Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

Review Article

MACHINE LEARNING APPROACHES FOR HEALTHCARE DATA ANALYSIS

Hemalatha Eedi[1], Morarjee Kolla[2]

[1]JNTUH College of Engineering Hyderabad, Department of CSE, Hyderabad, India.
[2]CMR Institute of Technology, Department of CSE, Hyderabad, India.

Received: 17.12.2019 *Revised: 20.01.2020* *Accepted: 22.02.2020*

Abstract
Breast cancer is the most common cancer in women worldwide and it remains the most common cause of cancer-related death in woman globally. Machine Learning techniques have been proven to be of great help in prognosis and diagnosis of various health related issues. This work constitutes a comparison of five machine learning (ML) algorithms: Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive-Bayes (NB), Decision Tree (DT), Random Forest (RF) on the Breast Cancer Wisconsin Diagnostic (BCWD) dataset. Features were extracted from the digitized images of FNA tests on a breast mass. Results show that Random Forests performs better among all the models across different classification metrics such as accuracy, precision, recall, and f1-score.

Keywords: Breast cancer, Decision Tree, K-Nearest Neighbor, Logistic Regression, Machine Learning, Naïve-Bayes, Random Forest.

DOI: http://dx.doi.org/10.31838/jcr.07.04.149

INTRODUCTION

Industry Giants like IBM, Microsoft, Google, Intel etc. are investing a lot for research in the usage of AI algorithms for healthcare. AI programs have been developed and applied to practices such as diagnosis, prognosis, drug development, patient monitoring, and personalized medicine. There is an increased usage of ML algorithms in the fields of radiology, imaging, tele-health, electronic health records. Better prognosis and diagnosis of diseases and decrease in medical costs are due to the usage of AI in health-care systems. Huge volumes and variety of health-care related data is generated at a fast velocity. It's beyond human potential to handle, interpret, analyse, and draw conclusions from this big data. Machine Learning on the other hand feeds on data, and efficient computational resources. The bigger the data, the better the machine learning model learn, and in turn yield better predictions. Classification performance metrics such as Accuracy, Precision, Recall, F1-score are being evaluated on different supervised machine learning models such as Logistic Regression, K-Nearest Neighbour, Naïve-Bayes, Decision-Tree, and Random Forest. Model Evaluation is done on Breast Cancer Wisconsin Diagnostic dataset from UCI Machine Learning repository. Model with best performance is chosen to aid in clinical diagnosis.

Data Insights

Data has become an essential commodity in everyone's life. Gaining insights into data equips an individual with better knowledge, and in due course lets the individual to better informed decisions [1]. The success level in competitive world depends upon the level of insights one extracts from the available data.

Think of an "insight" as anything that increases understanding of how the system actually works. It bridges the gap between how we think the system works and how it really works by analysing data [2].

Capability to draw actionable insights from data is key component for success of an organization. Ability to automatically draw insights from data and evolve with more available data, without programming, is the holy grail of business intelligence.

Machine Learning Insights

Machine learning is about the process of *programming to learn* instead of programming for a single output. It offers a capability that accelerates data-driven insights and knowledge acquisition. Although machine learning has been around for decades, it's the all-pervasive nature of data due to social networking, video-streaming, cloud computing etc., machine learning has become the pioneer of business intelligence [3].

With high computation power, processing and performing parallel complex mathematical operations on big data, has become a boon for resurgence of machine learning. IoT sensors, social media, mobile devices, content provider servers, cloud generates huge volumes of data. It's difficult to derive all the business value from this big data manually. Machine learning exactly serves the purpose in automate learning [4] about the data and drawing valuable business insights in no time, which enhances the productivity of an individual and organization in turn.

REVIEW OF LITERATURE

This chapter presents the fundamental understanding of three broad categories of machine learning algorithms such as supervised, un-supervised, and reinforcement learning algorithms. The study, then transitions to discussion of supervised machine learning algorithms such as Logistic Regression, Nearest Neighbors, Gaussian Naïve Bayes, Decision-Trees, and Random Forests [5]. The techniques enunciated above are used to evaluate their performance on Breast Cancer Wisconsin Diagnostic dataset [6].

S.M.M Hasan et al [5]focuses on removing the irrelevant attributes which are not highly correlated with other features with help of feature selection method and helps in predicting heart disease with minimal attributes. Error metrics are compared for different classification algorithms. Logistic Regression algorithm outperforms other classification algorithms. *P. Suryachandra et.al* [7] carried out research work says machine learning algorithm for cancer diagnosis has trained machine learning algorithm that predicts the severity and type of cancer that helps in treating the condition of the patient. The main drawback of classification algorithms is number of features far exceeds the number of patient cases. Here Random Forest

and Naive Bayes achieves performance efficiency using feature selection method. Random Forest is used to rank the feature importance and applied for relevant feedback. *Seigel R L et.al* [1] gives particulars about cancer for clinicians research on Wisconson dataset [2]. *Anusha Bharat et al* [8]presented Machine Learning Algorithms like KNN, Logistic Regression, Niave Bayes and Decision trees by performing Analysis Using Wisconsin Breast cancer Dataset [9]. Performance of each algorithm varies depending on parameter selection. Fine tuning of parameters gives better results for algorithms. KNN, Logistic Regression and Naive Bayes performed well [10]. This paper applies three machine learning techniques: Naïve Bayes, SVM and Random Forest to Wisconsin Breast Cancer Database. The three developed models predict whether the patients' trauma are benign or malignant. It aims at comparing the performance of these three algorithms through accuracy, precision, recall and f-measure. Results show that Random Forest yields the best accuracy of 99.42%, which is slightly better than both SVM and Naïve Bayes that have accuracies of 98.8% and 98.24% respectively. These results are very competitive and can be used for diagnosis, prognosis, and treatment. *L Liu* [11], worked on Logistic regression specifying a regularisation coefficient that determines better learning rate. Based on learning rate, lambda value is either increased or decreased yielding classification accuracy. Choosing the better Feature combination improves accuracy in prediction of breast cancer. Extracting the feature attributes of different types of tumours improves the classification accuracy.

A key aspect of machine learning is that it has huge business value as it does not require much explicit programming in advance to learn more about data and it simulates certain human learning methodologies. Once data is ready, models are evaluated, the learning system goes through the learning iterations on its own to uncover latent business value from data.
A wide variety of data can be chosen as input to learning system. Data could come from enterprise systems, mainframe databases, IoT edge devices etc. Learning processes used for business systems are either supervised or unsupervised which accomplish different goals. There are other learning methodologies such as *lazy* and *eager* learning methodologies which governs how to process data. Learning systems deliver output which is either predictive or prescriptive in nature which may be stored for analysis, delivered as reports or fed as input into other enterprise applications [12].

METHODOLOGY

ThisAs part of this work, Breast Cancer Wisconsin (Diagnostic) Dataset from UCI Machine Learning repository [2] is used for comparing performances of Supervised Machine Learning models like Logistic Regression, Nearest Neighbors, Gaussian Naïve Bayes, Decision Trees, and Random Forests
Accuracy, precision, recall, f1-score are the different classification metrics [5] considered. Goal of this analysis is to

choose better machine learning which better distinguishes a malignant from a benign tumour and aids in clinical diagnosis.

Dataset

Breast Cancer Wisconsin (Diagnostic) dataset is a multivariate dataset. It consists of 569 samples with 32 attributes with no missing values for any of the attributes. Except One attribute ('diagnosis' attribute which is categorical), the remaining 31 attributes are real valued.

System Design

Feature Engineering helps in extracting features from the digitized images of Fine Needle Aspirates (FNA) of a breast tumor cell mass. Breast Cancer Wisconsin dataset is generated after feature engineering.

BCWD dataset has 10 attributes and 30 features centered around those attributes. Id, diagnosis, Unnamed: 32 are the 3 other attributes present as part of dataset. So, pre-processing is required to cleanse data of unwanted features [1][7]. Normalisation helps in speeding up the learning process. Dataset once pre-processed and normalized is ready to be input into machine learning model.

Typical hyper-parameter for train-test split is *test size=.3*. Train data, then, is used for training the machine learning model. After learning parameters, the ML model evaluation is done upon test data and the classification results are reported.

IMPLEMENTATION

The three modules in this section are

a. Loading Breast Cancer Wisconsin Diagnostic dataset into IPython Notebook and perform cleansing and analysis on the dataset
b. Preprocess and Split dataset into Train, Test based on hyper-parameters
c. Evaluate the performance of models using classification metrics such as accuracy, precision, recall, f1-score and choose a better model.

Loading the Dataset

Breast Cancer Wisconsin Diagnostic dataset is downloaded from UCI Machine Learning repository [2] and copied into current working directory for model development and comparative analysis as shown in Fig. 1.

```
#Load Breast Cancer Wisconsin Diagnostic dataset from current directory
breastCancerDataset = pd.read_csv('breastCancerWisconsin(Diagnostic).csv')
```

```
#display first 5 elements of the BCWD dataset
breastCancerDataset.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 33 columns

```
#Determine the shape of the BCWD dataset
breastCancerDataset.shape
```

```
(569, 33)
```

Fig. 1: Loading Dataset

The dataset available in.csv format is loaded into python environment using pandas framework. The dataset being loaded has 569 samples and 33 features. The target label *'diagnosis'* is the second column and is categorical in nature.

Cleansing the Dataset

'Id', 'Unnamed: 32' are two unnecessary features which are dropped. After dropping the 2 samples, dataset consists of 569 samples and 31 features which includes *'diagnosis'* categorical target label.

Mapping Categorical to Numerical Type

A dictionary is defined which has numerical correspondence for every categorical value in the target label as shown in Fig. 2. The dictionary defined is then mapped onto every sample instance in the dataset.

```
#Define a map which maps categorical to numerical data
#Apply the defined map upon the 'diagnosis' feature in the modified dataset
classMap = {'B':0, 'M':1}
breastCancerDataset.diagnosis = breastCancerDataset.diagnosis.map(classMap)
```

```
breastCancerDataset.head()
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980

5 rows × 31 columns

Fig. 2: Data Preprocessing

Analyze the Dataset

The Fig. 3 gives an insight into the distribution of benign and malignant samples from the Wisconsin Diagnostic dataset.

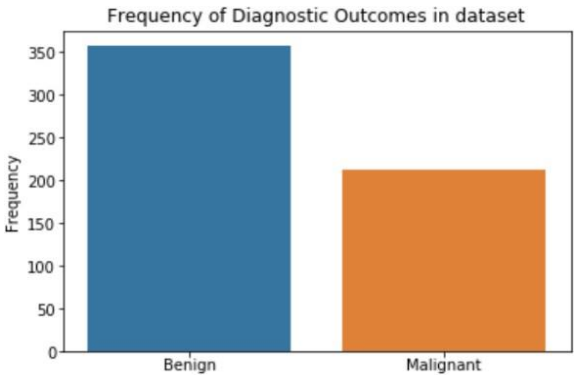

Fig. 3: Diagnostic Outcomes in Dataset

Tumor Features for Benign/Malignant Samples

Samples are segregated based on *benign* and *malignant* categories and the mean of features are listed out for a quick insight into the difference between *benign* and *malignant* tumors as shown below Fig. 4.

```
#Get a quick understanding of the tumor features (mean values) wrt diagnosis
breastCancerDataset.groupby('diagnosis').mean()
```

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	c p
0	12.146524	17.914762	78.075406	462.790196	0.092478	0.080085	0.046058	
1	17.462830	21.604906	115.365377	978.376415	0.102898	0.145188	0.160775	

2 rows × 30 columns

Fig.4: Dataset Samples Segregation

Single Feature Distribution of Benign/Malignant Tumors

The code-snippet gives an insight into the distribution of *radius_mean* featurearound 12 for *benign* samples and is around 17 for *malignant* samples. The Fig. 5 pictorially depicts that *benign* samples are tightly packed and has less *standard deviation* as opposed to *malignant* samples which are spread around the respective mean value.

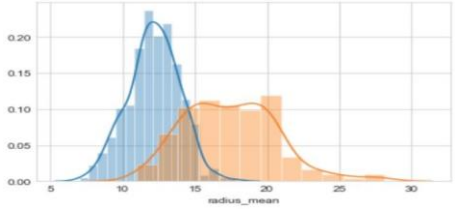

Fig. 5: Graph Representation of Radius_mean

Multiple Features Distribution of Benign/Malignant Tumors

Fig.6 shows the distributions of 10 features such as radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave_points_mean, symmetry_mean, fractal_dimension_mean.

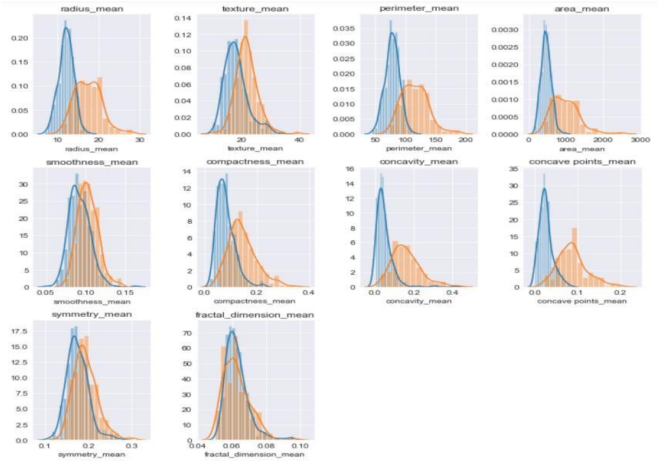

Fig. 6: Features Mean

The diagram shows that *radius_mean, perimeter_mean, concave_points_mean*features do not overlap much as opposed to other features which almost overlap.These set of features which do not overlap are used in Naïve Bayes which better classifies a test sample to either *benign* or *malignant* class once the model learns the distribution of the *benign* and *malignant* class with respect to the feature.

Heat Map for BCWD Dataset

The following Fig. 7 shows the heat map which depicts the correlation between various features. Heat map uses colored cells in a mono-chromatic scale in depicting relation between features.

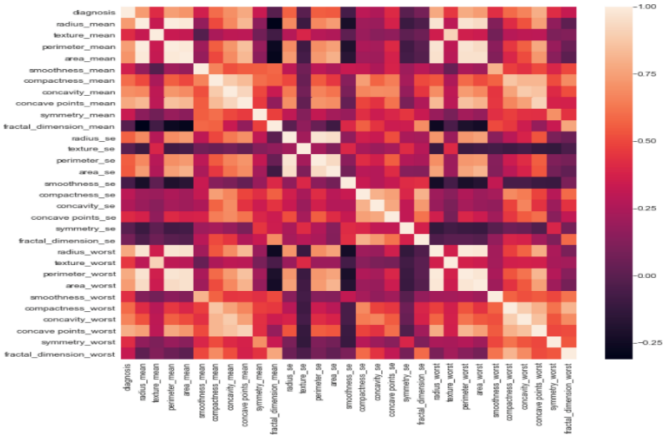

Fig. 7: Heat Map for BCW Dataset

Correlation gives an indication of the features fluctuate together. A positive correlation lets the features fluctuate together where as a negative correlation lets the features fluctuate in opposite direction. The higher the correlation between features, the more the scope for dimensionality reduction. Model development being done with 30 features can be done with 5-10features if there is a strong correlation between features. In absence of strong correlation between the features, every feature plays an equally important role in determining the target label. A visual inspection of the correlation between features helps data scientists get an insight into how the features are related and take decisions accordingly.

Normalizing the Dataset

Normalization brings features to same scale which induces better behavior of the gradients, in the process allowing the model to learn faster.

Train Test Split

Once normalized for faster training, the data is split into train, test groups by using hyper-parameter of 0.3 for test data.

Hyper-parameter Configuration

Logistic Regression, Nearest Neighbors, Gaussian Naive Bayes, Decision Trees, and Random Forests are the algorithms which are being modelled with different hyper- parameters.In logistic regression, the hyper-parameter C is inversely related to λ, the regularization regulator which helps in avoiding over-fitting. Ideal value for C is 10, which yields λ value of 0.1.

Model Segregation

Models with respective hyper-parameters are instantiated and segregated together in a single list as shown in code-snippet in next figure, makes the module development highly cohesive in nature which aids in better maintenance and debugging.

ANALYSIS AND RESULTS

The models being developed are trained on BCWD data and are evaluated on the same dataset.

Cohesive Module for evaluation

The estimator_function is a cohesive module that takes 2 parameters: models With Params which has all the models with their respective hyper-parameters, and scoring metric. The estimator_function returns mean and standard deviation of the scores obtained for different hyper-parameters for the chosen scoring metric.

Accuracy Metric Evaluation

The scoring metric under consideration is accuracy. The estimator_function shows that Random Forests perform better with 93.21% accuracy, Decision Trees stands next with 91.4% accuracy, followed by Logistic Regression with 90.96% accuracy, followed by Nearest Neighbors with 90.96% accuracy. Problem with NN is, there is actually no learning happening in NN. It's OK to work with NN for small datasets. But the advantage is lost when working with huge datasets. Decision Trees have an inherent disadvantage of not being robust to changed data which is not the case in real-world scenarios.

Random Forests are computationally intensive and n_estimators under consideration for huge data will be in 100's and 1000's which involves cloud computing.

Precision Metric Evaluation

The scoring metric under consideration is precision. Logistic regression out-performs all other models with highest mean of precision scores, reasonably less computationally intensive as compared to Decision trees and random forests, better learning model compared to NN which doesn't have learning feature.

Recall Metric Evaluation

The metric under consideration is recall. Random Forest perform better in recall scenarios outperforming all other models whereas Logistic regression gives least recall mean value. The mean value of the logistic regression model is being affected by performance with other hyper-parameters.

F1-score Metric Evaluation

The scoring metric under consideration is f1-score. F1-score gives a balanced view with respect to precision and recall. F1-score is harmonic mean of precision and recall. So, it is affected by the smallest of the two metrics under consideration.

Accuracy Metric Evaluation for Individual Hyper-parameters

The accuracy score for the various models with their respective different hyper-parameters. The accuracy for logistic regression is 94.9% for C = 10 which makes $\lambda = 0.1$. Nearest Neighbors show highest accuracy for n_neighbors=20 hyper-parameter. Random Forests gives an accuracy of 97.5% for n_estimators = 5.

Precision Metric Evaluation for Individual Hyper-parameters

The Logistic regression performs 100% for C=10. NN has precision of 96.8% for K = 20. Random Forests has a precision of 100% for n_estimators=5.

Recall Metric Evaluation for Individual Hyper-parameters

The regression has recall of 100% for C = 10 and is the same scenario with Random forests for n_estimators=5.

F1-score Metric Evaluation for individual hyper-parameters

The figure below shows that Logistic Regression has f1-score of 100% along with Random forests with n_estimators=5. Logistic regression gives the same performance as Random forests in most of the scenarios with least cost.

Summarized Performance

The different metric evaluations on various models clearly show that Logistic regression has an edge over other models owing to simplicity, learning capability and cost-effectiveness. Once Logistic regression is finalized as optimal model across different metric evaluation, the focus is being diverted to the optimal number of features required in almost equivalent prediction with 30 features.

Feature Importance Evaluation

The Fig. 8 shows that the following six features(texture_worst, perimeter_mean, area_mean, perimeter_worst, area_worst, texture_mean) are sufficient to learn the model with almost equal performance as the original model.Advantage with the above conclusion is that the model will be able to learn much faster, with less computation costs. Model development for the above is done using Principle Component Analysis.

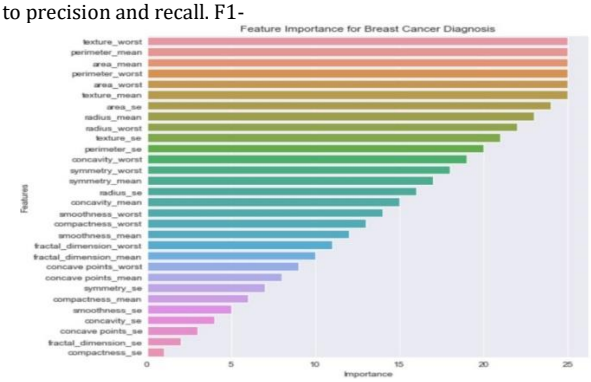

Fig. 8: Feature Importance Evaluation

Recursive Feature Elimination with Cross Validation on logistic regression model gives the *optimal number of features* with same validation scores as model with all features present. Fig. 9 below shows that, *optimal number of features* required for constant cross validation scores.

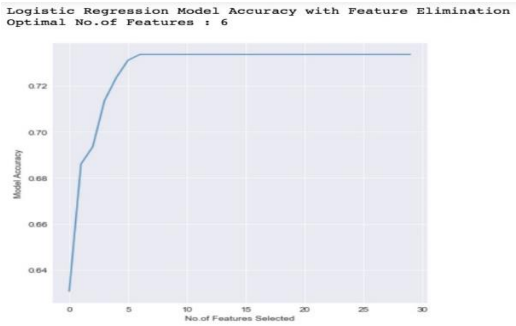

Fig. 9: Logistic Model Accuracy with feature elimination

CONCLUSION

Evaluation of classification metrics such as *accuracy, precision, recall, f1-score* is done on machine learning models such as *logistic regression, nearest neighbors, Gaussian naive bayes, decision trees, and random forests.* Apparently, random forests outperform all other models when mean scores of different metrics is considered. But, when a close observation is made in the respective metric scores with respect to different hyper-parameters, logistic regression performs on par with random forests. So, logistic regression model is chosen as better model among the lot owing to its simplicity and cost- effectiveness. Once the data is huge and complex in nature, the models under consideration may not perform well. In such scenarios, neural networks are default choice. Linear model such as logistic regression is very simple compared to complex and computationally intensive model like neural network.

REFERENCES

1. Seigel R.L, Miller K.D and Jemal A, "CA: *A Cancer Journal for clinicians*", Global Cancer Statistics, vol 67, pp 7-30,2017.
2. William H Wolberg, W Nick Street and Olvi L Mangasarian, "Breast Cancer Wisconson (diagnostic) data set", UCI Machine Learning Repository,1994.
3. P. Louridas and C.Ebert, "Machine Learning", IEE Software, vol. 33, no.5, pp. 110-115, 2016.
4. C.P. Utomo, A. Kardina, and R. Yuliwulandari, "*Breast cancer diagnosis using artificial neural networks with extreme learning techniques*," International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol.3, no.7, pp. 10-14, 2014.
5. Prajapati DS, Shah JS, Dave JB, Patel CN. "Therapeutic Applications of Monoclonal Antibodies." Systematic Reviews in Pharmacy 2.1 (2011), 37-42. Print. doi:10.4103/0975-8453.83437
6. William H Wolberg, W Nick Street and Olvi L Mangasarian, "Breast Cancer Wisconson (diagnostic) data set", UCI Machine Learning Repository,1994.
7. P. Suryachandra and P. VenkataSubba Reddy, "*Comparison of Machine Learning Algorithms for Breast Cancer*", ICICT, 2016.
8. Anusha Bharat, Pooja N and R Anishka Reddy, "Using Machine Learning Algorithms for breast cancer risk prediction and diagnosis", IEEE, 2018.
9. Rastogi, P., Singhal, R., Sethi, A., Agarwal, A., Singh, V.K., Sethi, R.Assessment of the effect of periodontal treatment in patients with coronary artery disease : A pilot survey(2012) Journal of Cardiovascular Disease Research, 3 (2), pp. 124-127.
   DOI: 10.4103/0975-3583.95366
10. Madeeh Nayer El-gedawy, "*Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes",* IJECS Volume 6 Issue 1 Jan., 2017 PageNo.19884-19889
11. L. Liu, "*Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning*," International Conference on Robots & Intelligent System (ICRIS*)*, Changsha, pp. 157-160, 2018.
12. Rashmi G D, A. Lekha and N. Bawane, "*Analysis of efficiency of classification and prediction algorithms (Naive Bayes) for Breast Cancer dataset*," *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, pp. 108-113,2015.