Indian Journal of Science and Technology, *Vol 11(3), DOI: 10.17485/ijst/2018/v11i3/119179,* January 2018

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Performance Analysis of Regression and Classification Models in the Prediction of Breast Cancer

Aritra Basu[1*], Rohit Roy[1] and N. Savitha[2]

[1]School of Electronics Engineering (SENSE), VIT University, Vellore – 632014, Tamil Nadu, India;
aritra.basu2014@vit.ac.in, rohit.roy@vit.ac.in
[2]School of Social Science and Languages (SSL), VIT University, Vellore – 632014, Tamil Nadu, India;
savitha.n@vit.ac.in

Abstract

Objective: To suggest an automated diagnostic system for the early detection of breast cancer. **Methods:** This problem has been addressed by making use of machine learning algorithms that can accurately classify a tumor as either malignant or benign by identifying the minimum number of image features. A comparative study on various classification approaches such as Decision Tree, Support Vector Machine, K-Nearest Neighbor and Random Forest have also been conducted with a focus on cross validation to identify the best performing model. **Findings:** The study shows that Random Forest classifier gives the maximum accuracy. It also highlights that cross validation and fine tuning are necessary to prevent over fitting of data. **Improvements:** It has been observed that the selection of parameters play a very important role in correct classification as multicollinearity among attributes can render classifier models ineffective.

Keywords: Breast Cancer, Classification, Cross Validation, Decision Tree, K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine

1. Introduction

Breast cancer accounts for the maximum number of cancer diagnosis among American women. The number of cases of breast cancer in the United States of America (USA) has already exceeded the 4 million mark with every 1 among 8 American women developing the invasive form of the disease during the course her lifetime[1]. It is estimated that about 30% of the cancer cases diagnosed in 2018 among women will be breast cancer. About 50,000 women in the USA are expected to die from breast cancer in 2018[2].

The average 5-year survival rate for people with breast cancer is 90%[3]. However this data is strongly affected by the metastasis of the disease which signifies the spread of the cancer cells in the body. An early detection of the disease is very crucial as it can lead to long term survival[4]. Hence we are trying to address this issue by identifying classifier models that can lead to faster and more accurate detection of breast cancer in the early stages.

The diagnostic approach of breast cancer involves the inspection of medical images by skilled doctors to detect the characteristic symptoms of the disease. This process is very time consuming and not all physicians are experts in identifying the symptoms. So there is an urgent need of a reliable and automatic diagnostic system for the precise prediction of tumors[5]. The data available for manual diagnosis are mostly noisy and raw and it must be preprocessed before a feature selection method can be applied to reduce the cost of management and error rate. The latest machine learning techniques can provide a solution to this problem as they can be used by life scientists to extract necessary information from the databases of tumor images. Supervised learning methods are the most

**Author for correspondence*

popular machine learning paradigm used in cancer diagnosis[6].

This paper is organized as follows: Section 2 describes the dataset used, Section 3 covers the methodology applied in solving the problem, Section 4 gives a detailed explanation of the results, and Section 5 concludes the paper.

2. Dataset

Here we have made use of the multivariate cross-sectional WDBC breast cancer dataset as available in the UCI machine learning repository. The data is collected from the records of patients in the USA as maintained in the General Surgery Department of the Clinical Sciences Center in Madison, Wisconsin[7]. The dataset contains 569 samples with 32 attributes each. We have utilized 80% of the instances for training purposes while the remaining 20% have been used for testing. These testing data are applied over the different classification methods to test the accuracy of the systems.

3. Methodology

Most of the databases are susceptible to noisy and inconsistent data because of their origin from miscellaneous sources. This makes data preprocessing an absolutely necessary step for the classification of data. Data preprocessing involves the cleaning of data, followed by dimensionality reduction and finally, data transformation. This ensures that the data is now fit for classification. Here we have used five different classification models: Decision Tree, SVM, K-Nearest Neighbor (KNN), Random Forest and LRM. Initially the dataset is divided into two parts. The first part is used to train the model while the second part tests for its accuracy. This is followed by the selection of important features using the Random Forest classifier. The performance of the prediction models have been compared using both the mean features and the worst features to predict which dataset would provide better results. Cross validation and fine tuning have also been carried out to prevent over fitting of data and give more accurate prediction results. Finally the confusion matrices and prediction accuracies have been compared to select the best model for the prediction of breast cancer data. Figure 1 demonstrates the workflow of the overall methodology.

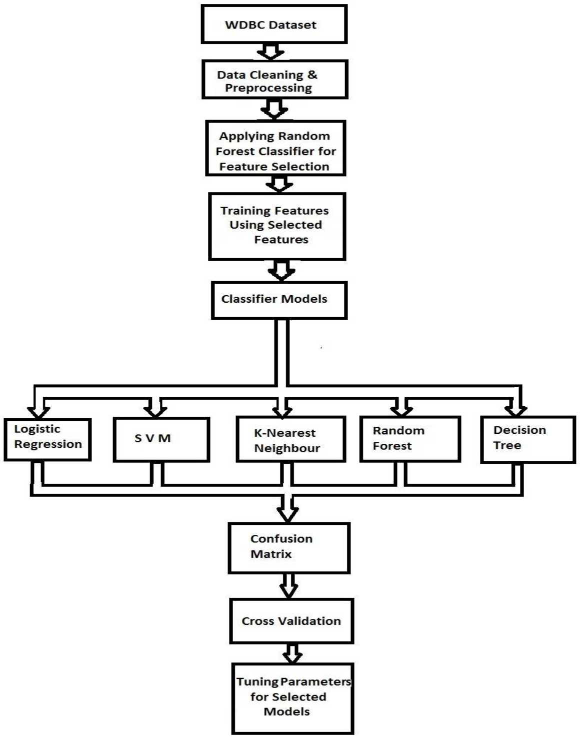

WDBC Dataset
Data Cleaning & Preprocessing
Applying Random Forest Classifier for Feature Selection
Training Features Using Selected Features
Classifier Models
Logistic Regression
S V M
K-Nearest Neighbour
Random Forest
Decision Tree
Confusion Matrix
Cross Validation
Tuning Parameters for Selected Models

Figure 1. Workflow of the overall methodology.

3.1 Pearson's Correlation Coefficient (PCC)

Pearson's Correlation Coefficient (PCC) is an extremely important and useful tool in statistics to measure the strength between values and relationships[8]. The strength

is given mathematically by $\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$, where

Cov uniquely denotes covariance and σ is the standard deviation corresponding to dataset x and y.

3.2 Decision Tree Classifier

Decision tree is a process of repeated division of the work-area into decisive subparts by identifying lines or classes. The decision tree induction algorithm works by recursively selecting the best attribute to split the data and expand the leaf nodes of the tree until a stopping criterion is met[9]. The choice of best split test condition is determined by comparing the impurity of child nodes using the following equation:

$$I_G(p) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^{\,2} = 1 - \sum_{i=1}^{J} p_i^{\,2} = \sum_{i \neq k} p_i p_k$$

3.3 Support Vector Machine (SVM)

SVMs excavate the idea of decision planes that define decision boundaries. The goal is to design a hyper-plane that efficiently classifies all training vectors into two classes[10]. The best choice in this case would be a hyper-plane that would leave maximum margin from both classes. The classifier looks at extremes and sets up margins, rather than training from examples already available.

3.4 K-Nearest Neighbor (KNN)

Basically, KNN is a method used extensively in pattern recognition to identify a particular class based on the closest training examples in the feature space. It is one of the simplest and most fundamental techniques when there is very less amount of knowledge about the data[11]. The excess risk in KNN classifier can be formulated as:

$$\Re_{\Re}(C_n^{wnn}) - \Re_{\Re}(C^{Bayes}) = (B_1 s_n^{\,2} + B_2 t_n^{\,2})\{1 + o(1)\},$$

where C_n^{wnn} denotes the weighted nearest classifier with weights $\{w_{ni}\}_{i=1}^{n}$ for constants B_1 and B_2 ,

where

$$s_n^{\,2} = \sum_{i=1}^{n} w_{ni}^{\,2}$$

and

$$t_n = n^{-2/d} \sum_{i=1}^{n} w_{ni} \{ i^{1+2/d} - (i-1)^{1+2/d} \}.$$

3.5 Random Forest Classifier

Random forest classifier is one of the most powerful and popular algorithms when it comes to prediction of a given dataset. It performs both classification and regression, and also handles missing values and accuracy of missing data[12]. It has the innate power to handle large and substantial datasets with higher dimensionality. The Random Forest classifier can be modeled as:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x'),$$

where, x' represents the unseen samples.

3.6 Logistic Regression Classifier

Logistic regression employs a non-linear function to describe the relation between the known input and expected output, as opposed to linear regression and can be used in complex, on-linear datasets. Input values are combined linearly using weights or coefficient values to predict an output value[13]. A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. The actual representation of the model that is stored in memory or in a file is the coefficients. These coefficients must be estimated from the training data. This is done by using maximum-likelihood estimation.

3.7 Cross Validation

A common practice in data science is to iterate over various models in order to find a better performing model. But this often leads to a contradicting scenario whereby it becomes difficult to distinguish whether this improvement in score is a result of improved capturing of the relationship or just over-fitting of the data[14]. The solution to this lies in the use of cross validation. Cross validation is a technique which reserves a portion of a data set, which is not used for training the model, but rather, is later utilized in testing the accuracy of the model. This method helps us in achieving more generalized relationships[15].

In figure 2, the first plot shows high error from training data points. This is an example of under fitting and this model will not perform well as it fails to capture the trend of the dataset. The second plot shows just the right relationship which refers to low training error and generalization of relationship. However, the third plot has almost zero training error. This is because the relationship is developed by considering each deviation in the data point which has rendered the model too sensitive by capturing every random pattern constituting the training dataset. This is an example of over fitting.

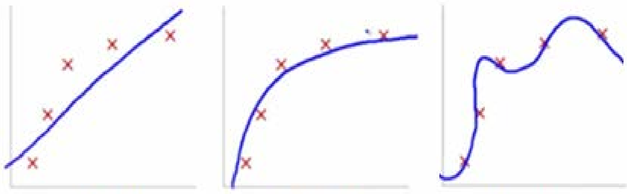


Fig. (a) Fig. (b) Fig. (c)

Figure 2. Need for cross validation.

Here we have made use of the k-fold cross validation technique. It ensures that the training model involves a large portion of the data so that the underlying trend of the dataset can be properly analyzed in order to avoid high bias. It also leads to a good ratio of testing data points which would otherwise lead to variance error. Finally it results in multiple iteration on the training and testing process by changing the training and test data set. The distribution of data over each iteration leads to improved effectiveness in the model validation.

4. Results

In this paper, we have conducted a comparative study on different classification techniques for the prediction of breast cancer. To begin with, the selection of important features using the Random Forest classifier is carried out as seen in Table 1 and 2. This is in accordance with Razo's Rule which states that although by taking all features, the model accuracy is increased; yet, it is not so much so as to rule out a simpler method in favour of a more complex one. The performance of the prediction models have been compared in Table 3 and 4 using both the mean features and the worst features to predict which dataset would provide better results. Figure 3 clearly shows that the mean features attribute provides better prediction accuracy as compared to the worst features. Finally, a correlation graph has also been plotted in Figure 4 so that we can remove multi collinearity. It basically refers to avoiding the use of multiple attributes which show a strong correlation as it would increase the complexity by making use of the same attribute twice in the prediction. As observed in Figure 5, with multi collinearity, the five most important features are concave points, perimeter, radius, area and concavity. However, upon removing multi collinearity, the five most important features are perimeter, compactness, symmetry, smoothness and texture.

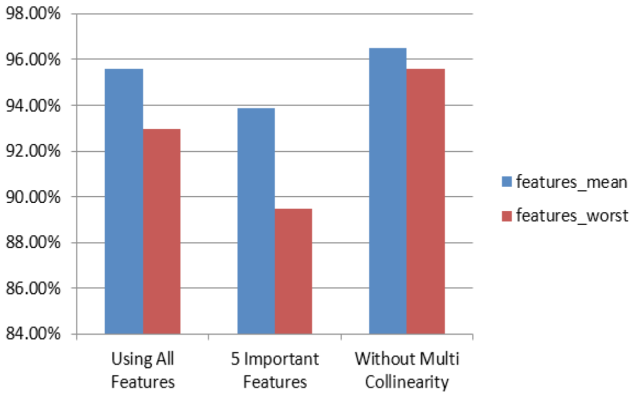


Figure 3. Attribute selection.

Table 1. Feature selection (features_mean dataset)

Features	Importance
concave_points_mean	0.319782
perimeter_mean	0.171614
concavity_mean	0.164598
area_mean	0.105712
radius_mean	0.090358
texture_mean	0.062389
smoothness_mean	0.031052
compactness_mean	0.025017
fractal_dimension_mean	0.014903
symmetry_mean	0.014574

Table 2. Feature selection (features_worst dataset)

Features	Importance
concave_points_worst	0.255380
perimeter_worst	0.193510
radius_worst	0.179010
area_worst	0.172068
concavity_worst	0.059352
texture_worst	0.042421
compactness_worst	0.034048
smoothness_worst	0.027344
symmetry_worst	0.018908
fractal_dimension_worst	0.017958

Table 3. Prediction accuracy (features_mean dataset)

Classifier Model	Using All Features	5 Important Features	Without Multi Collinearity
Support Vector Machine	63.158%	73.684%	90.351%
Random Forest Classifier	95.614%	93.859%	96.491%

Cross validation and fine tuning of the classification models have been carried out to prevent over fitting of data and give more accurate prediction results. The cross validation scores have been tabulated in Table 5. We have also made use of confusion matrices to describe the performance of the classification models based on a collection of test data whose responses were previously known.

Table 4. Prediction accuracy (features_worst dataset)

Classifier Model	Using All Features	5 Important Features	Without Multi Collinearity
Support Vector Machine	61.403%	59.649%	90.351%
Random Forest Classifier	92.982%	89.473%	95.614%

Table 5. Cross validation scores

Classifier Model	Score I	Score II	Score III	Score IV	Score V	Accuracy
Decision Tree Classifier	81.579%	85.965%	88.596%	89.254%	88.395%	100%
Support Vector Machine	74.561%	79.825%	84.211%	86.404%	85.052%	93.497%
K Nearest Neighbor	78.070%	83.772%	87.135%	88.377%	88.047%	92.970%
Random Forest Classifier	84.211%	88.596%	91.520%	92.544%	91.911%	100%
Logistic Regression Classifier	73.684%	79.386%	85.380%	87.281%	87.878%	89.807%

Table 6. Confusion matrix of different classification models

Decision Tree			**K Nearest Neighbor**		
	False	**True**		**False**	**True**
Benign	68	4	Benign	69	3
Malignant	2	40	Malignant	3	39
Support Vector Machine			Random Forest		
	False	True		False	True
Benign	69	3	Benign	70	2
Malignant	2	40	Malignant	0	42

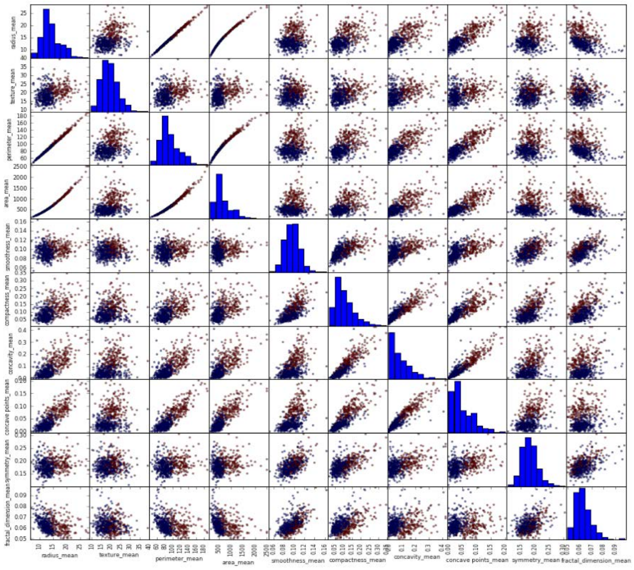
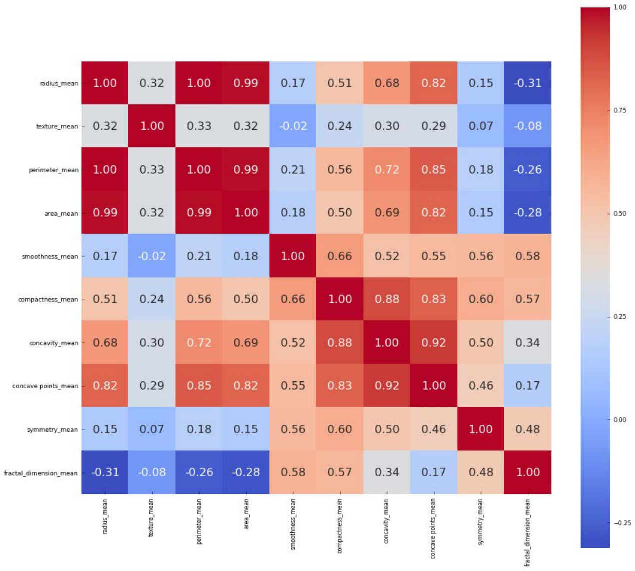


Figure 4. Correlation plot among mean features.



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1.00	0.32	1.00	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.32	1.00	0.33	0.32	-0.02	0.24	0.30	0.29	0.07	-0.08
perimeter_mean	1.00	0.33	1.00	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.99	0.32	0.99	1.00	0.18	0.50	0.69	0.82	0.15	-0.28
smoothness_mean	0.17	-0.02	0.21	0.18	1.00	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.51	0.24	0.56	0.50	0.66	1.00	0.88	0.83	0.60	0.57
concavity_mean	0.68	0.30	0.72	0.69	0.52	0.88	1.00	0.92	0.50	0.34
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1.00	0.46	0.17
symmetry_mean	0.15	0.07	0.18	0.15	0.56	0.60	0.50	0.46	1.00	0.48
fractal_dimension_mean	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1.00

Figure 5. Removing multi collinearity using heatmap

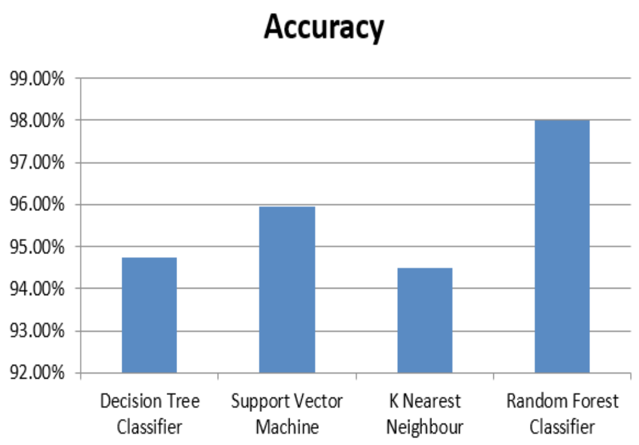


Figure 6. Accuracy of different classifiers.

These matrices have been tabulated in Table 6. Finally the accuracies of different classification models have been compared to select the most suitable one for the prediction of breast cancer data. Figure 6 clearly shows that Random Forest classifier gives the maximum accuracy.

5. Conclusion

This study makes it possible to detect the cases of breast cancer and classify it most accurately with greater precision and specificity. A variety of data processing techniques including data cleaning, feature selection, classification, cross validation and fine tuning have been used for ensuring maximum accuracy. The study reveals that Random Forest classifier gives the maximum accuracy with reduced subset of features. Support Vector Machine, K Nearest Neighbor and Decision Tree classifier have also shown reasonable performance in the diagnosis of breast cancer. It has also been observed that the selection of parameters plays a very significant role in correct classification as multi collinearity among attributes can render our model ineffective. Cross validation and fine tuning are also necessary to prevent over fitting of data.

6. References

1. Breast Cancer. Date accessed: 27/11/2017. https://www.medicalnewstoday.com/articles/37136.php.
2. Breast Cancer Statistics. Date accessed: 11/09/2017. https://ww5.komen.org/BreastCancer/Statistics.html.
3. Breast Cancer Figures. Date accessed: 12/13/2017. http://seer.cancer.gov/faststats/.
4. Survival Rates. Date accessed: 29/10/2017. https://en.wikipedia.org/wiki/Survival_rate.
5. Songa X, Mitnitskib A, Coxb J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. MEDINFO (IMIA International Marine and Industrial Applicators); 2004. p. 1–5.
6. Dubey SKD, Soni S. Predictive machine learning techniques for breast cancer detection, IJCSIT International Journal of Computer Science and Information Technologies. 2013; 4(6):1023–28.
7. Dataset. Date accessed: 29/09/2017. https://archive.ics.uci.edu/ml/machine-learning-databases/breast/cancer-wisconsin/wdbc/.
8. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility, Biometrics. 1989 Mar; 45(1):255–68. https://doi.org/10.2307/2532051.
9. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man, and Cybernetics. 1991 May; 21(3):660–74. https://doi.org/10.1109/21.97458.
10. Suykens JA, Vandewalle J. Least squares support vector machine classifiers, Neural Processing Letters. 1999 Jun; 9(3):293–300. https://doi.org/10.1023/A:1018628609742.
11. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm, IEEE Transactions on Systems, Man, and Cybernetics. 1985 Jul; 15(4):580–85. https://doi.org/10.1109/TSMC.1985.6313426.
12. Liaw A, Wiener M. Classification and regression by Random Forest, R News. 2002 Dec; 2(3):1–5.
13. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), The Annals of Statistics. 2000; 28(2):337–407. Crossref, Crossref.
14. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection, IJCAI International Joint Conference on Artificial Intelligence. 1995 Aug; 14(2):1137–45.
15. Improve Your Model Performance using Cross Validation. Date accessed: 18/11/2015. https://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-r/.