IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10306

Applying Machine Learning Techniques for Early Prediction of Breast Cancer

Ms. R. ManjulaDevi[1], Dr. Saranya C P[2]

Student, Department of Computer Science and Engineering, CIET, Coimbatore[1]

Assistant Professor, Department of Computer Science and Engineering, CIET, Coimbatore[2]

Abstract: Women are the major strengths for world and family. Around 15 percent of the women are affected by breast cancer in various stages of their lives. It has become very common to find women with breast cancer nowadays. Though there are various treatments for breast cancer, the treatment gets effective only if it is identified at an earlier stage. It will be better if the cancer is diagnosed at the initial stage. Machine Learning provides a way in which the disease can be predicted at an early stage. There are various classification algorithms which can be used to classify the type of cancer mainly in terms of medical diagnosis. Using Breast Cancer dataset, classification algorithms like Linear Discriminant Analysis, Logistic Regression, KNN, SVM and Naïve Bayes are applied, and the data is analyzed for prediction. The classification algorithm determines the accuracy, recall, precision and F1 Score. SVM and Logistic regression has better performance than all other classification algorithms.

Keywords: Classification, Machine Learning, KNN, Logistic Regression, SVM, Recall.

I.INTRODUCTION

Breast cancer is the on which is very common among women in the global world. In most of the cases the cancer is diagnosed at a very late stage where treatment is not applicable. If the disease is predicted at an early stage then lives of many can be saved. There are various methods to test breast cancer like mammography, biopsy and ultrasound scan. But these are not the only test which can be used to identify cancer there are various other tests too. The cells collected from the women body are subjected for test.

Machines are making humans life easier with their intelligence. Machine learning algorithms are used to automate human behaviors using acquired data. There are three main types of machine learning algorithms. They are Supervised learning, unsupervised learning and reinforcement learning. Supervised learning classifies the new data using the earlier learnt dataset. The outcome of the new problem is compared with the expected outcome. Unsupervised Learning predicts the outcome of the new data based on probability of occurrences. Semi-Supervised Learning uses both labelled data as well as unlabeled data for training purpose so that expense can be reduced by using unlabeled data. Reinforcement Learning is a trial and error learning method which plays a major role in artificial intelligence. The three main components used in reinforcement learning are learner, agent and actions of the agent. Collaborative Learning is also used for recommendation system where a preference should be chosen from a huge number of choices. Clustering is an unsupervised learning in which the data is clustered based on the input properties. It divides the data set into group of clusters. Classifications is a supervised learning algorithm which classifies the data based on the existing outcome.

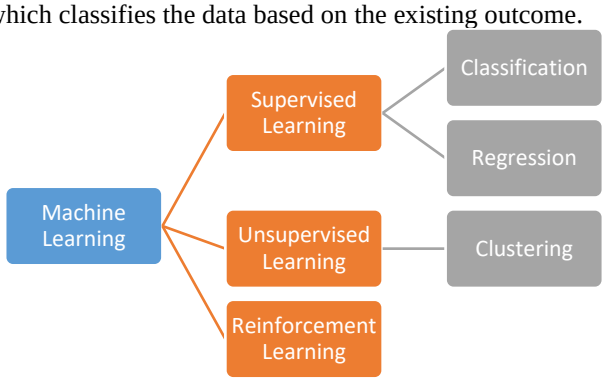

Figure 1: Machine Learning Techniques

II. LITERATURE REVIEW

Nonlinear algorithms like Random Forest, Naive Bayes, Support Vector Machine and K Nearest Neighbor are used for prediction of breast cancer for a comparative study. Authors used the Bioinformatics and Medical Science classification technique. This technique was based on the selection of best classifier, comparison of data mining algorithm was

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10306

performed to choose the most suitable algorithm for the prediction. SVM is found more accurate prediction algorithm than other algorithms which gave 97.9% accuracy.

To predict and detect breast cancer classification algorithms like Bagging Algorithm, IBk (Instance based learning with some parameters), Random Committee Algorithm, Random Forest Algorithm, Simple Classification and Regression Tree (Simple CART Algorithm) are used. the algorithm is implemented and tested using Antenna dataset. Weka was used to analyse the results. the authors proved that Random forest algorithm gives the best accuracy of 92.2%.

Breast prediction uses Gene Expression (GE) and DNA methylation data using Support Vector Machine (SVM), Decision Tree and Random Forest algorithm to classify nine models for the prediction of cancer. Weka and Spark tool was used to visualise the error rate and accuracy. Common genes are identified to exactly classify the presence of cancer in the cells. Naive Bayes, K Nearest Neighbors and J48 algorithm are used on the dataset collected from various doctors and hospitals. These datasets are used for predicting different types of cancer cells. The dataset contains 1059 medical records with 61 attributes each. Initially the symptoms are compared with the test results. Accuracy of the prediction of the algorithm was also determined. Accuracy of NB is 98.2%, KNN is 98.8% and J48 is 98.5%.

Support Vector Machine, a recursive feature elimination technique is used to develop a learning model. It chooses the correct features for benign and malignant records. The algorithm is evaluated, and the performance matrix is used to check the accuracy. This is checked on different kernels like linear, RBF and polynomial kernel where accuracy is 99%, 98%, 97% respectively.

Hyper Parameter Optimization is used to increase the accuracy of the classification model by forming different clusters. Dataset was collected from “National Cancer Institute” of Egypt. Clustering technique is used on the dataset collected to combine the data with similar properties. the features that are more likely for prediction process is identified.

III. METHODOLOGY

A.Feature Selection

Within the fields of machine learning high dimensional data analysis could be a challenge for re-searchers and engineers. Solving drawback by removing immaterial and redundant data through an efficient way provided by feature selection, which might cut back the computation time, improve learning accuracy, and facilitate a higher understanding for the learning model or data. During this study, we tend to discuss many frequently- used analysis measures for feature choice, and so survey supervised, unsupervised, and semi-supervised feature selection strategies, that are wide applied in machine learning issues, like classification and clustering. Variable selection or attribute selection is known as feature selection. Dimensionality reduction is completely different from feature selection. Each strategies request to scale back the quantity of attributes within the dataset, however a dimensionality reduction methodology do thus by making new combination of attributes, wherever as feature selection strategies embrace and exclude attributes present within the data while not ever-changing them. An accurate predictive model is created by feature selection methods. Helping in choosing features will provide best or better accuracy whilst requiring less data. Identifying and removing unneeded can be done by using the feature selection method. Filter methods, wrapper methods and embedded methods are the three feature selection algorithms.

Filter method

Statistical measure to assign evaluation to each feature applied by the filter feature selection methods. The features are selected are either used or removed from the dataset in a hierarchy manner. The methods are typically unilabiate and consider the feature severally, or with reference to the variable quantity.

Wrapper method

Wrapper ways think about the selection of a group of options as a search drawback, wherever completely different features are ready, evaluated and compared to different mixtures. A predictive model us accustomed valuate a mixture of combinations and assign a score supported model accuracy. The search method is also organized like a best-first search, it should random like a random hill-climbing formula, or it should use heuristics, like forward and backward passes to feature and take away options.

Embedded method

Embedded strategies learn that options best contribute to the accuracy of the model whereas the model is being created. the foremost common kind of embedded feature choice methods are regularization methods. Additional constraints into the optimization of a predictive algorithm is introduced by Regularization methods are also called penalization methods. That bias the model to-ward lower complexity.

B. Principal Component Analysis

Principal Component Analysis is used to consider data set with various dimensions and to eliminate the variables that are similar and retain the variables which show difference. These elements are called the principal elements that are used to remodel. With all the original elements the first element shows more variation. The dataset is scaled using PCA method and the results are subjected to change and summarizes information. Since similar properties are chosen redundancy is possible.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10306

C. Train-Test Split

Data, in machine learning, in most scenarios are split into training data and testing data (andsometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. Training dataset is a part of the actual dataset that we useto train the model. The model sees and learns from this data. Test data, on the other hand, isthe sample of data used to provide an unbiased analysis of a final model fit on the training dataset. The Test dataset provides the ideal standard used to evaluate the model. It is used once the model is completely trained. Splitting the dataset into training, validation testing sets can be determined on two categories. Firstly, it depends on how much the total number of samples in the data and second, on the actual model the user is training. Some models need efficient or large data to train upon, so in that case one could optimize for the larger training sets. Models with very few hyper parameters are estimated to be easy to validate and tune, so one can possibly reduce the size of your validation set. However, given the model has many hyper parameters, the user would want to have a large validation set as well.

We have split our dataset into 70%-30% ratio for training and test data (the first 400 instances for training while the next 169 instances for testing the model).Keeping in mind that training the model, making the machine learn, is vital, we have slotted70% of the dataset to training. Out of the 70% dataset for training, we are keeping 63 percentfor training and 7 percent for cross validation test. Cross-Validation is applied to a part of the training data and is validated with the remaining dataset. Several rounds of cross validation are performed to get various results for comparison.

IV. ALGORITHMS

A. Logistic Regression

After linear regression, logistical regression is the most famous machine learning algorithm.Linear regression and logistic regression are similar in many ways. But what they are used for is the biggest distinction. Algorithms for linear regression are used to predict values, but logistic regression is used for classification tasks. The rule is to train the model by considering the input variables and a target variable is known as the Logistic Rule. In logistical rule the output or target variable may be a categorical variable, in contrast to regression towards the mean, and is therefore a binary classification rule that categorizes a knowledge purpose to one of the categories of information. The general equation of logistic regression is:

loger (p) = bi0 + bi1X1 + bi2X2 + bik Xk, where p is the probability of presence of the characteristic of interest.

Logistic regression measures the link between the variable quantity, the output, and therefore the freelance variables, the input. By estimating chances exploitation its underlying supply perform. It uses 1.2 penalty for regularization. The expected worth is often anyplace between negative eternity to positive eternity. The resultant chances are then born-again to binary values zero or one by the supply perform, conjointly referred to as the sigmoid function. The Sigmoid perform takes any real-valued variety and maps it into a worth between the vary 0-1 excluding the bounds themselves. Afterwards, a threshold classifier transforms the result to a binary worth. One in every of the first assumption of supply regression is that the input options ought to be freelance of every alternative. One variable ought to have very little or no co-linearity with the opposite variable. Hence, PCA is dead on the info beforehand, to convert the related variables to a collection of unrelated variables. Logistic Regression is used to develop a model to predict breast cancer. The system has developed consists within the estimation of unknown dependencies in a very system from a given knowledge set to make a helpful and general model to analyze new incoming knowledge.

B. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning which addresses classification and regression challenges. However, it's principally utilized in classification issues. In this algorithmic rule, we plot each data item as a point in n-dimensional space where n is number of features one has with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes well shown in the figure below:

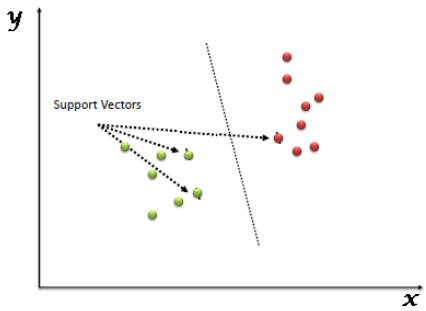

Figure 2: Support Vector Machine

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10306

Often researchers tend to plot every knowledge item as some extent in n-dimensional area with the worth of every feature being the worth of a selected coordinate. Then, to perform classification by finding the hyper-plane that differentiate the 2 categories fine. It is a non- probabilistic binary linear classifier, how-ever are often manipulated during a manner that it will perform non-linear and probabilistic classification also, creating it versatile algorithmic program. AN SVM model could be an illustration of the instances as points in area mapped, so they will be categorized and divided by a transparent gap. New instances are then mapped into the identical area and foreseen that within which class it would be supported which aspect of the gap they fall in. the most advantages of SVM is that the indisputable fact that it's effective in high dimensional areas.

Pseudo code for SVM
initialize y = Y for i I REPEAT
compute SVM solution w, b for data set with imputed labels
compute outputs fi = (w, xi) + b
for all xi in positive bags set y = sgn(fi)
for every i I, Y = 1
for (every positive bag BN)IF (PaI (1 + y)/2 == 0)
compute i = arg maxiI faset y = 1
END
END
WHILE (imputed labels have changed)OUTPUT (w, b)

C. K-Nearest Neighbors

The k-nearest neighbor's algorithmic program is one of the simplest machine learning algorithms. It has merely supported the concept that objects that are 'near' every alternative can additionally have similar characteristics. If it can recognize the characteristic options of one of the objects, it will be additionally predicted for its nearest neighbor. k-NN is associate improvisation over the nearest neighbor technique. It is based mostly on the plan that any new instance will be classified by the majority vote of its 'k' neighbors, wherever k is a positive number, sometimes a little variety.
It is known as Memory-Based Classification as the coaching examples must be in the memory at run-time. Once handling continuous attributes, the distinction between the attributes is calculated Euclidean distance. a serious drawback once dealing with the Euclidean distance formula is that the big values frequency swamps the smaller ones.

When KNN is employed for classification, the output is calculated because the category with the very best frequency from the K-most similar instances. Every instance in essence votes for their class and therefore the class with the foremost votes is taken for the prediction.
Class probabilities is calculated because the normalized frequency of samples that belong to every class within the set of K most similar instances for a new data instance. For instance, during a binary classification problem (class is zero or 1):

$$(\text{class}=0) = \text{count}(\text{class}=0) / (\text{count}(\text{class}=0)+\text{count}(\text{class}=1))$$

Pseudocode of K-Neighbors

1. Load the training and test data
2. Choose the value of K
3. For each point in test data:

- determine the Euclidean distance with respect to all training data points
- write the Euclidean distances in an array and sort it
- choose the first k points
- Based on Majority choose the particular class for the data points.

4. End
5.

V. RESULTS

	Accuracy	Precision	Specificity	Recall	F1 Score
K-Neighbours	0.887	0.877	0.692	0.974	0.923
Logistic Regression	0.876	0.838	0.65	1	0.912
SVM	0.899	0.869	0.696	1.000	0.93

Table: Scores of Accuracy, Precision, Recall, F1 Score and Specificity with PCA

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10306

VI. CONCLUSION

Using Breast Cancer dataset, classification algorithms like Logistic Regression, KNN, SVM are applied and the data is analyzed for prediction. The classification algorithm determines the accuracy, recall, precision and F1 Score. SVM and Logistic regression has better performance than all other classification algorithms. Other algorithms can also be used on the same dataset and prediction accuracy can be evaluated.

REFERENCES

[1] Breast cancer facts and figures 2003-2004 (2003). American Cancer Society.
[2] Stages | Mesothelioma | Cancer Research UK Breast cancer survival statistics September 26,2017
[3] Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta. Diagnosis of Breast Cancer usingDecision Tree Models and SVM (2016)
[4] Rohith Gandhi. Nearest Neighbor. Understanding Machine Learning (2018)
[5] Adi Bronshtein. Train/Test Split and Cross Validation in Python. Understanding MachineLearning (2017).
[6] C. desantis r. siegel, P. bandi and A.jemal, “Breast cancer Statistics”, A Cancer Journal for Clinicians, pp.2015
[7] Q.Su,’’A Cancer Gene Selection Algorithm Based on the K-S Test ans CFS’’, Biomed Research International,2017
[8] M.Morovvat and A.Osareh,”An Ensemble of Filters and Wrappers for Microarray Data Classification Machine Learning and Applications:An International Journal, June 2016
[9] Matamala N, Vargas MT, Gonzalez-Campora R, Minambres R et al. Tumor micro RNA expression profiling identifies circulating microRNAs for early breast cancer detection. Aug 2015
[10] Veer, L. J., Dai, H., and Vijver, M. J. Gene Expression Profiling Predict Clinical Outcome of Breast Cancer nature