SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

MULTIPLE LINEAR REGRESSION MODEL OF UNDERGRADUATE STUDENTS' ACADEMIC PERFORMANCE AT UEAB, KENYA

Bakker Daniel. K.
University of Eastern Africa, Baraton, Kenya, Department of Mathematics, Physics & Chemistry

ABSTRACT
Generally, the University students have a vast variety of subjects in the core area which is sometimes problematic for the students to comprehend. They have chosen social media or rather internet, where there is a lot of learning materials concerning the area of their specialty. Predicting the academic performance of the students helps the instructor to develop the virtuous understanding of student's community and take the fit procedures to make their learning comfortable. MLR was used to build a model which predicts the performance of the students in the discipline of Statistics. Based on the collected data, the predictors of this model focused on how many hours spent on the social media. The predicted Cumulative GPA at every academic semester period was the dependent variable. The results revealed that the predicted model scores gives a better accuracy with M_1 *through* M_3 *in the range of* $\pm 5\%$ *from the original scores. Therefore, the instructors can analyze the performance of the students and can also extemporize the teaching techniques used based on the result, of not only Statistics students, but also other areas they handle.*
KEY WORDS: *Multivariate Linear Regression, Grade Point Average, Students, Social Media, Academics.*

1. INTRODUCTION

Introduction to Probability theory & Statistics (STAT 150) is a compulsory first year course for all students undertaking science related courses at UEAB. Poor performance in this course has been over the years seen by students not only restricted to Kenya, but to Colleges and Universities of other countries, for example, USA [3]. Other subjects for which poor performance of students have been noted and have been therefore subject of research studies. Examples of such studies are found in [6, 8 & 9].

Almost every Science student is required to take the corresponding statistics course which is highly impacted over their area which they have chosen to pursue. These courses perhaps contain all the essential and may require some basic subjects from all the Statistics. So, there is a chance of these students using internet for the various purpose including the entertainment along with the studies.

Internet has been widely accessed as an innovative style of gaining information. The amount of information available therein exceeds that of any physical library. Despite many of the college and university students using the internet, academic purpose triumphs the highest desirable reason as far as their studies are concerned.

In statistical model, multivariate linear regression is one such commonly used in the study of prediction. It is easy to use and understand since it doesn't require complicated mathematical skills for the researchers to learn. There are also several tasks that can be used to build a prediction model which includes classification, regression and categorization and also generate rules for prediction [12]. Some of the algorithms are Decision tree, Naive Bayes, KNN [2,10&11] among others.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

1.1 Problem Statement

Generally, the university-going students have an immense variety of subjects in the core area which is sometimes difficult for the students to understand. This has made students to choose internet where there is a lot of information provided concerning the area of their specialty. Internet currently, plays a major role in and around the people and makes them depend on it for each and everything. This is mainly affected by student's community where they use internet for various purpose like getting notes for studies, doing the assignment and other related activity and also for communications purposes. Therefore, predicting the student's academic performance will help the instructor to develop the virtuous understanding of student's community and take the fit procedures to make their learning comfortable.

1.2 Objective

To observe the patterns of using the internet to develop a set of multivariate linear regression models to predict the academic performance of Statistics students at UEAB based on their CGPA category.

1.3 Significance

Prediction of student's academic performance has long been observed and considered to be an important research topic in many disciplines because it benefits the educator and learners. Educators can use the predicted results to identify the number of students who will do well, averagely or poorly in a particular class to take measures accordingly.

2. LITERATURE REVIEW

Regression is a statistical method to identify the relationship between the variables present in the data. It mainly focuses on the relationship between the dependent variable and independent variables which is otherwise called as predictors. It helps to understand the changes occur in the value of dependent variable when anyone of the independent variables is changed. By using the value of the independent variable, an equation is formulated which contains the independent variables along with some coefficients and the slope value. There are lot many types of regression techniques. One such is linear regression technique which is mainly used for prediction. The linear regression is used to examine the relationship between one dependent variable and one or more independent variables.

[11] proposed a study based on the prediction of student's academic performance using data mining techniques. The study identifies the relationship between the student's academic performance and their final scores. The model was built using the SVM technique and it was compared with other classification algorithms. The final result has shown that the accuracy obtained through SVM classification is much greater than the other algorithms.

A research project done by [2] from Bulgarian University mainly focused on the usage of data mining techniques for university management. The results achieved by selected data mining algorithms for classification doesn't reveal any worthy outcomes.

[13] explored the student's demographic attributes along with their corresponding study environment which is used for the analysis of these factors affecting their success rates in their course of the study. The results show that the important factors to distinguish between successful and unsuccessful students and for predicting the category of students, the CART algorithm is used which produce an overall percentage of 60.5%. It does not contain adequate information for distinguishing between successful and unsuccessful students.

Some authors explored the difference between data mining techniques [1,4,10] and explored the comparison of the methods for educational learners and provided a better predictive model among all the data mining techniques. Some researchers focused on the use of classification algorithms and provided the comparison of all the classifiers in their paper [5].

In all these works, the authors concentrated on the students' performance prediction using different data mining techniques to carry out the analysis but the work mainly focuses on building and validating the regression technique of undergraduate students of statistics stream who use the internet for various purposes.

The simple linear regression with one dependent variable and one independent variable is given in equation (2.1) as:

$$y = \beta_0 + \beta_1 x + \xi \qquad (2.1)$$

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

where y is the predicted dependent variable value , β_0 is a constant, β_1 represent the coefficient of the regression, and x is the independent variable and ξ is the error term. For our study, the multiple linear regression was used amidst several types of linear regression techniques like: Simple linear regression, Multiple linear regression, Logistic regression, Ordinal regression, Multinomial regression and Discriminant analysis. Multiple linear regression is used for one dependent variable and more than two independent variables containing dataset.

The multiple linear regression model is written as a straightforward extension of the simple linear model given in equation (2.1). The model is specified in equation (2.2) below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \xi \tag{2.2}$$

where y is the dependent variable; $x_j, j = 1,2,3,...,m,$ represent m different independent variables; β_0 is the intercept (value when all the independent variables are 0); $\beta_j, j = 1,2,3,...,m,$ represent the corresponding m regression coefficients and ξ is the random error, usually assumed to be normally distributed with mean zero and variance σ^2.

To estimate the regression coefficients, we use a set of n observed values on the $(m+1)-$tuple $(x_1, x_2, ..., x_m, y)$ and use the least squares principle to obtain the following equation (2.3) for estimating the mean of y.

$$\hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_m x_m \tag{2.3}$$

The least squares principle specifies that the estimates, $\hat{\beta}_i$'s, minimizes the error sum of squares

$$SSE = \Sigma\left(y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - ... - \hat{\beta}_m x_m\right)^2 \tag{2.4}$$

For convenience we redefine the model

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \xi \tag{2.5}$$

where x_0 is a variable that has the value 1 for all observations. Obviously, the model is not changed by this definition, but the redefinition makes β_0 look like any other coefficient, which simplifies the computations in the estimation procedure. The *ESS* to be minimized is now written as:

$$SSE = \Sigma\left(y - \hat{\beta}_0 x_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - ... - \hat{\beta}_m x_m\right)^2 \tag{2.6}$$

The solutions to these **normal equations** provide the least squares estimates of the coefficients, which we have already denoted by $\hat{\beta}_0, ..., \hat{\beta}_m$.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

$$\begin{array}{ccccccccccc}
\beta_0 n & + & \beta_1 \Sigma x_1 & + & \beta_2 \Sigma x_2 & + & \cdots & + & \beta_m \Sigma x_m & = & \Sigma y \\
\beta_0 \Sigma x_1 & + & \beta_1 \Sigma x_1^{\,2} & + & \beta_2 \Sigma x_1 x_2 & + & \cdots & + & \beta_m \Sigma x_1 x_m & = & \Sigma x_1 y \\
\beta_0 \Sigma x_2 & + & \beta_1 \Sigma x_2 x_1 & + & \beta_2 \Sigma x_2^{\,2} & + & \cdots & + & \beta_m \Sigma x_2 x_m & = & \Sigma x_2 y \\
\cdot & & \cdot & & \cdot & & \cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & & \cdot & & \cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & & \cdot & & \cdot & & \cdot & \cdot & \cdot \\
\beta_0 \Sigma x_m & + & \beta_1 \Sigma x_m x_1 & + & \beta_2 \Sigma x_m x_2 & + & \cdots & + & \beta_0 \Sigma x_m^{\,2} & = & \Sigma x_m y
\end{array} \qquad (2.7)$$

Because of the large number of equations and variables, it is not possible to obtain simple formulas that directly compute the estimates of the coefficients as for the simple linear regression model. In other words, the system of equations must be specifically solved for each application of this method. Although procedures are available for performing this task with handheld or desk calculators, the solution is almost always obtained by computers. We will, however, need to represent symbolically the solutions to the set of equations. This is done with matrices and matrix notation.

To show the solution procedure using matrix notation for the general case and numerically, define the matrices **X**, **Y**, **E** and **B** as follows:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \; Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \; E = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \cdot \\ \cdot \\ \xi_n \end{bmatrix}, \text{ and } B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix} \qquad (2.8)$$

Where x_{ij} represents the i^{th} observation, $i = 1, 2, ..., n$, of the j^{th} independent variable, $j = 1, 2, ..., m$. Using these matrices, the model equation for all observations,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \xi \qquad (2.9)$$

can be expressed as

$$Y = XB + E \qquad (2.10)$$

The first column of the matrix **X** is a column of ones, used as the "variable" corresponding to the intercept. Using matrix notation, we can express the normal equations as:

$$\left(X'X\right)\hat{B} = X'Y, \qquad (2.11)$$

where $\hat{B}$ is a vector of least squares estimates of **B**.

The solution to the matrix equation is written as:

$$\hat{B} = \left(X'X\right)^{-1} X'Y. \qquad (2.12)$$

These expressions are valid for a multiple regression with any number of independent variables. That is, for a regression with *m* independent variables, the **X** matrix has *n* rows and *(m + 1)* columns. Consequently, the matrices **B** and $X'Y$ are of order $\{(m+1)\times 1\}$, and $X'X$ and $\left(X'X\right)^{-1}$ are of order $\{(m + 1) \times (m + 1)\}$.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

The procedure for obtaining the estimates of the parameters of a multiple regression model is a straightforward application of matrix algebra for the solution of a set of linear equations. To apply the procedure, first compute the $X'X$ matrix as follows:

$$X'X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 & \cdots & \Sigma x_m \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 & \cdots & \Sigma x_1 x_m \\ \Sigma x_2 & \Sigma x_2 x_1 & \Sigma x_2^2 & \cdots & \Sigma x_2 x_m \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \Sigma x_m & \Sigma x_m x_1 & \Sigma x_m x_2 & \cdots & \Sigma x_m^2 \end{bmatrix} \quad (2.13)$$

that is, the matrix of sums of squares and cross products of all the independent variables.

2.1 Conceptual Framework

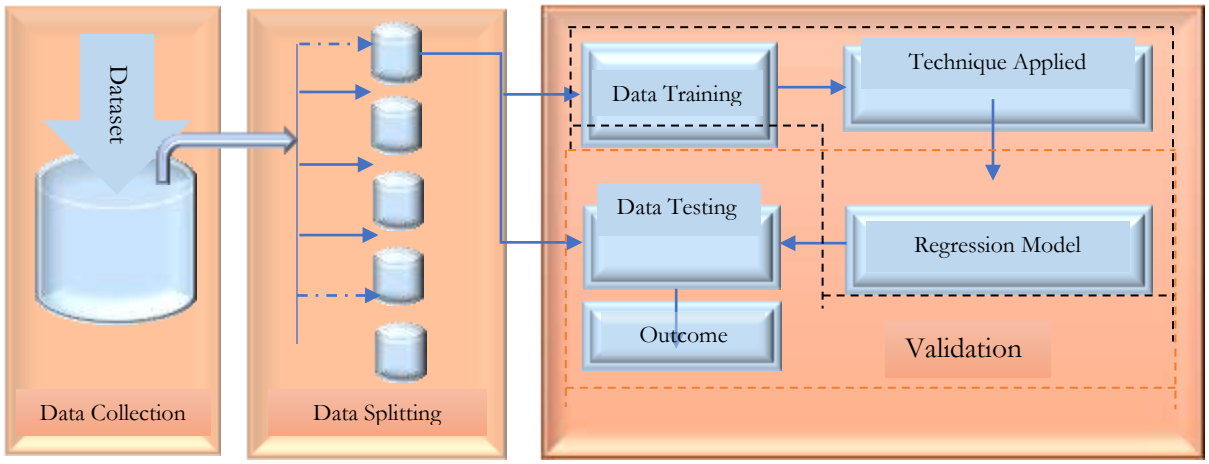

Figure 2.1: Predictive Model Framework

3. RESEARCH DESIGN

The questionnaire method coupled with extraction of information from the respective student's record in the department of Mathematics at the UEAB was the research instrument used. The test-re-test reliability of the questionnaire was tested for content validity before administering it to students. Data was then selected and split as training and testing datasets. After splitting the dataset, the multivariate linear regression technique was applied on the training dataset and this regression equation was used for prediction purposes by passing the test dataset. Based on the categories of training dataset, the models would be predicted as summarized in figure (2.1).

3.1 The Data and Variables used

The data used were derived from the departmental examination record office and the questionnaire administered to students using appropriate numerical codes. Understanding the collected data and performance of the analysis was done using the Minitab version 18 software. The dependent variable y is the CGPA. Independent variables were as follows: academic usage purposes, x_1; entertainment usage purposes, x_2; communication usage purposes, x_3; time period spent on social media, x_4 and cumulative intermate usage before end of semester was x_5.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

3.2 Target Population

The target population consists of 200 series level students of Mathematics and Statistics and a total of 120 students taking statistics courses were included and their corresponding semesters. The main study sample size was derived using the sample table guide for sample size decisions provided by [7]. The response rate is 100% because concerted efforts were made to retrieve the questionnaire from students. Demographically, 40.7% and 59.3% were the respondents per the study year. 16% had a CGPA between 3.33 and 4.00, 37.3% between 2.67 and 3.32 and 46.7% between 2.25 and 2.66. out of these 110 respondents, 33.3% were males while females were 66.7%.

4. ANALYSIS OF RESULTS

The multivariate linear regression technique was employed to develop three predictive models
based on the training dataset collected. The statistical formula of each predictive model from the table is expressed as follows:

Model 1 (M_1)		Model 2 (M_2)		Model 3 (M_3)	
CGPA	**B**	**CGPA**	**B**	**CGPA**	**B**
Constant	0.131	Constant	0.421	Constant	0.199
X_1	0.756	X_1	0.607	X_1	0.712
X_2	-0.100	X_2	-0.120	X_2	-0.103
X_3	-0.128	X_3	-0.141	X_3	0.009
X_4	0.011	X_4	0.122	X_4	0.210
X_5	0.152	X_5	0.201	X_5	0.103

Table 4.1: Model's Summary

Model # 1:

$$y_1' = 0.131 + .756x_1 - .100x_2 - .128x_3 - .011x_4 - .152x_5 \quad (4.1)$$

Model # 2:

$$y_2' = 0.420 + .607x_1 - .120x_2 - .141x_3 + .122x_4 + .201x_5 \quad (4.2)$$

Model # 3:

$$y_3' = 0.199 + .712x_1 - .103x_2 + .009x_3 + .210x_4 + .103x_5 \quad (4.3)$$

There is a variance with the presence of independent variables for each and every category of the students. To validate the model and check its accuracy, the dependent variable can be calculated, from the projected models, by passing the test data values (i.e. for sample data sets:

$M_1 : x_1 = 4; x_2 = 4; x_3 = 5; x_4 = 3; x_5 = 3$; $M_2 : x_1 = 7; x_2 = 4; x_3 = 3; x_4 = 2; x_5 = 1$

and $M_3 : x_1 = 4; x_2 = 2; x_3 = 1; x_4 = 3; x_5 = 2$), substituting these values into the prediction equation (4.1) above, we get:

$$y_1' = 0.131 + .756(4) - .100(4) - .128(5) - .011(3) - .152(3) = 2.604 \quad (4.4)$$

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

A deviation of 0.19 from the original observation of $y_1 = 2.794$. Similarly, following the same procedure as equation (4.4), a deviation of 0.21 from the original $y_2 = 3.03$ from equation (4.2) and lastly the original observed $y_3 = 3.701$ with a predicted value of $y_3' = 3.791$ from equation (4.3).

4.1 Model Evaluation

Each of the predictive model was evaluated by the RMSE. It gives the standard deviation of the model prediction error where a smaller value indicates a better model performance. The model's residuals are considered and the RMSE calculated for the three models were as shown in figure (4.1). From the figure, the values are slightly deviated from the original regression line. Therefore, any choice of the model can be used to test the data and predict the values.

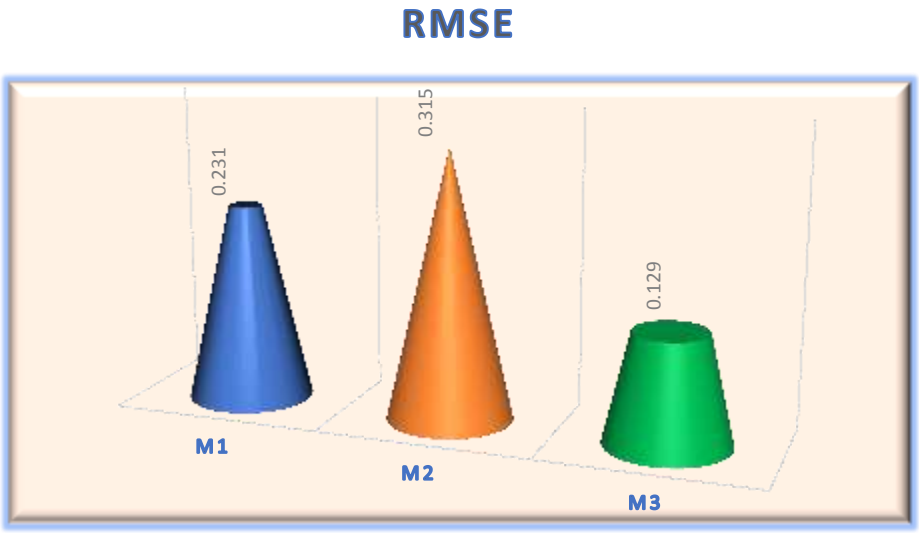

Figure 4.1: RMSE Evaluation

5. CONCLUSION AND RECOMMENDATIONS

Predicting student's performance would give teachers a better approach for teaching and advising the students and therefore boosting the results of they who are at risk of failure. Multivariate linear regression technique was used in building the model. From the model predicted, it shows how the value of the dependent variable differs based on the value of the independent variables. The educators can analyze the performance of the class and can also improvise the teaching techniques used based on the result of each category of the students. The results revealed that the predicted model scores gives a better accuracy with M_1 through M_3 in the range of $\pm 5\%$ from the original scores.

These predictive models developed in this research were based on the data collected at our private university. The developed models can be employed as a general tool to predict student academic performance in other courses especially the mathematics related ones, so they can benefit both teaching and learning. When extending the regression technique to another institution of higher learning, it is suggested to collect the data on student academic performance at that particular institution to develop a corresponding regression model. This will ensure that the regression model best represents teaching and learning at that particular institution.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 9 | September 2020 - Peer Reviewed Journal

REFERENCES

1. *Amirah M. S., W. H. (2015). A Review on Predicting Student's Performance using Data Mining Techniques . Elsevier , 414-422.*
2. *Dorina K. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms . International Journal of Computer Science and Management Research, 686-690.*
3. *Ferrini M. J. & Graham, K. G. (1991). An Overview of the Calculus Curriculum Reform Effort: Issues for Learning, Teaching & Curriculumn Development. The American Mathematical Monthly, 627-635.*
4. *Geraldine G., M. C. (2014). An application of classification models to predict learner progression in tertiary education . IEEE Explorer, 549-554.*
5. *Hilal A. (2017). Analysis of Students' Performance by Using Different Data Mining Classifiers. Modern Education and Computer Science, 9-15.*
6. *Iduseri, A. (2011). Descriptive Discriminant Analysis of English Language Proficiency and Academic Performace. Nigerian Annals of Natural Sciences, 11(1) 75-82.*
7. *Krejcie, R. &. (1970). Determining Sample Size for Research Activities. Educational and Phychological Measurement, 607-610.*
8. *Lauren, E. S. (2008). The Persistence of the Gender Gap in Introductory Physics. Physics Education Research Conference, 139-142.*
9. *Mamta, S. (2014). Students Enrolled in Biology Majors Pre-Requisite Courses: Effect of High School Academic Performance. Journal of Educational Research, 2(4) 183-188.*
10. *Norlida B., Usamah M., & Pauziah M. A., (2015). "Educational Data Mining for Prediction and Classification of Engineering Students Achievement . IEEE Explorer, 49-53.*
11. *Oloruntoba S. A., A. J. (2017). Student Academic Performance Prediction Using Support Vector Machine. International Journal of Engineering Sciences & Research Technology , 588-598.*
12. *Sivagowry .S, D. M. (2013). An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease. IEEE Explorer.*
13. *Zlatko J. K. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of InSITE, 647-665.*