1/11/2022

“Data visualization and Algorithms Techniques for Breast Cancer Prediction and Classification with High Accuracy Rates”

CSCI-208 Design and Analysis of Algorithms

Presented To:

Dr. Hala Zayed

Eng. Ammar & Tasneem

Team Names:

Shrouk Hesham 19106271

Asmaa Mohamed Elmaghraby 19200036

Hussin Fekry 19105777

Sherif Elrefaey 19100006

Abstract

Breast cancer is one of the leading causes of death in women worldwide. Around-one in 30 women is affected by breast cancer. It's better to have knowledge of the type of cancer in the preliminary stages because early diagnosis leads to better outcomes. Mammography has helped in detecting breast cancer in the early stages which have reduced mortality. The diagnosis of breast cancer is dependent on a variety of parameters. We aim to create the best model for predicting breast cancer through preprocessing, feature extraction, data visualization and prediction using breast cancer data. Various visualization techniques like violin plot, grid plot, swarm plot and heat plot were utilized for proper feature extraction which has improved the accuracy of our results. For the purpose of prediction, we have used algorithms like the random forest, decision tree with single and multiple predictors, along with the commonly used statistical model, logistic regression model.

Table of Contents

2.Introduction

2.1 Problem Definition:

Breast Cancer is one of a major issue that some of the women are facing today.

Earlier detection of cancer by performing detailed analysis based on the existing records which may assist the physicians in providing a better treatment to their patients.

2.2 Scope:

The diagnosis of breast cancer is dependent on a variety of parameters. We aim to create the best model for predicting breast cancer through preprocessing, feature extraction, data visualization and prediction using breast cancer data. Various visualization techniques like violin plot, grid plot, swarm plot and heat plot were utilized for proper feature extraction which has improved the accuracy of our results. For the purpose of prediction, we have used algorithms like the random forest, decision tree with single and multiple predictors, along with the commonly used statistical model, logistic regression model.

3. Literature Review and Background

3.1 Breast cancer background

Breast cancer is a form of cancer that develops in the cells of the breast in women. Breast cancer comes in a variety of forms. Breast cancer treatment numerous criteria, such as its stage and simultaneous the individual's operations and comorbidities. It is preferable. to be aware of the type of cancer that exists in one's body preliminary phases, as early detection correlates to better outcomes. The location of the tumor determines the diagnosis aggregation clusters on a large and tiny scale, which are essential in the detection of breast cancer at an early stage Microcalcification is a collection of mineral stores within the breast tissue that look as little white-hued spots that may be caused by malignancy.

Breast lumps include a wide spectrum of diagnosis, from benign to malignant tumors. Because of their unpretentious appearance and unclear boundaries, alterations from the standard from typical breast tissues are challenging to peruse in growth detection. Radiologists can employ automated instruments to detect breast cancer in its early stages. Tumors are classified as either benign or malignant. A benign tumor can only live in the local area and cannot spread through other mechanisms. A malignant tumor takes advantage of adjacent tissues by infiltrating blood vessels and invading nearby cells. Because of the high death rate associated with breast cancer, early identification can offer hopeful results. However, because it is still a difficult topic to overcome, a promising technique to diagnose breast cancer has yet to be explained in our scientific literature.

3.2 Literature review and background of breast cancer with algorithms

There are two basic data mining algorithms: neural networks and decision trees. The logistic regression model was used as the statistical model. Breast cancer diagnosis accuracy was also predicted using 10-fold cross-validation methods. The decision tree model (C5) was shown to be the best predictor of all three models, with an accuracy of. Artificial neural networks came in second, and logistic regression models came in third, with an accuracy of. A recent comparison investigation of all of the above models, as well as 10-fold cross-validation, provides insight into the relative prediction abilities of various data mining methodologies. The results (in light of normal exactness of Heart and Breast Cancer) showed that the Nave Bayes is the best indicator on the holdout test, while the RBF Network came in second.

Simple Logistic came in third place for exactness, J48 came in fourth place for accuracy, and Decision table models came in fifth place for accuracy. In their research, Williams et al. (2015) looked at two different data mining algorithms for predicting breast cancer and compared their performance to find the best classifier. For the estimations of exactness, review, accuracy, and mistake rates reported for the two models, the J48 choice trees is a superior model for the forecast of bosom disease hazards, with an accuracy of 94 percent compared to navies Bayes of 83 percent.

They used a training set of 500 records from an arbitrary example and then attached the arrangement control set to the data. Information on breast cancer in its entirety They were able to achieve an accuracy of 94 percent the training stage and a testing accuracy of 93 percent stage. They compared how C4.5 was executed. computation and other ways of arrangement Future Improvising on this work is part of the process of improving it. To improve the classification rate, the C4.5 algorithm was used. increase the prominence of accuracy. The proposed technique

has a 74.5 percent accuracy. The Simple Logistic may be used to reduce the measurement of highlight space, and the proposed Rep Tree and RBF Network model can be used to produce quick automatic diagnostic frameworks for additional infections, according to the study.

3.3 Algorithm and Comparative Analysis

The data consisted of 32 characteristics that were used to determine the type of cancer. Some of these characteristics were unimportant since they did not give us with enough information to diagnose a breast tumor. Algorithms and data visualization techniques such as the violin plot, swarm plot, heat plot, and correlation matrix reveal redundant and unimportant characteristics.

Various visualizations plots are available to choose only the dataset's most significant features. The accuracy of models (random forest, decision tree, and logistic regression model) was greatly enhanced as a result of this.

The selection of which features to utilize in the model when doing classification was the second factor. The decision tree model's accuracy would sometimes suffer the consequences of selecting one feature. When the decision tree model is used with predictors such as radius mean, perimeter mean, area mean, compactness means, and concave points mean, the accuracy is 100 percent, indicating overfitting. However, when the decision tree is classified with only the radius mean predictor, the accuracy is 97.236 percent, but the 5-fold cross validation model is poor.

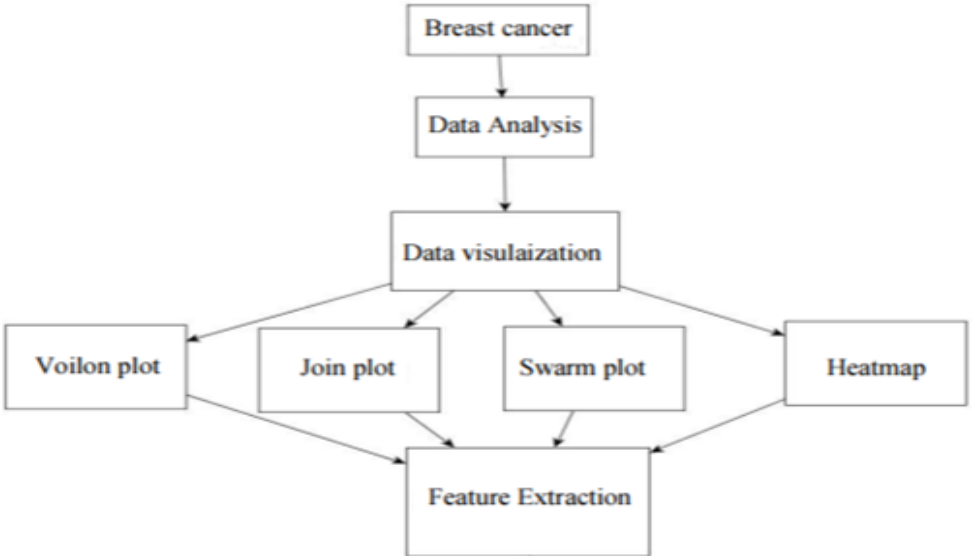
4. Methodology

The overall methodology process is done as follows. The dataset that we are going to use is taken from UCI machine repository. The next step is called the selection of the data. The data selected possess various attributes (around 32). Some of these may be redundant, while others may be lacking. There is also the possibility of noise in the data. As a result, the data must be cleaned before it can be processed further (data wrangling). They can be cleaned only after we explore or analyze the data. The attributes are viewed, and data is checked for what must be taken care-off. The data is then pre-processed to reduce noise and outliers. Because the data has 32 attributes, there's a good chance that just a small portion of them will be needed for data mining. These steps are taken in the Data Visualization process where data is visualized, and useful information can be taken out of it. The Violin plot, Swarm plot, join plot, and Heat map are used to visualize the data. This is the only way to determine which attributes are truly important for further processing. Then, we move to the feature extraction. Then, the data classification.

4. Methodology

4.1 Data visualization and Algorithms

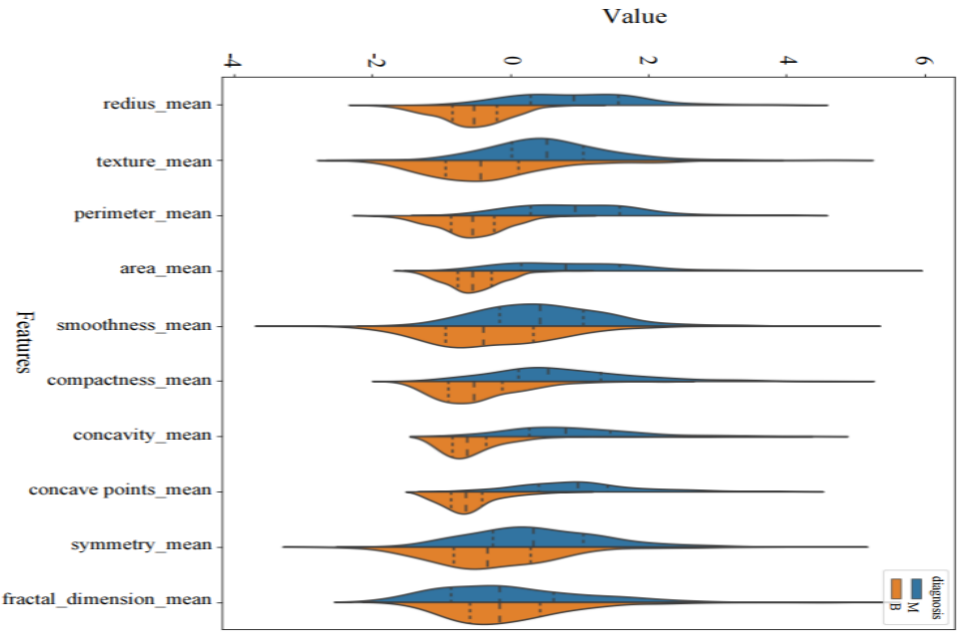
A basic data analysis is carried out to look at the pattern of data. Firstly, the mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in the classification of cancer. Larger values of these parameters tend to show a correlation with malignant tumours. Secondly, the mean values of texture, smoothness, symmetry or factual dimension do not show a particular preference of one diagnosis over the other.



4.1.1 Data visualization

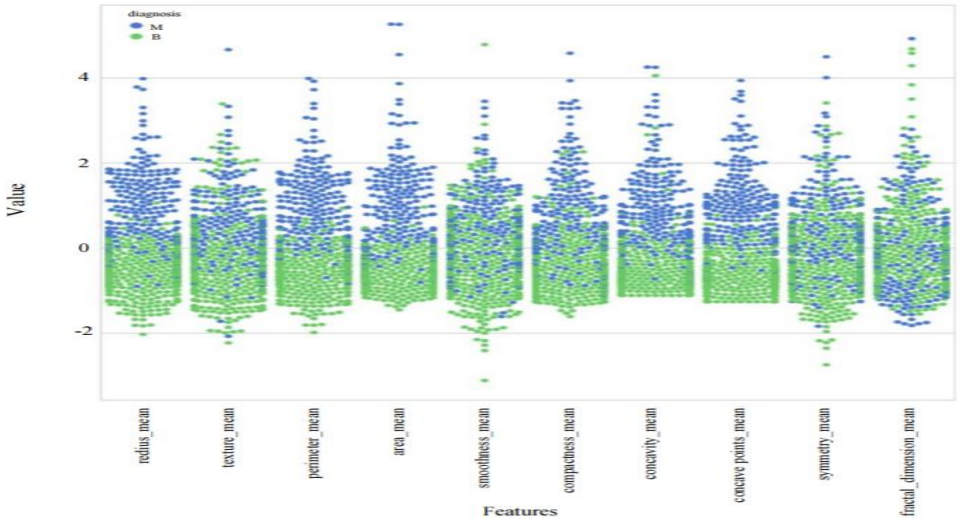
- Violin plot

in texture mean feature, the median of the Malignant and Benign looks like separated so it can be good for classification. However, in the fractal dimension mean feature, median of the Malignant and Benign does not look like separated so it does not give good information for classification.



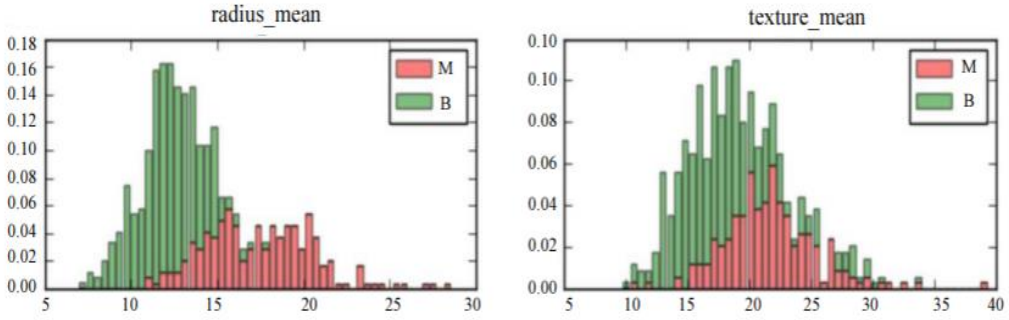
- Swarm plot

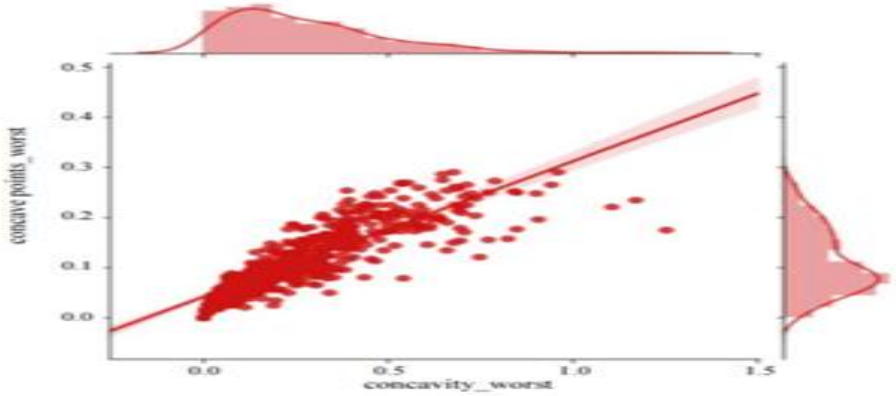
the variance can be seen more clearly. In this plot which feature looks like clearer in terms of classification. Here from the area worst feature in the last swarm plot looks malignant and benign are separated not totally but mostly



- Join plot

In the first join plot graph one can easily conclude that concavity worst and point worst predicted from violin plot turn out to be correlated. Hence, one feature is dropped instead of taking both the features. In order to compare two features deeper, we can use joint plot as shown in the next graph. Pearson value is correlation value and 1 is the highest. Therefore, 0.86 looks enough to say that they are correlated.

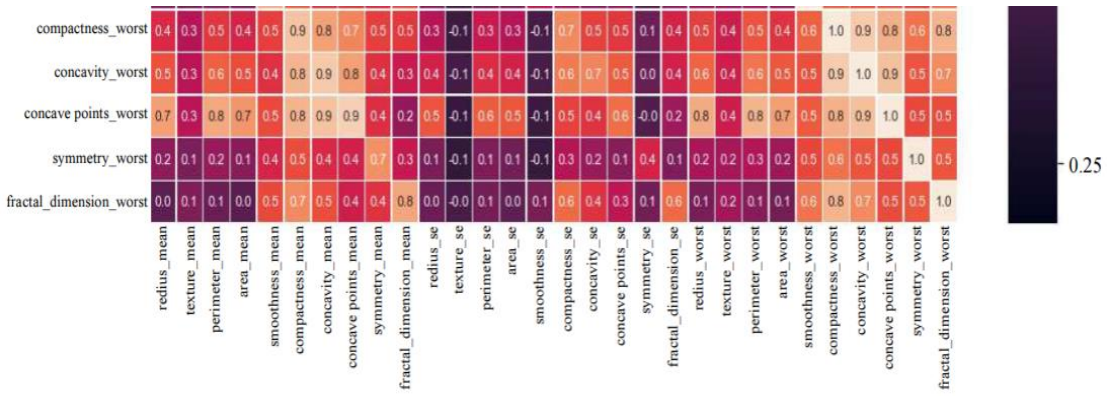




Correlation graph between concavity worst and concave point worst using the join plot

- Heat Map

shows all the correlation between features. Through which all the features which are not relevant are eradicated from this



4.1.2 Algorithms implemented

❖ **1. Logistic Regression Model**

We used a Logistic Regression which is a Machine Learning predictive analysis algorithm in order to classify our problem, and predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification

problems so we used it here in cancer detection. So here, based on the observations in the histogram plots, we can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. We can then perform a logistic regression analysis using those features as follows:

```
In [89]: predictor = ['radius_mean', 'perimeter_mean', 'area_mean', 'compactness_mean', 'concave points_mean']

In [90]: outcome = 'diagnosis'

In [91]: model = LogisticRegression()

In [92]: classification_and_fit_model(model, data, predictor, outcome)

Accuracy : 89.807%
Cross-Validation Score : 78.070%
Cross-Validation Score : 82.456%
Cross-Validation Score : 86.842%
Cross-Validation Score : 88.158%
Cross-Validation Score : 89.287%

In [93]: predictor1 = ['radius_mean']

In [94]: classification_and_fit_model(model, data, predictor1, outcome)

Accuracy : 87.873%
Cross-Validation Score : 74.561%
Cross-Validation Score : 78.947%
Cross-Validation Score : 84.795%
Cross-Validation Score : 86.184%
Cross-Validation Score : 86.469%
```

❖ **2. Decision Tree Model**

J4is was used Here, which is a simplistic learning supervised technique. This Algorithm employs the concept of ID3 to build a Decision tree by a top-down approach which is a greedy Search through training data where each and every Attribute is tested at each node while building a tree. It Uses Information gain which is a measure of the decrease in entropy of attribute after the split of the dataset.

```
[242]: model = DecisionTreeClassifier()

[243]: classification_and_fit_model(model, data, predictor, outcome)

Accuracy : 100.000%
Cross-Validation Score : 84.211%
Cross-Validation Score : 85.526%
Cross-Validation Score : 87.427%
Cross-Validation Score : 88.816%
Cross-Validation Score : 88.929%
```

```
[245]: classification_and_fit_model(model, data, predictor1, outcome)

Accuracy : 95.782%
Cross-Validation Score : 75.439%
Cross-Validation Score : 76.316%
Cross-Validation Score : 81.287%
Cross-Validation Score : 81.579%
Cross-Validation Score : 81.723%
```

❖ **3.Random Forest Model**

Is an ensemble of Decision tress, classifier creates a set of decision trees From a randomly selected subset of the training set. Using all the features improves the prediction Accuracy and the cross-validation score is great.

```
[246]: features_mean = list(X_train_df.columns[1:11])

[247]: model = RandomForestClassifier(n_estimators=100, min_samples_split=25, max_depth=7, max_features=2)

[248]: classification_and_fit_model(model, data, features_mean, outcome)

Accuracy : 96.485%
Cross-Validation Score : 85.965%
Cross-Validation Score : 89.035%
Cross-Validation Score : 91.813%
Cross-Validation Score : 92.544%
Cross-Validation Score : 92.973%
```

❖ **4.Support Vector Machine (SVM)**

96% accuracy on test data. We used it for solving classification problems are those that necessitate categorizing a given data set into two or more categories.

```
classification_and_fit_model(model, data, features_mean, outcome)
```

```
Accuracy : 96.485%
Cross-Validation Score : 85.965%
Cross-Validation Score : 89.035%
Cross-Validation Score : 91.813%
Cross-Validation Score : 92.544%
Cross-Validation Score : 92.973%
```

4.2 Dataset preparations & preprocessing

- ❖ The dataset that we use is taken from UCI Machine Learning
- ❖ Remove non-required columns such as id and nameless ones.
- ❖ Map the diagnosis column into numeric form; The diagnosis column is not numeric and must be transformed so that the algorithm can analyze and grasp the data.
- ❖ **Description of attributes in our dataset:**

Attribute Information: (1) ID number (2) Diagnosis (M = malignant, B = benign) 3-32) Ten real-valued features are computed for each cell nucleus: These attributes are radius which is mean of distances from center to points on the perimeter, texture which is standard deviation of gray-scale values, perimeter, area, smoothness which is local variation in radius lengths,

compactness which is perimeter^2/area - 1.0,concavity which is severity of concave portions of the contour, concave points which signifies number of concave

All feature values are recoded with four significant digits. Missing attribute values: None the class distribution contains 357 benign tumors and 212 malignant tumors.

id	diagnosis	redius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
82517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

5. Results

In our results, different models were used, trained and tested. The accuracy was predicted for all these models and k-fold cross-validation was performed on the training data to improve and correctly predict the accuracy. Out of all the results, random forest came out to be the best model in terms of classification with an accuracy score of high rates. The results shows that random forest turns out to be the most accurate model in terms of accuracy and 5-fold cross validation score.

```
classification_and_fit_model(model, data, predictor1, outcome)

Accuracy : 95.782%
Cross-Validation Score : 75.439%
Cross-Validation Score : 76.316%
Cross-Validation Score : 81.287%
Cross-Validation Score : 81.579%
Cross-Validation Score : 81.723%
```

```
classification_and_fit_model(model, data, predictor1, outcome)

Accuracy : 95.782%
Cross-Validation Score : 75.439%
Cross-Validation Score : 76.316%
Cross-Validation Score : 80.994%
Cross-Validation Score : 81.360%
Cross-Validation Score : 81.371%
```

This figure shows: With using predictor1 gives a better prediction accuracy.

```
classification_and_fit_model(model, data, features_mean, outcome)
```

```
Accuracy : 96.485%
Cross-Validation Score : 85.965%
Cross-Validation Score : 89.035%
Cross-Validation Score : 91.813%
Cross-Validation Score : 92.544%
Cross-Validation Score : 92.973%
```

```
classification_and_fit_model(model, data, predictor, outcome)
```

```
Accuracy : 93.849%
Cross-Validation Score : 82.456%
Cross-Validation Score : 86.842%
Cross-Validation Score : 90.643%
Cross-Validation Score : 90.789%
Cross-Validation Score : 91.216%
```

This Figure shows: Using all the features improves the prediction accuracy and the cross-validation score is great which gives a better prediction accuracy, but the cross-validation is not great.

6. Discussion& Conclusion

In conclusion, using data visualization to diagnose breast cancer can be used to exclude some features and determine which ones are significant. All of the models were evaluated in terms of accuracy and cross-validation. According to our findings, the Random-Forest Model with top 5 predictors 'concavepoints_mean','area_mean','radius_mean','perimeter_mean','concavity_mean' is the best model to apply for diagnosing breast cancer. As indicated in the results, it has a prediction accuracy of ~95% and a cross-validation score of ~93% for the test data. As a conclusion, it is easy to conclude that the Random-Forest classifier is the most accurate of all these models in terms of accurately diagnosing breast cancer with careful feature selection using data visualization for this data set.

7. Future work & Suggestions

In our project future work, we can develop a robust data analytical model with algorithm which can assist in better understanding of breast cancer survivability, providing better insights into factors associated with patient survival, and establishing cohorts of patients that share similar properties. In addition, the results of our analysis and algorithms implemented can be used to segment the historical patient data into clusters or subsets, which share common values.

Furthermore, providing enhanced computational visualization and algorithms and thus saving patients' lives.

8. References

- Williams, K., P.A. Idowu, J.A. Balogun and A.I. Oluwaranti, 2015. Breast cancer risk prediction using data mining classification techniques. Trans. Netw. Commun. DOI: 10.14738/tnc.32.662.
- Wolberg, W.H., W.N. Street and O.L. Mangasarian, 1994. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett., 77: 163-71. DOI: 10.1016/0304- 3835(94)90099-x.
- Hosmer, D., S. Lemeshow and R.X. Sturdivant, 2000. Applied Logistic Regression. 1st Edn., A WileyInterscience Publication, New York.
- Mangasarian, O.L. and W.H. Wolberg, 1990. Cancer diagnosis via linear programming. SIAM News, 23: 1-18.
- Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106. DOI: 10.1023/A:1022643204877.
- Rajesh, K. and S. Anand, 2012. Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. Int. J. Adv. Res. Comput. Commun. Eng., 1: 72-77.