


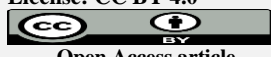


Determination of the Receiver Operating Characteristics (ROC) Curve of the Logistic Regression Model Accuracy Using Some Breast Measurements in the Presence of Multicollinearity

Uchenna P. Ogoke

Department of Mathematics and Statistics, University of Port Harcourt, Nigeria.

Correspondence: uchenna.ogoke@uniport.edu.ng; +2348035400184

Abstract	Article History
<p>This study was aimed at determining the Receiver Operating Characteristics Curve of the Logistic Regression Model accuracy using some breast measurements in the presence of Multicollinearity to detect the presence or absence of tumorous cell of cancer patients. A secondary data from the Breast Cancer Wisconsin (Diagnostic) was used for analysis. The data was cleaned for outliers, recoded numerically and tested for multicollinearity. The ROC curve for the logistic regression model also revealed a high sensitivity and high specificity of the presence of tumorous cells in the patients with a percentage of 95% which is extremely high indicating that the logistic regression model is good in predicting better diagnosis of tumor cell of cancer patients accurately when combined with ROC.</p> <p>Keywords: <i>Sensitivity; Specificity; Logistic Regression; Multicollinearity; Tumor</i></p>	<p>Received: 26 Nov 2022 Accepted: 12 Jan 2023 Published: 24 Feb 2023</p>
	<p>Scan QR code to view*</p> 
	<p>License: CC BY 4.0*</p>  <p>Open Access article.</p>
<p>How to cite this paper: Ogoke, U. P. (2023). Determination of the Receiver Operating Characteristics (ROC) Curve of the Logistic Regression Model Accuracy Using Some Breast Measurements in the Presence of Multicollinearity. <i>IPS Journal of Public Health, 3(1)</i>, 23–28. https://doi.org/10.54117/ijph.v3i1.11.</p>	

Introduction

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast. Breast cancer is a type of cancer brought on by a malignant tumor in the breast tissue's cells (Ruhil et al. (2013), A malignant tumor is a mass of cancer cells that has the potential to invade nearby tissue or metastasize to other parts of the body (Siegel et al. 2015). A mass of cancer cells called a malignant tumor has the potential to spread to other areas of the body or to infect surrounding tissue (Siegel et al. 2015). Due to its proximity to lung cancer, breast cancer is the second most common cause of mortality for women. Early detection can definitely lower the death rate for breast cancer, which is a fatal condition. The most recent data analysis revealed that the survival rate was 88% after five years of diagnosis and 80% after ten years of diagnosis (Heymach, 2018). According to estimates, there will be 60,290 new instances of non-invasive (in situ) breast cancer and 231,840 new cases of invasive breast cancer in women in the United States in 2015. (Breast Cancer, 2019). Except for lung cancer, breast cancer is the cancer that kills the most women in the US. In the world, breast cancer affects 25% of women and is the most frequent type of cancer in them.

Early indications of cancer include tumors, wounds that do not heal, unusual bleeding, persistent indigestion, and chronic hoarseness. For malignancies of the breast, cervix, mouth, throat, colon, rectum, and skin, early diagnosis is very important (UCI Machine Learning Repository, 2022). In statistical models, multicollinearity has always been a threatening presence. There have been much research that have looked at the effects of multicollinearity, and this is not by chance. The violation of no Multicollinearity can lead to large standard errors, inconsistencies in the estimated predictors, and possibly cause the solution for the model to not converge (Hill & Adkins, 2001; King, 2008).

Many tools have been developed to detect MCL. One of the most basic is the correlation matrix for a dataset, which provides the coefficient of simple correlation for each pair of variables (Tabachnick & Fidell, 2013; Kutner et al. 2005). However, MCL is not limited to two variables; this means that the simple correlation coefficients within the correlation matrix can be misleading when multiple variables are involved (Kutner et al. 2005). This led to the development of other methods to determine the extent of MCL. One such method is the Variance Inflation Factor (VIF). This allows one to see how much the variance is inflated compared to when the predictor variables are uncorrelated at all. A VIF of 1.0 means no MCL is present, and a VIF of 5 or even 10 or more is said to indicate the presence of MCL. Multicollinearity is so prevalent in statistical analysis alone that many methods have been devised to circumvent it. Some of these logistic regression methods consist of increasing the sample size, removing redundant variables, performing variable transformations, or doing nothing to report the MCL problem (Gujarati et al. 2009; King 2008; Belsley et al. 1980; Kleinbaum et al. 2014). This research is not restricted to only multicollinearity and logistic regression, however, other methods of diagnostic performance to measure different thresholds to distinguish diseased and non-diseased cases from normal cases with Receiver Operating Characteristics (ROC) Curve (Metz, 1978, Ogoke, et al. (2013)) are equally considered. This will go a long way to harmonizing the three methods together and bringing out the way they could yield a good result in terms of accuracy.

Methodology

Research Design

Data entry, computing and coding for the data sets were done using Microsoft office excel 2010. In order to determine whether or not the data sets contained multicollinearity, SPSS 25 was used to calculate Variance Inflation Factor (VIF). Data analysis on logistic regression was carried out using SPSS

*This work is published open access under the [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/), which permits free reuse, remix, redistribution and transformation provided due credit is given.

software. The degree of multicollinearity between models was evaluated using VIF. In order to determine the most effective model for dealing with multicollinearity to detect the presence or absence of the tumorous cell of cancer patients, the ROC and AUC curve were used, the data sets were fitted to logistic regression model.

The data were tested for the presence of Multicollinearity using VIF, outliers were tested for using Boxplot before proceeding to apply Logistic regression techniques to solve the problem of multicollinearity.

Multicollinearity

In checking for the presence of multicollinearity, SPSS was used to analyse the collinearity diagnosis using Variance Inflation Factor (VIF). Variance inflation factor is given by:

$$V(\hat{\beta}_j) = \frac{1}{1 - R^2_j} \tag{1}$$

where $1 - R^2_j$ is the tolerance and R^2 is the coefficient of determination. When $VIF > 5$, it indicates high multicollinearity for the regression model, when VIF values range between 1 and 5, it indicates medium level of collinearity, when VIF value is 1 it is non-collinear and considered to be negligible.

The Logistic Equation

The statistical model for logistics regression is $g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ (2)

let $\pi(x)$ be the conditional mean of y given x

where $\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$ (3)

Transforming $\pi(x)$ to linearize the model (i.e Logit transformation)

$$g(x) = \ln(e^{\beta_0 + \beta_1 X})$$

$$y = \beta_0 + \beta_1 X + \varepsilon \tag{4}$$

Receivers Operating Characteristics (ROC) curve

A ROC curve is a graph showing the performance of a classification model at all classification thresholds (Fig. 1). This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR)

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

$$FPR = \frac{FP}{FP+TN} \tag{6}$$

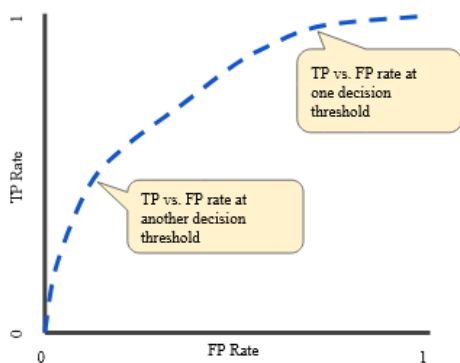


Figure 1: Receivers Operating Characteristics (ROC) curve

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figures 2-3 shows a typical ROC curve.

Area under the ROC Curve (AUC)

AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds.

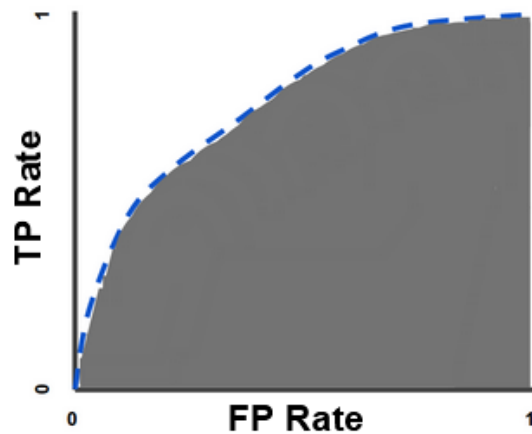


Figure 2: Area under the ROC Curve

Data description

This database can be found on UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29> which comprises: Attribute Information: The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits; Missing attribute values: none Class distribution: 357 absence of tumor, 212 presence of tumor.

ID number; Diagnosis (P = presence of tumor, A = absence of tumor); Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter); texture (standard deviation of gray-scale values); perimeter; area; smoothness (local variation in radius lengths); compactness (perimeter² / area - 1.0); concavity (severity of concave portions of the contour); concave points (number of concave portions of the contour); symmetry; fractal dimension ("coastline approximation" - 1).

Results and Discussion

The results obtained from the research is presented in tables 1-5 and figures 1-3. It revealed results on the tests for the presence of Multicollinearity using VIF, Outliers using Box plots (see appendix). It also revealed the result for the comparison between Logistic Regression using the Receivers Operating Characteristics (ROC) Curve and Area under the Receivers Operating Characteristics (AUC) Curve.

Logistic Regression Results

This is the Dependent Variable Encoding Table 2 displays how the values for Absence of Tumor (A) and Presence of Tumor (P) were coded. This is important for classification in the logistic regression.

ROC Curve

The ROC curve in figure 3 is a plot of the values of sensitivity vs 1-specificity as the value of the cut-off point moves from 0 to 1. A model with high sensitivity and high specificity will have a ROC curve that hugs the top left corner of the plot while a model with low sensitivity and specificity will have a curve closer to the 45-degree diagonal line. Thus figure 3 revealed that the ROC curve (the blue line) hugs the top left corner of the plot which indicates that the model does a good job of predicting better diagnosis of tumor cell of cancer patients accurately.

Table 1: Multicollinearity table

Coefficients ^a		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
Model		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4.022	.428		9.397	.000		
	radius_mean	.218	.174	1.586	1.255	.210	.000	3806.115
	texture_mean	-.005	.008	-.040	-.572	.567	.084	11.884
	perimeter_mean	-.024	.025	-1.192	-.946	.345	.000	3786.400
	area_mean	.000	.001	-.231	-.605	.545	.003	347.879
	smoothness_mean	-.085	2.017	-.002	-.042	.967	.122	8.194
	compactness_mean	4.222	1.334	.461	3.166	.002	.020	50.505
	concavity_mean	-1.398	1.046	-.230	-1.337	.182	.014	70.768
	concave_points_mean	-2.142	1.979	-.172	-1.082	.280	.017	60.042
	symmetry_mean	-.103	.743	-.006	-.138	.890	.237	4.221
	fractal_dimension_mean	-.033	5.572	.000	-.006	.995	.063	15.757
	radius_se	-.435	.310	-.249	-1.401	.162	.013	75.462
	texture_se	.007	.037	.008	.183	.855	.238	4.205
	perimeter_se	.023	.041	.094	.548	.584	.014	70.360
	area_se	.001	.001	.087	.660	.509	.024	41.163
	smoothness_se	-15.854	6.625	-.098	-2.393	.017	.248	4.028
	compactness_se	-.065	2.169	-.002	-.030	.976	.065	15.366
	concavity_se	3.565	1.301	.222	2.741	.006	.064	15.695
	concave points_se	-10.568	5.452	-.135	-1.938	.053	.087	11.521
	symmetry_se	-1.697	2.728	-.029	-.622	.534	.193	5.175
	fractal_dimension_se	7.146	11.676	.039	.612	.541	.103	9.718
	radius_worst	-.195	.058	-1.949	-3.367	.001	.001	799.106
	texture_worst	-.007	.007	-.091	-1.030	.303	.054	18.570
	perimeter_worst	.002	.006	.169	.410	.682	.002	405.023
	area_worst	.001	.000	1.190	3.163	.002	.003	337.222
	smoothness_worst	-.543	1.435	-.026	-.378	.705	.092	10.923
	compactness_worst	-.067	.383	-.022	-.175	.861	.027	36.983
	concavity_worst	-.381	.269	-.164	-1.419	.156	.031	31.971
	concave points_worst	-.464	.914	-.063	-.508	.612	.027	36.764
	symmetry_worst	-.557	.494	-.071	-1.126	.260	.105	9.521
	fractal_dimension_worst	-4.303	2.383	-.161	-1.806	.072	.053	18.862

a. Dependent Variable: diagnosis

Table 2: Dependent Variable Encoding

Original Value	Internal Value
P(presence of tumor)	1
A (Absence cancerous tumor)	0

AUC Results

The Area under the curve on Table 3 gives us an idea of how well the model is able to distinguish between positive and negative outcomes. The AUC ranges from 0 to 1. The higher the AUC, the better the model is at correctly classifying outcomes. Hence, in this case, the AUC of the binary logistic regression model is 0.955, which is extremely high. This indicates that the model is good in predicting better diagnosis of tumor cell of cancer patients accurately.

From Table 4, shows how effectively the model is able to predict the cancer category well when the predictors are included in the study. Overall PAC of 97.2% were correctly classified by the model.

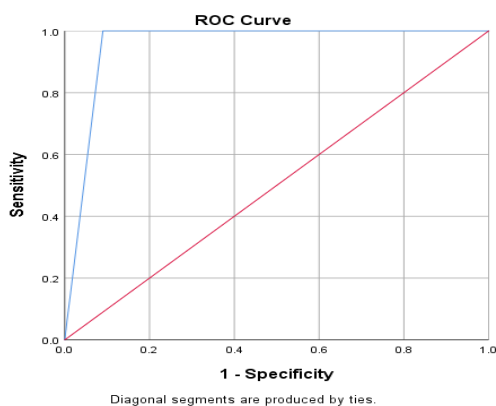


Figure 3: ROC curve for Logistic Regression

Table 3: Area under the Curve of the Logistic Regression

Area Under the Curve				
Test Result Variable(s): diagnosis	Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval
				Lower Bound Upper Bound
	.955	.011	.000	.933 .977

The test result variable(s): diagnosis has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption
 b. Null hypothesis: true area = 0.5

Table 4: AUC of Logistics Regression Models

Models	Observed	Predicted		AUC
		Absence(A)	Presence(P)	
Logistic Regression	Absence (A)	208	4	0.955
	Presence (P)	12	345	0.955

Table 5: Percentage Accuracy Classification (PAC) Table for Logistic Regression

	Observed	Predicted		Percentage Correct
		A	P	
Step 1	Y-diagnosis	208	4	98.1
	Overall Percentage	12	345	96.6

a. The cut value is .700

Conclusion

This study also revealed the presence of outliers in the dataset as shown graphically by the boxplot (Appendix) which was treated. This coincides with Osborne (2004) study on the power of outliers were he revealed that outliers may cause a significant impact on the mean and standard deviation hence decrease normality. The multicollinearity was also handled before the analysis Findings from comparing the Receiving Operating Characteristics (ROC) and Area Under Curve (AUC) of the logistic regression model shows a percentage of 95% which is extremely high indicating that the logistic regression model is good in predicting better diagnosis of tumor cell of cancer patients accurately when combined with ROC and AUC.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Detecting and Assessing Collinearity. *Reversion Diagnostics: Identifying Influential Data and Sources of Collinearity* 85-169. Hoboken: John Wiley & Sons, Inc.

Gujarati, D. N., Porter, D. C., & Pal, M. (2009). *The Nature of Multicollinearity*. Basic Econometrics. 315-345. McGraw Hill.

Heymach J, (2018). Clinical Carcinomas advances 2018: annual report on progress against Carcinomas from the American society of clinical oncology. *J Clin Oncol* 36(10):1020–44.

Hill, R. C., & Adkins, L. C. (2001). Collinearity. *A Companion to Theoretical Econometrics* 254-277. Blackwell Publishing. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

King, J. E. (2008). Binary Logistic Reversion. In J. W. Osborne, *Best practices in quantitative methods* (1st ed.). Sage Publications.

King, J. E. (2008). Collinearity. In J. W. Osborne, *Best practices in quantitative methods* (1st ed., 379-380. Sage Publications.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). *Collinearity. Applied reversion scrutiny and other multivariable methods*. 358-372. Cengage Learning.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Multicollinearity and Its Effects*. In *Applied Linear Statistical Prototypes* (pp. 278-289). New York: McGraw-Hill.

Metz, C.E (1978). Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, 8, 283-298.

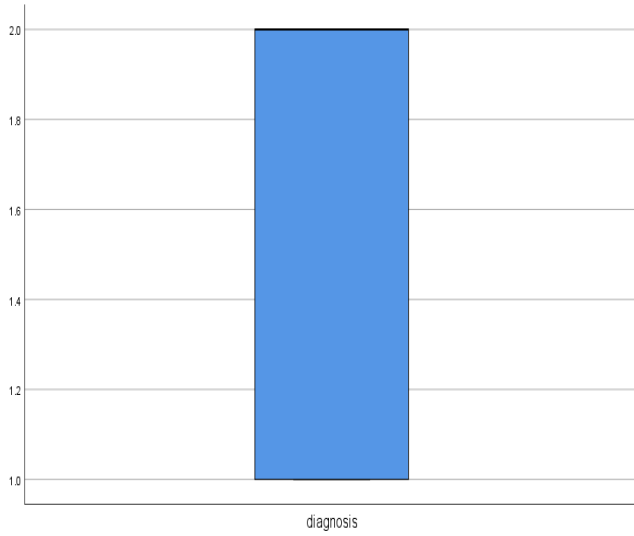
Ogoke, U.P., Nduka, E.C., Biu, O.E., and Ibeachu C. (2013). A Comparative Study of Foot Using Receiver Operating Characteristics (ROC) Approach. *Journal of Pure & Applied Sciences (Scientia Africana)* 12(1):76-88.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1), 6.

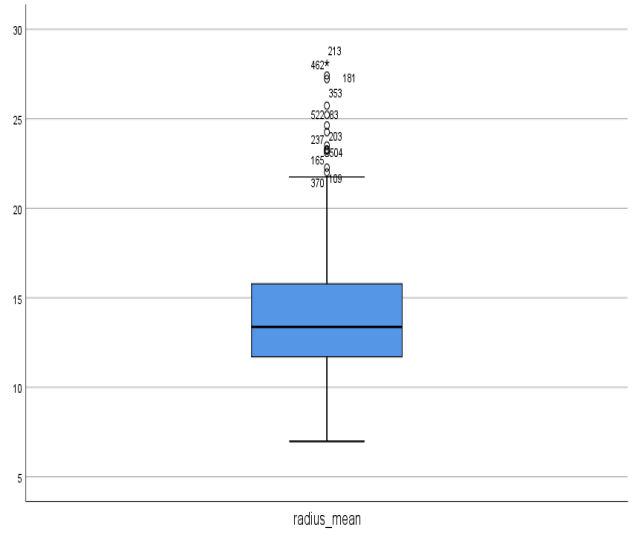
Ruhil, A. P., Raja, T. V., & Gandhi, R. S. (2013). Preliminary study on forecast of body weight from morphometric measurements of goats through ANN prototypes. *Journal of the Indian Society of Agricultural Statistics*, 67(1), 51-58.

Siegel RL, Miller KD, & Jemal A. (2015). Carcinomas statistics, 2015: Carcinomas statistic, CA. *Carcinomas J. Clin.* 65(1):5–29.

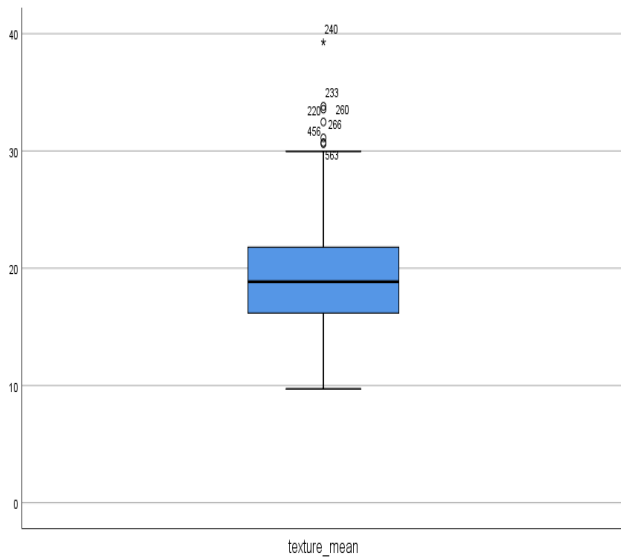
Appendix: Selected Outputs of some Test for Multicollinearity using VIF



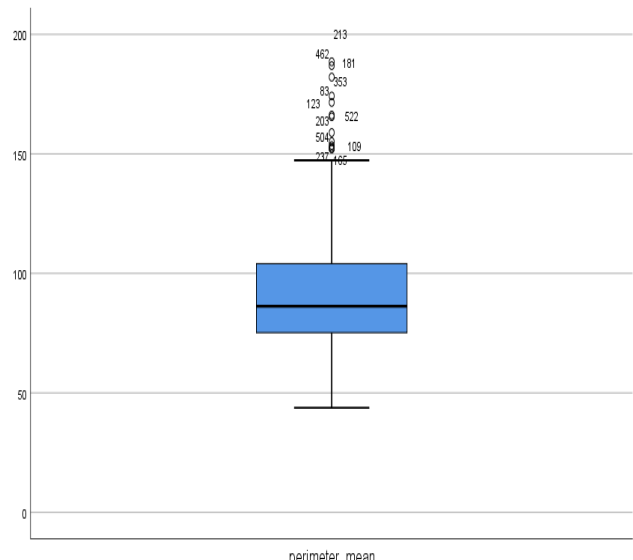
Boxplot for Diagnosis



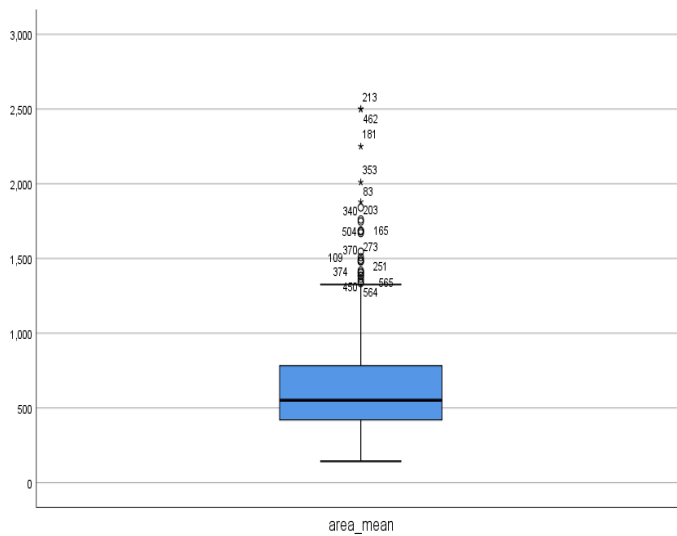
Boxplot for Radius Mean



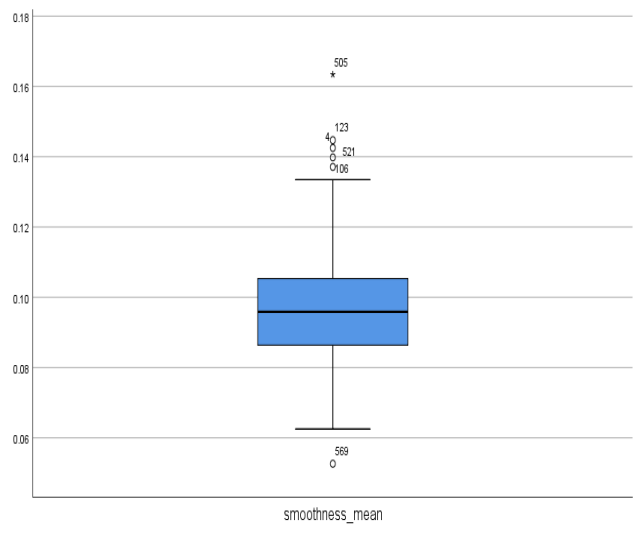
Boxplot for texture mean



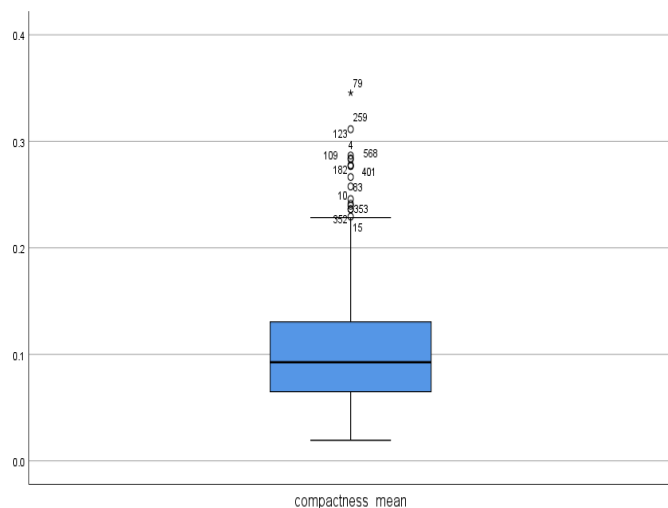
Boxplot for perimeter mean



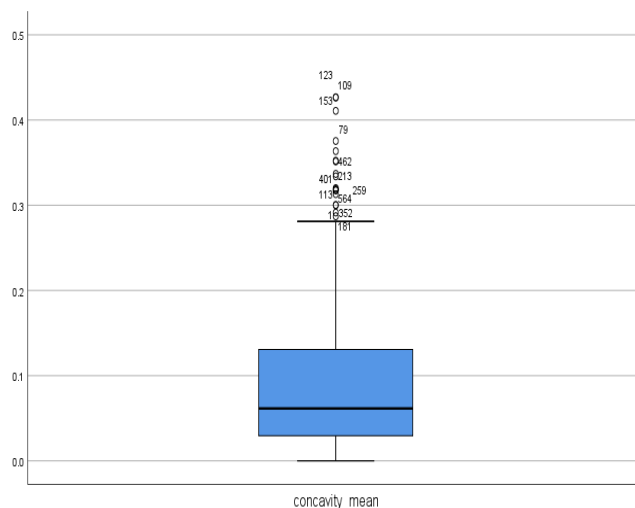
Boxplot for area mean



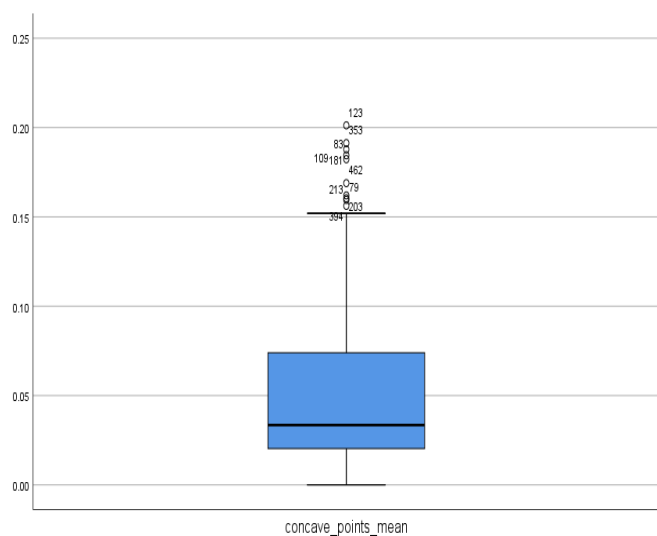
Boxplot for smoothness mean



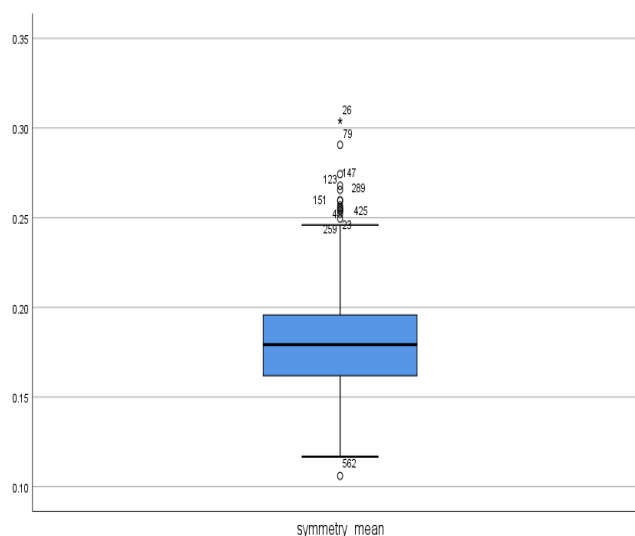
Boxplot for compactness mean



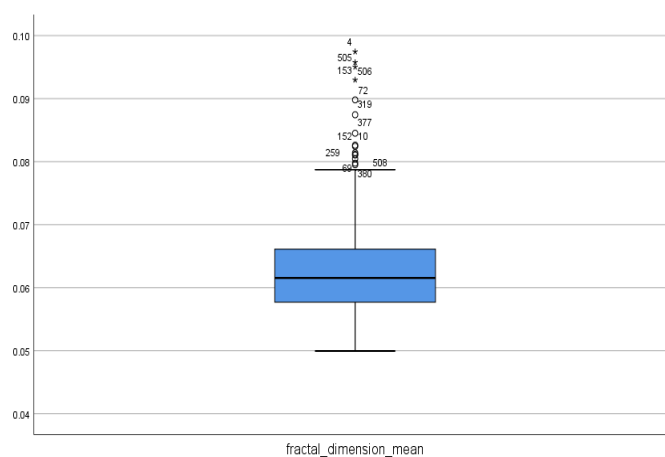
Boxplot for concavity mean



Boxplot for concave point mean



Boxplot for symmetry mean



Boxplot for fractal dimension mean

• Thank you for publishing with us.