniformity of Cell Size
iii. Uniformity of Cell Shape
iv. Marginal Adhesion
v. Single Epithelial Cell Size
vi. Bare Nuclei
vii. Bland Chromatin
viii. Normal Nucleoli
ix. Mitoses
x. Class

The ID Number attribute contains the patient identifiers. The qualities of the cell tests from every patient are contained in attribute Clump Thickness to Mitoses. The values are evaluated from 1 to 10, with 1 being the nearest to begin. the class field contains the conclusion, as affirmed by isolated clinical procedures, on whether the tests are begins (value = 2) or malignant (value = 4).

Table I. shows the statistics of classes in the dataset.

Class	Instances	% Distribution
Begin	458	65.52
Malignant	241	34.48
Total	699	100

IV. LITERATURE REVIEW

Benbrahim et al. [3] use classification experimentation to call attention to that the most straightforward accuracy inside the paper was accomplished by the Neural Network calculation, which had, in its best configuration, 96.49% of exactness.

Deepika et. al. [4] uses two classification algorithms Naive Bayes and Multi-Layer Perceptron and after analyzing the performance of both algorithm found that Naive Bayes gives the more accurate results.

Mariam et. al. [5] uses two different classifiers namely Naive Bayes and K Nearest Neighbors for breast cancer classification on comparing accuracy using cross-validation and KNN achieved that 97.51% accuracy with lowest error rate then Naive Bayes Classifier 96.19% accuracy.

Aruna et al. [6] uses three different classifiers namely Naive Bayes, Support Vector Machine, and Decision Tree to classify a Wisconsin breast cancer dataset and

got the best outcome by utilizing a support vector machine with an accuracy score of 96.99%.

Chaurasia et al. [7] looked at the performance of supervised learning classifiers by utilizing a Wisconsin breast cancer growth dataset and Naive Bayes, Support Vector Machine, Neural Networks, Decision Tree techniques applied. reliable with the investigation results, the Support Vector Machine gave the chief the exact outcome with a score of 96.84%.

V. OBSERVATION

Confusion matrix is a table that's frequently wont to depict the performance of a classification model on a gathering of test information that truth values are known.

TABLE II
CONFUSION MATRIX

		Predicted Class	
Actual Class		Class=Yes	Class=No
	Class=Yes	True Positive (TP)	False Negative (FN)
	Class=No	False Positive (FP)	True Negative (TN)

In Table II TP and FP are the observations that are accurately predicted and hence shown in blue shading. we might want to decrease false positives and false negatives all together that they have appeared in red shading.

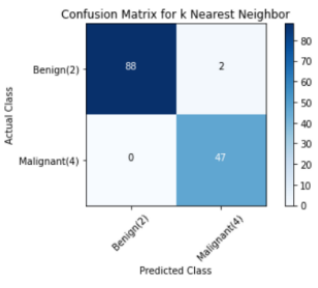

Figure 2(a)

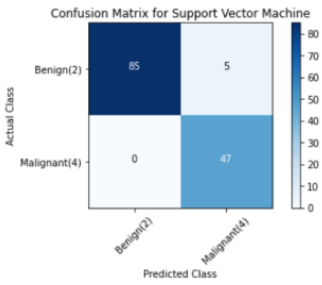

Figure 2(b)

Figure 2(c)

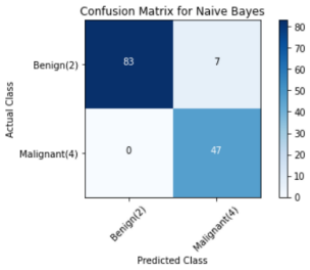

Figure 2(d)

Figure 2. Graphical Representation of Confusion Matrix

A. Accuracy

The classifier exactness is a proportion of how well the classifier can accurately predict cases into their right classification. it's the number of right forecasts separated by the whole number of instances within the data set. it's significant that the accuracy is extremely reliant on the edge picked by the classifier and may, hence, change for different testing sets. Along these lines, it's not the ideal technique to check various classifiers but rather may give a rundown of the classification. Hence, accuracy are often calculated using the following equation:

Accuracy = (TP+TN) / (TP+FP+FN+TN)

Table III shows the accuracy values for all four machine learning algorithms.

TABLE IIII. ACCURACY VALUES

Algorithms	Accuracy
kNN	0.99
SVM	0.96
LR	0.97
NB	0.95

B. Recall

Recall, likewise generally referred to as sensitivity, is that the pace of the positive predictions that are effectively predicted as positive. This measure is attractive, particularly within the clinical field because of what level of the observations are accurately analyzed. during this examination, it's progressively imperative to appropriately recognize a threatening neoplasm than it's to inaccurately distinguish a considerate one.

Recall = TP / (TP+FN)

Table IV shows the recall values for all four machine learning algorithms.

TABLE IV. RECALL VALUES

Algorithms	Begin	Malignant	Average
kNN	0.98	1.00	0.99
SVM	0.94	1.00	0.97
LR	0.96	1.00	0.98
NB	0.92	1.00	0.96

C. Precision

Precision, additionally generally referred to as confidence, is that the pace of both true positive and true negative that are distinguished as obvious positive. This shows how well the classifier handles the positive observations however doesn't say a lot of regarding the negative ones.

Precision =TP / (TP+FP)

Table V shows the precision values for all four machine learning algorithms.

TABLE V. PRECISION VALUES

Algorithms	Begin	Malignant	Average
kNN	1.00	0.96	0.98
SVM	1.00	0.90	0.95
LR	1.00	0.92	0.96
NB	1.00	0.87	0.94

D. F1-Score

F1-Score is the weighted harmonic mean of Precision and Recall. Subsequently, this score takes both false positive and false negative into thought.

F1-Score = 2*(Recall * Precision) / (Recall + Precision)

Table VI shows the f1-score values for all four machine learning algorithms.

TABLE VI. F1-SCORE VALUES

Algorithms	Begin	Malignant	Average
kNN	0.99	0.98	0.98
SVM	0.97	0.95	0.96
LR	0.98	0.96	0.97
NB	0.96	0.93	0.95

E. Jaccard Index

Jaccard Index likewise referred to as the Jaccard similarity score is a measurement used in understanding the similarities between test sets. The estimation underscores the similarity between limited test sets and is officially defined because of the fact that the size of the crossing point partitioned by the size of the union of the test sets.

Table VI shows the Jaccard index for all four machine learning algorithms.

TABLE IIII. JACCARD INDEX VALUES

Algorithms	Jaccard
kNN	0.98
SVM	0.96
LR	0.97
NB	0.94

VI. RESULTS AND DISCUSSION

Table VII shows all five parameter values for all four machine learning algorithms.

TABLE IVI. PARAMETER VALUES

Algorithms	Accuracy	Precision	Recall	F1-Score	Jaccard
kNN	0.99	0.98	0.99	0.98	0.98
SVM	0.96	0.95	0.97	0.96	0.96
LR	0.97	0.96	0.98	0.97	0.97
NB	0.95	0.94	0.96	0.95	0.94

In Table-VII k nearest neighbor accomplishes the critical performance as far as accuracy, precision, recall, f1-score and Jaccard index are 0.99, 0.98, 0.99, 0.98, and 0.98 respectively. Logistic Regression accomplishes the second performance as far as accuracy, precision, recall, f1 score and Jaccard index are 0.97, 0.96, 0.98, 0.97 and 0.97 respectively. Support Vector Machine accomplishes the third performance as far as accuracy, precision, recall, f1 score and Jaccard index are 0.96, 0.95, 0.97, 0.96 and 0.96 respectively. Naive Bayes accomplishes the fourth performance as far as accuracy, precision,

recall, f1 score and Jaccard index are 0.97, 0.96, 0.98, 0.97 and 0.97 respectively.

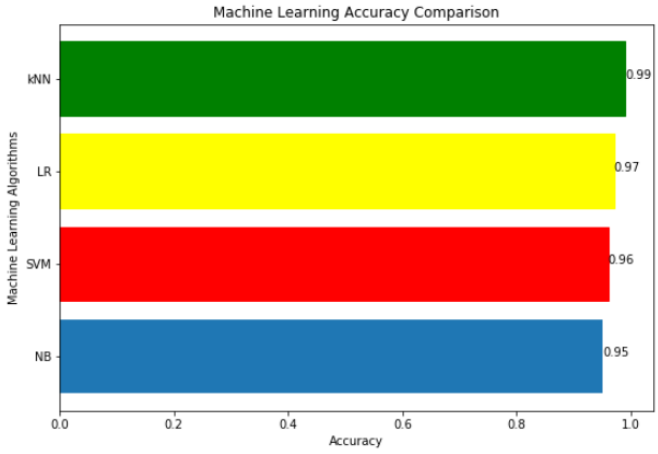

Figure 2. Graphical Representation of Accuracy Comparison

Figure 2 shows that k Nearest Neighbor gives the more accurate algorithm having classified the samples with 99% accuracy in conventional validation. Logistic Regression, Support Vector Machine and Naive Bayes comes second, third, and fourth respectively in classification accuracy.

This comparative investigation shows that the classification accuracy, precision, recall, f1-score and Jaccard index of k Nearest Neighbor is above Support Vector Machine, Logistic Regression and Naive Bayes classification algorithm within the predictive breast cancer data from the UCI Machine Learning Repository Wisconsin breast cancer dataset. we have seen that k Nearest Neighbor gives critical performance classification algorithm as far as accuracy, precision, recall, f1-score, and Jaccard index.

The limitation of this analysis is that the size of the information used. the amount of samples used for training and testing is low. The analysis of information with respect to the clinical settings should be administered with a bigger dataset.

VII. CONCLUSION

In this paper, we have compared the classification parameters as far as four Machine Learning algorithms, in particular, k Nearest Neighbor, Support Vector Machine, Logistic Regression, and Naive Bayes available on UCI Machine Learning Repository Wisconsin Wisconsin breast cancer dataset. the target of this comparative analysis was to search out the foremost accurate machine learning algorithm which will act as a tool for the diagnosis of breast cancer. consistent with the prediction results, k Nearest Neighbor has the very best accuracy for the given dataset. This shows k Nearest Neighbor is regularly better for the prediction of breast cancer as compared with Support Vector Machine, Logistic Regression, and Naive Bayes.

VIII. ACKNOWLEDGEMENT

I have finished this work under the guidance of Dr Pankaj Agarwal(Professor & head) & Ms Sapna Yadav(Assistant Professor), Department of Computer Science and Engineering at IMS Engineering College, Ghaziabad, Uttar Pradesh. I am doing an online Summer Internship on Machine Learning where I have learn various Machine Learning algorithm from both of my mentors as a course instructor. This paper has been assigned as a project assignment for us.

I would like to express my special thanks both of my mentors for inspiring us to complete work and write a paper. Without their active guidance, help cooperation & encouragement, I would not lead way in writing the paper. I am extremely thankful for their valuable guidance and support on completion of this paper.

I extend my gratitude to "IMS Engineering College, Ghaziabad, Uttar Pradesh" for giving me this opportunity. I also acknowledge with a deep sense of reverence, my gratitude towards my friends, parents and member of my family, who always supported me morally, mentally as well as economically.

Any omission in this brief acknowledgement does not mean a lack of gratitude.

IX. REFERENCES

[1] WHO – Breast Cancer https://www.who.int/cancer/

[2] Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin, USA http://archive.ics.uci.edu/ml/datasets.php

[3] Benbrahim H., Hachimi H., Amine A. (2020) Springer, Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset.

[4] Deepika Verma and Nidhi Mishra,"Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques" 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).

[5] Mariam Amrane, Saliha Oukid, Ikram Gagaoua and Tolga Ensari, "Breast cancer classification using machine learning"2018 Electric Electronics, Computer Science, Biomedical Engineerings, Meeting (EBBT).

[6] Aruna S, Rajagopalan S and Nandakishore L, "Knowledge based analysis of various statistical tools in detecting breast cancer"2011 Computer Science Information Technology.

[7] Chaurasia V and Pal S,"Data mining techniques: To predict and resolve breast cancer survivability" Int. J.Computer Science 2014.

About Author

Gaurav Singh is B.Tech student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. His areas of interest is Programming in Python, Data Science and Machine Learning.

Cite this article as :

Gaurav Singh, "Breast Cancer Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 278-284, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206457
Journal URL : http://ijsrcseit.com/CSEIT206457