Analysis of Breast Cancer dataset using Supervised Machine Learning Classifiers

Parshavi Bolya, Divya Jain

Abstract: ***We Have Extracted Our Dataset From Kaggle. Our Study Is About Breast Cancer Diagnosis Based On 31 Input Attributes To Produce One Output Attribute That Is The Type Of Breast Cancer. Our Analysis Is On Two Major Aspects That Are Malignant And Benign On The Basis Of 10 Attributes That Is Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Symmetry, Fractal Dimension, Concave Points And Radius.***

Keywords: ***Breast Cancer, Malignant, Benign, Fractal Dimension.***

I. INTRODUCTION

When the cells in the breast grow abnormally it transforms to a disease called breast cancer [1]. The point to be taken into consideration is that mostly breast lumps are Benign and not Malignant (cancer) [2]. Symptoms for breast cancer can include – breast lump, change in size, change in skin, redness etc. [3]. To analyze the type of breast cancer biopsy is performed that further indicates the value of fractal dimension, radius, texture, perimeter, area, compactness, concavity, smoothness, concave points, and symmetry [4]. Benign and Malignant are basically breast tumors where Benign is not so harmful towards surrounding tissues whereas malignant are cancerous and can damage the surrounding tissues [5].

II. METHODOLOGY

The dataset for the project is taken from the Kaggle. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Our dataset consists of 569 instances and 31 input attributes plus 1 diagnosis attribute. The result is based on the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension factors.

For our study we have used Weka 3.8.4 tool to classify data for diagnosis of breast cancer in terms of Benign and Malignant. On the basis of above mentioned attributes we determine the type of breast tumor. Also Microsoft Excel has been used for graphical representation of the same..

Revised Manuscript Received on March 5, 2020.

Parshavi Bolya*, Techno India NJR Institute of Technology, Udaipur (Raj)-313003. **parshavibolya@gmail.com**

Divya Jain, Techno India NJR Institute of Technology, Udaipur (Raj)-313003.

III. EXPERIMENTAL ANALYSIS AND RESULTS

Various data classification algorithms were used to determine the breast cancer in patients. Data set consists of 569 instances breast cancer factors. To find the best fit of the data set following table was generated:

Table 1: Classification results using Weka 3.8.4

C	CC	MAE	RMSE	RRSE	KS	CCI
bayes.BayesNet	0.95248	0.054	0.218	45.0832	0.8984	542
functions.Logistics	0.945518	0.0548	0.2335	48.291	0.8845	538
functions.MultilayerPerception	0.966608	0.0337	0.1709	35.3401	0.9277	550
functions.SGD	0.980667	0.0193	0.139	28.7567	0.9584	558
functions.SimpleLogistics	0.97118	0.0444	0.1408	29.1122	0.9395	553
functions.SMO	0.97891	0.0211	0.1452	30.0354	0.9545	557
lazy.IBK	0.961336	0.0405	0.1963	40.591	0.9171	547
meta.AdaBoostM1	0.952548	0.0569	0.1943	40.1811	0.8976	542
meta.LogitBoost	0.959578	0.0576	0.1732	35.8215	0.9133	546
meta.RandomCommitee	0.968633	0.0696	0.1766	36.5191	0.9323	551
rules.Jrip	0.952548	0.0668	0.2092	43.2667	0.8986	542
trees.RnadomForest	0.96608	0.0758	0.1715	35.4701	0.9282	550
trees.LMT	0.97118	0.0444	0.1408	29.1122	0.9395	553

Where,
MAE: Mean Absolute Error
RMSE: Root Mean Squared Error
CC: Correlation Coefficient
RRSE: Root Relative Square Error
KS: Kappa Statistics
CCI: Correctly Classified Instances

The best suitable algorithm with the highest accuracy is found out to be SGD algorithm with correlation coefficient equal to 98.0667% which also include Kappa Statistics equals to 0.9584 and Mean Absolute Error 0.0193.

IV. GRAPHICAL REPRESENTATION OF ANALYSIS

A. Texture Mean:

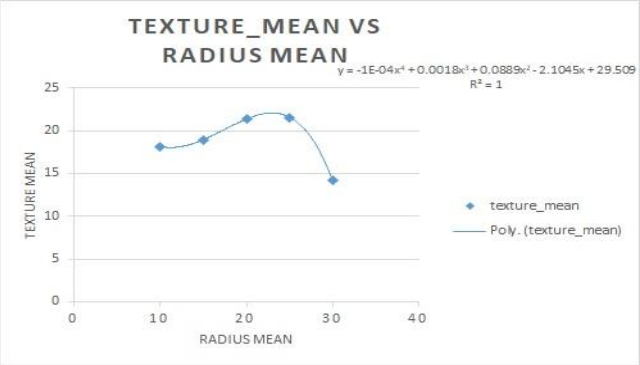

Fig 1: Relationship between texture mean and radius mean

The above graph evaluates that the maximum value of Texture mean that is found in the range 20-25 of Radius mean that is 21.5687.

The equation acquired from the analysis is

$y = -1E-04x^4 + 0.0018x^3 + 0.0889x^2 - 2.1045x + 29.509$

and the corresponding trend line is of 4 degrees polynomial.

B. Perimeter Mean:

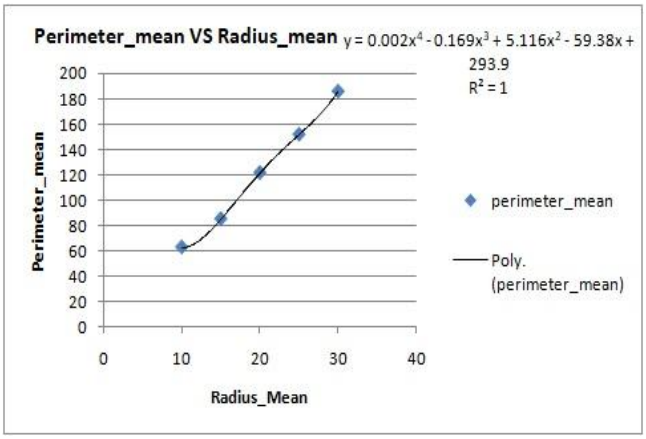

Fig 2: Relationship between perimeter mean and radius mean

The above graph evaluates that the maximum value of Perimeter mean that is found in the range 26-30 of Radius mean that is 185.833.

The equation acquired from the analysis is

y = 0.002x4 - 0.169x3 + 5.116x2 - 59.38x + 293.9

and the corresponding trend line is of 4 degrees polynomial.

C. Area Mean:

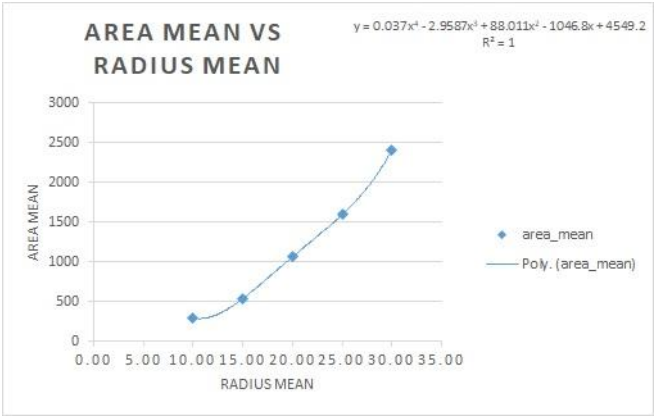

Fig 3: Relationship between area mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 26-30 of Radius mean that is 2416.667.

The equation acquired from the analysis is

$y = 0.037x^4 - 2.9587x^3 + 88.011x^2 - 1046.8x + 4549.2$

and the corresponding trend line is of 4 degrees polynomial.

D. Smoothness:

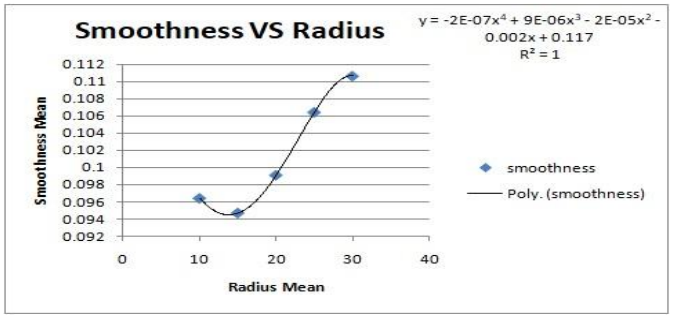

Fig 4: Relationship between smoothness mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 26-30 of Radius mean that is 0.110667.

The equation acquired from the analysis is

$y = 0.002x^4 - 0.169x^3 + 5.116x^2 - 59.38x + 293.9$

and the corresponding trend line is of 4 degrees polynomial.

E. Compactness:

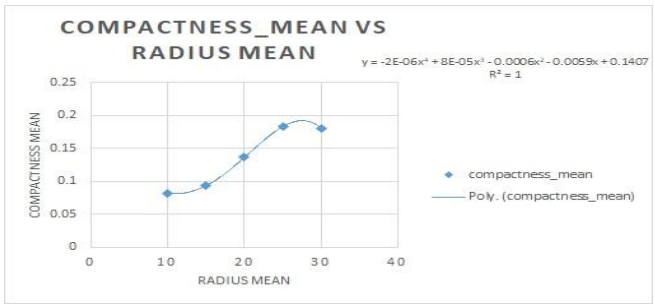

Fig 5: Relationship between compactness mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 21-25 of Radius mean that is 0.183154

The equation acquired from the analysis is

$y = -2E-06x^4 + 8E-05x^3 - 0.0006x^2 - 0.0059x + 0.1407$

and the corresponding trend line is of 4 degrees polynomial.

F. Concavity:

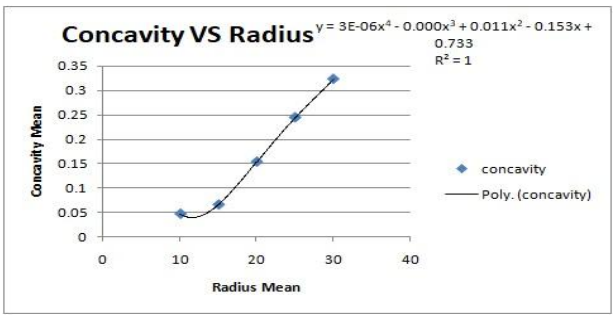

Fig 6: Relationship between concavity mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 26-30 of Radius mean that is 0.323567

The equation acquired from the analysis is

$y = 3E\text{-}06x^4 - 0.000x^3 + 0.011x^2 - 0.153x + 0.733$

and the corresponding trend line is of 4 degrees polynomial.

G. Concave Points:

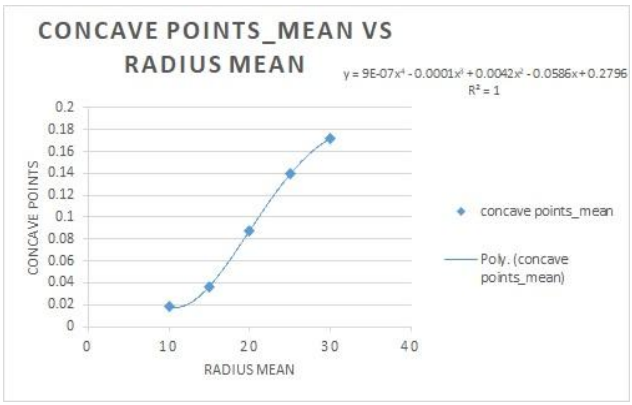

Fig 7: Relationship between concavity mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 26-30 of Radius mean that is 0.172067.

The equation acquired from the analysis is

$y = 9E\text{-}07x^4 - 0.0001x^3 + 0.0042x^2 - 0.0586x + 0.2796$

and the corresponding trend line is of 4 degrees polynomial.

H. Symmetry:

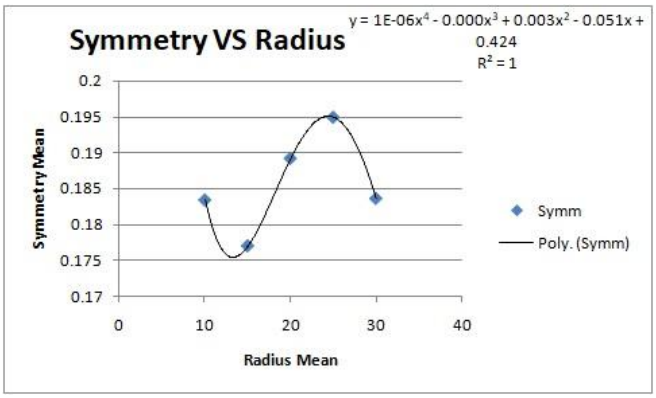

Fig 8: Relationship between symmetry mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 21-25 of Radius mean that is 0.194947.

The equation acquired from the analysis is

$y = 1E\text{-}06x^4 - 0.000x^3 + 0.003x^2 - 0.051x + 0.424$

and the corresponding trend line is of 4 degrees polynomial.

I. Fractal Dimension:

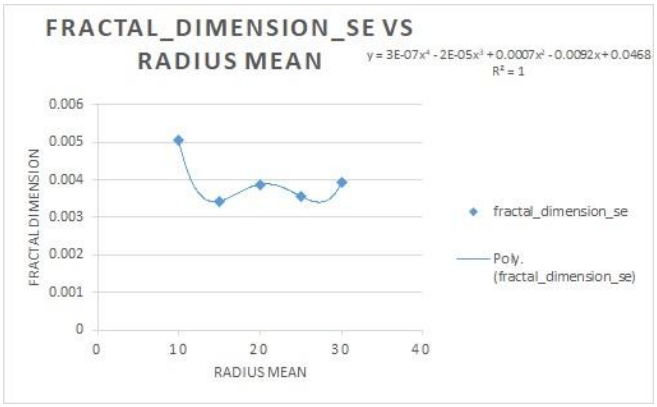

Figure 9: Relationship between fractal dimension mean and radius mean

The above graph evaluates that the maximum value of Area mean that is found in the range 6-10 of Radius mean that is 0.005045

The equation acquired from the analysis is

$y = 3E\text{-}07x^4 - 2E\text{-}05x^3 + 0.0007x^2 - 0.0092x + 0.0468$

and the corresponding trend line is of 4 degrees polynomial.

V. CONCLUSION

The above analysis represents that the radius of the breast cells in range 21-30 gives the maximum values of texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry which suggests that these are mostly the case of malignant breast cancer. Also, the best accuracy for the above analysis was seen at SGD algorithm.

REFERENCES

1. Division of cancer prevention and control, centers for disease control and prevention.
2. The American Cancer Society medical and editorial content team.
3. Breast Cancer care at Mayo Clinic.
4. National Breast Cancer Foundation.
5. Differences between a malignant and benign tumor by Lisa Fayed(verywellhealth.com)

AUTHORS PROFILE

Parshavi Bolya Department of Computer Science & Engineering, Techno India NJR Institute of Technology, Udaipur 313003

Divya Jain Department of Computer Science & Engineering, Techno India NJR Institute of Technology, Udaipur 313003