Prediction and Diagnosis of Breast Cancer Using Machine Learning Algorithms

Goutam R, Sachitha S
Assistant Professor, Student
Computer Science and Engineering,
Atria Institute of Technology, Bangalore, India.

Abstract: Machine learning is as often as possible utilised in clinical applications, for example, discovery of the kind of dangerous cells. Breast malignant growth speaks to one of the ailments that causes a high number of deaths consistently. It is the most widely recognised sort of malignant growth and the fundamental driver of women demises around the world. The harmful cells are of two types Benign (B) or Malignant (M). There are numerous algorithms for characterisation and expectation of breast cancer: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbors (kNN). In this venture, Support Vector Machine (SVM) on the Wisconsin Breast Cancer dataset is utilised. The dataset is likewise prepared with the different algorithms: KNN, Naives Bayes and CART and the exactness of expectation for every algorithm is looked at.

IndexTerms - Breast Cancer, kNN, Naives Bayes, CART, SVM.

I. INTRODUCTION

Breast cancer disease is a kind of malignant growth that happens generally in females and is the main source of women's demise. These deaths can be decreased by early discovery of the harmful cells. Malignant cells are identified by performing different tests like MRI, mammogram, ultrasound and biopsy. The dataset utilised right now includes those that are figured from a digitized picture of a fine needle aspiration (FNA) biopsy of a breast mass. They depict qualities of the cell cores present in the picture. Determination of breast cancer is finished by identifying the tumor. Tumors can be either benevolent or harmful. Threatening tumors are more hurtful than the favorable. Sadly, not all doctors are masters in recognizing between the favorable and dangerous tumors and the identifying of tumor cells may take as long as 2 days. Machine Learning calculations are utilised to anticipate the sort of carcinogenic cells productively and precisely. Machine learning is a use of man-made consciousness (AI) that gives frameworks the capacity to naturally take in and improve from experience without being expressly modified. Machine learning centers around the advancement of computer programs that can get to information and use it to learn for themselves. The various algorithms utilised are: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbors (kNN).

A. K – Nearest Neighbor (KNN)

KNN makes forecasts utilising the preparation dataset legitimately. Forecasts are made for another example (x) by looking through the whole preparing set for the K most comparative examples (the neighbors) and abridging the yield variable for those K cases. For regression this may be the mean yield variable, in grouping this may be the mode (or generally normal) class value. To figure out which of the K occurrences in the preparing dataset are generally like another info a separation measure is utilised. For real-valued information factors, the most well known separation measure is Euclidean separation. Euclidean separation is determined as the square base of the total of the squared contrasts between another point (x) and a current point (xi) over totally input traits j. Euclidean Distance(x, xi) = sqrt( sum( (xj – xij)^2 )

The preparation models are vectors in a multidimensional highlight space, each with a class name. The preparation period of the calculation comprises just putting away the element vectors and class names of the preparation tests. In the arrangement stage, k is a client characterised consistently, and an unlabelled vector(an inquiry or test point) is arranged by relegating the name which is generally visited among the k preparing tests closest to that question point.

B. Naives Bayes

Naive Bayes classifiers are an assortment of characterisation algorithms dependent on Bayes' Theorem. It's anything but a solitary algorithm however a group of algorithms where every one of them share a typical guideline, for example each pair of highlights being grouped is free of one another. Bayes' Theorem is expressed as: P(h/d) = (P(d|h) * P(h))/P(d)

Naive Bayes is a grouping algorithm for double (two class) and multi-class characterisation issues. The system is most effortless to comprehend when portrayed utilising paired or clear cut information esteem. Expect that we have a dataset with two classes of information inside. We have a condition for the likelihood of a bit of information having a place with Class 1:p1(h,d), We have a condition for the class having a place with Class 2:p2(h,d). To group another estimation with highlights (h,d), we utilise the accompanying rules:If p1(h,d) > p2(h,d), at that point the class is 1. If p2(h,d) > p1(h,d), at that point the class is 2.

C. Classification and Regression Trees (CART)

A Classification and Regression Tree (CART), is a prescient model, which clarifies how a result variable's qualities can be anticipated dependent on different qualities. A CART yield is a choice tree where each fork is a part in an indicator variable and each end hub contains a forecast for the result variable. The portrayal for the CART model is a parallel tree. Each root hub speaks to a solitary input variable (x) and a split point on that factor (accepting the variable is numeric). The leaf hubs of the tree contain a yield variable (y) which is utilised to make a expectation.
D. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a regulated machine learning algorithm which can be utilised for both arrangement and what's more, relapse difficulties. Be that as it may, it is for the most part utilised in grouping issues. Right now, plot every datum thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each component being the estimation of a specific arrangement. At that point, we perform grouping by finding the hyper-plane that separates the two classes quite well. Support Vectors are essentially the coordinates of individual perception. Support Vector Machine is a wilderness which best isolates the two classes (hyper-plane/line). At first SVMs map the information vector into an element space of higher dimensional and recognize the hyper plane that isolates the information focuses into two classes. The peripheral separation between the choice hyper plane and the occurrences that are nearest to limit is expanded. The coming about classifier accomplishes significant generality also, can in this manner be utilised for the dependable order of new examples. It is important that probabilistic yields can likewise be obtained for SVMs; the figure underneath outlines how a SVM may work so as to order tumors among generous and dangerous dependent on their size and patients' age. The recognised hyper plane can be thought as a choice limit between the two groups. Clearly, the presence of a choice limit takes into consideration the identification of any misclassification created by the strategy.

II. LITERATURE SURVEY

From [1] the Wisconsin Breast Cancer dataset was gotten. In proposed a Support Vector Machines (SVMs) based classifier in correlation with Bayesian classifiers and Artificial Neural Networks for the prognosis and analysis of breast cancer disease. The paper gives the execution subtleties alongside the relating results for all the evaluated classifiers. A SVM model is executed for the breast malignancy determination and forecast issue utilising the Wisconsin Diagnostic Breast Cancer (WDBC) just as the Wisconsin Prognostic Breast Cancer (WPBC) datasets. The enhanced SVM calculation performed brilliantly, displaying high estimations of exactness (up to 96.91%), explicitness (up 97.67%) and affectability (up to 97.84%). The counterfeit neural system has been the most generally utilized prescient strategy in clinical forecast, despite the fact that its structure is hard to comprehend. The paper records out the advantages and impediments among different machine learning strategies, for example, Decision trees, Naïve Bayes, neural systems and SVM. In [2], every calculation acts in an alternate way depending on the dataset and the parameter determination. For a by and large approach, the KNN system has given the best outcomes. Naive Bayes and logistic regression have additionally performed well in the conclusion of breast cancer. But SVM is the most fit method for recurrence/non-recurrence forecast of breast disease.

III. MATERIALS AND METHODOLOGY

Materials that we have utilised include: Python programming for coding and breast malignant growth information from UCI safe. Our procedure includes utilization of Machine Learning systems for example, SVM, KNN, decision trees and Naives bayes.

A. Dataset

The Wisconsin Diagnostic Breast Cancer dataset was acquired from the UCI Machine Learning safe (accessible at: http://archive.ics.uci.edu/ml). The dataset contains 357 instances of considerate breast malignancy and 212 instances of threatening breast malignancy. The dataset contains 32 sections, with the primary segment being the ID number, the second segment being the determination result (kind or dangerous), followed by the mean, standard deviation and the mean of the most exceedingly worst estimations of ten highlights. There were no missing qualities. The highlights are acquired from a digitized picture of a fine needle aspiration biopsy of the tumor. These highlights depict the cores of the cell. The unique highlights are as appeared:
B. Methodology

The dataset is partitioned into a preparing set and testing set. 80% of the information is utilized to prepare the framework and the remaining 20% is utilized for testing. From the dataset, we break down and manufacture a model to foresee if a given arrangement of manifestations lead to breast disease. The machine learning algorithms are prepared on the preparation information, and tried on the undeveloped information. In the event that the model is too intricate, for example, having an excessive number of parameters, it is probably going to prompt the issue of overfitting. Similarly, if the model is too straightforward that can't catch the fundamental pattern of the information, underfitting happens. Both overfitting and underfitting lead to poor prescient execution. There are a few methods to survive overfitting, for example, cross validation, regularization and drop out. One of the most ordinarily utilized strategies is k-fold cross validation, where the first information is haphazardly divided into k equivalent estimated subsamples. Out of the k subsamples, one subsample is accustomed to testing the model, and the rest of the k-1 subsamples are utilized to prepare the model. The k results are then arrived at the midpoint of to produce one single estimation. One favorable position of k-fold cross approval is each testing subsample is utilized precisely once. Support vector machine (SVM), a twofold classifier, looks at the hyperplane leaving the biggest conceivable portion of purposes of the equivalent class on a similar side, while boosting the separation of each class from the hyperplane. SVMs are a later approach of machine learning techniques applied in the field of malignant growth forecast/anticipation. At first SVMs map the information vector into a component space of higher dimensionality and distinguish the hyper plane that isolates the information focuses into two classes. The minor separation between the choice hyperplane and the examples that are nearest to limit is augmented. The subsequent classifier accomplishes impressive generalizability and can along these lines be utilized for the dependable order of new examples.

<table>
<thead>
<tr>
<th>TABLE I. FEATURES USED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Radius</strong></td>
</tr>
<tr>
<td><strong>Texture</strong></td>
</tr>
<tr>
<td><strong>Perimeter</strong></td>
</tr>
<tr>
<td><strong>Area</strong></td>
</tr>
<tr>
<td><strong>Smoothness</strong></td>
</tr>
<tr>
<td><strong>Compactness</strong></td>
</tr>
<tr>
<td><strong>Concavity</strong></td>
</tr>
<tr>
<td><strong>Concave points</strong></td>
</tr>
<tr>
<td><strong>Symmetry</strong></td>
</tr>
<tr>
<td><strong>Fractal dimension</strong></td>
</tr>
</tbody>
</table>
IV. FLOWCHART

The information got from the patients is utilized to frame a dataset. The dataset is partitioned into preparing and testing information, what's more, it is guaranteed that the dataset has no fragmented qualities. The four machine learning algorithms are applied and the precision of expectation of every calculation is analyzed. The calculation with best precision is utilized for expectation.

![Flowchart](image)

Fig 1. Flowchart

V. RESULTS AND DISCUSSIONS

A. DATA EXPLORATION

The conveyances of the mean, standard blunder and most exceedingly worst normal of the 10 highlights removed from the fine needle aspiration slides show that smallness, concavity, fractal measurement, smoothness and balance each have generally little qualities for the estimation. Border, range and surface each have generally enormous qualities for the estimation, with territories that show the biggest estimation worth and measure of variety for every one of the three estimations. From the dissemination perception, we can see by malignant analysis class has moderately higher mean for all the qualities.

B. Correlation

Among the mean estimation of the 10 properties, we can see a few of them are exceptionally connected between each other. The red around the slanting recommends that traits are associated with one another. The yellow and green patches propose some moderate relationship and the blue boxes appear negative relationships.

![Correlation graph](image)

Fig 2. Correlation graph

C. Count of Benign(B) and Malignant(M)

From the dataset we can see that the quantity of BENIGN patients are more than the quantity of MALIGNANT patients and this is demonstrated by the visual diagram underneath got in our jupyter notebook.
Malignant: 212
Benign: 357

Fig 3. Count of patients diagnosed with benign and malignant cancer

D. Performance Comparison
From the underlying run, it would appear that Gaussian NB, KNN and CART played out the best given the dataset (all above 92% mean exactness). Support Vector Machine has a shockingly terrible execution here. Be that as it may, in the event that we institutionalize the info dataset, it's presentation will improve.

Fig 4. Box plot before standardizing
In the wake of institutionalizing the dataset, precision of SVM radically improves as demonstrated as follows:

Fig 5. Box plot after standardizing

E. Calculation of Accuracy:
At the point when we compute exactness we get the results to be as demonstrated as follows:
Accuracy score 0.991282
VI. CONCLUSION AND FUTURE SCOPE
Every algorithm acts in an alternate manner relying upon the dataset and the parameter determination. For generally speaking procedure, KNN strategy has given the best outcomes. Naive Bayes and logistic regression have likewise performed well in determination of breast malignant growth. SVM is a solid system for prescient investigation and attributable to the above mentioned discovering, we infer that SVM utilizing Gaussian part is the most fit system for recurrence/non-recurrence forecast of breast malignant growth. The SVM that is utilized in the investigation right now as it were appropriate when the quantity of class variable is paired for example we can't have multiple classes. To take care of this issue researchers have thought of multiclass SVM. Further investigate right now as the production of SVM classes like LIBSVM has occurred. Adjusting of parameters utilized in algorithms can bring about better exactness. Moreover, this can likewise be actualized on a cloud stage for simplicity of use.

REFERENCES

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>75</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>1.00</td>
<td>0.99</td>
<td>39</td>
</tr>
<tr>
<td>Average / Total</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>114</td>
</tr>
</tbody>
</table>