Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 6, July 2021: 6163- 6168

Research Article

A Comparative Study on Classification Algorithms

Santhoshini Banda[a], Nadia Anjum[b]

[a] Assistant Professor, Dept of CSE, Stanley College of Engineering and Technology
[b] Assistant Professor, Dept of CSE, Stanley College of Engineering and Technology

***Corresponding author:** [a] santhoshini.banda@gmail.com, [b] nadiaanjum@stanley.edu.in

Abstract

The word big data records series of facts that is substantial in length and still developing with time. Breast cancer takes the second position in dangerous diseases. Breast cancer occurs in women which is said to be the most dreadful disease. Around hundreds of thousands of cases are being recorded inside the world each year. It remains an awful lot more usual place in high-profits countries. However, it is now growing at fast in center and low-benefits nations which includes within Africa, America and Asia etc... In this paper, we present a comparison between different classification algorithms. This paper implements popular records mining algorithms (Support vector machine, simple logistic regression, decision tree and random forest) on wiscosin breast cancer dataset. The algorithms are compared based on the accuracy achieved, precision, and recall. The output proves that the most classification accuracy of 97% is achieved by Random forest, Support Vector Machine

Keywords: breast cancer, decision tree, simple logistic regression, random forest, decision tree, support vector machine

1. Introduction

The dimensions of databases taking down clinical realities are expanding quickly. Clinical information made from estimations, assessments, remedies, etc are saved in exceptional databases on an uninterrupted procedure. The substantial quality of statistics runs over the potential of conventional strategies to test and search for exciting examples and records which covered up in them. Consequently, new methodologies and tools to find new strategies have become more demanding.

Data mining has played an essential role inside the area for finding out hidden patterns in huge amount of data sets. Data Mining is one of the most critical and essential region of studies with the objective of locating meaningful statistics from big data sets. The reason of records mining intends to extract useful information from enormous databases or data warehouses. Data mining programs are used for business and scientific sides. Data mining intends to mining or deriving information from massive measure of information or databases. The improvement of locating useful styles or importance in raw information has been called knowledge discovery in databases KDD.

2 Literature Survey

Breast cancer growth is one in all the threatening diseases due to which many women fall to death every year. To identify the tumors existing inside the breast which are harmful cannot be alone done by big data. By using different techniques of data mining along with big data, we can build efficient algorithms for detecting this cancer in the early stages.

In artificial intelligence (AI), classification ought to be the important directive. Specialists have just done parcel of explores by applying AI calculation on clinical dataset for order and information mining calculation to

discover an example in dataset for quicker figuring and expectation. A considerable lot of the methodologies give great exactness what's more, result.

Dr. G. R. Sakthidharan et al[2018][1]: The purpose of this network is to receive visual imagery accurately. Generally, we start with some numbers of filters for low-level feature identification of a disease. The deeper we go in the study of Convolution neural networks, the more filters we will identify high-level features. Feature detection is predicated on studying the input with the filter of a given size and to use matrix calculations to retrieve a feature map. It's useful within the classification of benign and malignant tumors.

Peter Adebayo Idowu et al[2015][2]: Breast malignancy is certainly be included in the top cancers that arise in women. Classification is extensively necessary to differentiate tumors. This technique can detect the similarities or differences that a human analyst cannot notice. Therefore, the decision tree creates and introduces more accurate information. The decision tree algorithm analyses and classifies the breast cancer dataset for disease prediction.

Alireza Osareh et al [2010][3]: The detection and treatment of breast cancer in beginning stages may reduce the death rate in women. Some specialized reasons, that square measure associated with image quality and human mistakes, increase the improper diagnosing of breast cancer by doctors. Due to these reasons Computer-aided detection systems (CADs) were created to defeat these limitations. Also, they have been concentrated in many imagery procedures for breast cancer identification in recent years. The CAD systems improve the accuracy of diagnosing and evaluation risk.

T. L. Octaviani et al[2019][4]: To examine the clinical information, numerous techniques for information mining and AI are accessible. One of the most significant difficulties in AI and information mining zones intend to manufacture the precision and calculatedly productive classifiers considering clinical applications. The random forest classifier (RF) is a calculation that shapes a group of characterization techniques that rely upon a blend of a few decision trees. The data set being processed using random forest classifier gives 99% accuracy.

Ch shravya et al [2019][5]: In this paper, the researchers have chosen three data mining classifiers to choose the best model that gives more accuracy on predicting breast cancer. Three kinds of classifiers are taken and are run on weka machine. Support vector machine (SVM) has given an perfection of 92%.

Madhu kumaria et al[2018][6]: In this paper, the researchers have taken wisconsin based breast cancer data set is taken in which there are more number of predictive values. It is observed that the K nearest neighbour (KNN) classifier has given the more accurate results. This classifier reduces the treatment cost.

3. Methodology

In this paper, comparison is done between different classification algorithms which are implemented on the Wisconsin breast cancer dataset which is a diagnostic dataset. The dataset has 33 distinct features and 569 instances. The different classes are malignant and benign. The different algorithms that have been implemented are Logistic regression, Support vector machine, SVM with best parameters, Random forest and Decision tree algorithm.The training and testing ratios are changed to observe how the models behave for different ratios. The data is split in the ratio 75/25 i.e. 75% of instances for training the model and the remaining 25% of instances for validating. The other train and test ratios implemented are 70/30, 80/20 respectively, then finally results were evaluated and compared. The performance of the models is evaluated in terms of accuracy, precision, recall and F-score. Finally, the results are tabulated, compared and visualized to see which classifier performs best.

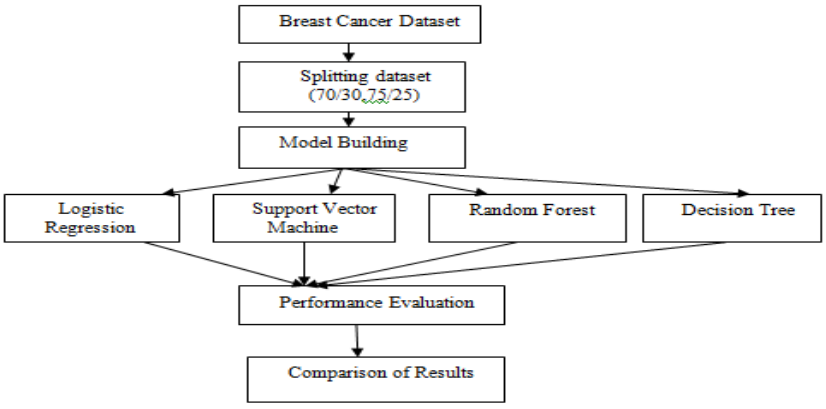

Fig 1: Methodology of the proposed system.

A. Data representation

Data representation is an important part of information science. These makes a difference in one towards grasping on and furthermore pass on the information to other individual in a significant way. It helps to better understand the data and the patterns hidden in it.

B. Working methodology

The proposed work has 3 phases.

- In phase 1 consists of model building, where the models are trained using different train ratios.
- In phase 2, test data set is practically applied on the Model acquired from trained dataset and furthermore the outcomes are secured.
- In phase 3, the results acquired are assessed in the terms of accuracy, precision, recall, F1- Score and Support. The best working model is recognized based on highest achieved accuracy.

4. Implementation

The research work has been implemented in python programming language gaining knowledge from [7] and [8]. Google colab has been used for carrying out this research.

The different algorithms implemented are:

1) Logistic Regression

2) Support Vector Machine

3) SVM with best parameters

4) Random Forest

5) Decision Tree

5 Results and Conclusion

In this paper a comprehensive study has been carried on exclusive classification techniques and provided a basis for evaluation among them in terms of accuracy, precision, recall and F1-Score.

Results

Results can be better understood by confusion matrix. The performance of a classifier is based up on "confusion matrix" through which real values can be known. The accuracy of a classification model is computed with the number of correct and incorrect instances in the data set.

Train and Test ratios: 75/25

Table 1: Performance of Logistic Regression

	Precision	Recall	Accuracy	F1 Score	Support
0	97	98	97	97	91
1	96	94	97	95	52
Avg	96.5	96	97	96	

Table 2: Confusion Matrix for LR

	Predicted	
	89	2
Actual	3	49

Table 3. Performance of SVM

	Precision	Recall	Accuracy	F1 Score	Support
0	97	98	97	97	91
1	96	94	97	95	52
Avg	96.5	96	97	96	

Table 4: Confusion Matrix for SVM

	Predicted	
	89	2
Actual	3	49

Table 5. Performance of SVM with best parameters

	Precision	Recall	Accuracy	F1 Score	Support
0	97	98	97	97	91
1	96	94	97	95	52
Avg	96.5	96	97	96	

Table 6. Confusion Matrix for SVM

	Predicted	
	89	2
Actual	3	49

Table 7. Performance of Random Forest

	Precision	Recall	Accuracy	F1 Score	Support
0	98	98	97	98	91
1	96	96	97	96	52
Avg	97	97	97	96	

Table 8. Confusion Matrix for RF

	Predicted	
	89	2
Actual	2	50

Table 9. Performance of Decision Tree

	Precision	Recall	Accuracy	F1 Score	Support
0	95	97	94	96	91
1	94	90	94	92	52
Avg	94.5	93.5	94	94	

Table 10. Confusion Matrix for DT

	Predicted	
	88	3
Actual	5	47

The following graphs verify excessive performance of the simple logistic regression, Decision tree and random forest classifier in sense of the breast cancer cases database.

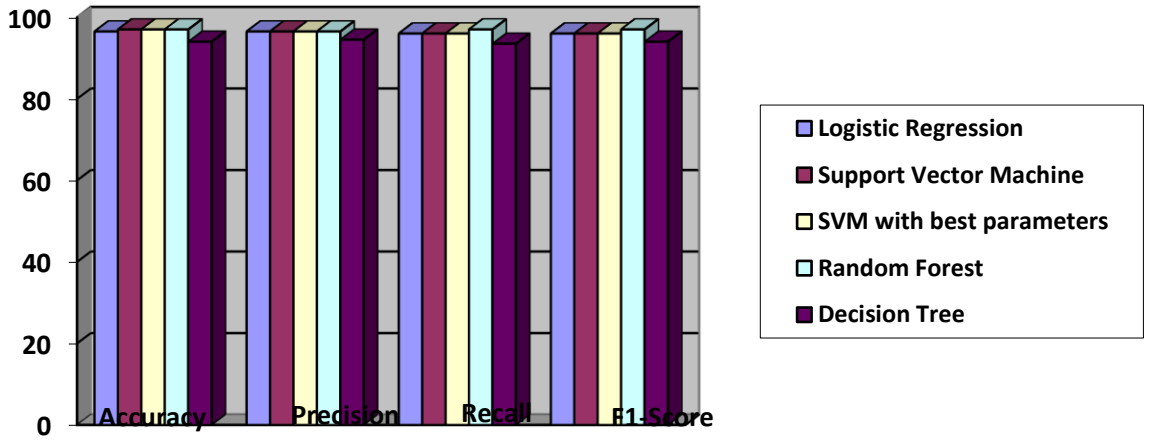

Fig 2: Visualization of performance of classifiers (75/25) ratio

Train and Test ratios: 70/30

Table 11: Performance of Logistic Regression

	Precision	Recall	Accuracy	F1 Score	Support
0	95	99	96	97	106
1	98	91	96	94	65
Avg	96.5	96	96	96	

Table 12: Confusion Matrix for LR

	Predicted	
	105	1
Actual	6	59

Table 13. Performance of SVM

	Precision	Recall	Accuracy	F1 Score	Support
0	96	98	96	97	106
1	97	94	96	95	65
Avg	96.5	96	96	96	

Table 14: Confusion Matrix for SVM

	Predicted	
	104	2
Actual	4	61

Table 15. Performance of SVM with best parameters

	Precision	Recall	Accuracy	F1 Score	Support
0	95	99	96	97	106
1	98	92	96	95	65
Avg	96.5	96	96	96	

Table 16. Confusion Matrix for SVM

	Predicted	
	105	1
Actual	3	62

Table 17. Performance of Random Forest

	Precision	Recall	Accuracy	F1 Score	Support
0	97	99	98	98	106
1	100	95	98	97	65
Avg	98.5	97	98	98	

Table 18. Confusion Matrix for RF

	Predicted	
	105	1
Actual	3	62

Table 19. Performance of Decision Tree

	Precision	Recall	Accuracy	F1 Score	Support
0	96	98	96	97	106
1	97	94	96	95	65
Avg	94.5	93.5	96	96	

Table 20. Confusion Matrix for DT

	Predicted	
	104	2
Actual	4	61

The following graph verify excessive performance of the random forest classifier in sense of highest accuracy achieved i.e., 98% .

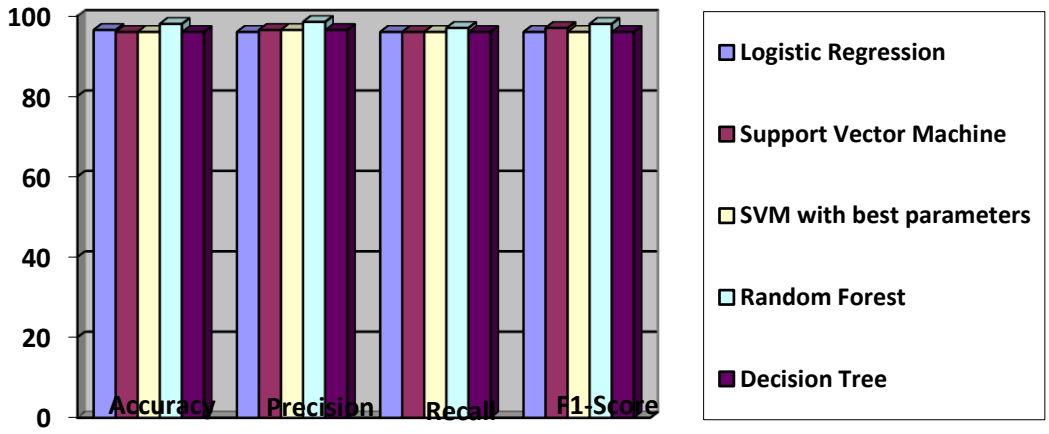

Fig 3- Visualization of Performance of Classifiers (70/30) ratio

The final comparison between the different combinations of algorithms that we have implemented can be tabulated as the following:

Table 21- Performance Comparison of Classifiers

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	97	96	96	97
	96	96.5	96	96
Support Vector Machine	97	96.5	96	96
	96	96.5	96	96
SVM with best parameter	97	96.5	96	96
	96	96.5	96	96
Random Forest	97	96.5	97	97
	98	98.5	97	98
Decision Tree	94	94.5	93.5	94
	96	96.5	96	96

6 Conclusion and Future Works

Contrasting with every single other malignant growth, breast cancer is one of the primary reasons of loss of life in ladies. In this way, the early detection of breast most cancer is required in lessening life misfortunes. In this paper, we have utilized machine learning algorithms to predict the breast cancer accurately. Our work reveals that random forest when implemented with train and test ratio as 70/30 achieves the highest accuracy i.e., 98% and the least accuracy achieved is by decision tree when train/test ratio are 75/25 i.e., 94%.

In the future, convolution neural networks can be implemented on a more versatile dataset.

References

[1] Dr. G. R. Sakthidharan, Dr. P. Chandra Sekhar Reddy, Dr. S. Govinda Rao, “Detection and Prediction of Breast Cancer Using CNN-MDRP Algorithm in Big Data and Machine Learning: Study and Analysis”, 2018.

[2] M.deepika, L. Mary Gladence, R.Madhu Keerthana, “A review on prediction of breast cancer using various data mining techniques”, 2016, pgno.808.

[3] Alireza Osareh, Bita Shadgar, “Machine Learning Techniques to Diagnose Breast Cancer”, april 2010.

[4] T. L. Octaviani, and Z. Rustam, “Random forest for breast cancer prediction”, Nov 2019.

[5] Ch. Shravya, K. Pravalika, Shaik Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques”, volume 8, April 2019.

[6] Madhu Kumaria, Vijendra Singh, “Breast Cancer Prediction system”, 2018.1

[7] Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org

[8] G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995