ijar
ISSN NO. 2320-5407

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/9856
DOI URL: http://dx.doi.org/10.21474/IJAR01/9856

RESEARCH ARTICLE

TO IMPROVE THE ACCURACY IN IDENTIFYING BREAST CANCER USING VARIOUS TECHNIQUES OF BIG DATA ANALYSIS.

Aurobind Ganesh[1], K. Vijiyakumar[2], K. Premkumar[2] and P. Mathivanan[2].

1. M.Tech, MLIS, Senior System Analyst, NIMHANS Digital Academy, Dept. of Psychiatry, NIMHANS, Bengaluru, INDIA.
2. M.Tech, Assistant Professor, Department of Information Technology, Manakula Vinayagar Institute of Technology, Puducherry, INDIA.

Manuscript Info

Manuscript History
Received: 08 August 2019
Final Accepted: 10 September 2019
Published: October 2019

Key words:-
Big data, Naive Bayes, Support Vector Machine, Decision Tree Classifier, Random Forest, K-Nearest Neighbor.

Abstract

In present, breast cancer in women is most is the prominently discovered life-threatening cancer in women and took over too many life's of women all around the world. This project deals with the Breast Cancer Wisconsin dataset to compare the accuracy of various machine learning algorithm in predicting the breast cancer in women. The various classification model are built and trained with the Wisconsin dataset. The different classifiers that are used to construct the model are Naive Bayes, Support Vector Machine, Regression Tree, Random Forest and K-Nearest Neighbor. The efficient working of these model are assessed by estimating the accuracy score of each model with the usage of unstandardized and standardized dataset. Once the performance of the model are evaluated, the optimal working algorithm are used to identify the type of breast cancer in the patient entry.

Introduction:-

Breast cancer is the prominent disease in women, which took the life of too many women at present. After skin tumor, breast cancer is the most general tumor diagnosed in women in India. Each woman is at the threat for breast cancer,

When she is in the early 80's, It is said that It is said that one in sixth chance (10%) to get affected by breast cancer at least once in their lifespan. In 2010, breast tumor was graded as the seventh leading reason of loss in the (KSA), Pervious that, in 2008, it was estimated that there was 1,408 fresh breast tumor cases, representing 20% of register fresh cancer cases within the Saudi women. Breast cancer could occur in both man and women, but it often occur in women's. Now-a-days, the advanced treatment and the early prediction has made diagnosis easier.

In 2019, a prediction of 268,400 of new occurrence of breast tumor disease is anticipated in women in the U.S. It is also important that fatness , delay at child birth ,young age of puberty, and an insanitary routine; and environmental ,ethnic and tribal character are the main threat factors that's adding to the cause of breast disease. Early investigation indicate that the reason of this syndrome are highly fierceness, poor clinical treatment, research reported to the advanced phase of breast tumor infection is common in the women who are in their early 45's , rather in elder women with in the age of 65 years . This project deals with the Breast Cancer Wisconsin dataset to compare the

Corresponding Author:-Aurobind Ganesh.
Address:-M.Tech, MLIS, Senior System Analyst, NIMHANS Digital Academy, Dept. of Psychiatry, NIMHANS, Bengaluru, INDIA.

accuracy of various machine learning/algorithm in predicting the breast cancer in women. The various classification model are built and trained with the Wisconsin dataset. The different classifiers are used to predict the breast cancer are Naive Bayes, Support Vector Machine, Regression Tree, Random Forest and K-Nearest Neighbor.

This paper is organized in the following manner, initially the list of various classifiers and their score in classifying the type of breast cancer is done. The standardization technique is included with these classifiers and the change in the accuracy score are evaluated. The best performing classifiers are used to produce the output to the patient entry. Then the problem statement is discussed. In Section IV, the proposed system are discussed and the architecture diagram is discussed and explanations are given and then we have discussed about the various classifier's algorithm and its accuracy level and to end with section V, we have concluded the paper along with the classification technique and its accuracy. The major reason behind this study is to evaluate the performance of various classifiers in classifying the type of breast cancer in women.

Related works

This segment presents overall review and planning on categorization of tumors in the lungs. CAD tool is used for mechanical recognition of lung module, which helps radiologist to analysis the disease. Here we use different classification technique to get the result more accurate. Among all the classifiers algorithm CNN gives more accurate result.

Our base paper is evaluate the performance of the three main classifiers by comparing the accuracy of the three models. To increase the efficient performance of the system they have used a technique in preprocessing like feature selection. The Particle Swarm Optimization (PSO) algorithm is used for the feature selection process and compared the performance between include and not include feature selection algorithm. Paper concludes to with PSO and without using PSO the naïve bayes establish a better performance [1].

In Genetically Optimized Neural Network(GONN) the prediction of the breast cancer in women is done and with the GONN the classification problems were sorted out. This system involves in the optimizing the performance of the model by constructing the number of hidden layer to the GONN algorithm. The system concentrates in the classification of breast cancer whether the patient is affected with malignant or benign type of breast cancer. To demonstrate the results, they had taken the Women breast cancer dataset from the UCI public repository and assessed the efficient working of each algorithm on their performance basis. In this paper the algorithm shows . accuracy 98.24%, and 99.63% of sensitivity. These results are obtained by providing 50-50, 60-40, 80-20, 90-10 ratios of test and training dataset. The paper concluded by declaring that with GONN the system shows better performance in classification. [2].

The machine learning algorithms are used for the prediction of breast cancer and diabetes disease in human. classification technique are used to classify the type of breast cancer and diabetes, for the working of the model they have used WEKA tool to build and run the system. It is said that the WEKA tool give the performance percentage of 74.28% [10].

The Breast Cancer Detection is done in this paper using the machine learning algorithm, where the system uses the feature selection technique that can play a dramatic change in the accuracy of the system in this build model. The accuracy of the machine learning algorithm are increased by embed the preprocessing technique to the systems, the technique like feature selection are used where the attributes that influence the output are alone given as an input to the constructed model. The paper discussed the use of deep CNN which help in detection of the breast cancer in women in more efficient manner. In the deep CNN they have segregated into two stages like mass detection and mass diagnosis. The mass detection deals with detecting the BC in patient and mass diagnosis helps in the further steps to be taken to cure the disease. [11].

Proposed System

Our project is concentrating on analyzing the text format reports of breast cancer, to compare the accuracy of the various classifiers. At first the data are preprocessed and the accuracy of various model are compared. The models used in our project are Support Vector Machine, Naive Bayes, Decision Tree, Random Forest and K-Nearest Neighbor. The performance of these model are evaluated by comparing the accuracy score of each model with the usage of unstandardized and standardized dataset. Once the performance of the model are evaluated, the optimal working algorithm are worn to classify the breast cancer type in the patient entry.

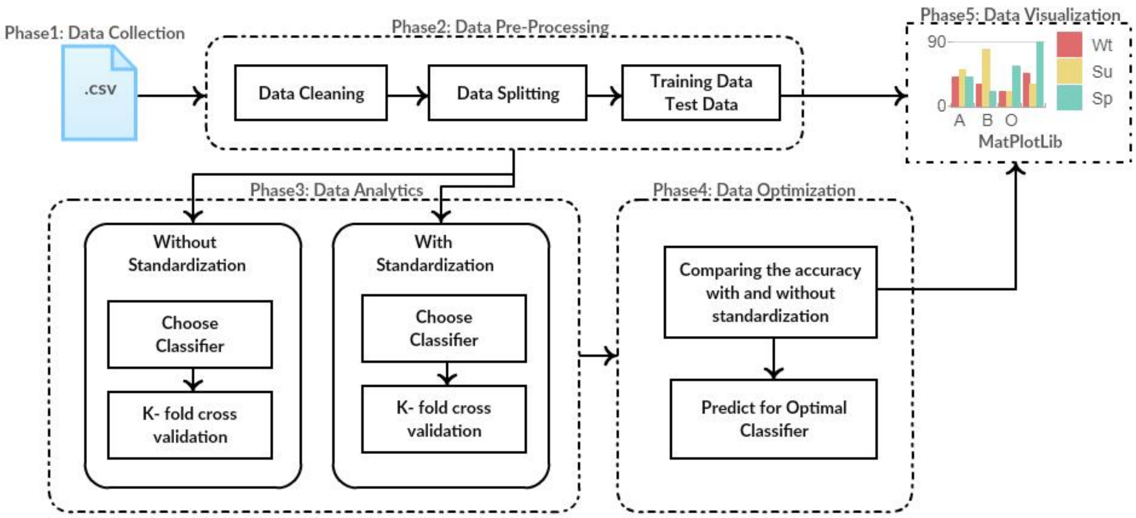


Fig 1:-System Architecture

Modules

Data Collection

We have collected from the public UCI repository that consist of the Breast Cancer Wisconsin Dataset. The Breast Cancer Dataset is presented with 569 instances and 33 attributes in it.

Data Pre-Processing

Data processing involves in the collection and manipulation of items of data to produce meaningful information. It is the process of converting the data into usable and desired form.

Data Cleaning

The data cleaning phase under goes the following:

1. The process of removing the null valued attributes in the dataset.
2. Replacing the null values with the related or the constant value.

Data Spliting

1. Once the data are cleaned, The data undergoes splitting process where the data are spilt into test and train data set. The train dataset are used to train the constructed module and the test dataset are used test the module for its correctness.
2. The half of the test dataset are used for validation where we can able to verify the output correctness with the help of these dataset.

Data Analytics

1. The various machine learning models are constructed and trained with k-fold sets of input for the efficient working of the system.
2. In this phase the model takes two types of dataset like the standardized dataset and the unstandardized dataset.
3. These standardization process done to eliminate the outliners present in the dataset which can increases the performance of the classifiers.

Data optimization

In this phase the optimal working algorithm is identified and the classification is done to desired patient entry. The optimal algorithm working algorithm are fed with the standardized scale input set for the best classification output of the patient entry.

Data Visualization

Data visualization involves in the analyzing and evaluating the dataset with the visual representation of data. It is used to communicate information clearly and efficiently, data visualization are used to identify the outliers present in the dataset. The performance of the different module are plotted in the graph to compare the accuracy of the different module for the different input dataset.

Algorithms

Support vector machine

Support Vector Machine (SVM) is used for the classification purpose where this algorithm use the discriminative classification. The algorithm works by separating the hyper plane in order to form the cluster of similar property data. Once the model is constructed they are given with the training input set that undergoes the hyper plane division and the test data are given for the evaluation of the constructed model and patient entry are given classified.

$f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} yici(\Psi\left(\vec{x_i}\right) \cdot \psi(\overline{x_j}))y_j c_j$

…………(1)

Where,

x - Input (x)

xi - Support vector

Navie Bayes

Bayesian classification comes under the supervised learning technique where the system is provided with the labeled dataset. The model works on the probabilities of the attribute occurrence of each attributes in the dataset. The outcome of the system is influenced by the probabilities of occurrence of the attribute in the dataset. Navie Bayes classifiers are categories on the probability of the attributes in the dataset, they produce the high efficiency when they are fed with the text format dataset. The classifier is capable of solving both the predictive and classification system models. Bayesian classifier assumes the probabilistic model which enables us to capture the uncertainty in the system constructed. The system exhibits a robust characteristic to the noisy input dataset.

$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ …………..(2)

Where,

$P(x|c)$ – Likelihood

$P(c) -$ class prior probability

$P(c|x) -$ Postirior probability

$P(x) -$ Predictor prior probability

Decision Tree

A decision tree is a tree that are constructed by the classification on the attributes, they form a tree like a structure and classify the attributes on their similar characteristics. The decision tree consist of the control and conditional statement which are used to classify the patient dataset, it is one of the powerful tool that is used for the classification and prediction. Decision tree construct a flow chart like structure where the internal nodes in the tree structure indicate the test in the attributes and the braches indicate the outcome of the each test performed in the node. Once the model is constructed they are fed with the test and training dataset input for the evaluation of the model performance for a single input and they are scored for their performance.

$H(s) = \sum_{c \in C} - P(c)\log_2 P(c)$ ……..(3)

Where,

S - Current data set for which entropy being calculated

C - Set of classes in S

P(c) – Proportion of the number of elements in class c to number of elements in set S

Random Forest

Random forests or random decision forests are an ensemble learning method which has the number of decision tree model constructed and the random set of input are given to these decision tree. The random forest corrects over fitting error that might occur in the decision tree when they are not provided with the depth value for the decision tree. Input to the model are provided to the system by dividing the given input randomly for the efficient working of the algorithm. These algorithm can be used for the both prediction and regression systems.

$H(s) = \sum_{c \in C} - P(c)\log_2 P(c)$ ………..(4)

K-Nearest Neighbor

K-nearest neighbors (KNN) comes under the supervised learning technique where the classifier use the clustering technique to classify the characteristics of the attributes. The K value determines the output of the model, the system works such a way that the characteristics of the k nearest neighbor of the patient input and hence decision that the patient displays the same character as its k neighbor. The patient entry are plotted and the distance between the various data's that are plotted and the output is determined by its k number of neighbor.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}2 \text{ ....(5)}$$

Where,

N – Number of nearest neighbor.

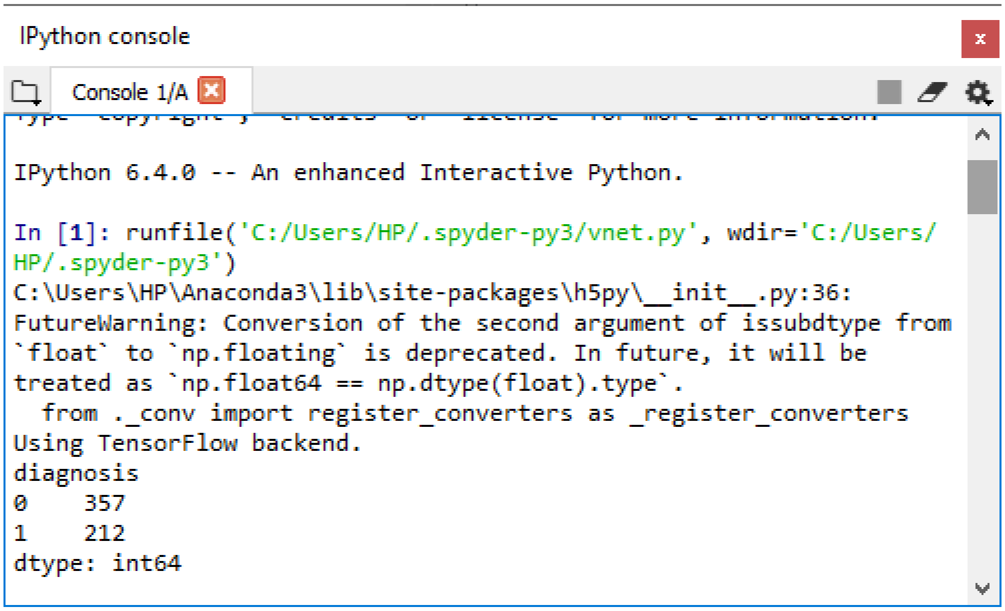
b, a – Measure the hamming distance

Implementation

The system can be implemented through downloading the anaconda environment, which is to be installed to run the python platform in it. Once the installation of the environment are done the spyder IDE are installed through the anaconda environment. The modules like scikitlearn, matplotLib, pandas and cmode are installed for the construction of the machine learning model. The output are visualized through matplotLib.

Snapshots

The data's from the csv file are extracted in the very first stage and the shape of the data set are identified. The unnamed or empty attribute set are deleted and the string or the character values in the dataset are changed to a 0 or 1, as you could see in the output Fig.2, that shows the number of benign[0] and malignant[1] type of cancer patients count are displayed.



```
Type "copyright", "credits" or "license" for more information.

IPython 6.4.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/HP/.spyder-py3/vnet.py', wdir='C:/Users/
HP/.spyder-py3')
C:\Users\HP\Anaconda3\lib\site-packages\h5py\__init__.py:36:
FutureWarning: Conversion of the second argument of issubdtype from
`float` to `np.floating` is deprecated. In future, it will be
treated as `np.float64 == np.dtype(float).type`.
  from ._conv import register_converters as _register_converters
Using TensorFlow backend.
diagnosis
0    357
1    212
dtype: int64
```

Fig 2:-Attributes Transformation

The density of the each attribute are plotted in the graphical format as shown in Fig.3, and the correlation map is also plotted in this system, these map help us to identify the relation and the influence of one and other attributes.

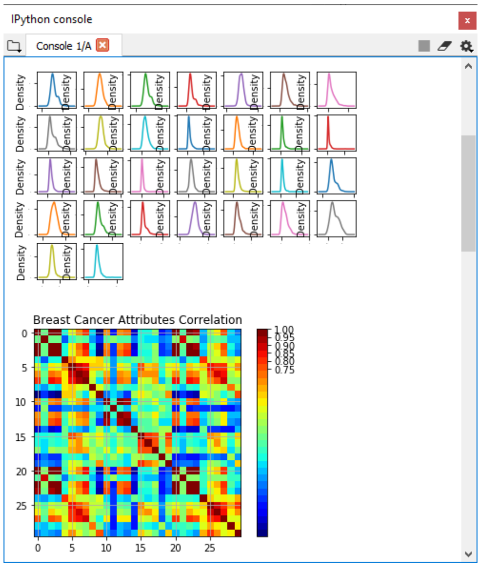
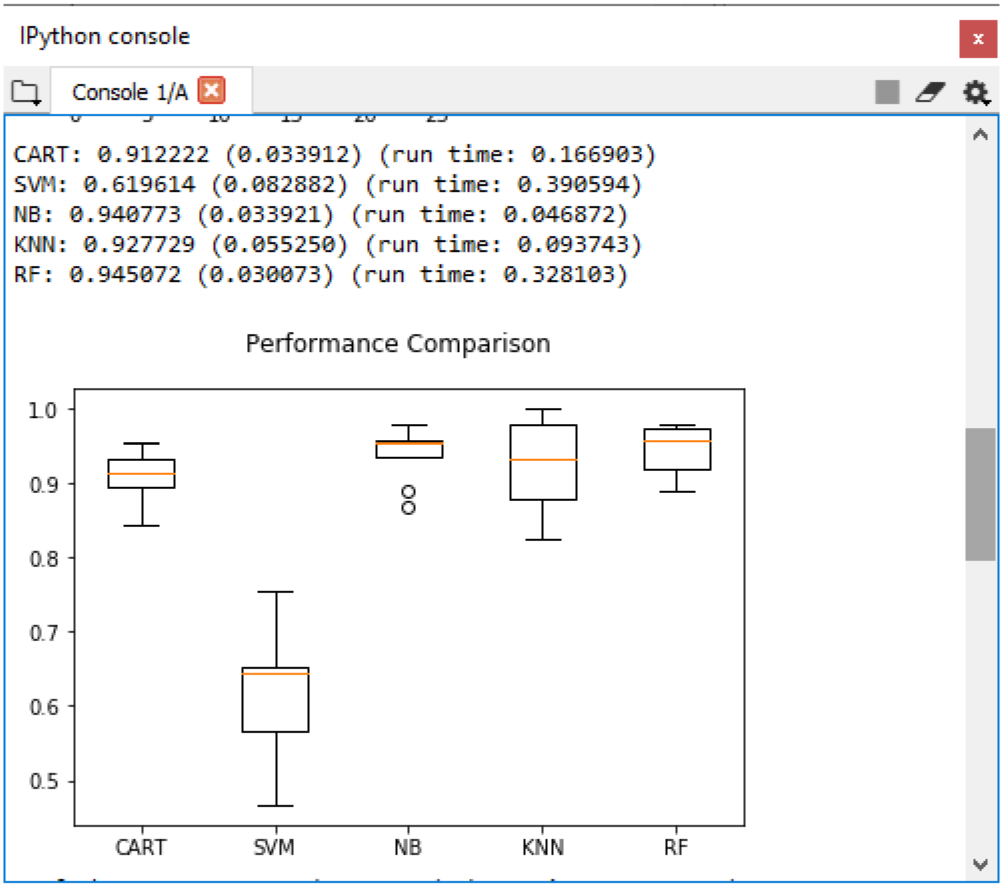


Fig.3:-Density and Correlation of attributes

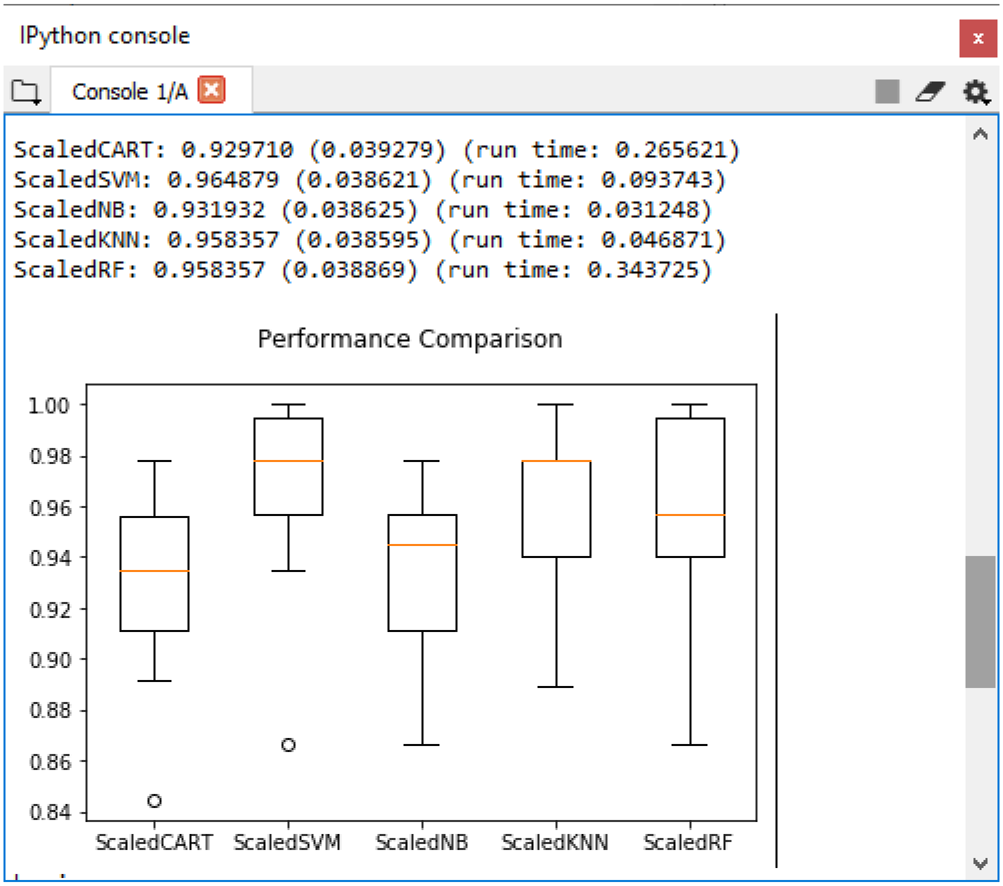
Once the data are cleaned and pre-processed they are fed into the five constructed modules namely Decision Tree, Support Vector Machine, k-Nearest Neighbor and Random Forest. The performance of these modules are stored in the result list and are displayed as shown in the Fig.4. The score of these modules are plotted as a graph and displayed.



```
CART: 0.912222 (0.033912) (run time: 0.166903)
SVM: 0.619614 (0.082882) (run time: 0.390594)
NB: 0.940773 (0.033921) (run time: 0.046872)
KNN: 0.927729 (0.055250) (run time: 0.093743)
RF: 0.945072 (0.030073) (run time: 0.328103)
```

Fig 4:-Performance comparison of the classifier before standardization

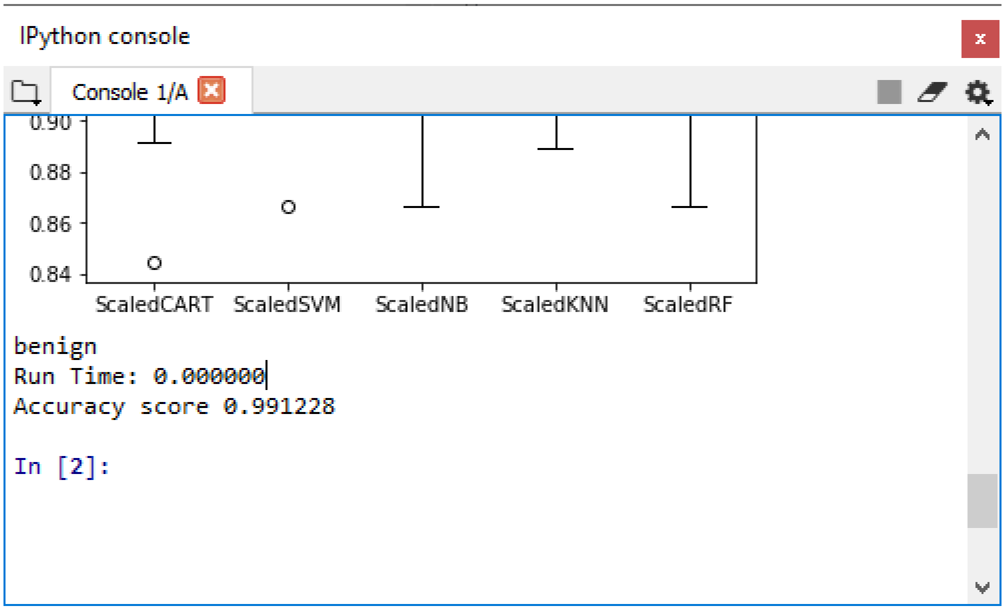
The performance of each classifier are identified and viewed, Now the performance of these classifier are improved by using the standardization technique to it. The process of standardizing the dataset help us to avoid analyzing the outlier in the module, with this the performance of the each algorithm can improve in a dramatic way. The algorithms performance score after implementing these modules using the standardized dataset are given in the Fig.5.



```
ScaledCART: 0.929710 (0.039279) (run time: 0.265621)
ScaledSVM: 0.964879 (0.038621) (run time: 0.093743)
ScaledNB: 0.931932 (0.038625) (run time: 0.031248)
ScaledKNN: 0.958357 (0.038595) (run time: 0.046871)
ScaledRF: 0.958357 (0.038869) (run time: 0.343725)
```

Fig. 5:-Performance comparison of the classifier after standardization

The performance of the classifier are compared by giving two set of dataset as an input. Now the optimal working algorithm out of all the five is Support Vector Machine is identified and the standard scalar function is applied to this algorithm to further enhance the performance of the algorithm. As you could see in the Fig.6, where the accuracy score of the algorithm is improved from 96.4% to 99.1%. The SVM algorithm is now used to classify the patient input data set whether the patient is affected with benign or malignant type of cancer.



```
benign
Run Time: 0.000000
Accuracy score 0.991228

In [2]:
```

Fig 6:-Classifying patient data

Conclusion:-

In this paper, the various classifiers are used in the classification of breast cancer and the standardization technique is used to increase the performance of the constructed model using the scalar function in it. We have presented the five models accuracy score and the influence of standardization in the improvement of these classifiers accuracy. Finally the patent entry are given and the type of breast cancer is identified using the optimal performing classifier. The support vector machine has the higher efficiency in classifying the type of breast cancer

References:-

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
4. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.
5. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” IEEE Trans. Antennas Propagat., to be published.
6. J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” IEEE J. Quantum Electron., submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),” IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
9. M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
10. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
11. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
12. (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given Available: http://www.(URL).