International Journal on Emerging Technologies **10**(3): 55-63(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis

Tina Elizabeth Mathew
Research Scholar, Faculty of Applied Science and Technology,
University of Kerala, Thiruvananthapuram, Kerala, India.

(Corresponding author: Tina Elizabeth Mathew)
(Received 05 June 2019, Revised 20 August 2019 Accepted 30 August 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Cancer also called malignant tumor or malignant neoplasm is one of the world's deadliest diseases. It is the abnormal growth of cells which has the capability of spreading to other parts of the body. Not all tumors are malignant some are benign and they do not invade surrounding cells. Cancer of the Breast is a deadly disease in women. Logistic Regression a binary classifier is used to predict breast cancer. Feature selection methods are employed to find whether reduction of the number of features of the dataset are effective in prediction of Breast cancer. Recursive feature elimination helps in ranking feature importance and selection. The optimal feature sets are selected for building the model using recursive feature elimination with and without cross validation. Recursive feature elimination is used to show the accuracy in prediction when different combinations of features are used based on their ranks. The study shows that reduction of features using RFE helps in improving prediction accuracy.

Keywords: Data Mining(DM), Recursive Feature elimination (RFE), Logistic Regression(LR), Recursive Feature elimination with cross validation (RFECV), Wisconsin Breast Cancer Database(WBCD).

I. INTRODUCTION

Breast Cancer is the second most leading malignancy in the world. It is now the most common cancer in cities and stands second in rural areas in India. Early detection plays a key role in the diagnosis, treatment, prognosis and survivability of the disease. Data Mining(DM) defined as extraction of information from large data sets and part of knowledge discovery in databases(KDD) involves exploring and analyzing large quantities of data and identifying new, valid and useful information from repositories[16]. It has wide applications in market analysis, anomaly detection, medical diagnosis, business analysis and many more. Data mining adopts its techniques from statistics, machine learning, database systems, rough sets, visualization and neural networks. Data mining and statistical techniques can be applied to find useful patterns to help in the important tasks of medical diagnosis and treatment [8].

Data mining strategies are categorized into supervised and unsupervised learning. In supervised learning models, values of inputs are used to make predictions about a target variable with known values. Unsupervised learning models help in predicting on data for which the target variable has unknown values.

Data mining models are classified into predictive and descriptive models [16]. Predictive models help in making inferences and forecasting. Descriptive models help to reveal patterns by grouping events, identifying relationships and finding links between events. The tasks included in the Predictive data mining models are prediction and classification Prediction algorithms help to predict continuous or discrete target values from given input data. Prediction models decide the future outcome rather than existing behavior. Some Predictive models using supervised learning are Regression, Neural Networks, Decision trees, memory based reasoning, and Support Vector Machine. Comparison of these models [25] in heart disease prediction show the efficacy of these models. Similar studies in breast cancer diagnosis using Support Vector Machines [26] also show promising results in prediction when feature elimination was done.

Logistic Regression has been used in many studies in diagnosing breast cancer. In this study, Feature selection and elimination of irrelevant features is applied with Logistic Regression and the significance of using Logistic Regression models alone and in combination with recursive feature elimination techniques with and without Cross validation for breast cancer prediction is analyzed. In section II, the first segment illustrates the related work in this field followed with the materials and. methodology used and Section III explains the experimental setup along with the results, Section IV and V gives the Conclusion and Future Scope respectively.

II. MATERIALS AND METHODS

A. Related Work

Ahmed *et al* [1] used Logistic Regression to predict breast cancer. The model selected variables with least correlation and used it to build the LR model. Pearson and deviance statistics were used to measure how closely the model fits the observed data. The model gave an accuracy of 98.9%. Wang *et al* [2] used logistic regression to identify significant factors in hypertension, A neural network with back propagation algorithm was

developed using these significant factors to predict hypertension. This model was seen effective in the prediction of the disease. Haq *et al* [3] used three feature selection algorithms, Relief feature selection, minimal redundancy maximal relevance feature selection algorithm and Least absolute shrinkage and selection operator on seven classifiers, Logistic Regression, KNN and SVM and studied the impact of each methods. It was concluded that feature election helped in classification accuracy and reduced execution time. Choudhury *et al* [4] used Logistic Regression for diagnosis of early stage symptoms of mesothelioma disease and found it to be effective in prediction giving a prediction accuracy of 81.4% in the training set and 63.46% in the testing set. Leopord *et al* [5] in their work used data mining techniques to predict disease outbreak. They suggested that hybrid methods provided better prediction compared with individual classification and regression methods. Bhatti et al[6] in their study found that Logistic Regression was an effective method in predicting risk of ischemic heart disease. It was used to assess the risk factors that enhanced the disease risk. Sultana *et al* [7] in their work compared the efficiency of different classifiers, Simple Logistic regression, MLP, Multi-Class Classifiers, DT, REP tree, K-star, IBK, Decision table, PART and Random Forest. Results concluded that Simple Logistic regression method gives the best model in predicting breast cancer. Results indicated that Simple Logistic regression obtained best performance in general compared to the other classifiers in terms of classification accuracy, RMSE, specificity and sensitivity, F-measure, ROC curve area, time taken to build the model and Kappa Statistics. Chang *et al* [9] used Bayesian LR to predict breast cancer. Since the dataset has multicollinear variables, they were avoided based on scores of the variance inflation factor(VIF). Variables with high VIF values were avoided. The model produced an F1 score of 0.95. Mythili *et al* [10] proposed that combinations of support vector machines, logistic regression, and decision trees helped in an accurate prediction of heart disease. Hasan *et al* [11] evaluated the prediction performance of neural networks using different techniques. These when compared with the performance of Logistic Regression it was seen that neural networks performed better. Rahimloo *et al* [12] in their work used Artificial Neural Networks and Logistic Regression models to evaluate performance in predicting diabetes. A hybrid model was later constructed using ANN and Logistic regression and it was seen that the error in prediction was less in the hybrid model than that of individual models. Johnson *et al* [13] used Genetic algorithms to find the best feature set which gave better accuracy in prediction of Alzheimer's disease using logistic regression. Yadav *et al* [14] compared three methods Logistic Regression, Multi-layer Perception and Sequential Minimal Optimization algorithms for predicting heart disease. Logistic regression was seen to give the best F measure. Rajbharath *et al* [15] in their work proposed a hybrid of Random Forest and Logistic Regression algorithms for building a breast cancer survivability prediction model. The Random Forest Technique is used to perform a preliminary screening of variables and to rank them. Then, the new data set based on the top-k important predictors and is used as input into the Logistic Regression model for predicting breast cancer survivability. Rani *et al* [17] in their paper used Logistic regression to preprocess data and eliminate outliers. This helped in increasing the prediction accuracy of heart disease. Sharanyaa *et al* [18] in their paper concluded that Logistic Regression gave an accuracy of 82% in predicting Parkinson's disease over 71% of Random Forest and 94% of K-Nearest Neighbor models. Javali *et al* [19] in their paper compared multiple logistic regression models to evaluate the effectiveness of the models in identifying risk factors for dental caries and periodontal disease. Using a reduced set of risk factors in the logistic model was found to give better predictions. Shravya *et al* [20] compared various models and found that Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) for Breast cancer prediction and found that SVM model gave the best prediction accuracy. Gai *et al* [21] in their paper found that Logistic Regression model had good prediction accuracy, satisfactory accuracy and strong robustness in diagnosis of Hepatobiliary Disease. Liang *et al* [22] in their work compared diagnostic performance between back propagation artificial neural networks (BP ANNs) and Logistic regression (LR) models in predicting the prognosis of acute ischemic stroke. Both methods were found to be promising, while ANN's showed better performance comparatively. Yusuf *et al* [23] applied Logistic regression analysis on the variables from the mammogram results and found the variables and combination of variables that had an impact on identifying breast cancer. Leena *et al* [24] in their survey show cased the necessity of combining two or more data mining methods for better performance in disease prediction.

B. Dataset

The Wisconsin Breast Cancer Database (WBCD) obtained from Dr. William H. Wolberg of Wisconsin University Hospitals, Madison is used. The Original data set contains 699 instances with 11 attributes each. The first attribute the ID of an instance, is discarded as it has no role in prediction, and the next 9 represent different characteristics of an instance. These are the cytological characteristics of the breast fine needle aspiration(FNA) test. The instances all have a value between 1 and 10. 1 for benign and 10 for the most malignant. The diagnosis made is the last attribute. Each instance belongs to one of the 2 possible classes, benign with value 2 or malignant with value 4. The 9 attributes that are used in the prediction process are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The database has 699 instances with 458 benign cases - 65.5% and 241 malignant cases- 34.5%. Sixteen instances were avoided due to missing values and 683 instances were taken for the study, of which 444 belong to benign class and 239 to malignant class.

C. Methodology Used

Regression. Regression is a machine learning technique that determines the relationship between the dependent variable, the target variable whose value is to be predicted, and one or more independent variables. There are three types of regression models: linear, polynomial, and logistic regression. Linear and Polynomial regression uses numeric continuous variables. Logistic regression makes use of categorical dependent variables [11].

Logistic Regression(LR). Logistic regression introduced by David Cox in 1958, is used in predicting binary problems. Consider a dataset containing *N* points. Each point *i* consists of a set of *m* input variables $x_{1,i}$ ... $x_{m,i}$ which are called independent variables (or predictor variables, features, or attributes), and a binary outcome variable Y_i (or dependent variable, response variable, output variable, or target class). It can assume only the two possible values 0 for failure or 1 for success. A sigmoid function is used in logistic regression to squash the values within the range of [0,1]. When the value is greater than a threshold value it is assigned label 1, otherwise it is assigned label 0. The goal of logistic regression is to use the dataset to create a predictive model of the outcome variable. The logistic function is defined in equation 1

$$\sigma(t) = \frac{1}{1+e^{-t}} \quad \text{or} \quad ln(\frac{P}{(1-P)}) = t \qquad (1)$$

Logistic regression gives probability value, Y=1 if malignancy is diagnosed and gives value Y= 0 for a benign condition. The conditional probability or likelihood that a person has the disease can be computed as P(Y=1|X) Where X represents the set of attributes,{x_0, x_1,x_2,x_3.. x_n} that are used in diagnosis and the equation can be given as a linear combination of the inputs as

$$\log\ (\frac{P(X)}{1-P(X)}) = x_0+a_1x_1+a_2x_2+..+a_nx_n \qquad (2)$$

where a_1,a_2,,,a_n are the coefficients of the attributes x_1,x_2, x_n and act as weights that imply significance.

Feature Importance and Selection. A dataset has lots of features however, not all features, contribute to the prediction variable. Removing features of low importance improves accuracy, reduces both model complexity and over fitting and also training time of large datasets. The 9 features of the WBCD data set are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature or features and keeps repeating the process with the remaining features until the specified number of features is attained or exhausted. It removes features, builds a model using the remaining attributes and calculates model accuracy. Features are ranked relatively according to the order of elimination. A significance level is chosen and the model is fit with all attributes. The attributes with highest *p*- value is selected and if the *p*-value is greater than the significance level it is discarded. The model is again built over the remaining attributes. This is repeated till the removal of an attribute affects the accuracy of the built model. Ranking is done based on the coefficient values of the attributes. Higher the coefficient value better is its rank. RFE is able to work out the combination of attributes that contribute mainly to the prediction of the target variable or class.

In this study the value of each attribute is computed using RFE and the features are ranked according to importance. These were explored to find the most prominent and dominating features. The higher the score the more important are the attributes. Studies show that a training partition between 40% to 80 % gives good results in precision and accuracy. The attribute ranking in these training data portions is done to observe whether ratio of training data used has an effect on the relevance of attributes and selection. The values of the features obtained from the classifier is shown in Table 1. The study among partitions suggests slight rank variations but that most prominent features seen in almost all cases are uniformity in cell shape and bare nucleoli, and least being epithelial size.

Table 1: Attribute ranking.

Sl . No	Attributes	Feature Ranks based on RFE in				
		70-30 partition	60-40 partition	50-50 partition	80-20 partition	No partition
1	clump_thickness	5	5	6	3	5
2	size_uniformity	1	8	9	1	8
3	shape_uniformity	2	1	1	5	1
4	marginal_adhesion	4	2	2	7	7
5	epithelial_size	9	9	5	9	9
6	bare_nucleoli	3	3	3	2	2
7	bland_chromatin	6	6	7	4	4
8	normal_nucleoli	7	4	4	8	6
9	mitoses	8	7	8	6	3

III. RESULTS AND DISCUSSION

The data is partitioned into training and testing sets feature selection is done and the logistic regression model is used. Feature selection is done in two ways using the RFE and RFECV methods. In RFE, the preferred number of subset of attributes required are selected at first and it recursively selects subsets of features based on importance. The estimator is first trained on the initial set of features and the importance of each feature is obtained. The least important features are pruned from current set of features and the procedure is recursively repeated on the pruned set until

the specified number of features to be selected is eventually reached.

RFECV does feature ranking with recursive feature elimination and 10 fold cross-validation and selects the optimal number of features and builds the model based on this optimal subset of features. 10 fold Cross-validation divides the samples into a training set and a testing set. The algorithm learns from the training set to constitute the classification rules, the samples of the testing data are used to measure the performance of the classification rules created. All the samples are randomly divided into 10-folds. A fold of the data is used as the testing data and the remaining 9 folds are used as the training set. The step is repeated 10 times, and each testing set validates the classification rules learnt from the corresponding training set to achieve an accuracy rate. The average of the accuracy rates of all 10 testing sets can be used in the final evaluation results. The subset of features are used in building the model. The model is trained on the training sets and then tested on the testing set. Different number of features are used to build the LR-RFE models and the prediction accuracy of each is assessed . The dataset is partitioned into four groups of training - testing sets in the ratios 70%-30%, 60%-40% 50%-50%, 80%-20% and the performance is accessed. The features are in the order 'clump_thickness', 'size_uniformity', 'shape_uniformity', marginal_adhesion', 'epithelial_size', 'bare_nucleoli', bland_chromatin', 'normal_nucleoli', 'mitoses. The work was done using Python Programming using Scikit learn packages. The block diagram in figure 1 represents the working of the model.

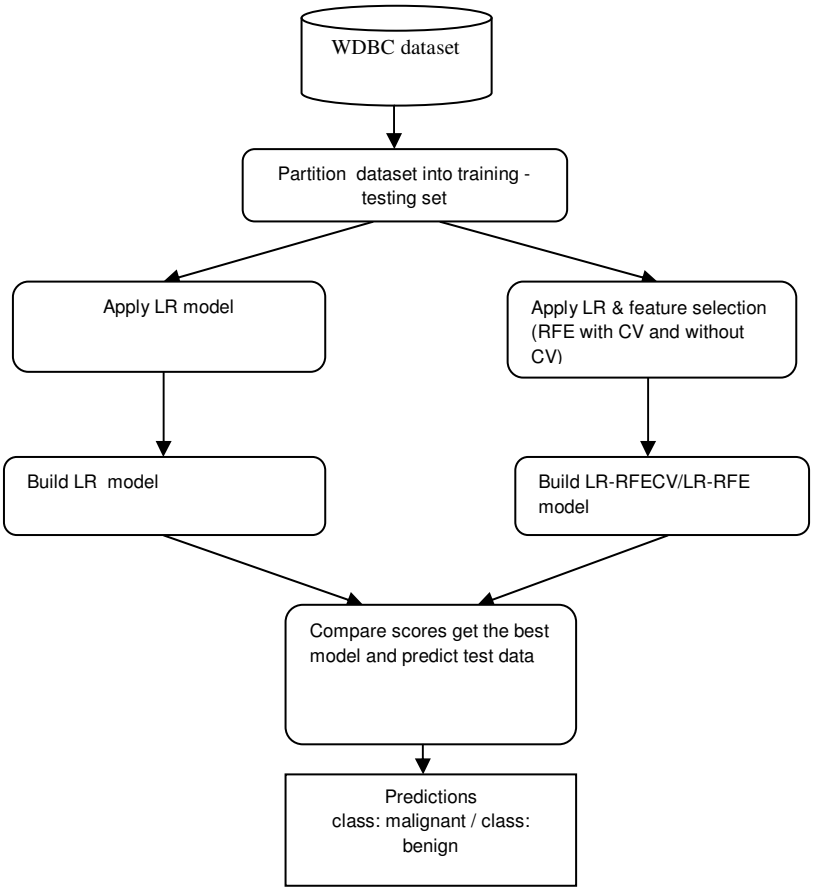

Fig. 1. Block diagram of the model.

A. RFE with cross validation (RFECV)

RFE with 10 fold cross validation(RFECV) was used and the optimal number of features are selected. Graphs for all training- testing partitions were produced and as seen in the concerned figures the optimal number of features selected against cross validation is shown. The model is built with these selected optimal features.

For the training-testing partition of 70-30 the optimal features were 4 and the graph of figure 2 shows the cross validation scores plotted against number of features selected. The graph has its peak values when 4 attributes are selected.

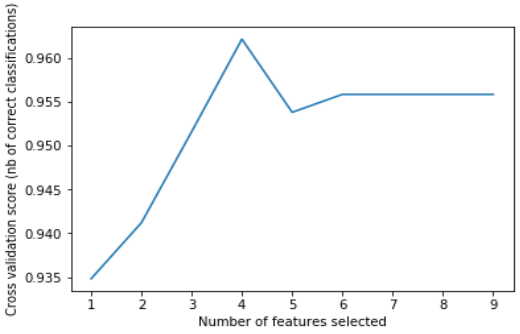

Fig. 2. CV vs Number of Features selected.

For the training-testing partition of 50-50 the optimal features were 3 and the cross validation scores plotted against number of feature selected was attained as shown in Fig. 3. The model is built using these three features.

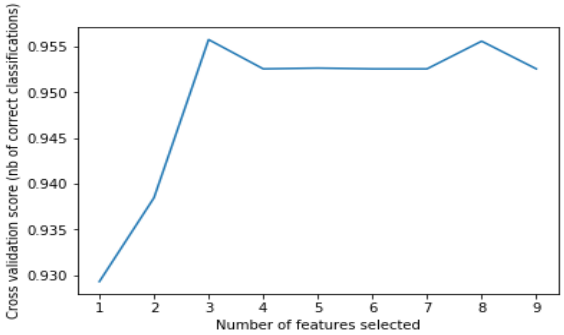

Fig. 3. CV score vs Number of Features selected.

For the training-testing partition of 60-40 the optimal features were 8 and the graph in Fig. 4 shows the cross validation scores plotted against number of feature selected. The model is then built using these 8 features.
For the training-testing partition of 80-20 the optimal features were 7 and the graph in Fig. 5 shows the cross validation scores plotted against number of feature selected. Model with the selected 7 features are built. It lies within the range[-1,1]. A -1 value indicates wrong classifications by classifier.

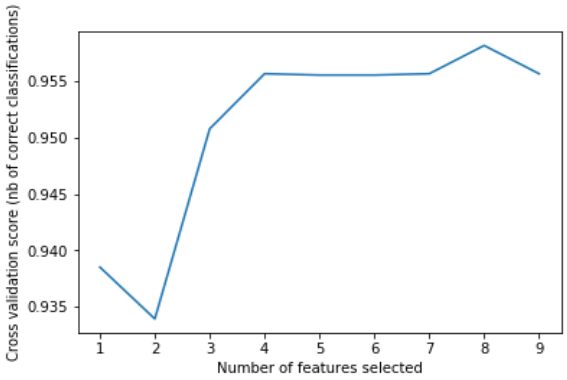

Fig. 4. CV score vs Number of Features selected.

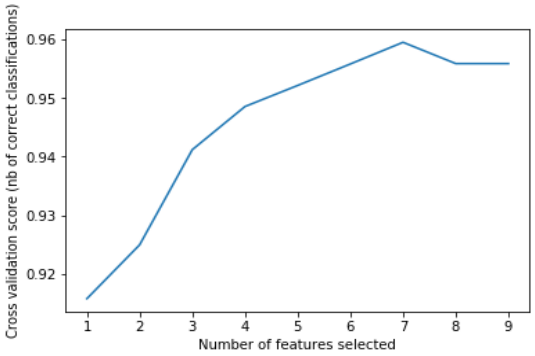

Fig. 5. CV score vs Number of Features selected.

Table 2: Performance scores in 70-30 partition.

Sl No	No. of Features used in the model	Features used	Precision =TP/(TP+FP)	Sensitivity / Recall	F1 score	Confusion matrix	MCC	No of Instances in each set(70-30) class 2 - benign and 4-malignant
1	1	[5 1 2 4 9 3 6 7 8]	0.93	0.93	0.92	[[130 0] [15 60]]	0.847	2-130
2	2	[4 1 1 3 8 2 5 6 7]	0.92	0.92	0.91	[[129 1] [16 59]]	0.825	4-75
3	3	[3 1 1 2 7 1 4 5 6]	0.95	0.95	0.95	[[130 0] [10 65]]	0.917	
4	4	[2 1 1 1 6 1 3 4 5]	0.95	0.95	0.95	[[129 1] [9 66]]	0.923	
5	5	[1 1 1 1 5 1 2 3 4]	0.96	0.96	0.96	[[129 1] [7 68]]	0,952	
6	6	[1 1 1 1 4 1 1 2 3]	0.97	0.97	0.97	[[129 1] [6 69]]	0.966	
7	7	[1 1 1 1 3 1 1 1 2]	0.97	0.97	0.97	[[129 1] [5 70]]	0.980	
8	8	[1 1 1 1 2 1 1 1 1]	0.97	0.97	0.97	[[129 1] [5 70]]	0.980	
9	9	[1 1 1 1 1 1 1 1 1]	0.97	0.97	0.97	[[129 1] [5 70]]	0.980	

A +1 value indicates a perfect classification and value near 0 indicates random predictions. Matthews Correlation Coefficient(MCC) is calculated and taken into consideration. Maximum MCC value attained here is 0.98 for 7, 8 and 9 feature sets.
Table 3 gives the performance of the 60-40 partition datasets. In the 60-40 partition the following results were seen. The model with 8 attributes and 9 attributes gave an F1 score of 0.97. The models using 4, 5, 6, 7 features respectively had an F1 score of 0.96 but the number of false negatives varied. MCC value is maximum at 0.963.
The results of the 50 -50 partition is seen in table 4. The 50-50 set showed the following results. The highest F1 score was attained when 4 features were used. But best MCC value was obtained for model with 5 and 7

features. Table 5 is the result for the 80-20 partitioned dataset. In the 80-20 training-testing partition set the models using 8 and all 9 features gave an f1 score of 0.99 and MCC of 1. Comparing the ratio of the testing datasets for the four partitions it is seen that performance scores when feature selection is applied on each testing set varies when the percentage of the testing dataset is altered. This is due to the nature of the dataset.

Table 3: Performance scores in 60-40 partition.

Sl No	No. of Features used in the model	Features used	Precision =TP/(TP+FP)	Sensitivity/ Recall	F1 score	Confusion matrix	MCC	No of Instances in each ser(60-40) class 2 - benign and 4-malignant
1	1	[5 8 1 2 9 3 6 4 7]	0.91	0.91	0.90	[[165 5] [21 83]]	0.799	2-170
2	2	[4 7 1 1 8 2 5 3 6]	0.92	0.91	0.91	[[167 3] [21 83]]	0.811	4-104
3	3	[3 6 1 1 7 1 4 2 5]	0.95	0.95	0.95	[[169 1] [12 92]]	0.913	
4	4	[2 5 1 1 6 1 3 1 4]	0.96	0.96	0.96	[[169 1] [10 94]]	0.933	
5	5	[1 4 1 1 5 1 2 1 3]	0.96	0.96	0.96	[[169 1] [11 93]]	0.923	
6	6	[1 3 1 1 4 1 1 1 2]	0.96	0.96	0.96	[[169 1] [9 95]]	0.943	
7	7	[1 1 1 1 3 1 1 2 1]	0.96	0.96	0.96	[[169 1] [9 95]]	0.943	
8	8	[1 1 1 1 2 1 1 1 1]	0.97	0.97	0.97	[[169 1] [8 96]]	0.953	
9	9	[1 1 1 1 1 1 1 1 1]	0.97	0.97	0.97	[[169 1] [7 97]]	0.963	

Table 4: Performance score in 50 -50 partition.

Sl No	No. of Features used in the model	Features used	Precision =TP/(TP+ FP)	Sensitivity/ Recall	F1 score	Confusion matrix	MCC	No of Instances in each set(50-50) class 2 - benign and 4-malignant
1	1	[6 9 1 2 5 3 7 4 8]	0.92	0.92	0.92	[[206 5] [22 109]]	0.834	2-211
2	2	[5 8 1 1 4 2 6 3 7]	0.93	0.92	0.92	[[209 2] [25 106]]	0.825	4-131
3	3	[4 7 1 1 3 1 5 2 6]	0.96	0.95	0.95	[[210 1] [15 116]]	0.908	
4	4	[3 6 1 1 2 1 4 1 5]	0.96	0.96	0.96	[[210 1] [14 117]]	0.916	
5	5	[2 5 1 1 1 1 3 1 4]	0.95	0.95	0.95	[[210 1] [16 115]]	0.900	
6	6	[1 4 1 1 1 1 2 1 3]	0.95	0.95	0.95	[[210 1] [17 114]]	0.893	
7	7	[1 3 1 1 1 1 1 1 2]	0.95	0.95	0.95	[[210 1] [16 115]]	0,900	
8	8	[1 2 1 1 1 1 1 1 1]	0.95	0.95	0.95	[[210 1] [17 114]]	0.893	
9	9	[1 1 1 1 1 1 1 1 1]	0.95	0.95	0.95	[[210 1] [17 114]]	0.893	

A comparison of LR-RFE and LR- RFECV models are shown in table 6, the models show a F1 score varying between the range of 0.95 to 0.98. Precision, Recall and F1 score and MCC of each model is shown in table 6. The recursive feature elimination method ranked features according to importance and it was seen that in all cases the features having the most impact on predictions were the attributes uniformity of cell shape. marginal adhesion and bare & normal nucleoli. MCC scores for LR-RFE was better than LR-RFECV and the F1 score of LR-RFECV was better than LR-RFE. RFECV method uses the number of optimal features identified by it during cross validation of the data, whereas, the RFE method takes the optimal features as half of the number of features available in the dataset unless specified otherwise. Results in the previous tables 2-7, confirm that better accuracy is obtained by using varied number of features than exactly using half of the available features.

Table 5: Performance score in 80 -20 partition.

Sl No	No. of Features used in the model	Features used	Precision =TP/(TP+FP)	Sensitivity/ Recall	F1 score	Confusion matrix	MCC	No of Instances in each set(80-20) class 2 - benign and 4- malignant
1	1	[3 1 5 7 9 2 4 8 6]	0.95	0.94	0.94	[[95 0] [8 34]]	0.864	2-95
2	2	[2 1 4 6 8 1 3 7 5]	0.97	0.97	0.97	[[94 1] [3 39]]	0.980	4-42
3	3	[1 1 3 5 7 1 2 6 4]	0.98	0.98	0.98	[[95 0] [3 39]]	0.991	
4	4	[1 1 2 4 6 1 1 5 3]	0.97	0.97	0.97	[[95 0] [4 38]]	0.966	
5	5	[1 1 1 3 5 1 1 4 2]	0.98	0.98	0.98	[[95 0] [3 39]]	0.991	
6	6	[1 1 1 2 4 1 1 3 1]	0.97	0.97	0.97	[[95 0] [4 38]]	0.966	
7	7	[1 1 1 1 3 1 1 2 1]	0.98	0.98	0.98	[[95 0] [3 39]]	0.991	
8	8	[1 1 1 1 2 1 1 1 1]	0.99	0.99	0.99	[[95 0] [2 40]]	1.0	
9	9	[1 1 1 1 1 1 1 1 1]	0.99	0.99	0.99	[[95 0] [2 40]]	1.0	

Table 6: Comparison of LR-RFECV and LR-RFE models.

Sl No	Training-Testing Set Ratio	LR- RFECV					Confusion matrix	LR- RFE					Confusion matrix
		Precision	Recall	F1-Score	No. of features used	MCC		Precision	Recall	F1-Score	No. of features used	MCC	
1	70-30	0.95	0.95	0.95	4	0.903	[[129 1] [9 66]]	0.95	0.95	0.95	4	0.923	[[129 1] [9 66]]
2	50-50	0.96	0.95	0.95	3	0.903	[[210 1] [15 116]]	0.96	0.96	0.96	4	0.916	[[210 1] [14 117]]
3	60-40	0.97	0.97	0.97	8	0.934	[[169 1] [8 96]]	0.96	0.96	0.96	4	0.933	[[169 1] [10 94]]
4	80-20	0.98	0.98	0.98	7	0.949	[[95 0] [3 39]]	0.97	0.97	0.97	4	0.966	[[95 0] [4 38]]

Table 7: Comparison of LR and LR-RFE models.

		LR without feature elimination(9 attributes used)					LR-RFE					
Sl No	Training-Testing partition ratio used	Precision	Recall	F1 score	Confusion Matrix	MCC	Training-Testing partition ratio used	No. of attributes used in LR-RFE model	Precision	Recall	F1-Scores	Confusion Matrix	MCC
1	70-30	0.97	0.97	0.97	[[129 1] [5 70]]	0.937	70-30	7	0.97	0.97	0.97	[[129 1] [5 70]]	0.980
2	50-50	0.95	0.95	0.95	[[210 1] [17 114]]	0.893	50-50	4	0.96	0.96	0.96	[[210 1] [14 117]]	0.916
3	60-40	0.97	0.97	0.97	[[169 1] [7 97]]	0.963	60-40	8	0.97	0.97	0.97	[[169 1] [8 96]]	0.953
4	80-20	0.99	0.99	0.99	[[95 0] [2 40]]	1.0	80-20	8	0.99	0.99	0.99	[[95 0] [2 40]]	1.0

LR (without feature elimination) using all 9 features and LR-RFE models are also compared in table 7. It was seen that LR -RFE obtained the same Precision, Recall and F1 scores for a reduced set of attributes when compared with the scores of LR without feature elimination. In the 70-30 partition, an F1 score of .97 was obtained with LR-RFE using 7 attributes instead of all 9. In 50-50 ratio LR-RFE got a better F1 score of .96 with 4 features. Similarly, 60-40 and 80-20 partitions also showed an F1 score for 0.97 and 0.99 respectively for reduced feature set of 8 attributes. MCC values were better for LR with feature elimination than the individual LR model in all partitions as shown in table 7. Thus supporting the fact that feature elimination and reduction of attributes enhances prediction accuracy. The receiver operating characteristics curve (ROC) is a performance measurement curve which plots sensitivity against specificity. Area Under Curve(AUC) has a value range between 0.5 to 1, where 0.5 denotes a bad classifier and 1 denotes a good classifier. The ROC and AUC for the various models are given as follows

The ROC of the LR-RFECV model is shown in figure 6 and the Area under curve calculated has value 1.

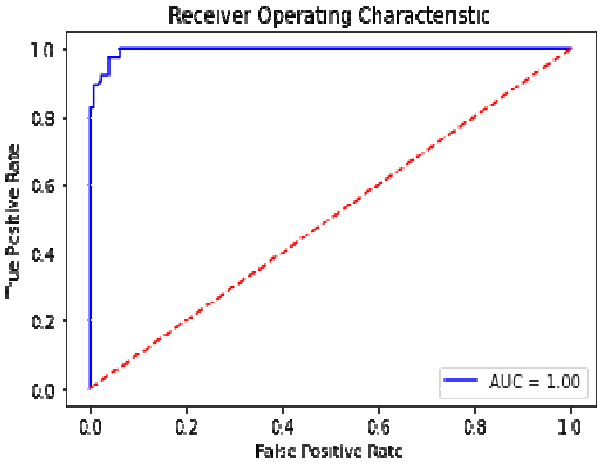

Fig. 6. ROC curve of LR-RFECV.

The ROC for LR-RFE model is shown in figure 7 and it can be seen that the Area Under Curve has value 1.

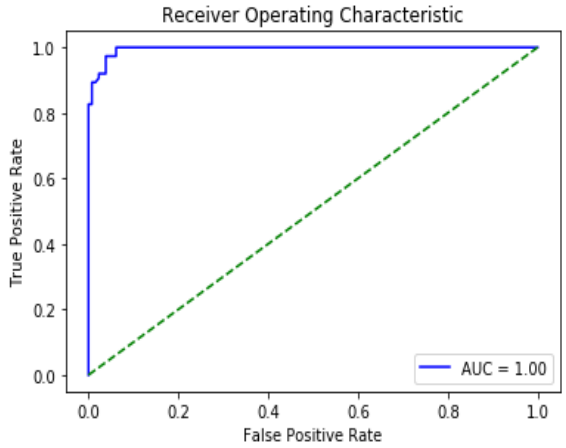

Fig. 7. ROC for LR-RFE.

IV. CONCLUSION

From the work it can be concluded that LR models with feature elimination methods provide better performance than when using the same model without feature elimination. Reduced feature set helps in improving model accuracy. Logistic regression deals effectively with outliers. The study highlights the importance of feature elimination for performance enhancement, in terms of accuracy, in supervised data mining models, hence aiding medical practitioners in easy and quick diagnosis of the disease.

V. FUTURE SCOPE

The future work will be to apply other feature reducing methods on different classifiers and to combine various data mining methods to evaluate their impact and performance enhancement on prediction accuracy in breast cancer diagnosis.

Conflict of Interest. Nil

ACKNOWLEDGMENT

The author is thankful to Dr AnilKumar K S, Associate Professor and Research guide in Technology Management, University of Kerala for his encouragement and support for the work. The breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

REFERENCES

[1]. Ahmed F. S, & Shawky, D.M., (2015). Logistic Regression Model for Breast Cancer Automatic Diagnosis, *SAI Intelligent Systems Conference 2015* November 10-11.

[2]. Wang, A., An, N., Xia, Y., Li, L., & Chen, G., (2014). A Logistic Regression and Artificial Neural Network-based Approach for Chronic Disease Prediction: a Case Study of Hypertension, 2014 IEEE International Conference on Internet of Things (iThings 2014), Green Computing and Communications (GreenCom 2014), and Cyber-Physical-Social Computing (CPSCom 2014).

[3]. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*, 1-21.

[4]. Choudhury, A. (2018). Identification of Cancer-Mesothelioma Disease Using Logistic Regression and Association Rule. *arXiv preprint arXiv:1812.10384*.

[5]. Leopord, H., Cheruiyot, W. K., & Kimani, S. (2016). A survey and analysis on classification and regression data mining techniques for diseases outbreak prediction in datasets. *Int. J. Eng. Sci*, *5*(9), 1-11.

[6]. Bhatti, I. P., Lohano, H. D., Pirzado, Z. A., & Jafri, I. A. (2006). A logistic regression analysis of the ischemic heart disease risk. *Journal of Applied Sciences*, *6*(4), 785-88.

[7]. Sultana, J. & Jilani, A.K., (2018). Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers, *International Journal of Engineering & Technology, 7*(4.20): 22-26.

[8]. Mangasarian, O. L., & Wolberg, W. H. (1990). *Cancer diagnosis via linear programming.*

University of Wisconsin-Madison Department of Computer Sciences.
[9]. Chang, M., Dalpatadu, R.J., Phanord, D., & Singh, K.A., (2018). Breast Cancer Prediction Using Bayesian Logistic Regression, *Open Access Biostatistics & BioInformatics,* Vol. *2,* Issue 3, 1-5.
[10]. Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*, *68*(16): 11-15.
[11]. Hassan, M., Butt, M. A., & Baba, M. Z. (2017). Logistic Regression Versus Neural Networks: The Best Accuracy in Prediction of Diabetes Disease. *Asian Journal of Computer Science and Technology,* Vol. *6* No. 2, 2017, pp.33-42.
[12]. Rahimloo, P., & Jafarian, A. (2016). Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. *Bulletin de la Société Royale des Sciences de Liège*, *85*, 1148-1164.
[13]. Johnson, P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P., ... & Rowe, C. C. (2014). Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC bioinformatics*, *15*(16), S11.
[14]. Yadav, P. K., Jaiswal, K. L., Patel, S. B., & Shukla, D. P. (2013). Intelligent heart disease prediction model using classification algorithms. *IJCSMC*, *3*(08), 102-107.
[15]. Rajbharath, R., & Sankari, L., (2017). Predicting Breast Cancer using Random Forest and Logistic Regression. *International Journal of Engineering Science and Computing*, Vol. *7*, issue 4, pg. 10708-10712.
[16]. Data Mining Group, http://dmg.org/pmml/v2-0/Regression.html
[17]. Rani, S., K., Manoj, S., M., & S Mani, S.,G., (2018). A heart disease prediction model using Logistic Regression. *International journal of Trend in Scientific Research and Development,* Vol. *2*, Issue 3, 1463-1466.
[18]. Sharanyaa, S., Gunavathiel, M.A., Abitha. P, & Sangeetha, K., (2018). Classification of Parkinson's Disease Using Logistic Regression. *International Journal of Pure and Applied Mathematics*, Volume *118* No. 18, 1587-1593.
[19]. Pandit, P. V., & Javali, S. B. (2012). Multiple logistic regression model to predict risk factors of oral health diseases. *Romanian Statistical Review*, *5*, 1-14.
[20]. Shravya, Ch., Pravalika, K., & Subhani, S., (2019) Prediction of Breast Cancer Using Supervised Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* Vol. *8*(6): 1106-1110.
[21]. Gai, X., & Zhang, Y. (2019). Diagnosis of Hepatobiliary Disease Based on Logistic Regression Model. In *IOP Conference Series: Materials Science and Engineering* (Vol. *490*, No. 6, p. 062084). IOP Publishing.
[22]. Liang, Y., Li, Q., Chen, P., Xu, L., & Li, J. (2019). Comparative study of back propagation artificial neural networks and logistic regression model in predicting poor prognosis after acute ischemic stroke. *Open Medicine*, *14*(1), 324-330.
[23]. Yusuff, H., Mohamad, N., Ngah, U., & Yahaya, A. (2012). Breast cancer analysis using logistic regression. *International Journal of Research and Reviews in Applied Sciences*, *10*(1), 14-22.
[24]. Sarvaiya, L., Yadav, H., & Agrawal, C., (2019). A Literature review of Diagnosis of Heart Disease using Data Mining Techniques, *International Journal of Electrical, Electronics and Computer Engineering, 8*(1): 40-45.
[25]. Basha, S.M., Bagyalakshmi, K., Ramesh, C., Rahim, R., Manikandan, R. & Kumar, A. (2019). Comparative Study on Performance of Document Classification Using Supervised Machine Learning Algorithms: KNIME. *International Journal on Emerging Technologies, 10*(1): 148-153.
[26]. Mathew, T.E. (2019). A Comparative Study on the Performance of different Support Vector Machine Kernels in Breast Cancer Diagnosis. *International Journal of Information and Computing Science,* Volume *6*, Issue 6, 432-441.

How to cite this article: Mathew, T.E. (2019). A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. *International Journal on Emerging Technologies*, **10**(3): 55–63.