ISSN 2347 - 3983

Volume 8. No. 9, September 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter191892020.pdf

https://doi.org/10.30534/ijeter/2020/191892020

Breast Cancer Prediction using Classification Techniques

Dr. R.Vijaya Kumar Reddy[1], Dr. Shaik Subhani[2], Dr. G. Rajesh Chandra[3], Dr. B. Srinivasa Rao[4]

[1]Assistant Professor, Department of IT, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, A.P, India. Vijayakumarr285@gmail.com

[2]Associate Professor, Department of IT, Sreenidhi Institute of Science and Technology, (Autonomous),Hyderabad. shaiksubhani@sreenidhi.edu.in.

[3]Professor & HOD, Department of Computer Science and Engineering, SVR Engineering College, Nandyal, A.P, India. grajeshchandra@gmail.com

[4]Professor & HOD, Department of IT, Lakireddy Bali Reddy College of Engineering, Mylavaram, A.P, India. buragasrinivasarao@gmail.com.

ABSTRACT

Now a day's most of woman affected Breast at life of different stages. This disease affected rate reduced year by year using different procedures of medical treatment. Among these processes, early identification of breast cancer generates good results. Classification methods can support to decrease negative decisions. Novel techniques such as like knowledge discovery in database have become well-liked research tool from medical research. With the help of these techniques extract the patterns and find out relationship between large numbers of objects. Based on historical datasets of corresponding cases stored in data bases, these results compare with fresh outcomes. In this paper, implement a new model based on classification techniques for analyzing breast cancer data. The statistical results show the effectiveness of techniques and compare these techniques based on accuracy, sensitivity, and specificity.

Key words: KDD, LR, SVM, KNN, Breast Cancer, Prediction Analysis

1. INTRODUCTION

After lung cancer, major reason for women's death is breast cancer. According to Indian Council for Medical Research of new cancer cases is to be 14.5 lakhs in 2016. Every year this number increased up to 17.3 lakhs in 2020. Different organizations play crucial role for care of breast cancer in India for reduced the number of patients. The technical support gives boost to predict the breast cancer in early stages. Different types of technology like Data Mining, Big Data, Data Analytics, Machine Learning and Artificial intelligence. In this paper major concentrate on three classification techniques and approach is shown in Figure1. It also reduces the cost and time of prediction of results. Based on this technology to enhance the healthcare system in presence of quality and reducing the cost of treatment. In real-time scenario, front line warriors making strategic decisions to save people's lives. Classification is most essential technology in different fields for retrieve the accurate results in less time.

Most of the researchers used machine learning techniques in different areas in different aspects.

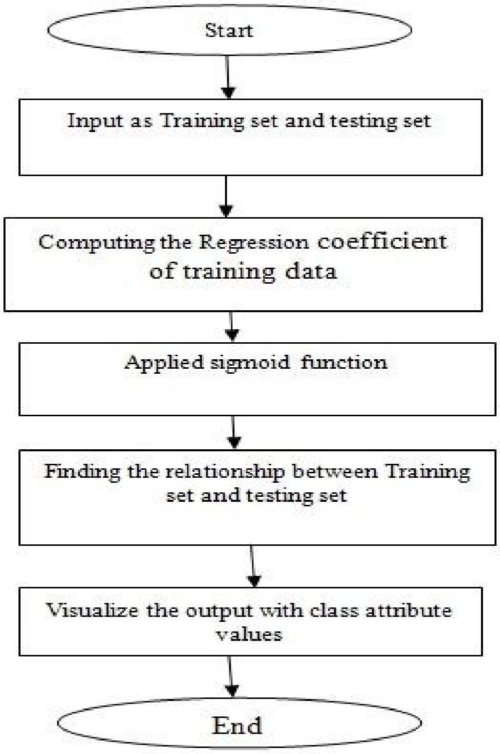

Start
Input as Training set and testing set
Computing the Regression coefficient of training data
Applied sigmoid function
Finding the relationship between Training set and testing set
Visualize the output with class attribute values
End

Figure 1: Diagrammatic representation of our proposed system

We shall explain Dataset description in Section 2, the methodology in Section 3. Visualization of results in section 4, and finally conclusion in last section.

2. DATASET DESCRIPTION

In this dataset description, the purpose of diagnosis we are using a FNA technique. It is a simple technique giving best performance in rapid way. It deletes cells from breast lesion with a fine needle. In our research used Wisconsin Dataset, it is obtained from Kaggle website [6]. Actually this data set was developed by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. At the time of dataset creation used uid samples, it was received from breast cancer patients to use graphical computer program called Xcyt. The Xcyt using for analysis of cytological attributes based on a digital scan. To compute ten features from each cell in the sample collected from patients; in this process calculate the mean, extreme value and standard error of every feature for the image. The following algorithm developed for this process.

Step 1: ID: 2, Diagnosis, M = malignant, B = benign, 3-32)
Step 2: To compute ten features from each cell in the sample.
Step 3: Calculate the mean
Step 4: SD of gray scale value.
Step 5: Calculate the perimeter
Step 6: Calculate the area
Step 7: Observe the variation in mean
Step 8: perimeter^2 / area - 1.0
Step 9: Calculate the concavity
Step 10: Number of concave portions of the contour
Step 11: Symmetry of cells
Step 12: Finally fractal dimension

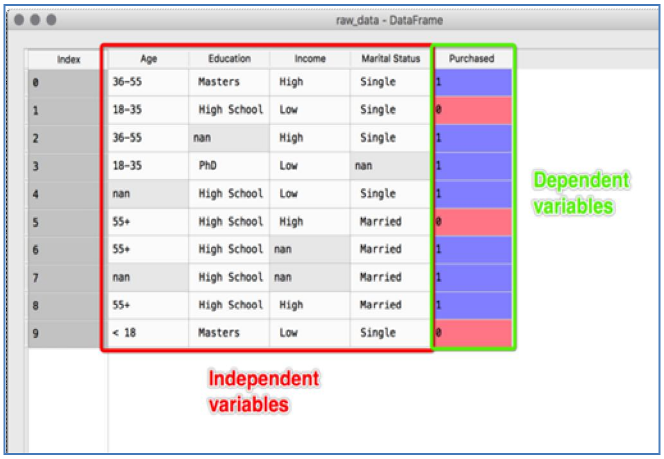
The mean, standard error and “worst” or largest of these features were calculated for every image, it generates 30 features from image. This analysis helpful for select the best features for cancer model.

2.1 Loading the dataset

The **Python** API provides for CSV model and method reader () support for loading CSV files. After that convert the CSV data to a NumPy array for machine learning. Here, we used read_csv method of pandas to import the dataset. Before importing the dataset, check the current working directory and set it to the directory where the data is available.

2.2 Dividing the dataset

The dataset is divided into two parts based on types of variables, dependent and independent variables. All the 30 independent variables are set into X and the dependent variable diagnosis is set to Y data is shown in Figure 2. This is done using iloc method of pandas; it divides the dataset based on indices.



Index	Age	Education	Income	Marital Status	Purchased
0	36-55	Masters	High	Single	1
1	18-35	High School	Low	Single	0
2	36-55	nan	High	Single	1
3	18-35	PhD	Low	nan	1
4	nan	High School	Low	Single	1
5	55+	High School	High	Married	0
6	55+	High School	nan	Married	1
7	nan	High School	nan	Married	1
8	55+	High School	High	Married	1
9	< 18	Masters	Low	Single	0

Figure 2: Segmentation of datasets

2.3 Dealing with missing values

In real time databases, most of the data having missing values, this type of data not comfortable for analysis in research. Different types of strategies available for filling the missing values in manual or automation (filtering data). We follow the process of Mean, Median and Mode for purification.

2.4 Encoding Categorical data

Categorical variables are often called nominal type of data. Most of the machine learning algorithms is not access label data directly. It accepts numeric values as input and output. So it is mandatory to convert categorical data into numeric format. Then only to use this data into different types of application in real time.

2.5 Splitting the Dataset

A machine learning algorithm works in two stages-the testing and training stage. The training dataset is the initial dataset used to train an algorithm to understand how to apply technologies like neural networks. It includes both input data and the corresponding expected output. The purpose of the training dataset is to provide your algorithm with “ground truth” data. The test dataset is used to assess train dataset in algorithm. You can’t simply reuse the training dataset in the testing stage because the algorithm will already “know” the expected output, which defeats the purpose of testing the algorithm shown in Figure 3.

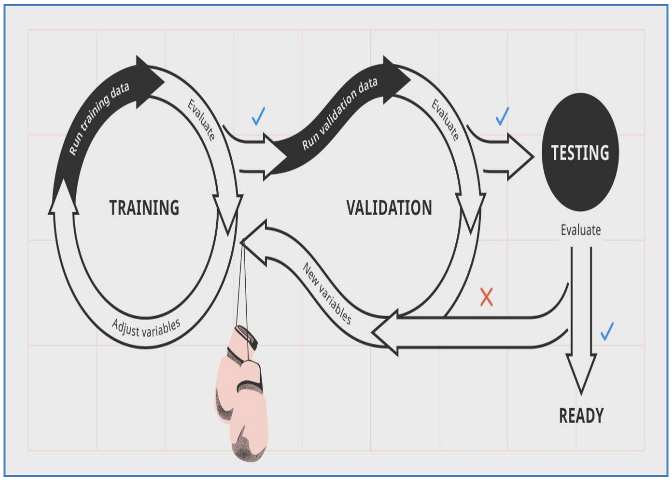


Figure 3: Training and validation of Data

2.6 Feature Scaling

Feature scaling is known for data pre processing and also called normalization. In normalization, most of classifiers work out the distance between two points by using Euclidean distance. Based on this formula easily identify the outliers. Feature standardization used for identify the value of every feature based on zero mean and unit variance. This process is frequently used in machine learning algorithms. The following equation is general method for calculating the feature value.

X_{stand}=mean(x)/standard deviation(x) (1)

In our model we have used Standard Scaler class sklearn. The Pre-processing module for normalize the range of the self-governing variables. This is the last step of data preprocessing. Once we are done with it, we can move further.

2.7 Dimensionality Reduction

In machine learning, dimensionality reduction is the process of reducing the number of random variables based on different techniques. Principal component Analysis and transform techniques are used for dimensionality reduction process. This process divided into feature selection and extraction [8][10].

3. METHODOLOGY

In methodology, accuracy is most crucial work to success the process with good outcome. Naturally accurate results generated methods avoid over fitting and cut the processing time also. Training different models is great understanding between machine learning algorithms. The training time close to accuracy of final results. Few machine learning algorithms more sensitive in presence of data points than other algorithms. The training time also based on size of dataset, it is limited to selection of machine learning algorithm.

Linearity followed by most of machine learning algorithms. In linear classification data points can be separated by straight line. For separation of data points, we can use different types of algorithms like logistic regression, support vector machine and linear regression are shown in Figure 4.

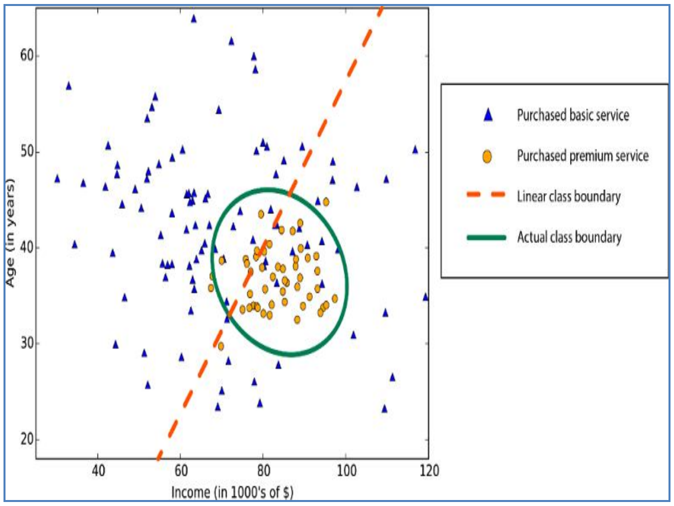


Figure 4: linear regression followed by straight line representation

A linear classification algorithm generates result in low accuracy in some times. It based on different size of datasets, type of data and selection of algorithm.

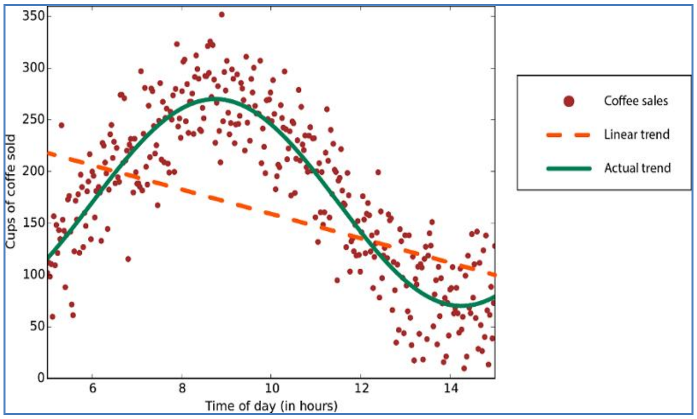


Figure 5: Time of Delay representation

Non linear regression suitable for some times, it based on different size of datasets, type of data. Most of the researchers prefer the linear regression because simple and fast training data. Number of parameters changes the behavior of algorithm. It affects the error tolerance and time. The large numbers of parameters require the more trial and error for finding the good combination data to algorithm is mention in Figure 5.

3.1 Principle Component Analysis

The principle component analysis using for classification and dimensionality reduction and it mapping linearly for low dimensional space. It finds out sum of variances of all personage principal components. This technique compare between the total variance with principle components. This classification is unsupervised technique to find meaningful data based on covariance matrix [8].

3.2 Confusion Matrix

This matrix used for measure the throughput of a machine learning algorithm, it is a supervised learning [7]. In this matrix row represents the actual class instance and column represents predicted class instance. It finds the confusion occurs in our algorithm. Labeling of a class not create confuse but mislabeling create the confuse, then confusion matrix incremented by one.

3.3 Logistic Regression

Logistic regression also using for classification, it borrowed by machine learning from the domain of statistics. This algorithm, falls under Supervised Machine Learning. It solves the problems of Classification. It is used to predict binary outcomes for a given set of self-governing variables. The dependent variable's outcome is discrete [9].

3.4 Support Vector Machine

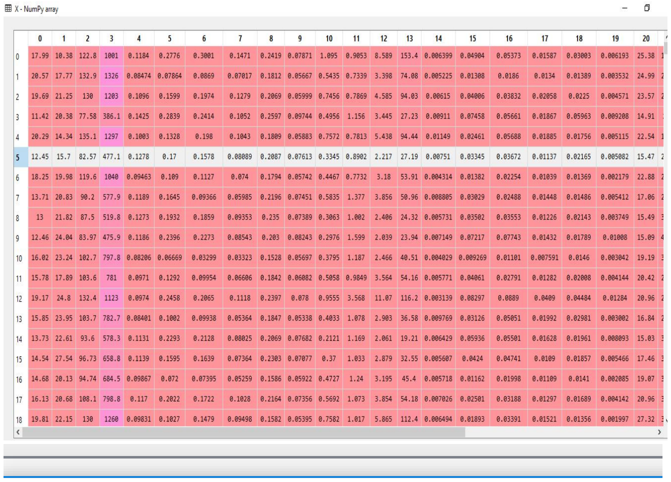
In machine learning domain, SVM are supervised learning models that analyze data used for classification and regression analysis [7][11]. These training algorithms construct a model that assign another category. This model as points in space, mapped to the separate categories is divided by a clearly.

3.5 k-Nearest Neighbor

The *k*-nearest neighbor's algorithm is a non-parametric method used for regression and classification. The input consists of the *k* close to the training examples in the feature space. The outcome depends on *k*-NN used for classification or regression. It find the distance between the neighbors using Euclidean distance. If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor [9][12].

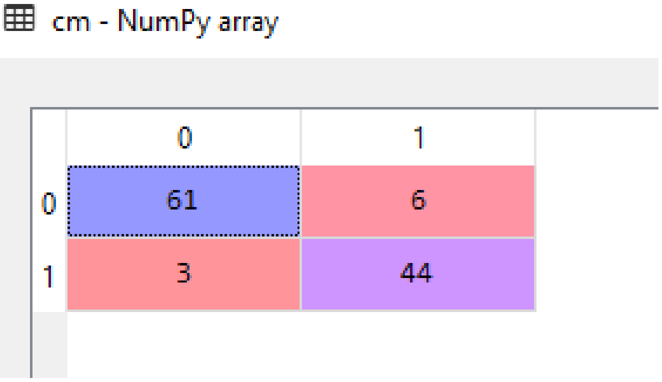
4. RESULTS AND ANALYSIS

The statistical results showing is important skill in machine learning. Statistics focus on quantitative description and estimation of data. It is easily understand by end users and helpful know a finding patterns, corrupt data and outlier detection, etc. data visualization can be used to display major relationship in plots and charts. It create mesh grid out of an array values of x and y. Each integer value between 0 and 4 in x and y directions. If you want to create rectangular grid, then pass the x and y data points. Classification Techniques and its results of breast cancer data is shown in Figure 6.



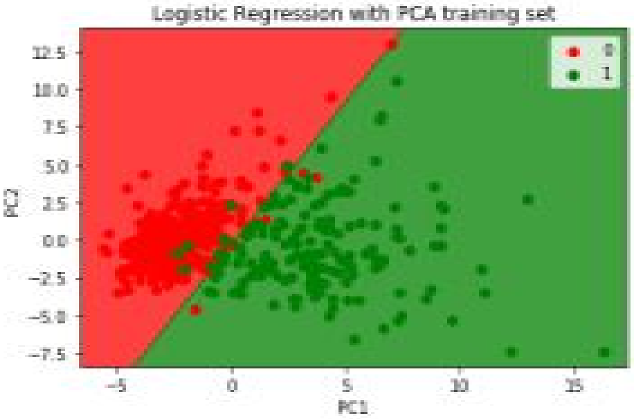
X - NumPy array

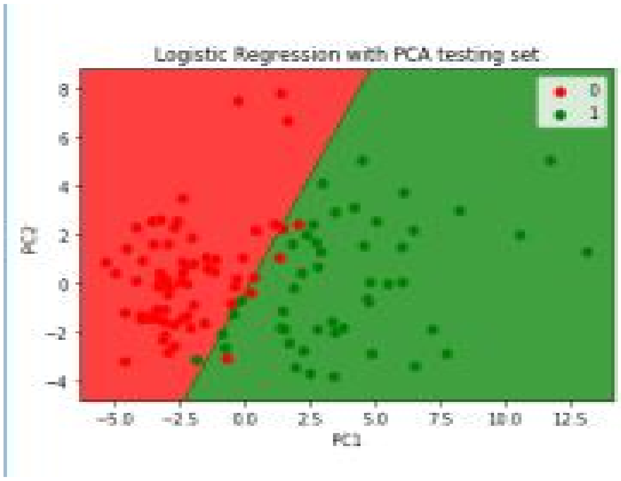
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38
1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99
2	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57
3	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91
4	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54
5	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47
6	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88
7	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06
8	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49
9	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09
10	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19
11	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42
12	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96
13	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84
14	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03
15	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46
16	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07
17	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96
18	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32

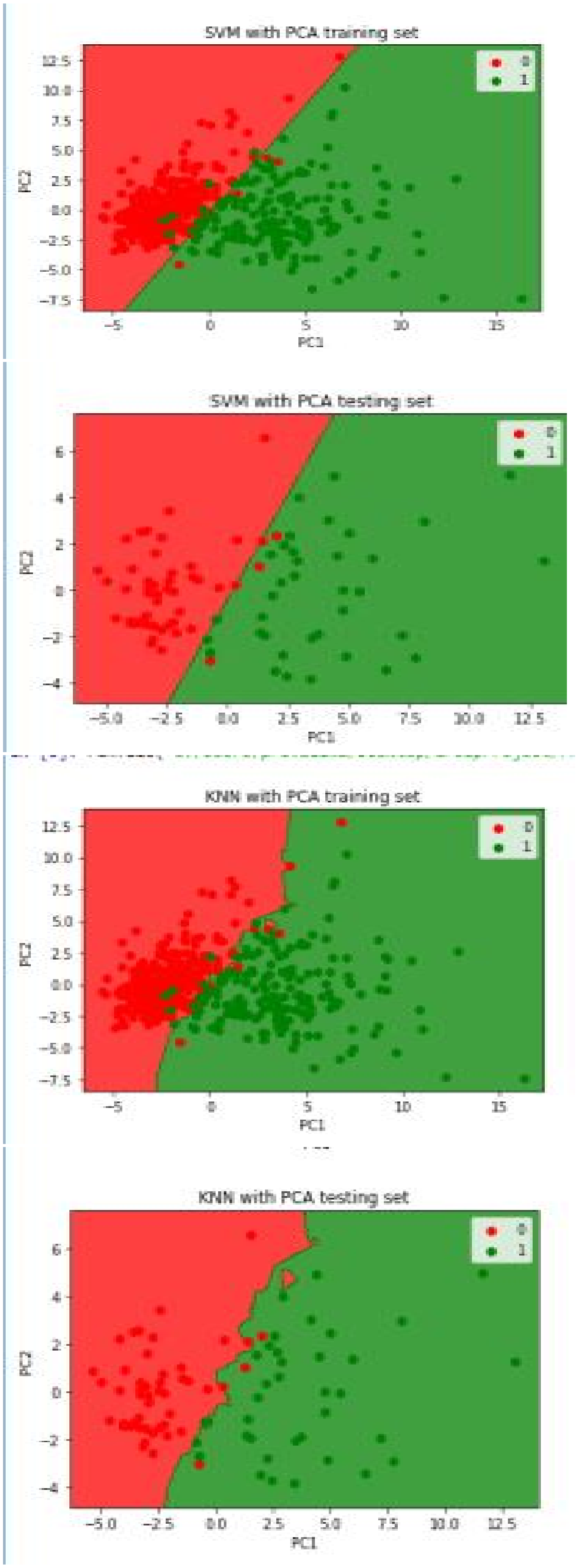


cm - NumPy array

	0	1
0	61	6
1	3	44







SVM with PCA training set

SVM with PCA testing set

KNN with PCA training set

KNN with PCA testing set

Figure 6: Classification Techniques of breast cancer data

5. CONCLUSION AND FUTURE SCOPE

The SVM (SMO) is used for analysis in this paper is only applicable when the number of class variable is binary. The SVM (SMO) is only applicable when the number of class variable is binary i.e. we can't have more than two classes. To solve this problem scientists have come up with multiclass SVM. Novel techniques such as like knowledge discovery in database have become well-liked research tool from medical research. With the help of these techniques extract the patterns and find out relationship between large numbers of objects. Based on historical datasets of corresponding cases stored in data bases, these results compare with fresh outcomes. In this paper, implement a new model based on classification techniques for analyzing breast cancer data. The statistical results show the effectiveness of techniques and compare these techniques based on accuracy, sensitivity, and specificity

REFERENCES

1. National Breast Cancer Foundation Inc.,http://www.nationalbreastcancer.org/about-breast-cancer.
2. T. Subashini, V. Ramalingam, and S. Palanivel,—**Breast mass classification based on cytological patterns using RBFNN and SVM**, Expert Systems with Applications, Voume.36,issue.3: p. 5284-5290,2009.
3. Ahmad LG*, Eshlaghy AT, Poorebrahimi A, EbrahimiM and Razavi AR **"Using Three Machine LearningTechniques for Predicting Breast Cancer Recurrence"** ,J Health Med Inform 2013,4:2,http://dx.doi.org/10.4172/2157-7420.1000124.
4. Xiaowei Songa, Arnold Mitnitskib,c, Jafna Coxb,Kenneth Rockwood – **"Comparison of Machine LearningTechniques with Classical Statistical Models in Predicting Health Outcomes"**, MEDINFO 2004 M. Fieschi et al. (Eds)Amsterdam: IOS Press © 2004 IMIA.
5. Tarigoppula V.S Sriram, M. Venkateswara Rao, G VSatya Narayana, DSVGK Kaladhar, T Pandu Ranga vital **"Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms",** International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3,pp.212-215 ,September 2013
6. https://www.kaggle.com/junkal/breast-cancer-prediction-using-machine-learning.
7. Ch. Shravya, Pravallika and Dr. Shaik Subhani, **"Prediction of Breast Cancer Using Supervised Machine Learning Techniques",** International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.
8. P. Santosh and Dr. Shaik Subhani, **"Heart disease prediction with PCA and SRP"**, International Journal of Engineering and Advanced Technology," Volume-8, Issue-4,pp.1279-1282, April 2019.

9. Shiva Keertan J and Dr. Shaik Subhani, **“Machine Learning Algorithms for Oil Price Prediction”**, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, pp.958-963, June 2019.
10. R.V Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha, P. Ravi Prakash, **“Prediction of Heart Disease Using Decision Tree Approach”**, IJARCSSE, Volume 6, Issue 3,pp.530-532, March 2016.
11. Swarna Kuchibhotla et al**.,"Autism Detection and Subgrouping using Machine Learning Algorithms"** International Journal of Emerging Trends in Engineering Research, Volume 7, No.11, pp.659-663,November 2019.
12. Anilkumar B and Dr.P.Rajesh Kumar **"Tumor Classification using Block wise fine tuning and Transfer learning of Deep Neural Network and KNN classifier on MR Brain Images"**, International Journal of Emerging Trends in Engineering Research, Volume 8, No.2, pp. 574- 583,February 2020.