

**Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"**

sponsored by

Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at **Department of English**,
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



Predict Whether Breast Cancer Is Benign Or Malignant Using Machine Learning

Radhika Sreedharan¹, Archana Praveen Kumar²
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, Karnataka, India, 576104

Abstract

Cancer occurred when changes called mutations had taken place in genes that regulate cell growth. The mutations let the cells divide and multiply in an uncontrolled, chaotic way. The cells kept multiplying, producing copies that got progressively more abnormal. In most cases, the cell copies eventually form a tumor. Breast cancer is cancer that develops in breast cells.

Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer also occurred in the fatty tissue or the fibrous connective tissue within breast. The uncontrolled cancer cells invaded other healthy breast tissue and could travel to the lymph nodes under the arms.

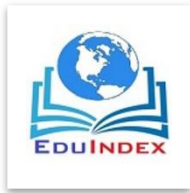
The lymph nodes were a primary pathway that helped the cancer cells to move to other parts of the body. Features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They described characteristics of the cell nuclei present in the image.

The main aim of this topic is to predict whether the cancer is benign or malignant using machine learning techniques. The prediction of survivability by treatment of breast cancer is a very challenging task.

Keywords: Breast cancer, survivability, machine learning

1. INTRODUCTION

Everyone in the world is curious to know about Cancer survival forecasting and medical care. It was designated by The World Health Organizations that the second noted source of death is cancer in the world. Surgery is one of the treatments for breast cancer. Chemotherapy, hormonal adjuvant therapy and radiation therapy are some of the treatments for Cancer. If the investigation of cancer is combined with early medical care in the beginning, then prognostications of cancer can be improved. Prognostication also counts on the cancer advancement, where cancer propagates to drainage areas of lymph node. TNM(Tumor,Node,Metastasis) analysis cancer staging system is limitless utilized for scrutinizing the cancer's position. Consequences of



**Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"**

sponsored by

Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at **Department of English,**
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



treatment procedures will differ depending upon cancer types and stages. It was noticed that breast cancer, bounded to only primary level, 96% of patients will survive in five years after investigation. The possibility of medical care and the high rate on occurrence made survival become a subject of much curiosity to health professionals and researchers. If the tumor is not risky to health and not deliberated as cancerous, it is called benign else it is malignant. Cells of benign tumor are almost normal to look; they develop gradually, do not infect near tissues or spread to other body parts. If tumors are cancerous, then they are malignant. The goal of this project is the application of several data mining and machine learning techniques to classify whether the tumor mass is benign or malignant in women residing in the state of Wisconsin, USA. This will help in understanding the important underlying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes. Through the understanding of nature of attributes in cancer prediction, the healthcare community can perform additional research corresponding to these attributes to help prevent pervasion of breast cancer into the population of USA. The scope of the project is confined only to prediction of breast cancer to be malignant or benign for the women of Wisconsin.

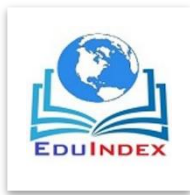
II.LITERATURE REVIEW

SEER(Surveillance, Epidemiology, and End Results) breast cancer set of data was investigated by Umesh and Ramachandra. They have encompassed an extraneous peculiarity "Menopause" which is not documented in the SEER databaset by assuming that a female more than 50 years of age signifies menopause. Misplaced value with mean value in their data pre-processing was compensated by Rathore et al.. Misplaced values of pertinent subjects were removed in a study from Boughorel. Boughorel et. al handled survivability forecasting as long as of 2,5, 8 and 11 yea

Their objective value for forecasting is treated as the double value and four partitioned study were handled. Jatalian et al reviews extent article on computer-aided discovering and problem identification for breast cancer by utilizing ultrasound and mammography sets of data. Devi et al. supported a survey of various kinds of ANNs used for cancer examination and forecasting. Civic et al. suggested utilization of mutli-stable biological neural networks in micro-coagulation discovery in initial investigation of breast cancer. Their process was used on MIAS (Mammogragraphic Image Analysis Society) database with the understanding of correctness, sensitivity and precision value.

III.METHODOLOGY

The technique of establishing computer algorithms that give computers the capability of acquiring information from scrutiny is known as machine learning. Adhoc procedures frequently separate evidences based on individual attributes such as gender of race, the projected path estimates companion with respect to every attributes. The implication of attribute merger is effortlessly unnoticed and hard to take by improvised procedures while the SOM method can take association of any precise mix of attributes. The



**Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"**

sponsored by

Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at **Department of English,**
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



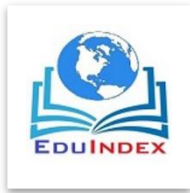
prospective techniques begin with a data pre-processing step which includes (i) a data-driven way to choosing records of patients and variables of data for analysis and (ii) balancing chosen variables of data in such a way by which every variables have a unit norm, which is suitable for distance evaluation.

Self-Organizing Maps(SOM): The SOM is set-up after pre-treatment approach. The data is mapped on to the low-dimensional array space by SOM. By that approach, it incorporates comparable fragments into thick groups. The SOM algorithm hence assists to model clusters although the algorithm cannot label or excerpt the clusters. Clustering process like k-means or DBSCAN are usually utilized to excerpt clusters from a SOM. Since we cannot assume that the clusters of patient cohorts are globular (round) in shape and hence we use DBSCAN with our approach.

DBSCAN: DBSCAN (Density-based spatial clustering of applications with noise) determines, excerpts inconsistent build clusters and needs little user-defined parameterization. These clusters are utilized to perform MLP based survivability forecasting investigation after they are recognized. Hence, MLP is utilized to check whether forecasting acts are revised if MLP-classifiers are experienced on separate clusters (or patient disciples). The logic of the way is that the establishment of models which are skilled in forecasting survival of patients that part correspondence this should enhance on general forecasting correctness when correlated to a distinct integrated classifier. We have adopted a new strategy of substituting lost values of a patient with the most persistent variable value in a patient cluster which are recognized by by SOM-DBSCAN.

K-nearest neighbor: K is number of nearest neighbors that the classifier will use to make its prediction. KNN creates forecastings on the basis of results of the K neighbors near to that point. The most well-known alternatives to measure this distance is Euclidean. To determine k, we made use of kNN algorithm to the training sets of data across different values of from 1 to 30 and based on the mis- classification rate choose the optimal value of k.

Random Forests: They are also called as random decision forests, are a well-known altogether process that can be utilized to frame forecasting models for issues of classification and regression. Ensemble methods utilize multiple learning models to achieve superior forecasting outcomes. The model develops complete forest of random uncorrelated decisions to achieve best response. The random Forest begins with a prevalent machine learning technique called a "decision tree". In a decision tree, an input is listed at the finest and as it goes over down the tree the data gets deposited into tinier and tinier sets. The random Forest proceeds this belief till later



Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"

sponsored by

Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at Department of English,
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



level by connecting trees with the idea of an ensemble. Therefore, the trees become weak learners and the random Forest becomes a strong learner.

Tuned random forest: The error rate for original random Forest Model is 4.04%. The process of calculating random forests' prediction error by making use of bootstrap aggregating to sub-sample data which can be utilized for training is called out of bag (OOB) error. OOB is the average forecasting error on every sample x of training, utilizing only those trees which didn't have x in their bootstrap sample. Sub-sampling grants one to describe an out-of-bag estimate of forecasting conduct enhancement by checking out prediction on those investigations which were not utilized in next base learner building. Out-of-bag estimates assist the requirement for self-sufficient proof data-set, however, usually underestimates real conduct enhancement and optimum number of redundancies.

Support Vector Machine: SVM replies upon supervised learning models and skilled by learning algorithms. An SVM creates parallel divisions by creating two parallel lines. It detaches the space in a single pass to generate flat and linear partitions. Divide the 2 categories by a clear gap that should be as wide as possible. Di-visioning can be done by hyper-plane.

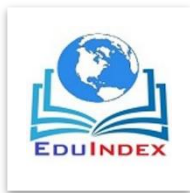
Hyper-planes which have the maximum limit in a high-dimensional space to isolate given data into classes are prepared by SVM. The limit among 2 classes produces high separation among those classes' adjacent data points.

Regression coefficient: The quantity which represents a change in subordinate variable which results from change in predicted value, when all variables are unvarying is known as regression coefficient. The regression coefficient between worst radius and mean perimeter and also between worst radius and worst area was computed and they were positive.

Decision tree algorithm: With the help of decision tree algorithm, by inputting certain parameters, it can predict whether breast cancer is benign or malignant based on given data set.

IV. RESULTS AND DISCUSSIONS

The coefficient correlation between radius_worst and perimeter_mean was positive. Also, the correlation between area_worst and radius_worst was positive. The sets of data attained from the



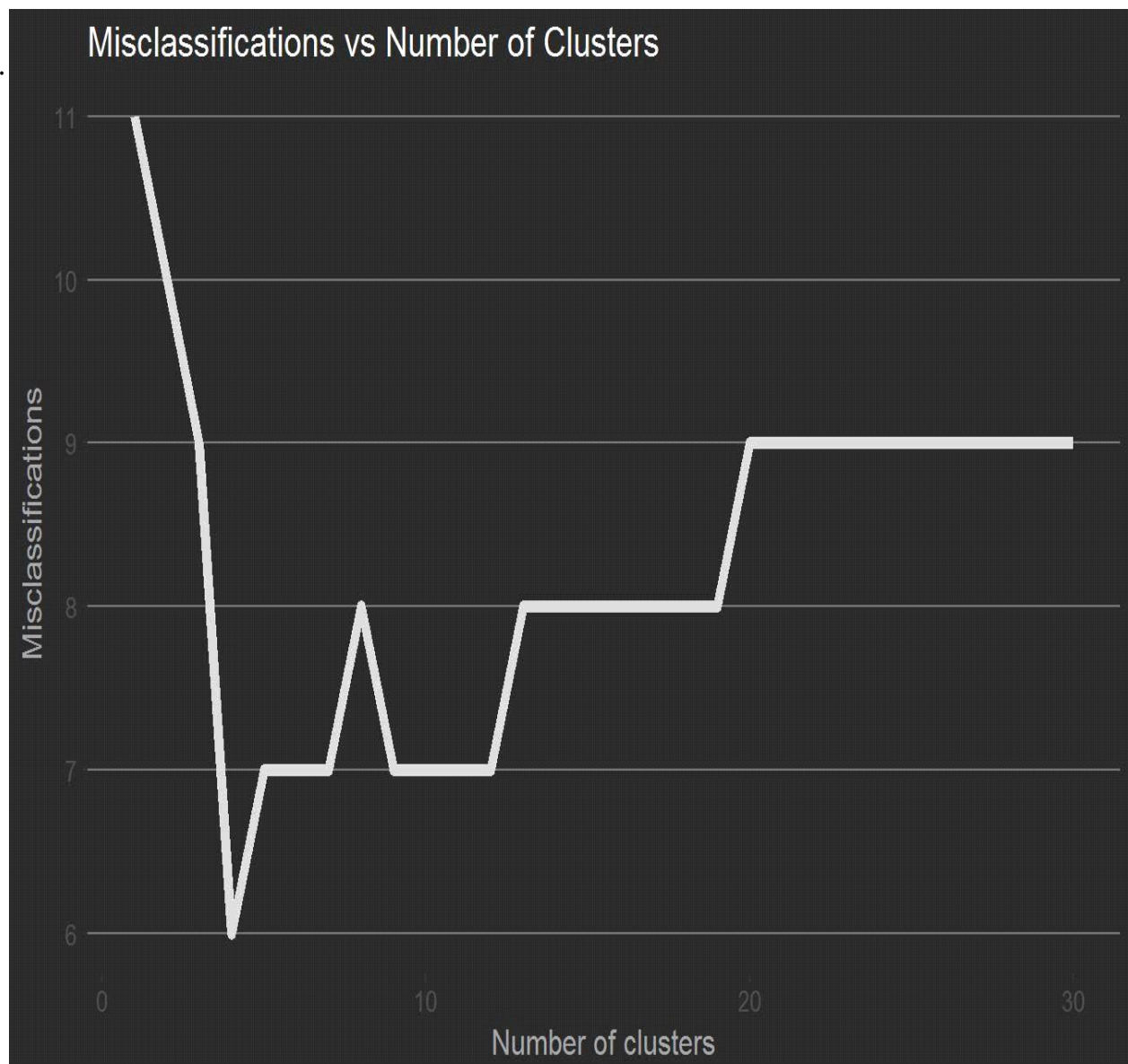
Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"

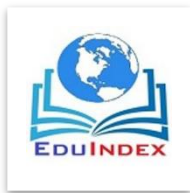
sponsored by

Indian Council of Social Science Research, New Delhi
 on 23rd and 24th September 2019 at **Department of English,**
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



step of pre-processing are utilized for self-organizing map (SOM) training. SOM is beneficial in current status as the variables of data may be non-linearly linked to each other. The mappings attained by a SOM are beneficial in visualizing complex relationship among data variables. The missing values in data set are saved and hidden by SOM amidst variables of data. TSOM hides and saves missing values of sets of data. Below plot displays the model prediction accuracy along with the corresponding value of k. The optimal value of K is 11 for k-Nearest Neighbor classifier and it gives an accuracy of 98.23% on test data





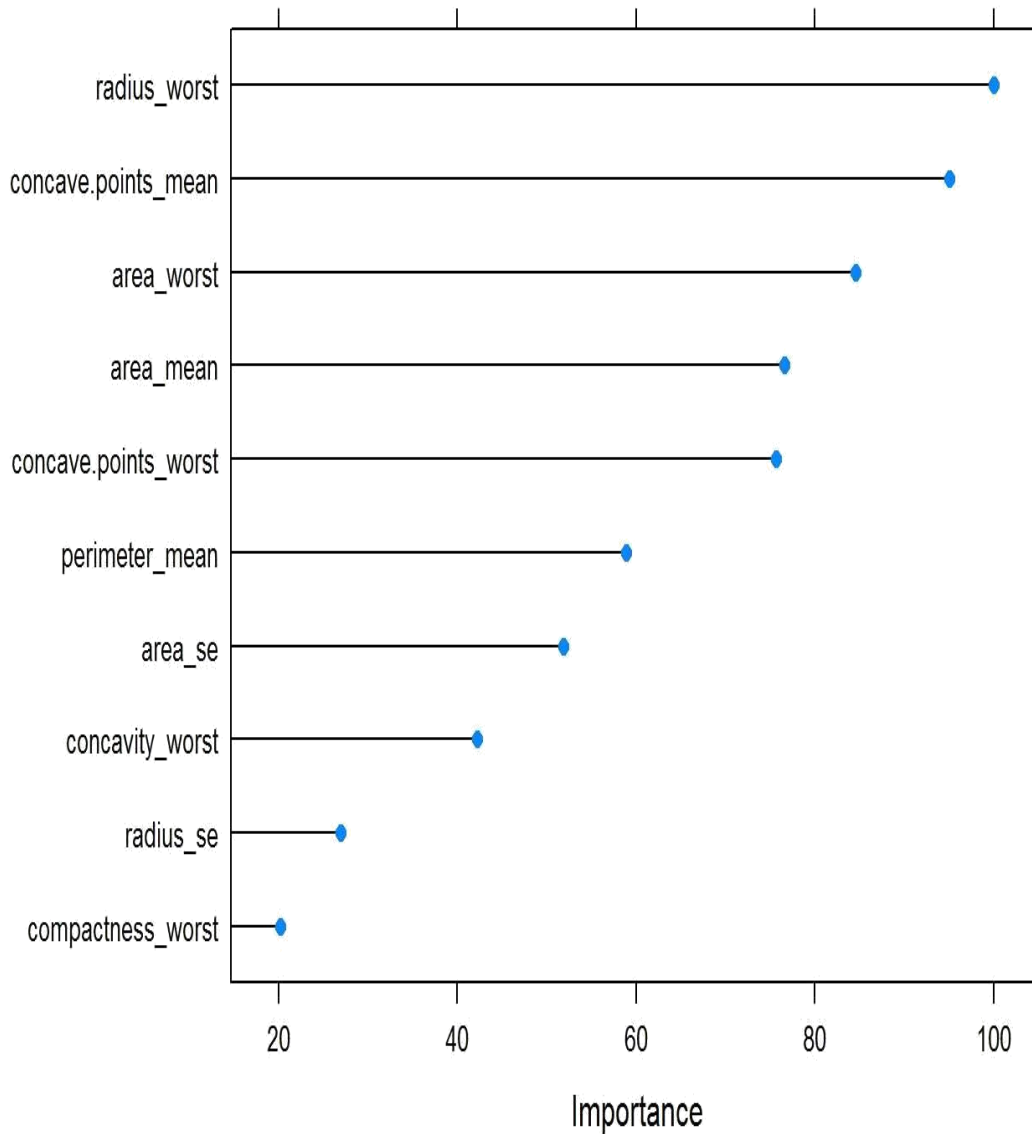
**Two-Day National seminar on
"Artificial Intelligence, Multimedia and ICT in Teaching,
Learning and Acquisition"**

sponsored by

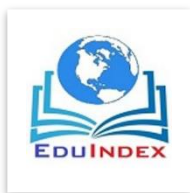
Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at **Department of English,**
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



Random forest

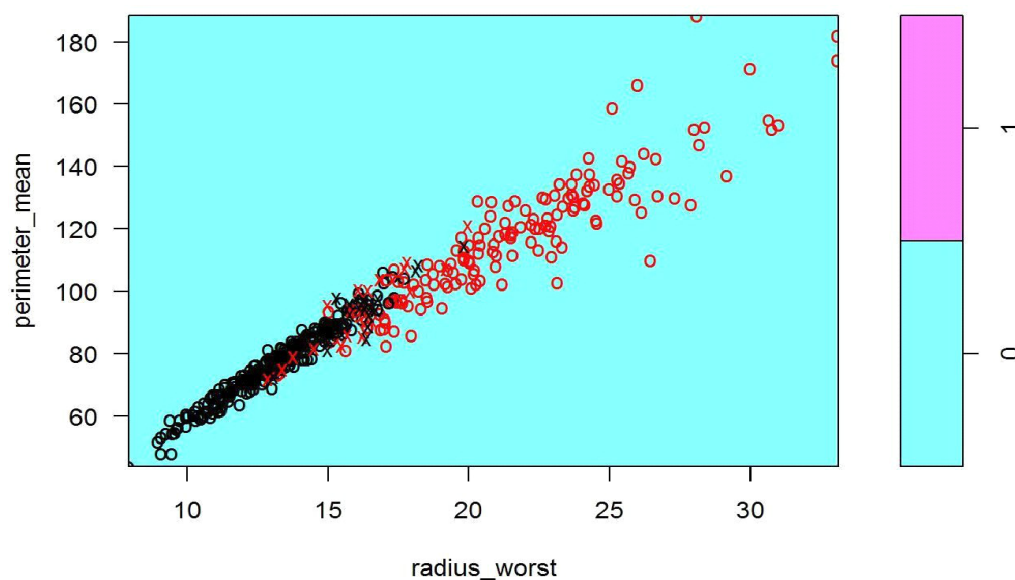


For the entire breast cancer data-set, we build a random Forest model till 500 decision trees.

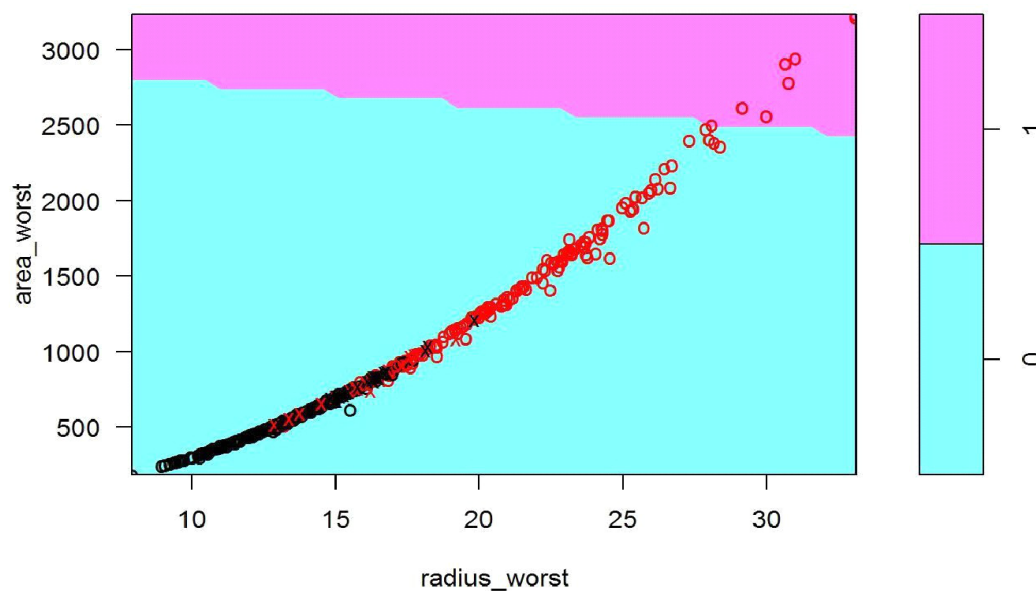


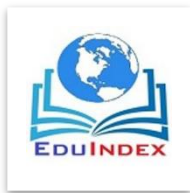
For random Forest model, error rate is 4.04%. After tuning the model based on random variable numbers utilized in each tree (mtree), the error rate decreases to 3.87%. Top5 predictor variables for classification according to variable importance of random Forest model are Radius worst, Concave points worst, Area worst, Perimeter worst and Concave points mean. Support Vector Machine model correctness for classification was 95.58%.

SVM classification plot

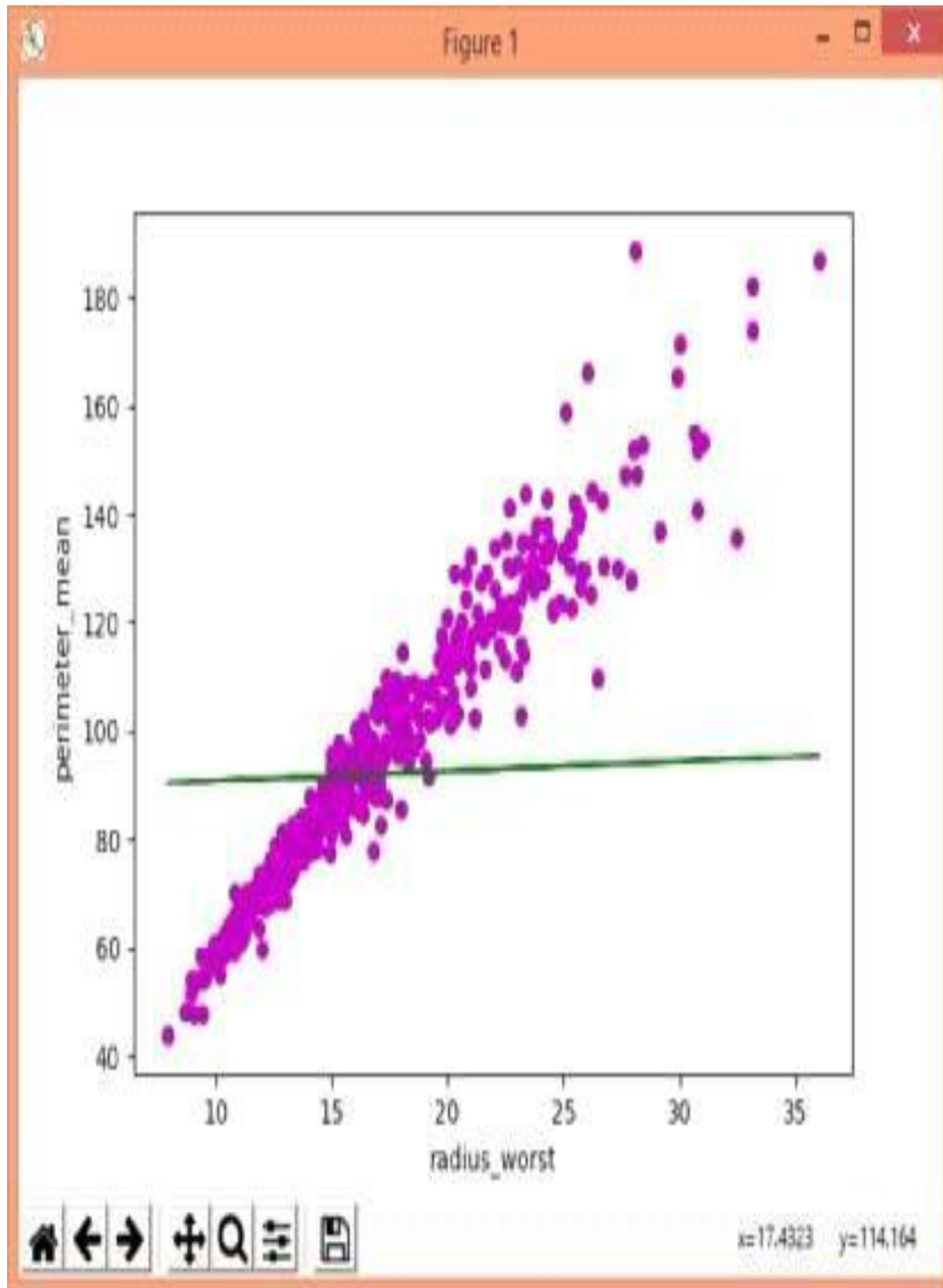


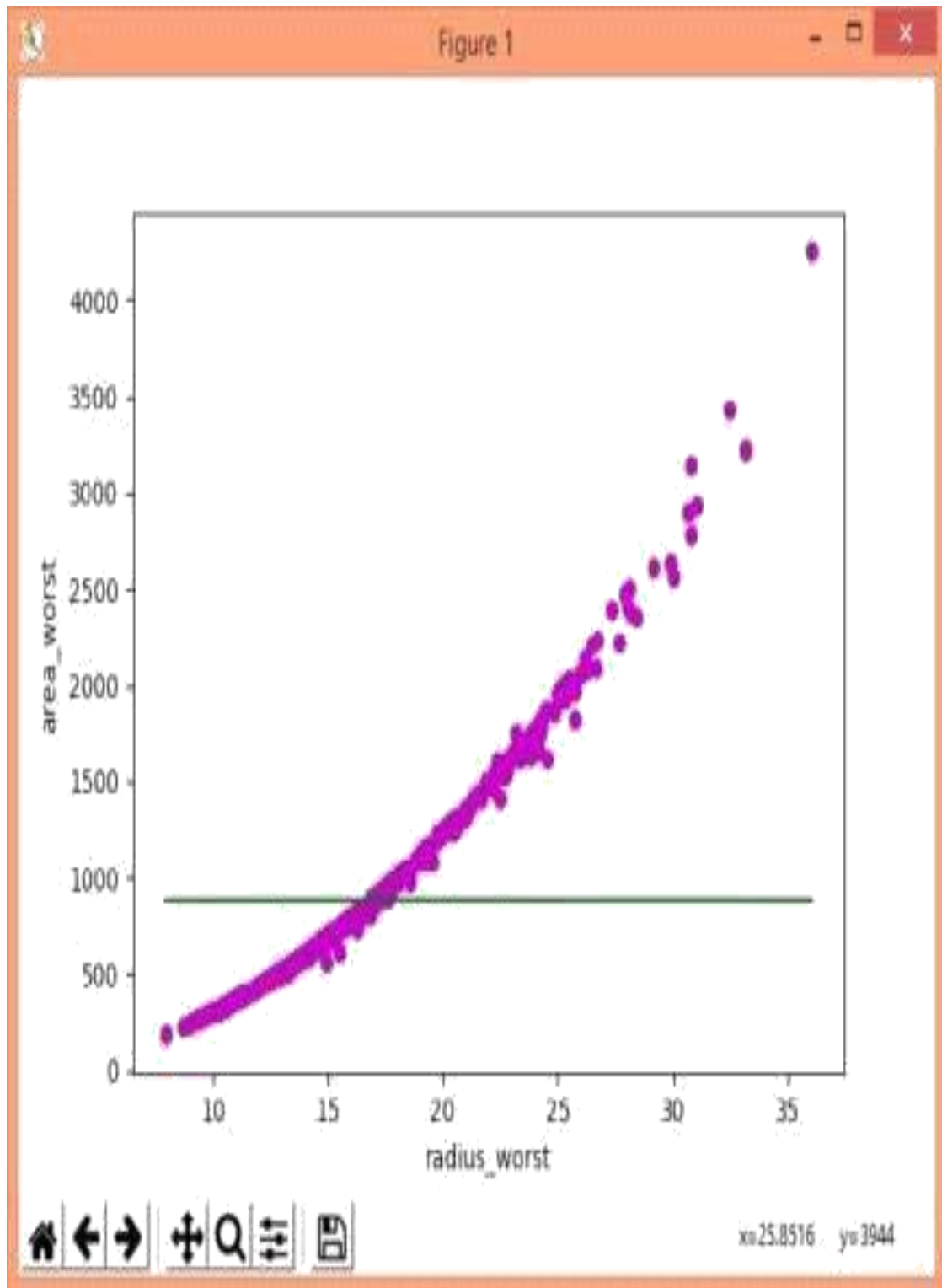
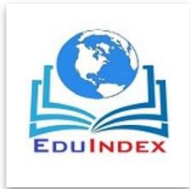
SVM classification plot





The regression coefficient between radius_worst and perimeter_mean and also between area_worst and perimeter_mean was positive.







Think India (Quarterly Journal)

ISSN: 0971-1260 Vol-22, Special Issue-07

Two-Day National seminar on "Artificial Intelligence, Multimedia and ICT in Teaching, Learning and Acquisition"

sponsored by

Indian Council of Social Science Research, New Delhi
on 23rd and 24th September 2019 at **Department of English,**
SRI S. Ramasamy Naidu Memorial College, Sattur, Tamilnadu, India



V. CONCLUSION AND FUTURE ENHANCEMENTS

The machine learning algorithms like SVM, k-nearest neighbor classifier gave good accuracy. However, in future, we need to develop model which can give 100% accuracy. Random forest model needs to be developed with 0% error rate.

REFERENCES

- [1] Shukla, N., Hagenbuchner, M., Win, K.T. Yang, J Computer Methods and Programs in Biomedicine 155 (2018) 199-208
- [2] C.C. Madubata, Y.Liu, M.S.Goodman, S.Yun, J.Yu, M.Lian, G.A.Colditz, Comparing treatment and outcomes of ductal carcinoma in situ among women in Missouri by race, Breast Cancer Res. Treat (2016) 1-10
- [3] O.O. Shiyabola, B.L. Sprague, J.M.Hampton, K.Dittus, T.A.James, S.Herschon, R.E.Gangnon, D.L. Weaver, A.Trentham-Dietz, Emerging trends in surgical and adjuvant radiation therapies among women diagnosed with ductal carcinoma in situ. Cancer 122(2016) 2810-2818
- [4] M.A. Devi, S.Ravi, J.Vaishnavi, S.Punitha, Classification of cervical cancer using artificial neural networks, Proc. Comput.Sci. (2016) 465-472