International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 01 | Jan - 2022 ISSN: 2582-3930

MACHINE LEARNING ALGORITHMS TO PREDICT BREASTCANCER IN WOMENS

Archana Rana

Amity University, Haryana

Guided By:-

Dr. Ashima Narang

Assistant Professor, CSE

Amity University, Haryana

Abstract

Each year number of deaths is increasing extremely because of breast cancer. It is the most successive sort of all cancers and the significant reason for death in ladies around the world. Any advancement for prediction and finding of malignant growth illness is capital significant for a sound life. Subsequently, high precision in disease forecast is critical to refresh the treatment viewpoint and the survivability standard of patients. Machine Learning procedures can bring a huge contribute on the course of forecast and early finding of bosom malignant growth, turned into an examination area of interest and has been demonstrated as a solid method. In this review, two of various Machine Learning algorithms are discussed: Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). India has witnessed 30% of the cases of breast cancer during the last few years and it is likely to increase. Breast cancer in India accounts that one woman is diagnosed every two minutes and every nine minutes, one woman dies. Early detection and diagnosis can save the lives of cancer patients. This paper presents a novel method to detect breast cancer by employing techniques of Machine Learning.

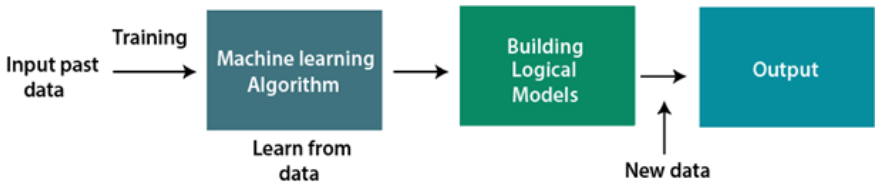
1. Introduction

1.1 Machine Learning

In reality, we are encircled by people who can gain everything from their encounters with their learning ability, and we have PCs or machines which work on our directions. Be that as it may, would a machine be able to likewise gain from encounters or past information like a human does? So here comes the job of Machine Learning.

Machine Learning is said as a subset of Artificial Intelligence that is principally worried about the improvement of calculations which permit a PC to gain from the information and previous encounters all alone. The term machine learning was first presented by Arthur Samuel in 1959. A Machine Learning framework gains from chronicled information, assembles the expectation models, and at whatever point it gets new information, predicts the result for it. The precision of anticipated result relies on how much information, as the colossal measure of information assists with building a superior model which predicts the result even more precisely.

Assume we have a mind-boggling issue, where we want to play out certain expectations, so rather than composing a code for it, we simply need to take care of the information to nonexclusive calculations, and with the assistance of these calculations, machine constructs the rationale according to the information and foresee the result. AI has changed our perspective with regards to the issue. The underneath block graph clarifies the working of Machine Learning calculation:



1.2 Applications of Machine learning

Machine Learning is a trendy expression for the present innovation, and it is becoming quickly step by step. We are utilizing machine learning in our regular routine even without knowing it like Google Maps, Google colleague, Alexa, and so on. The following are some most moving genuine uses of Machine Learning:

- **Image Recognition:** Image Recognition is one of the most widely recognized uses of machine learning. It is utilized to recognize objects, people, places, advanced pictures, and so forth the famous use instance of Image Recognition and face discovery is,
  Automatic companion labeling idea: Facebook gives us an element of auto companion labeling idea. At whatever point we transfer a photograph with our Facebook companions, then, at that point, we consequently get a labeling idea with name, and the innovation behind this is AI's face identification and acknowledgment calculation.
- **Traffic prediction:** If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:
  - ✓ Real Time location of the vehicle form Google Map app and sensors

- ✓ Average time has taken on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

- **Self-driving cars:** One of the most thrilling utilizations of machine learning is self-driving vehicles. AI assumes a critical part in self-driving vehicles. Tesla, the most famous vehicle fabricating organization is chipping away at self-driving vehicle. It is utilizing unaided learning technique to prepare the vehicle models to distinguish individuals and items while driving.
- **Medical Diagnosis**: In medical science, machine learning is utilized for illnesses analyze. With this, clinical innovation is becoming extremely quick and ready to construct 3D models that can anticipate the specific place of injuries in the mind. It helps in finding mind cancers and other cerebrum related illnesses without any problem.
- **Speech Recognition:** While utilizing Google, we get a choice of "Search by voice," it goes under speech recognition, and it's a well-known use of machine learning.
  Speech recognition is a process of changing over voice directions into message, and it is otherwise called "Text to Speech", or "Computer Speech Recognition". As of now, machine learning calculations are broadly utilized by different utilizations of speech recognition. Google aide, Siri, Cortana, and Alexa are utilizing speech recognition innovation to adhere to the voice guidelines.

As of now, we have discussed about machine learning and its applications. Now, lets discuss the broad use of machine learning in the field of medical science for detecting many harmful diseases mainly cancer.

2. Introduction of Machine Learning in Healthcare

Throughout the most recent many years, a consistent advancement connected with disease research has been performed [1]. Researchers applied various techniques, like separating beginning phase, to observe kinds of disease before they cause indications. In addition, they have grown new systems for the early expectation of malignant growth treatment result. With the appearance of new innovations in the field of medication, a lot of disease information have been gathered and are accessible to the clinical examination local area. Nonetheless, the exact forecast of an infection result is quite possibly the most intriguing and testing undertakings for doctor. Accordingly, Machine Learning strategies have turned into a famous instrument for clinical analysts. These methods can find and distinguish examples and connections between them, from complex datasets, while they can viably foresee future results of a malignant growth type. Given the meaning of customized medication and the developing pattern on the use of ML strategies, we here present an audit of studies that utilize these techniques with respect to the disease expectation and anticipation. In these examinations, prognostic and

prescient highlights are viewed as which might be free of a specific therapy or are coordinated to direct treatment for disease patients, separately [2]. Moreover, we examine the kinds of ML techniques being utilized, the sorts of information they incorporate, the general execution of each proposed conspire while we additionally talk about their advantages and disadvantages. An undeniable pattern in the proposed works incorporates the mix of blended information, for example, clinical and genomic. In any case, a typical issue that we saw in a few works is the absence of outer approval or testing with respect to the prescient presentation of their models.

The use of ML strategies could work on the precision of disease helplessness, repeat, and endurance forecast. In view of [3], the exactness of disease forecast result has fundamentally improved by 15%-20% the last years, with the utilization of ML procedures. A few investigations have been accounted for in the writing and depend on various techniques that could empower the early malignant growth determination and guess [4-7]. Despite the fact that quality marks could essentially further develop our capacity for anticipation in malignant growth patients, helpless headway has been made for their application in the centers. Be that as it may, before quality articulation profiling can be utilized in clinical practice, studies with bigger information tests and more satisfactory approval are required. In the current work just examinations that utilized ML procedures for demonstrating disease conclusion and forecast are introduced.

Albeit a few strategies have been presented, the methods are not ready to give an exact and dependable result. Every one of the modalities are associated with specialists or doctors or other clinical staffs. In this way, a framework which can work with next to no clinical gears and clinical staffs might prompt a proper arrangement. We acquaint an imaginative methodology with order the information credits as indicated by the presence or nonappearance of harmless or threatening sorts of bosom malignant growth. We have utilized two administered Machine learning procedures named as Support Vector Machine and K-Nearest Neighbors which relate to learning calculations that examine information used for relapse examination and characterization to recognize the bosom malignant growth.

Machine Learning Algorithms

We are going to discuss two of the most efficient machine learning algorithms. First is Support Vector Machine (SVM) which is utilized for a linear problem and The K-Nearest Neighbors (K- NN) is used for the nonlinear problem.

A. Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best

line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

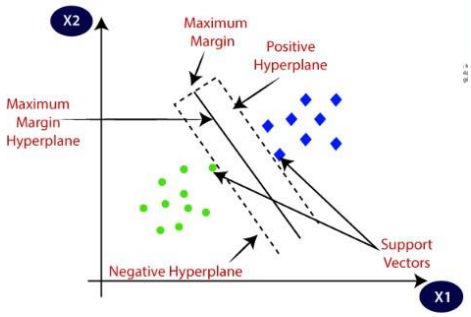


Fig. SVM Classifier Model

Support Vector Machine is a discriminative classifier that can be defined by a separating hyperplane. It is the generalization of maximal margin classifier which comes with the definition of hyperplane. In an n-dimensional space, the hyperplane is of (n-1) dimension with flat subspace that need not pass through the origin. The hyperplane is not visualized in higher dimension but the notion of an (n-1) dimensional flat subspace still applies [10]. If there doesn't exist any linearly separable hyperplane for any dataset, linear classifier can't be formed in that case. Kernel tricks must be applied to maximum-margin hyperplanes to develop nonlinear classifier. According to this, nonlinear kernel function will be applied to the hyperplanes in replacement of dot product. Cubic, quadratic or higher-order polynomial function, Gaussian Radial basis function or Sigmoid function are forms of nonlinear kernel function. In p-dimensions, a hyperplane is described as follows.

$$\beta 0 + \beta 1 X1 + \beta 2\, X2 + ... + \beta p\, X\, p = 0 \qquad (1)$$

where β0, β1, β2 …βp are the hypothetical values and Xp are the data points in sample space of p dimension.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.

- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities)

B.K-Nearest Neighbors (K-N N)

K-Nearest Neighbors (K-NN) algorithm is another supervised machine learning technique used for classification and regression. K-NN doesn't make any assumptions on the fundamental data distribution. It performs great in pattern recognition and predictive analysis. For any new data point, firstly K-NN gather data points that are close to it. Any attributes that can vary on a large scale may have effective impact on the distance between data points [10]. The algorithm then sorts those closest data points in terms of distance from the arrival data point. This distance can be measure in various way but Euclidian distance is the suggested one by experts. Next step is to take a specific number of data points whose distance are lesser among all and then categorize those data point. In KNN, the number of closest data points are usually chosen as an odd number if the number of classes is 2. The category with highest number of data point will be the category of the new data point.

The advantages of support K-NN Classifier include:

- The training process is speedy, and the cost is zero.
- Simple and fast to execute.
- It copes with the noise data.
- And convincing if the training data are immense.
- The algorithm is successful in computing more than one class label for an unknown instance.

The disadvantages of K-NN Classifier include:

- Computationally, it's very costly.
- Simple and fast to execute.
- It's very sensitive to unrelated features.
- It's a lazy algorithm that needs more time to run.
- It takes a considerable amount of memory that stores all the training examples.

- The estimated cost is high as the distance of each instance to all training tests is needed for a computer, and the value of K must be determined

The Performance Measure Indices

The performance of machine learning techniques is measured in terms of some performance measure indices. 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)

An existing system is being taken here to show the efficiency on different parameters of SVM and K-NN.

Parameters	Training Phase		Testing Phase	
	SVM	K-NN	SVM	K-NN
Accuracy (%)	99.68	98.25	98.57	97.14
Sensitivity (%)	99.76	99.26	100	100
Specificity (%)	99.54	96.40	95.65	92.31
Geometric Mean of Sensitivity and Specificity (%)	99.65	97.83	97.83	96.16
False Discovery Rate (%)	0.24	1.94	2.08	4.35
False Omission Rate (%)	0.46	1.38	0	0
Matthews Correlation Coefficient	0.99	0.96	0.97	0.94

Table1.Performance measure indices.

CONCLUSION

Breast cancer prediction is very significant in Medicare and Biomedical. In this paper we discussed on building a classifier which aims at predicting the most severe cancer known as breast cancer. Breast cancer is a remarkably risk disease that causes a lot of death for numerous ladies all over the world. So, early detection of this cancer can save a lot of valuable life. We algorithms for detecting breast cancer based on Support Vector Machine and K-Nearest Neighbors. In existing system, SVM has been implemented by the Python to be the most effective in classifying the diagnostic data set into the two classes in view of the seriousness of the cancer. We end up with an accuracy of 99.68% in SVM in training phase. The classifier obtained by supervised machine learning techniques will be very supportive in the field of medical disorders and proper diagnosing.

REFERENCES

[1] Breast cancer statistics. [Online]. Available:http://www.wcrf.org/int/cancer-facts-figures/data-specific-ancers/breastcancer- statistics, accessed on: Aug. 25, 2017.

[2] Arbab Masood Ahmad, Gul Muhammad, Khan, S.Ali Mahmud, Julian F.Miller, ‘‘Breast Cancer Detection Using Cartesian Genetic Programming evolved Artificial Neural Networks, ’’Philadelphia, Pennsylvania, USA,GECCO ’12, July 7–11, 2012.

[3] Ahmad Taher Azar, Shaimaa Ahmed El-Said, ‘‘Probabilistic neural network for breast cancer classification, ’’ Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.

[4] Warner E, Messersmith H, Causer P et al, ‘‘Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer, ’’Annals of Internal Medicine,148(9):671–679,06 May, 2008.

[5] Emina Alic ̆kovic ́, Abdulhamit Subasi, ‘‘Breast cancer diagnosis using GA feature selection and Rotation Forest ’’Neural Computing and Applications, Springer, Volume 28, issue 4, pp 753–763, April 2017.

[6] Fadzil Ahmad, Nor Ashidi Mat Isa, Zakaria Hussain, Siti Noraini Sulaiman, ‘‘A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis ’’, Neural Computing and Applications, Springer, Volume 23, Issue 5, pp 1427–1435 ,October 2013,.

[7] M. K. Hasan, M. M. Islam and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, pp. 574-579, 2016.

[8] H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, pp. 144-149, 2017.