Classification of Categorical Outcome Variable Based on Logistic Regression and Tree Algorithm

Pratibha V. Jadhav, Vaishali V. Patil, Sharad D. Gore

Abstract: ***Logistic regression is most popular techniques incorporated in traditional statistics. Usually, this regression is applicable when the dependent variable is of categorical binary in nature. In the field of Statistics and Machine learning, classification of data is critical to discriminate to which set of clusters a new observation belongs, in the base of training set of a data containing observation whose group relationship is known. In this paper, we are focusing on the concepts of Logistic regression and classification tree. A large data taken from UCI (Machine learning Repository) incorporated for this research work. The aim of study is to distinguish the results obtained from Logistic regression and decision tree. At the end, decision tree gives better results than Logistic regression.***

Keywords :Multiple logistic regression, CART, Misclassification error.

I. INTRODUCTION

Logistic Regression:

Logistic regression is most popular techniques used in conventional statistics. Generally, Linear regression is used if the response or dependent variable is occurring continuously and the residual errors are normally distributed. If the dependent or response variable is not continuous then we use Logistic regression [1]. This paper presents a regression model for categorical variable or dichotomous. For.eg whether a particular plant lives or dies, a student will be admitted or not and so on. A Logistic regression is also called as logistic or logit model; it analyzes the correlation among more than one independent variable and categorical dependent variable. It evaluates the probability of happening of event by fitting a logistic curve. There are two types of logistic regressions namely; Binary logistic regression where dependent variable is dual in nature and other is Multinomial logistic regression where response variables are more than two categorical. In logistic regression, the most desirable response considered as success and other is failure. Generally, Logistic regression predicts the possibility of success.

This probability is restricted to the interval [0, 1] and response variable is converted into ratio of odds as $\frac{p}{1-p}$.Here, odds are defined as the ratio of probability of success with the probability of failure. These odds are determined from probabilities and it lies between 0 to ∞. This odds ratio is transformed into logarithm function which is given in equation (1)

$$\log it(y) = \ln(odds) = \ln\frac{p}{1-p} = \alpha + \beta_1 X_1 + \beta_2 X_2 . + ...+ \beta_k X_k ....(1)$$

where p is the probability of desirable outcome, $X_1, X_2, ..., X_k$ is independent variable, α is intercept and $\beta_1, \beta_2, ..., \beta_k$ are regression coefficients [2] [3].

There are some assumptions of Logistic regression as follows:

1. Errors are distributed independently.
2. If dependent or response variable is in binary in nature then Binary logistic regression is required.
3. It can hold non-linear relations among the all variables such as independent and dependent variables as it applies non-linear log conversion to the linear regression.
4. This regression requires a big sample size as compared with linear regression as maximum likelihood approximations have low power for small samples.
5. If there is multicollinearity in logistic regression then it is handled by similar way like linear regression [4].

Classification tree:

The Classification and Regression tree algorithm (CART) was firstly announced by Breiman et al. in 1984 [5]. This has divide and conquer approach for classification. It is used for finding a feature and extracts some useful information from a large database. It is very essential for discrimination and prediction modelling. The decision tree structure is a classification method that make the beneficial methodology for chemical and biochemical applications [6]. Decision tree algorithm is a most popular technique for problems arised in classification and regression. There are different methods in data mining for discovering the useful information from huge data houses. When decision problem is used to classify the data then referred as classification tree and if it is applied for regression purpose then referred as Regression tree [7][8].

Manuscript published on January 30, 2020.
* Correspondence Author

Mrs. Pratibha Vijay Jadhav, Pursuing Ph.D. in Statistics from J.J.T.University, Jhunjhunu, Rajasthan, India.

Dr. Vaishali Vilas Patil, Assistant Professor in TC College Baramati Pune Maharashtra India.

Dr. Sharad Damodar Gore, Professor in JJTU Rajasthan India.

In this tree structure, each node of tree shows either a leaf node or decision node.

All these decision nodes have splits and testing the values of some functions of data attributes. Every branch of the decision node gives a different outcome of the test. Each leaf node has a class label attached to it [8]. These trees always remain easy to explain, interpret and visualization of results. It is very popular method to compare the results obtained by Multiple and Logistic regression. There are three different methods for impurity of node namely Gini Index, misclassification error and cross-entropy or deviation. In this paper, the methodology explains to find final prediction model using logistic regression by stepwise backward regression, splitting rule explaining to reach the decision at leaf node as probability of patient having breast cancer or not and distinguish the results of logistic regression and decision tree in terms of Misclassification error, Accuracy and Deviation.

II. DATA DESCRIPTION AND MODELLING

We have incorporated data Set from Breast Cancer Coimbra UCI (machine learning repository). In this data, Clinical structures were inspected or measured for breast cancer of 116 patients having 52 healthy controls (without breast cancer) and 64 patients (having breast cancer) . In the datasets, there are 10 independent variables namely, Age (years), Body Mass Index (BMI) (kg/m2), Glucose (mg/dL), Insulin (µU/mL), HOMA, leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), Monocyte Chemoattractant Protein 1(MCP-1) (pg/dL) and the dependent variable having labels 1 as healthy controls and label 2 as patients or having breast cancer [9]. These parameters are used to predict the person having breast cancer or not [10].

In this paper, response or dependent variable taken as categorical as labels 1 or 2. In the model p value indicates probability of success which is label 2. All the independent variables are continuous in nature therefore it is necessary to detect exploratory analysis which involves mean and standard deviations of each variable. A traditional logistic regression is applied on dataset with all 9 independent variables and binary classification variable means label 1 or label 2 as dependent variable. In this model p is the probability of success. The analysis is using logistic regression is obtained with performance of measurements such as Misclassification error, Accuracy and Deviation.

- **Performance Measure:**

The performance of measured by using the values of Misclassification error, Accuracy and Deviation. It is calculated from confusion matrix. A confusion matrix is generally used for describing the performance of a classification model. This is a matrix which contains the elements as True Positive (TP), True Negative (TN), False Positive and False Negative. In this matrix, True value a taken as having breast cancer patient and False value taken as Having healthy control. As in this study label 1 as Healthy control or failure and label 2 as patient or success.

True Positive (TP): This is the cases where classifier predicted value as TRUE (having patient) and actual class value is also TRUE (having patient).

True Negative (TN): This is the cases where classifier predicted value as FALSE (having healthy control) and actual class value is also FALSE (having healthy control).

False Positive (FP): When classifier predicted value as TRUE (having patient) but actual value is False (having healthy control). This is also called as Type-I error.

False Negative (FN): When classifier predicted value as FALSE (having healthy control) but actual value is TRUE (having patient). This is also called as Type-II error. The following Table-1 shows the confusion matrix including True Positive (TP), True Negative (TN), False Positive and False Negative values as shown below,

Table-1 shows the confusion matrix.

Confusion matrix		Actual class	
		Negative	Positive
Predicted class	Negative	TN	FN
	Positive	FP	TP

TN: True Negative FN: False Negative FP: False Positive TP: True Positive

The Accuracy and Misclassification Error is calculated by,

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} ....(2)$$

$$Misclassifcation\ Error = \frac{(FP + FN)}{(TP + TN + FP + FN)} ....(3)$$

Null Deviance: This shows the response is predicted by a model without intercept. The minimum value shows the best model.

Residual Deviance: In this, the response is predicted from the independent variables by the model. The minimum value of residual deviance shows best fit of the model.

Akaike Information Criteria (AIC):

This is comparable performance measure in logistic regression. This is nothing but of adjusted R^2 .This is the amount of fit which disciplines the model for the coefficients of model including intercept. The AIC is minimum for the best model. Generally, AIC is calculated with the help of software. The basic formula for AIC is given by,

$$A.I.C. = 2N - 2(\log likelihood)..........................(4)$$

Where,

N is the total number of parameters involved in model with intercept.

Log-likelihood is measure of fit models. The model is fit when this number is higher and is obtained in software.

In this study, the p values of the variables such as Age, Insulin, HOMA, Leptin, Adiponectin and MCP-1 are observed and it is relatively high. To obtain final model, a variable is selected for elimination whose p value is uppermost and also greater than 0.05 and it is removed from analysis. Here, Adiponectin is removed from analysis because it has highest p value. The analysis is carried out with rest of the independent 8 variables. Repeat this process till independent variables with p value to be less than 0.05. This process is called as Backward Stepwise regression [11].

The resultant variables are said to be statistically significant as their p value is less than 0.05 and formed a final logistic regression model. Table-4 represents the result of AIC values of each iteration of step wise regression and Table-5 represents the result of final logistic regression model with significant variables. Misclassification error, Accuracy and Deviation was calculated by using equation (2) and equation (3) for final regression model with significantly independent variables.

In this paper, outcome variable or dependent variable is categorical so that we are refer classification tree. A decision tree is used as classification tree. The following figure shows the structure of Decision classification tree.

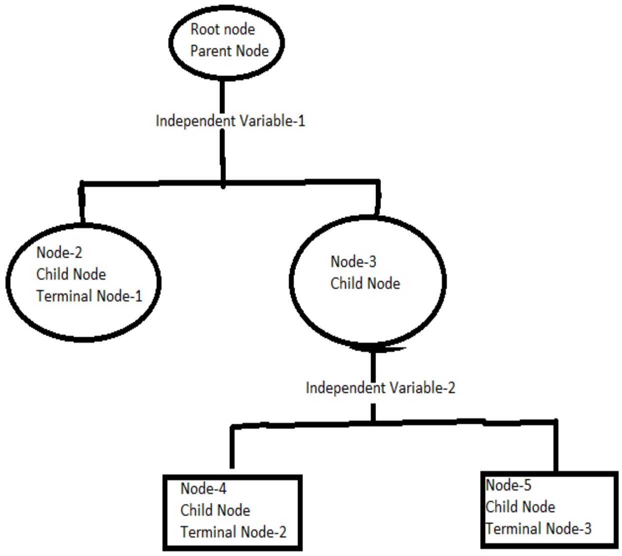

Fig.1 Structure of decision tree [11].

The classification tree is starts with root node which contains whole samples, it is also called as Parent node which is demonstrated as Root Node-1 in Figure 1. This algorithm observes all independent variables and selects that in binary groups in order to dependent variables. This algorithm divides Node-1 as root node into Node-2 and Node-3 according to independent variable then Node-2 and Node-3 becomes child node of Root node-1. Node 3 divides into Node-4 and Node-5 according to independent variable then Node-3 becomes parent node for Node-4 and Node-5. Node-4 and Node-5 are becomes child node for Node-3. This procedure remains over every branch of tree algorithm is touched and at this point terminal node will create. In figure1, Node-2, Node-3 and Node-5 are becomes terminal nodes. These Terminal nodes are occurring common exclusively and exhaustive subclasses in nature [11]. This tree algorithm has been growned by using caret package in R software [12].

III. RESULTS AND DISCUSSION

The model was executed in R software and results were obtained. Table.2 shows the exploratory analysis of independent variables using R software version 3.5.3

Table: 2 Exploratory analysis of Independent Variables.

Independent variables	Mean	Standard Deviation(S.D.)
Age	57.31	16.11
BMI	27.59	5.02
Glucose	97.80	22.53
Insulin	10.01	10.06
HOMA	2.70	3.65
Leptin	26.62	19.19
Adiponectin	10.18	6.85
Resistin	14.73	12.40
MCP.1	534.65	345.91

This table gives the mean and standard deviation of each variables as these variables are in continuous in nature. The logistic regression method is applied on dataset using R software the result is as follows,

Table: 3 Analysis using logistic regression method.

Independent variables	Estimates	Standard Error	Z-value	P-value
Regression Coefficients (β)				
Age (years)	-0.0234	0.0156	-1.495	0.13498
BMI (kg/m2)	-0.1501	0.0675	-2.224	0.02613<0.05
Glucose (mg/dL)	0.1056	0.0348	3.034	0.00242<0.05
Insulin (µU/mL)	0.2072	0.2630	0.788	0.43081
HOMA	-0.5979	1.0899	-0.549	0.58332
Leptin (ng/mL)	-0.0102	0.0173	-0.589	0.55582
Adiponectin (µg/mL)	-0.0053	0.0376	-0.140	0.88858>0.05
Resistin (ng/mL)	0.0586	0.0299	1.961	0.04982<0.05
MCP-1(pg/dL)	0.0007	0.0008	0.865	0.38730
Intercept (α)	5.6512	3.3580998	-1.683	0.09240
Null deviance	159.57 on 115 degrees of freedom			
Residual deviance	111.73 on 106 degrees of freedom			
AIC	131.73			
Accuracy	79.32			
Misclassification Error	20.68			

The logistic regression model becomes,

$$\ln\frac{p}{1-p} = 5.6512 - 0.0234 * Age - 0.1501 * BMI$$

$$+ 0.1056 * Glucose + 0.2072 * \mathrm{Insulin} - 0.5979$$

$$* \mathrm{HOMA} - 0.0102 * \mathrm{Leptin} - 0.0053 *$$

$$\mathrm{Adiponectin} + 0.0586 * \mathrm{Resistin} + 0.0007 * \mathrm{MCP\text{-}1}$$

The Accuracy and misclassification error of logistic regression model is calculated from confusion matrix, a confusion matrix is a matrix consisting of column values projects actual values, row values projects predicted values and elements projects the parameters in terms of True positive, True negative, False positive and False negative which is given by,

Table-4 Confusion matrix for logistic regression.

Confusion matrix		Actual class	
		Negative	Positive
Predicted class	Negative	41	13
	Positive	11	51

Accuracy of model is calculated by using equation (2)
Accuracy of model = [(41+51)/ 116] = 0.7932
Accuracy of model in percentage=0.7932*100=79.32%
Misclassification Error is calculated by using equation (3)
Misclassification Error =0.2068

Misclassification Error of model in percentage=0.2068*100=20.68%

From the above table, the BMI, Glucose and Resisitin variables are statistically significant because their P values are less than 0.05. Here p value indicates that value of significance for the analysis [13]. We eliminate that variable those having P value is greater than 0.05 [14].

The variable Adiponectin having highest p value is 0.88858 which is greater than 0.05 value. Firstly, it is removed from the analysis and complete the analysis with remaining variables and repeat the process till all variables get all values are less than 0.05 value. This is called Backward Stepwise regression. The following Table- shows the result of backward stepwise regression with AIC, Null deviance, Residual deviance, Misclassification error and Accuracy.

Table-5 shows the result of backward stepwise regression with AIC, Null deviance, Residual deviance, Misclassification error and Accuracy.

Sr.no.	AIC	Null deviance	Residual deviance	Misclassification error	Accuracy
1	131.73	159.57	111.73	20.68	79.72
2	129.75	159.57	111.75	20.68	79.72
3	128.01	159.57	112.01	19.82	80.18
4	126.48	159.57	112.48	20.68	79.72
5	125.57	159.57	113.57	20.68	79.72
6	125.54	159.57	115.54	25	75.00
7	127.50	159.57	119.50	26.72	73.68

Finally, using backward stepwise regression method, last iteration gives final result with all variables are significant that is all p values are less than 0.05, this is the final outcome which is further used for find final logistic regression model. The logistic regression model with significant variables becomes final model of our dataset. The following table-6 represents the result of final logistic regression model,

Table:6 Final Logistic regression model with significant variables

Independent variables	Estimates	Standard Error	Z-value	P-value
Regression Coefficients (β)				
BMI (kg/m2)	-0.1314	0.0468	-2.812	0.0049
Glucose (mg/dL)	0.0867	0.0206	4.210	0.000025
Resistin (ng/mL)	0.0702	0.0305	2.302	0.0214
Intercept (α)	-5.2851	2.0346	-2.598	0.0093
Null deviance	159.57 on 115 degrees of freedom			
Residual deviance	119.50 on 112 degrees of freedom			
AIC	127.5			
Accuracy	73.28%			
Misclassification Error	26.72%			

The final Logistic regression model becomes,

$$\ln \frac{p}{1-p} = -5.28518 - 0.13144 * BMI + 0.08676 * Gluc$$
$$+ 0.07023 * \text{Resistin}$$

Similarly, Accuracy and misclassification error of final logistic regression model is calculated from confusion matrix is given by,

Table-7 Confusion matrix for Final logistic regression model.

Confusion matrix		Actual class	
		Negative	Negative
Predicted class	Negative	39	18
	Positive	13	46

Accuracy of final model is calculated by using equation (2)

Accuracy of final model = [(39+46)/ 116] = 0.7328

Accuracy of final model in percentage=0.7328*100=73.28%

Misclassification Error of final model is calculated by using equation (3)

Misclassification Error =0.2672

Misclassification Error of final model in percentage=0.2672*100=26.72%

On other side, we obtained Decision tree with 4 terminal conditional trees as decisions which is shown in following graph.

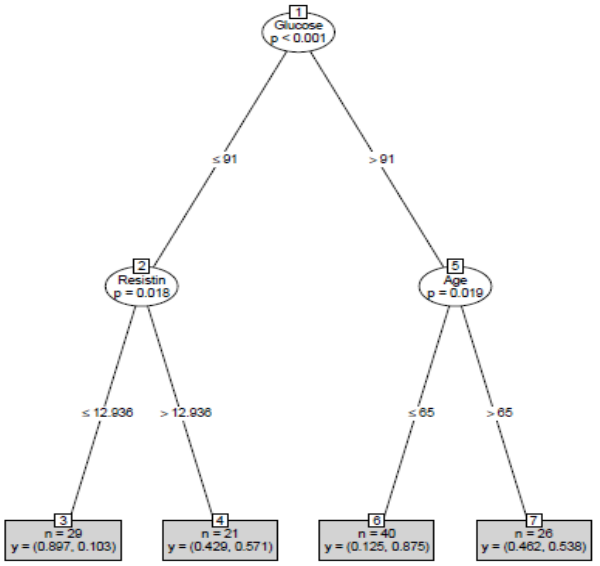

Figure.2 Decision tree with 4 terminal nodes.

The figure 2 indicates a decision tree with 7 nodes. This decision tree has total 7 nodes out of these node-1 shows root node or parent node as Glucose and it is splitted on if p <0.001 here p is the probability of patient having breast cancer.

From above figure.2 we got Glucose as root node or parent node, Resistin and Age as child node. The tree grows when Glucose is <=91 and Glucose >91.If Glucose <=91 then tree reaches at child node -2 with Resistin variable with p=0.018. If Glucose <=91and Resistin <=12.936 then tree reaches at terminal node 3. Similarly, If Glucose <=91and Resistin >12.936 then tree reaches at terminal node 4. If Glucose >91 then tree reaches at child node as Age variable with p=0.019. If Glucose >91 and Age<=65 then tree reaches at terminal node -6. Similarly, If Glucose >91 and Age>65 then tree reaches at terminal node-7.

After termination of this tree algorithm, we got 4 Terminal nodes with splitting constraints and tree algorithm ends. We obtained 4 splitting rules for each terminal node. These splitting rules are based on splitting constraints or conditions.

we got the splitting rule for Node-3 as when (Glucose<=91) and (Resistin <=12.936) then Node-3 declared that the persons are healthy or having healthy control. For Node-4, it is when (Glucose<=91) and (Resistin >12.936) then Node-4 declared as patients are having breast cancer as there are more number or proportion having breast cancer patients. For Node-6,it is when (Glucose>91) and (Age<=65) declared as patients are having breast cancer as there is high proportion of breast cancer patients. For Node-7, it is when (Glucose>91) and (Age>65) declared as the patients are having breast cancer. From the root node we can declared that when (Glucose>91) and (Age<=65) or (Age>65) then patients are having breast cancer. From this classification tree, we got single node that is Node-3 as healthy control persons or patient.
The splitting rule can be summarized in following table,

Table-8 Splitting rule for terminal nodes

Sr. No.	Node	Splitting rule	Decision
1	3	When (Glucose<=91) and (Resistin <=12.936)	Healthy or having healthy control
2	4	when (Glucose<=91) and (Resistin >12.936)	Patients are having breast cancer
3	6	when (Glucose>91) and (Age<=65)	Patients are having breast cancer
4	7	when (Glucose>91) and (Age>65)	Patients are having breast cancer
5	-------	when (Glucose>91) and (Age<=65) or (Age>65)	Patients are having breast cancer

Accuracy and the misclassification error for decision tree is calculated by using confusion matrix and it is given by,

Table-9 Confusion matrix for decision tree

Confusion matrix		Actual class	
		Negative	Positive
Predicted class	Negative	26	3
	Positive	26	61

Accuracy of decision tree is calculated by using equation (2)
Accuracy of decision tree = [(26+61)/ 116] = 0.75
Accuracy of decision tree in percentage=0.75*100=75%
Misclassification Error of tree is calculated by using equation (3)
Misclassification Error =0.25
Misclassification Error of tree in percentage=0.25*100=25%
The Accuracy of Logistic regression is 73.28 % and misclassification error is obtained as 26.72% whereas decision tree gives the accuracy as 75.00 % and misclassification error is obtained as 25.00%.

IV. CONCLUSION

Logistic regression is an important classification technique used in many applications considering variable dataset. In this work, we have taken substantially big dataset of breast cancer from UCI (Machine learning Repository). The p value greator than 0.05 is insignificant for further calculations so we have considered all the values below 0.05. Finally, all significant variables are achieved by using step-wise regression and final logit regression model is obtained. Accuracy, Misclassification error, AIC, Deviation are measured. Decision tree approach has been used for classification purpose. Logistic regression is a conventional approach which is based on computation of equation while decision tree model is an data driven and nonparametric approach. It has been observed that decision tree model gives better result than logistic regression in terms of misclassification error and accuracy.

REFERENCES

1. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
2. Wu, H. (2009). Comparative Analysis of Logistic Regression, Support Vector Machine and Artificial Neural Network for the Differential Diagnosis of Benign and Malignant Solid Breast Tumors by the Use of Three-Dimensional Power Doppler Imaging. 10(2), 464–471.
3. Christensen, R. (2006). Log-linear models and logistic regression. Springer Science &Business Media.
4. Gadekar, Kirti, and Sharad Gore. 2019. “Kirti Gadekar and Sharad Gore.” 06(1): 42–46.
5. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth & Brooks. Cole Statistics/Probability Series.
6. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.
7. Franjkovic, J. (2017). Shopping intention prediction using decision trees. (October). https://doi.org/10.29352/mill0204.01.00155
8. Li, Linna, and Xuemin Zhang. "Study of data mining algorithm based on decision tree." In Computer Design and Applications (ICCDA), 2010 International Conference on, vol. 1, pp. V1-155.IEEE, 2010.
9. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin , glucose , age and BMI to predict the presence of breast cancer. 1–8. https://doi.org/10.1186/s12885-017-3877-1
10. Crisóstomo J, et al. Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. Endocrine. 2016;53(2):433-42.
11. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of behavioral medicine, 26(3), 172-181.
12. Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28(5), 1-26.
13. Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (Vol. 279). John Wiley & Sons.
14. Dey, S., & Raheem, E. (2016). Multilevel multinomial logistic regression model for identifying factors associated with anemia in children 6–59 months in northeastern states of India. Cogent Mathematics, 3(1), 1159798.

AUTHORS PROFILE

Mrs. Pratibha Vijay Jadhav, has Completed M.Sc in Statistics in 2005 and now pursuing Ph.D. in Statistics from J.J.T.University, Jhunjhunu, Rajasthan. Area of research is Data Science, Machine Learning. She has published 2 papers in International journals.

Dr. Vaishali Vilas Patil, has completed her MSc, MPhil, Ph.D. in Statistics. Her area of research is Design of Experiment, Data Science, and Machine Learning. She has published 6 papers in reputed journals. She is working as Assistant Professor in TC College Baramati Pune Maharashtra India.

Dr. Sharad Damodar Gore, has completed his M.Sc, Ph.D. in Statistics. His area of research is Multivariate Statistics, AI, Applied statistics, environmental Science, Design of Experiment, Statistical modeling, Statistical Analysis, Data Science and Machine Learning. He has published papers more than 100 papers in reputed journals and his citation index is 432. He has worked as Faculty Pennsylvania State University Department of Statistics. He was Department Head of Statistics and Computer Science at Savitribai Phule Pune University. Currently he is working as Professor in JJTU Rajasthan India.