Data Analysis and Data Classification in Machine Learning using Linear Regression and Principal Component Analysis

Lokasree B S [a]

[A] EEE, CMR Institute of Technology, Benguluru, India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: In this paper step-by-step procedure to implement linear regression and principal component analysis by considering two examples for each model is explained, to predict the continuous values of target variables. Basically linear regression methods are widely used in prediction, forecasting and error reduction. And principle component analysis is applied for facial recognition, computer vision etc. In Principal component analysis, it is explained how to select a point with respect to variance. And also Lagrange multiplier is used to maximize the principle component function, so that optimized solution is obtained.
Keywords*:* Linear Regression, Least squares, Principal Component analysis (PCA), unsupervised learning algorithm, Covariance, Lagrange multiplier.

1. Introduction

Linear Regression is the simplest method which models the relation between dependent and independent variables. Here in this models it is considered as one dependent and one independent variable for the analysis of linear regression, hence it is a simple linear regression. If we use multiple data sets for analysis then such models is called as multiple linear regression model (Mohan, 2019).

Principal component analysis is a method for dimensionality reduction and for feature extraction and it's an unsupervised learning algorithm, it is in some sense, the mostly used unsupervised learning algorithm in the field of machine learning and artificial intelligence. Actually PCA is a linear transformation which transforms the given data from n-dimensional space to another space with the same number of dimensions, but despite the fact that the original space, the original data has some correlations among the space variables or dimensions or features (Arunkarthikeyan, 2021; Pavan, 2020). Always the resulting data is the target space which will have variables which are completely uncorrelated to each other and they are mutually orthogonal or they are mutually perpendicular to each other (Wang, 2008). On using statistical methods, for studying multivariate problems, n number of variables will increase the problem complexity and amount of calculation. The principal component analysis method is to use the dimension reduction method to transform the multi index into a small number of mutually independent comprehensive indexes, and then the data for further analysis (Zhao, 2014; Garikipati, 2021). Let's see how PCA does this work and how it can be used to reduce the number of dimensions and reduce dimensionality of given data.

2. Procedure for linear regression

The vital feature of linear regression is that use of least square method. "standard" Least Square Error (LSE) methods fitting data to a function $Y = f(x)$, where x is an independent variable and y is a measured or given value, "orthogonal" Total Least Square Error (TLSE) fitting data to a function f(x)=0, i.e. fitting data to some d-1 dimensional entity in this d-dimensional space, e.g. a line in the E^2 space or a plane in the E^3 space (Groen, 1980; Balamurugan, 2018; Arunkarthikeyan, 2020; Huffel, 1991)"orthogonally Mapping" Total Least Square Error (MTLSE) methods for fitting data to a given entity in a subspace of the given space. However, this problem is much more complicated. As an example, we can consider data given in and we need to find an optimal line in E^d, i.e. one dimensional entity, in this d-dimensional space fitting optimally the given data. Typical problem: Find a line in the E^d space that has the minimum orthogonal distance from the given points in this space. This algorithm is quite complex and solution can be found in (Skala, 2015). Here in this paper least square method is used for most quantitative prediction. Estimates for the parameters are obtained by minimizing the sum of squares of differences between the observed values and predicted values (Zahedan , 2015; Deepthi, 2019).A simple straight line equation is considered where equation of straight line is:

$$y = mx + c \quad \text{---------- (1)}$$

Where, m=slope of the line

C= y intercept

x and y are two inputs.

Prerequisites to perform linear regression are Basic level of statistics, Line equation (which is given already), Steps to apply linear regression.

Consider a function

$$f(x_i) = a\,x_i + b \quad \text{---------- (2)}$$

Here b is intercept.

i varies from 0 to n.

Minimize the error,

$$E = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \quad \text{---------- Least Squares (Kevin, 2012)}$$

$$= \sum_{i=1}^{n} [y_i^{\,2} - 2a\,y_i\,x_i - 2b\,y_i + (a\,x_i + b)^2] \quad \text{-- (3)}$$

Apply partial differentiation

$$\frac{\partial E}{\partial a} = 0, \frac{\partial E}{\partial a} = 0 \text{---------- (4)}$$

$$(2\,x_i + 2b - 2\,y_i)\,x_i = 0 \quad \text{------ (5)}$$

$$-y_i + a\,x_i + b = 0 \quad \text{---------- (6)}$$

From eq. (5)

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i\,y_i \text{ ---------- (7)}$$

From eq. (6)

$$\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \text{ ---------- (8)}$$

Now form a matrix which is in the form,

$$AX = B$$

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & b \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i\,y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \text{ ------- (9)}$$

$$X = A^{-1}B \text{---------- (10)}$$

Here, by using all the required formulas, mathematical modeling of linear regression can be done, which is given in PCA section.

2.1 Mathematical Modelling

Let us consider data of x and y, which is given below, where x is number of labors involved and y, is number of shoes produced. Here x is independent variable and y is dependent variable.

Given x=[3,4,5,6,7,8,9,10,12,13,14]

y=[5,7,9,11,13,15,17,19,21,23,25]

n=11

From eq. (9), calculate all the elements in matrix and form reduced matrix.

$$\begin{bmatrix} 889 & 91 \\ 91 & 11 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1009 \\ 165 \end{bmatrix} \text{---------- (11)}$$

Above matrix elements can be written in the form of equations and after solving values of a & b are obtained. Where, a=+1.792, b=+0.178

Table 1. Table for error at each point

X	Y	$y^{\wedge}$	$y - y^{\wedge}$
3	5	5.554	-0.554
4	7	7.346	-0.345
5	9	9.136	-0.136
6	11	10.93	0.072
7	13	12.722	0.2804
8	15	14.511	0.489
9	17	16.303	0.697
10	19	18.1	0.905
12	21	21.68	-0.678
13	23	23.47	-0.47
14	25	25.262	-0.262

The next step is to calculate $y^{\wedge}$, where it can be calculating using formula $y^{\wedge} = ax + b$

From the above table, calculate total error using formula

$$E_1 = \sum_{i=1}^{11} \left(y_i - f(x_i)\right)^2 \text{---------- (12)}$$

Error at data point (x,y)=(true y)-(predicted y)

$$= y - y^{\wedge}$$

The total error obtained for a straight line equation using linear regression model is, E_1=2.821.

2.2 Flow Chart of Linear Regression

Step by step procedure to be followed for performing linear regression is explained through flow chart concept which is given in figure (1) Here input data x and y is considered for a given straight line equation, first find sum of x and sum of y, sum of squares of x. Next step is to find sum of product of x and y, where n is total length of x. next step is to calculate slope, M. calculate y-intercept. Residuals is obtained using the data x and y, which gives error occurred between actual and predicted values of a straight line equation.

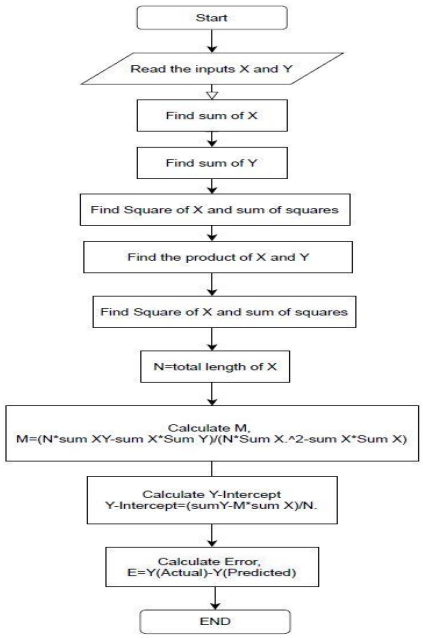

Figure 1. Flow chart of Linear Regression

3. Principal Component Analysis

Simple description of idea behind PCA, assume that you have some data points in a 2-dimensional space with two featured variables X & Y. for example if you are studying some patients data, X can be BP and Y can be the heart rate. We have these two variables representing our data points in 2-dimensional space. We have few samples and from that if you are asked to select just one of these variables X or Y, then which one of them are we going to select, X or Y. we need to select variable with Higher variance and here it seems that X has higher variance. Because, the variance of a variable is proportional to its range of variation. To perform PCA first calculate data covariance matrix of the existing data and then calculate Eigen values and Eigen vectors of the original data.

In the flow chart of basic steps of PCA, first step is to form the data as p*q matrix, where p is number of attributes and q is number of samples. Now optimize the formed data and calculate correlation matrix. From correlation matrix find Eigen values and Eigen vectors. Plot the transformed values of Principle components (Wold, 1987; Bhasha, 2020; KanagaSubaRaja, 2015; Jayanthiladevi, 2018; Sampathkumar, 2020).

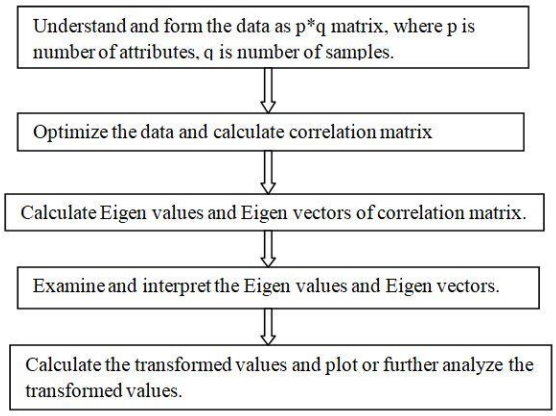

Figure 2. Flow Chart of Basic Steps of PCA

From fig 3, it seems that selecting x, will result in a higher variance and it will be a wiser choice, but if it is not limited for selecting just a single variable and instead of that we are allowed to form a new variable named Z, which is a linear combination of X and Y, then new equation becomes

$$z = C_1 x + C_2 y \text{ ---------- (13)}$$

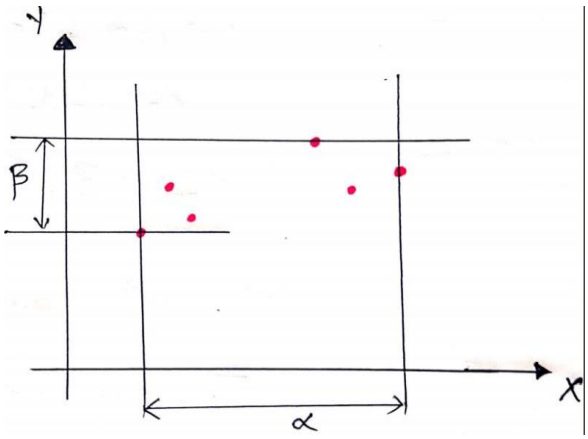

Figure 3. Plot with data points

Choose any value for c1 and c2 but with some restrictions with their sizes for which value of c1 and c2 we will have highest possible variance, for variable z and we know that selecting X & Y is a special case of general combination.

Eg: If C1=1,C2=0 then consider Z=X

And also, if C1=0, C2=1 then consider Z=Y

But question is which values of C1 and C2 result in highest possible variance for valuable Z.

First PCA problem has to be converted to optimization problem [9].

Rename, X=X1, Y=X2, C1=U1, C2=U2

$$\underset{x}{\longrightarrow} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \underset{u}{\longrightarrow} = \begin{bmatrix} u_1 \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \text{---------- (14)}$$

$$z = u.x = u^T x = x^T u \text{ ---------- (15)}$$

And problem here is to find out u1, u2---un, to maximize the variance of Z.

$$x = \left\{ \underset{x_1}{\longrightarrow}, \underset{x_2}{\longrightarrow}, \underset{x_3}{\longrightarrow} - - - - \underset{x_n}{\longrightarrow} \right\} \text{--- (16)}$$

$$\underset{x}{\longrightarrow}_i = \begin{bmatrix} \underset{x_{i1}}{\longrightarrow} \\ \cdot \\ \cdot \\ \underset{x_{iD}}{\longrightarrow} \end{bmatrix} = R^D \text{ ---------- (17)}$$

$$\underset{x}{\longrightarrow} = \frac{1}{N} \sum_{i=1}^{N} \underset{x_i}{\longrightarrow} \text{ ---------- (18)}$$

Covariance can be calculated using formula,

$$c = conv(x) = \frac{1}{N} \sum_{i=1}^{N} x_i * x_i^T \text{ ---------- (19)}$$

Equation of variance is,

$$\text{var}(z) = \frac{1}{N}\sum_{i=1}^{N}(z_i - z)^2 \text{ ---------- (20)}$$

Since, $Z_i = u^T x_i$

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N} x_i u^T \text{ ---------- (21)}$$

$$\text{var}(z) = \frac{1}{N}\sum_{i=1}^{N}\left(u^T x_i - u^T \bar{x}\right)^2 \text{ ---------- (22)}$$

$$\text{var}(z) = u^T cu \text{ ---------- (23)}$$

Maximize,$u^T cu$ subject to $u^T u = 1$. To do this we need to go for optimization technique. In this paper Lagrange multiplier is used for optimization.

$$j = u^T cu - \lambda(u^T u - 1) \text{ ---------- (24)}$$

For calculating maximum optimization condition, differentiate eq. (24) with respect to u^T and equate it to zero. Maximum condition obtained is

$$u\lambda = cu \text{ ---------- (25)}$$

Here u=Eigen vector

λ=Eigen value

Eq. (25) is the Eigen vector of covariance matrix for a sample set. But which one of the vector is the solution can be calculated by simplifying the objective function, Maximize, $var(Z) = u^T cu$

$$\text{var}(z) = u^T cu = u^T \lambda u = \lambda u^T u \text{ ---------- (26)}$$

Since, $u^T u = 1$, $var(Z) = \lambda$

Set $\lambda = \lambda_{max}$, and take Eigen vector u. if$Z = u^T x$, is a transformation function considered,

$$z = u^T \vec{x} \text{ ---------- (27)}$$

X=D*1 matrix, where D varies from 1 to D.

Z=m*1 matrix, where m varies from 1 to m.

U=C*m matrix, where c varies from 1 to c.

By choosing u, transforming vectors or columns of this transformation matrix as Eigen vectors and according to descending order of Eigen values then we ensure that two components of Z have the largest possible variances among other possible combinations.

This is the idea behind principle component analysis. So we implement PCA step by step using MATLAB code.

First creating random data points, two dimensional space and then find principal components of data generated by random generator.

An example of digits.csv data set is considered and used that data for finding principle components, x and y is the input data which we get from digits.csv, first mat lab has to read the data from file, then principle component analysis method is applied to for reducing the dimensionality of the data and summarize the larger data into smaller data sets.

3.1 Flow Chart for PCA

From the flow chart of PCA it shows steps to perform PCA. Where it is considered to be input data as digits data set. System reads input data X and Y from digits.csv file. Principle Components of inputs are calculated by

finding covariance of principle components. To check whether the points are satisfying, compute Eigen values and Eigen Vectors. After finding cumulative sum of Eigen values, plot them. Here after implementing principle component analysis to the digits data set, it is obtained that dimensionality of data is reduced and large data is summarized into smaller data sets. For example, same code can be used for implementing PCA for random numbers. By considering keyword it will read some random data. Now, find covariance of matrix of the random data and finally Eigen values and Eigen vectors are also obtained. The results of random data are shown in results and discussions.

In this paper it is also considered for iris data set and same code can be used for performing PCA for iris data set also. Using keyword, iris dataset from mat lab, it will be able to read iris data. The results of PCA using iris data set is shown in fig.5.

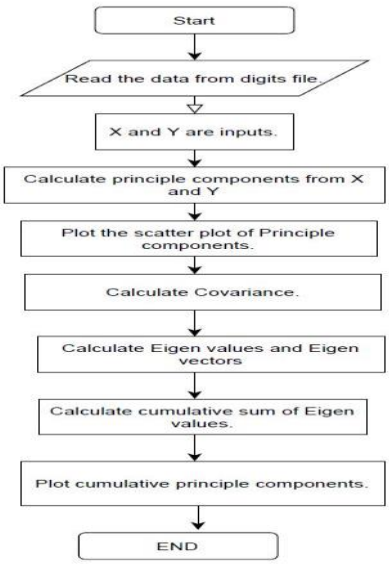

Figure 4. Flow Chart for PCA

4. Results and Discussion

Output for linear regression using least square method is shown in fig. 5. Here it is clearly shown that best fitting line using least square method will fall in maximum data points that are considered during mathematical modeling. Here error obtained is 2.821 which is very less.

All the points touching thick line, is of the original data. We get largest possible variance from data.

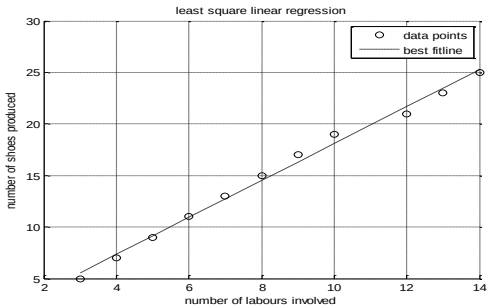

Figure 5. Scatter diagram and regression line for number of labors and shoes produced

Case 1: In the principle component analysis, three examples are taken explained how large group of data is put into one small set of data.

Fig.6, shows the plot between random data of x and y-axis, where blue dots gives original data which in large group and scattered. Green line gives how entire random original data can be put into one set of data. This result is obtained from mat lab.

Case 2: An example of iris data set is considered to perform PCA. Scattered plot of iris data is shown in fig.7. In the iris data set, 150*4 double data set is considered in x-axis and in y-axis 150*1 double data set is considered. With this we get a new variable z whose dimension becomes 150*2. So in figure it is shown with species1 (Blue), species2 (Green), species3 (Red) of an iris flower[18-21]. Figure 5 gives 2D view of pc1 and pc2, it is clearly

shown that species1 is completely separated from species2 and species3. But separating species2 and species3 is harder.

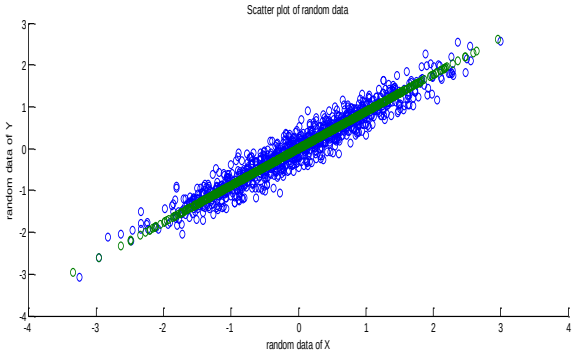

Figure 6. PCA using random data

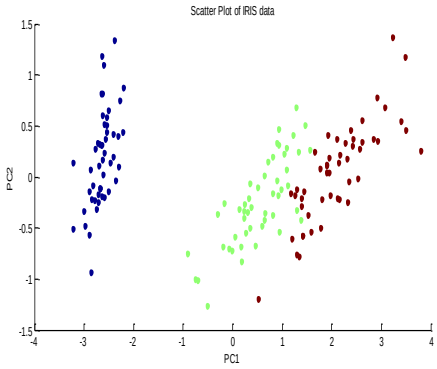

Figure 7. Scatter plot of IRIS data set

Case 3: Fig.8 shows the plot between Eigen vector components and ranked Eigen values of a digits data set which is of 1795*65 double data set. Different numbers of data are represented with different colours and a colour bar is used to show the difference.

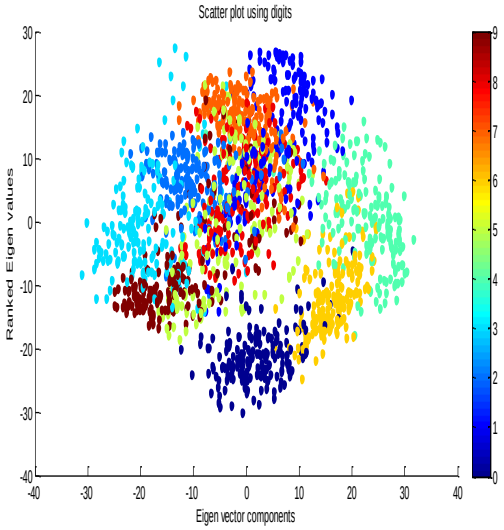

Figure 8. Scatter Plot using digits data set

By increasing the number of Eigen values and using more principal components, we can preserve more data. For example, if you want to preserve 80% of data then we should use 13 principal components and that will preserve almost 80% of original data, using 30 Eigen values and principal components, we will have 95% of original data and this is how principal component analysis works and this is how it reduces the amount of data needed to work on data set.

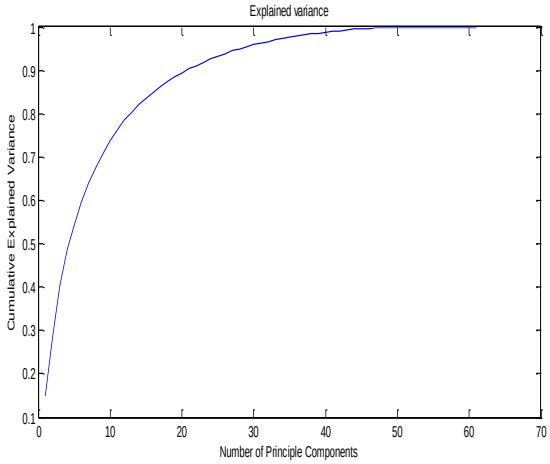

Figure 9. Plot of Cumulative PCA using Variance

5. Conclusion

So, in this paper linear regression using least square method is explained and how data fitting is done for straight line equation. By applying linear regression model for a particular data of a shoe production company, the error obtained for actual and predicted Y is 2.821 which is very less. In this paper Principle Component analysis is discussed with three different cases. In first case example is considered as random inputs, output is clearly shown in figure.6 where thick straight line is best fitting line for PCA. In second case example considered is iris data set, plot for first two components is shown and all three different species are classified, which is shown in fig.7. In third case, example of digits data set is considered, where by selecting 13 principle components it can preserve 80% of original data which is shown in figure.9. It is also shown that best principle components can be obtained by selecting Eigen values of large variance.

References

1. Arunkarthikeyan K, and Balamurugan K. (2021) Experimental Studies on Deep Cryo Treated Plus Tempered Tungsten Carbide Inserts in Turning Operation. In: Arockiarajan A., Duraiselvam M., Raju R. (eds) Advances in Industrial Automation and Smart Manufacturing. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-4739-3_26
2. Arunkarthikeyan, K. and Balamurugan, K., (2020) July. Performance improvement of Cryo treated insert on turning studies of AISI 1018 steel using Multi objective optimization. In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE) (pp. 1-4). IEEE.
3. Balamurugan K, Uthayakumar M, Sankar S, Hareesh US, Warrier KG. (2018) Preparation, characterisation and machining of LaPO4-Y2O3 composite by abrasive water jet machine. International Journal of Computer Aided Engineering and Technology, 10(6), pp.684-697.
4. Bhasha, A.C., Balamurugan, K. (2020) End mill studies on Al6061 hybrid composite prepared by ultrasonic-assisted stir casting. Multiscale and Multidiscip. Model. Exp. and Design, https://doi.org/10.1007/s41939-020-00083-1
5. Groen. P. (1996) *An introduction to total least squares", Nieuw Archief voor Wiskunde,Vierde serie, deel 14, 237–253.*
6. Garikipati P, and Balamurugan K. (2021) Abrasive Water Jet Machining Studies on AlSi 7+ 63% SiC Hybrid Composite. InAdvances in Industrial Automation and Smart Manufacturing, pp. 743-751, Springer, Singapore.
7. Huffel, S. V., Vandewalle, J. (1991*) The total least squares problems: computational aspects and analysis", SIAM Publications, Philadelphia PA.*
8. Jeffers, J. N. R., (1967) *"Two Case Studies in the Application of Principal Component Analysis" Journal of the Royal Statistical Society. Series C (Applied Statistics),16 (3) pp. 225-236*.
9. Jayanthiladevi, A., Murugan, S., Manivel, K. (2018). *Text, images, and video analytics for fog computing. In Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science (pp. 390-410). IGI Global.*

10. Khabbaz, F., Albertsson, A.-C., Karlsson, S. (1999) *Chemical and morphological changes of environmentally degradable polyethylene films exposed to thermo-oxidation", Polym. Degrad. Stab., 63, pp. 127-138,*

11. KanagaSubaRaja, S., Balaji, V. (2015) *'Smart Robot for Disaster Detection Using Zigbee Technology', International Journal of Applied Engineering Research, ISSN 0973-4562, 10 (10), pp. 27311-27320.*

12. Kevin P., Murphy. (2012) *Machine Learning A Probabilistic Perspective", ISBN 978-0-262-01802-9, Palo Alto, California,*

13. Mohan, S., Acharya, Asifa Arman, Aneeta, S. (2019) *A Comparison of Regression Models for Prediction of Graduate Admissions", Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019),* ***USB ISBN:*** *978-1-5386-9470-1.*

14. Pavan MV, and Balamurugan K. (2020) Compressive Property Examination on Poly Lactic Acid-Copper Composite Filament in Fused Deposition Model–A Green Manufacturing Process. Journal of Green Engineering.10, pp.843-852.

15. Skala,V (2008) *Barycentric Coordinates Computation in Homogeneous Coordinates", Computers & Graphics, Elsevier, , 32 (1), pp.120-127.*

16. Wold, S., Esbensen, K., Geladi. P. (1987) *Principal Component Analysis" Chemometrics and Intelligent Laboratory Systems, 2, 37-52.*

17. Sampathkumar, A., Murugan, S., Ahmed A. Elngar, Lalit Garg, Kanmani, R., Christy Jeba Malar, A. (2020) *A Novel Scheme for an IoT-Based Weather Monitoring System Using a Wireless Sensor Network." In Integration of WSN and IoT for Smart Cities, pp. 181-191. 2020.*

18. Wang, Z.Y., Chen, L.E. (2008) *Application of principal component analysis in influencing factor weighting", Mathematics in Practic,38 (4): 26-29.*

19. Zhao, J., .Li, L.M. (2014) *Comprehensive Evaluation of Robotic Global Performance Based on Principal Component Analysis and Kernel Principal Component Analysis", Journal of Beijing University of Technology, 40 (12), 763-1768 :2014*

20. Zahedan, Mahsa Salajegheh. (2015) *Study of Linear Regression Based on Least Squares and Fuzzy Least Absolutes Deviations and its Application in Geography", 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS).*