Application Of Machine Learning (Linear Regression Model) To Predict Students Enrollment Among Senior High Schools In Ghana

Osei Wusu Brempong Jnr

Abstract: The mission of the Ghana Education Service (GES) [1], is to ensure that all Ghanaian children of school-going age are provided with inclusive and equitable quality formal education and training through effective and efficient management of resources to make education delivery relevant to the manpower needs of the nation. The GES uses a computerized school selection and placement system that assigns senior high schools to students based on their test scores from previous Junior secondary schools. The computerized school selection and placement system (CSSPS) uses a deferred acceptance algorithm for each school assignment. Under these procedure students are ranked according to their priority levels (that is Test scores in the case of the CSSPS) [2]; they are then proposed as a match to their first-choice school in order of their test score rankings. Students are assigned to their first choice if there is a space available in the schools. What the CSSPS failed in doing is to determine the average number of students that can be allocated to each school during the placement selection process. The objective of this research is to use machine learning (linear regression) to predict the increase in student enrollment for schools in each region based on the school's average test score performance i.e. average GPA from the previous year. Firstly, we investigate the relationship between increase in student's applications for each school and the school average GPA (test score) from previous year. A sample dataset from 10 senior high schools in the capital region of Ghana was used for this research. Supervise machine learning models and associated algorithm (simple linear regression) to analyze data for regression helped in training the model to predict the increase in student's enrollments for each school based on the school average GPA (test scores)

Key words: Machine learning, simple linear regression, prediction, enrollment, GPA

1 INTRODUCTION

In 2017 Ghana introduced a free senior high school (free SHS) policy. The policy core seeks to provide quality and equal free secondary education for all to fulfill the United Nations modified sustainable development Goals, where member countries amalgamate free quality and equality in their educational systems to certify adequate learning experiences for students. Goal 4 of the United Nations Sustainable Development Goals (SDGs) states: "By 2030, ensure that all girls and boys complete free equitable and quality primary and secondary education leading to relevant and effective learning outcomes." [3]. The Ghana government and ministry of education prioritized ensuring that education is made free from basic to secondary to afford more children in Ghana the opportunity to access quality education. After the implementation of free SHS in 2017, that year, there was an 11% increase in enrolment, breaking records from previous years. In the 2017/18 academic year, a new record was set with the highest enrolment ever seen in the country: over 470,000 students enrolled in senior high school. The policy has provided hope to children who otherwise could not have furthered their education after junior high school mostly due to cost barriers. In the 2016/2017 academic year, the number of children who qualified and had been placed in senior high school but could not enroll stood at 62,453 out of a total of 420,134 who were placed. However, in 2017/18 academic year, following the introduction of the policy, this dropped to 11,366 out of 424,224 who were placed in the senior high schools. Therefore, more children were accessing senior high school. The number of increased students enrolled each year to access policy is projected to skyrocket, and this intends to drain the efficiency of the CSSPS (computerized school selection and placement system) in Ghana.

- *Osei Wusu Brempong Jnr is currently pursuing PhD degree program in Computer Science and Technology in Dalian Maritime University, in Dalian China. kobeoseijnr@hotmail.*

The scramble for best students from junior schools by the SHS schools within the regions is exposing the CSSPS efficiency as the systems fail to predict the average increase in enrolled into the SHS. Student's selection for their specific SHS is based on the performance of the SHS previous academic year (average test scores) [4]. This means that if school 'X' performed better than school 'Z' and 'Y' during the previous academic year exams, more students will be attracted to school 'X', new students will be encouraged to choose school 'X' as their first choice for SHS. This as a result will automatically increase the number of student enrollment for school 'X' for that year. If school 'X' can predict the average number of student's enrollment expected base on its previous year's average test score, it will prepare well in advance in terms of infrastructure and teaching equipment's to accommodate for the increase in students, hence maintaining the quality of education. Machine learning [5] a subset of Artificial Intelligence provides systems the ability to automatically learn and improve from experience without actually being programmed. ML also focuses on the development of computer programs that can access the data and use it to learn by itself. ML algorithms is classified mainly into 3 categories. 1. Supervised machine learning algorithm 2. Unsupervised machine. learning algorithm 3. Semi-supervised machine learning algorithm. With supervised, learning we use label data (classified dataset) to infer a learning algorithm. This data is used as basis for predicting the classification of other unlabeled data through the use of machine learning algorithm. Linear regression is a machine learning algorithm based on supervised learning. It performs regression tasks. Regression models a target prediction value based on independent variables. It's mostly used for prediction and finding the relationship between variables and forecasting.

2 LITERATURE SURVEY

Ashutosh Nandeshwar and Subodh Chaudhari, 2009 researched on enrollment prediction models using data mining. The research was to build models to predict enrollment

using the student's admissions and to evaluate the model using cross-validation, win-loss table, and quartile charts. CRISP-DM version 1.0 created by Daimler Chrysler, SPSS, NCR in 1996 was used for the research. CRISP is a non-property, free available application standard for data mining. CRISP was the standard used as the base of the research and they created data mining models using Weka, which is an open-source software collection of machine learning algorithms for data mining tasks. In addition, they used MS-Access to import the flat files in database format, modifying and creating new fields and converting Access tables to ARFF using VBA. They concluded after their research overall financial aid was the most important factor that attracted students to enroll. They suggested students enrolled at institutes only if they received some sort of financial aid regardless of the school average GPA scores. The disadvantage of their research was the CRIPS software they used, which has been proven to be an Iterate failure. This is major factor because as the model ages, they need to be kept up to date to be valuable. The Data patterns that drove the model can also change and undermine the value of the model. Failure to Iterate of CRIPS software is a disadvantage to this researchIn 2014, Anal Acharya and colleagues proposed using classification; a machine learning technique for early prediction of student's performance. The Data set of the study research was derived from sets of students majoring in computer science. The course curriculum was divided into six semesters. In each semester a student has to appear in two examinations: a mid-semester examination and a term-end examination. Since the objective of the study was to perform early prediction of student performance, semester1 results was used for prediction purposes. Their research aim was to apply machine learning algorithms for the prediction of student results. For this purpose, a suitable representative from five classes of MLA were first trained and then tested. They were found to yield useful results. They noted that the developed research methodology is general in nature: and it could be applied to full-time as well as distance education courses including the courses where web based learning is used. The training set used contains 309 instances whereas the testing set contains 104 instances. After detailed analysis, it was found that DTs are the most convenient algorithm to generate the set of production rules. Accordingly, C4.5 was used to generate the decision tree. F-Measure and Kappa Statistic were used to determine the efficiency of the prediction algorithm. The average F-Measure value for the training dataset was found to be 0.79 whereas for the testing dataset was found to be 0.66. Based on their research, the proposed methodology used needed major improvement on several accounts. Firstly, several students who took admission in a course appeared in the mid-semester exams but due to certain reasons was unable to appear in the term end exams. These students should not have been considered for prediction as they contain missing attributes. Also, the efficiency of prediction could have been increased by Combining Multiple Classifiers (CMC). Genetic algorithms may also be applied for this purpose

2.1 WORK – METHODOLOGY

SIMPLE LINEAR REGRESSION MODEL

Simple linear regression model is a machine learning supervised model that models with a single regressor x which has a relationship with response y that is a straight line. [6]

This simple linear regression coefficient bˆ

$$y = \beta_0 + \beta_1 + \epsilon$$

Where the intercept β_0 and the slope β_1 are unknown constants and ε is a random error component. The errors are assumed to have a mean zero and unknown variance of ∂^2. We also assumed that the errors are uncorrelated which means the value of one error does not depend on the value of any other error. The simple linear regression equation is also called the least squares regression equation. It tells the criterion used to select the best fitting line, namely the sum of the squares of the residuals should be the least. The least-square regression equation is the line for which the sum of squares residuals is a minimum

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$\sum_{i=1}^{n}(y_i - \hat{y}_i)$ is a minimum.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

2.2 Collection of Data – Retrospective study

A retrospective study based on historic data analysis was used for data collection. The GES (Ghana Education Service) recorded test score archives for each Senior High School in the 16 regions of Ghana. Out of the 16 regions, 26 schools were recorded. The model sampled 26 schools from 8 regions that all benefitted from the free senior high school policy. The school ranking, average increase enrollment as well as average GPA scores from 2017-2019

Figure 1. GES records of school's GPA and average enrollment

2.3 Model formation and estimation

The least-squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve Suppose for any observation,

$$(x_1, y_1,), (y_2 x_2), \ldots,, (x_i y_i)$$

be a pair of random variables. To predict y, the parameters β_0 and β_1 minimize the expression[7]

$$\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_{1x_i}])^2$$

straight line formula maybe written as:

$$y = \beta_0 + \beta_{1x_i} + \varepsilon_i\,, i = 1,2,\dots,n \quad (1)$$

the coefficient that minimize the square of the distance between the line and the point is given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \quad (2)$$

Where

$$\beta_{1=} \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \quad (3)$$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ AND $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Are the averages for y_i and x_i respectively. Therefore, β_0 and β_1 are the least square estimators of the intercept and slope. The residuals ε are the difference between the observed and the predicted values

$y_i - \hat{y}_{i.\ i=1.2\dots n}$. The fitted simple linear regression line is then given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \quad (4)$$

In addition to estimating the β_0 and β_1, an estimate of the ∂^2 is made to test hypothesis and interval estimate of the regression model. The estimation of ∂^2 is obtained from the residual or error sum of squares

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \quad (5)$$

The coefficient of determination which is in percentage form, indicates how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted. The coefficient of determination R^2 is given by

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

Where $0 \le R^2 \le 1$. R^2 is often called the proportion of variation explained by the regressor x. Values of R^2 that are closed to 1 imply that most of the variability in y is explained by the regression model.

The correlation coefficient r measures the strength of the relationship between the two variables. [8] It evaluates the goodness of the fitting of data considered and the standard of error measured, s is calculated. The correlation coefficient value can vary in the range -1 and +1, for the strong correlation between the variables x and y. If the value is zero, then there is no linear correlation between the variables. The calculation of r and s is as follows:

(7)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(8)

$$s = \frac{\sum_{i=1}^{n} y_i^2 - \beta_0 \sum_{i=1}^{n} y_i - \beta_i \sum_{i=1}^{n} x_i y_i}{n}$$

.

2.4 Model evaluation and prediction

Two alternative methods are used for testing whether a linear association exists between the predictor x and response y in this simple linear regression model. First to estimate the regression line and use the t-test to determine if the slope β_1 of the population regression line is 0. Alternatively, perform an (analysis of variance) F-test.

$$H_0: \beta_1 = 0 \; versus \; H_A: \beta_1 \neq 0$$

An α-level hypothesis test for the slope parameter β_1 will follow the standard hypotheses test procedures in computing a hypothesis test for the slope β_1, null and alternative hypotheses is specified: [9]

NULL hypothesis $H_0: \beta_1$ = some number β

Alternative hypothesis $H_A: \beta_1 \neq$ some number β

This means that it can test whether or not the population slope takes on any value. The value of the test statistics is calculated as follows:

(9)

$$t^* = \frac{b_1 - \beta}{\left(\frac{\sqrt{MSE}}{\sqrt{\Sigma(X_i - \bar{x})^2}}\right)} = \frac{b_1 - \beta}{se(b_1)}$$

The result of the test statistic calculates the P-value. The P-value is determined by referring to a t-distribution with n-2 degrees of freedom. The decision finally made if the P-value is smaller than the significance level α, we reject the NULL hypothesis in favor of the alternative. If the P-value is larger than the significance level α, we fail to reject the NULL hypothesis. In addition to t-test, we may also obtain confidence interval to estimate parameters of β_1, β_0, δ^2. The width of these confidence intervals is a measure of the overall quality of the regression line, this results don't only give us range of values that is likely to contain the true unknown value of β_1, it also allows us to estimate if the predictor x linearly related to the response y. If the confidence interval for β_1 contains 0, then it can be concluded that there is no evidence of a linear relationship between the predictor x and y response in the population. But if the confidence interval for β_1 does not contain 0, then conclude that there is evidence of a linear relationship between the predictor x and the response y in the population. Formula for confidence interval is as follows: [10]

$$b_1 \pm t_{\left(\frac{\alpha}{2},n-2\right)} * \left(\frac{\sqrt{MSE}}{\sqrt{\sum(X_i-\bar{x})^2}}\right) \quad (10)$$

The confident value for a new response μy when the predicted value is x_h is given as follows:

$$\hat{y}_h \pm t_{\left(\frac{\alpha}{2},n-2\right)} * \sqrt{MSE * \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)} \quad (11)$$

$\hat{y}_h$ is the predicted value of the response when the predictor is x_h,

$t_{\left(\frac{\alpha}{2},n-2\right)}$ is the t-multiplier. The t –multiplier has n-2 degrees of freedom, because the prediction interval uses the mean square error (MSE) whose denominator is n-2.

$\sqrt{MSE * \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$ is the standard error of the prediction, which depends on the mean square error (MSE), the sample size (n), how far the squared units the predictor value is from the average of the predictor values $\bar{x}$, and the sum of squared distances of the predictor values x_i from the average predictor values $\bar{x}$.

3 RESULTS AND DISCUSSION

Based on the records of GES, the senior high schools average increase in enrollment due to the average GPA performance recorded for each school between 2017-2019, a simple linear regression can be modeled to estimate the relationship between the school average GPAs / enrollment and construct future prediction. Chart of GES records indicating senior school enrollment and GPA is shown in figure 1 and figure respectively.

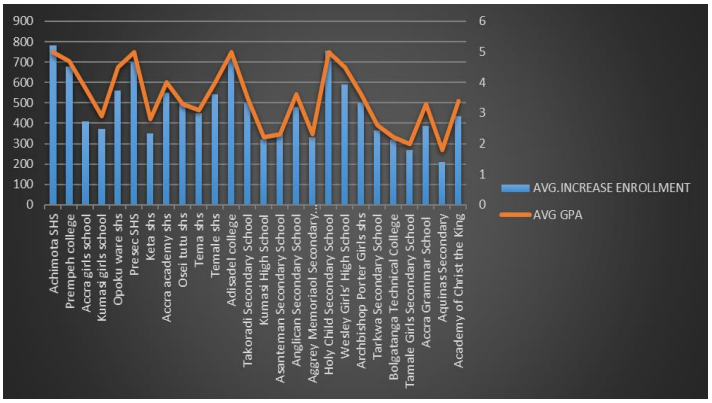

FIGURE 2. GES RECORDS FOR 2017-2019

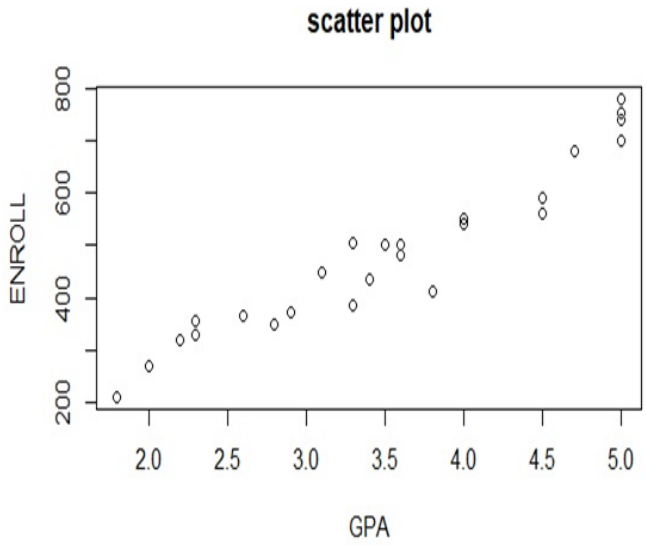

Figure 3. scatter plot of enrollment vs GPA

This dataset of size n=26 is sampled from the 120 senior high schools in 4 regions in Ghana. The variables are y = average increase in enrollment from 2017-2019 and x = west Africa senior school certificate examination from 2017-2019 converted to average GPA. The WASCE is a type of standardized test in West Africa, mandated for every senior high school in Ghana. The plot data above (enroll on the y) shows a generally linear relationship on the average, with a positive slope. As the average GPA scores for each school increases, the average increase in enrollment for each year rises

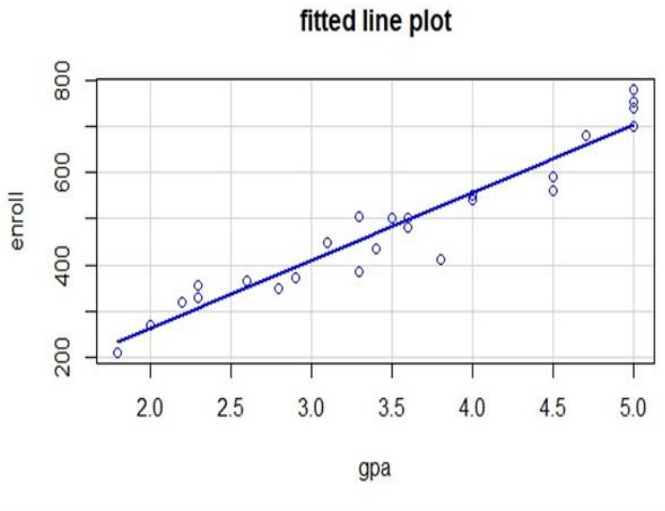

Figure 4, fitted line on scatter plot

The equation of the fitted regression line is given as enroll = -31.940 +146.885 GPA. This equation should indicate that if is for the "average" enrollment or "predicted" enrollment it would be okay. Because since a regression equation describes the average value of y as a function of one or more x variables. Equation: y=-31.940+146.885xThe interpretation of the slope (value = 146.885) is that from 2017-2019 the enrollment rate increased by 146.885 students, on average for each 1.0 GPA increase of school average GPA score. The interpretation of the intercept (value = -31.940) if there is a school with average GPA = 0, the predicted enrollment for that school will be -31.90 students. Since no school is likely to have an average GPA = 0, this interpretation of the intercept is not practically meaningful. In the fitted line plot graph, s= 44.2 and r^2 = 92.3%. The value of s tells us roughly the standard deviation of the differences between the y-values of individual observations and predictions of y based on the regression line. The r^2 can be interpreted to mean that GPA rates 'explain' 92.3% of the observed variation in an average increase in enrollment of students between 2017-2019 due to increase GPA.

Assumption and Model Adequacy

Estimating the model adequacy and assumption of the regression analysis gives a residual analysis defined as

$$e_i = y_i - \hat{y}_i, i = 1,2, \dots n$$

Where y_i is an observation and $\hat{y}_i$ is the corresponding predicted values. Analysis of residuals is an effective method for discovering several type of model deficiencies.

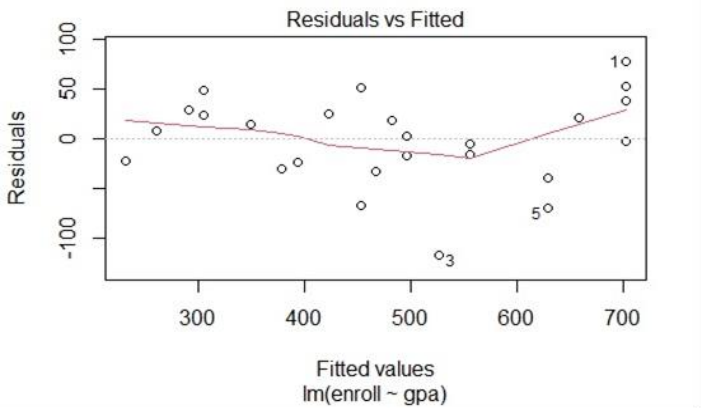

Figure 5, graph of residual plot

The above plot of residuals versus fitted values after a straight –line model was used on data y= enroll and x = GPA, for n= 26 senior high schools. This looks okay in that the variance is roughly and almost the same all the way across and there are no worrisome patterns except for the 3 outliers which can be ignored or deleted for better interpretation.

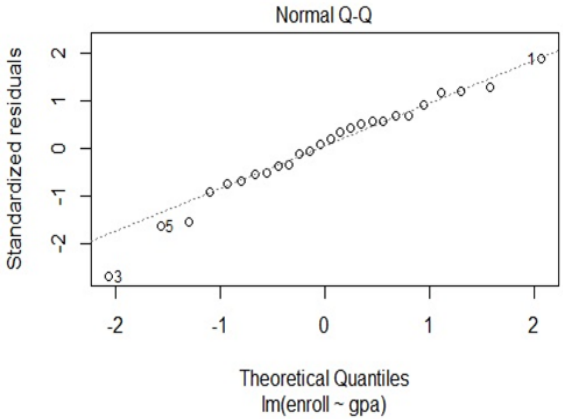

Figure 6. normal QQ plot

Above is a normal probability plot, the pattern of the normal probability plot is almost a straight line which gives evidence that it is reasonable to assume the errors have a normal distribution. Therefore, residual plots are satisfactory and the model is adequate.

4 CONCLUSION

According to the above results, this simple linear regression model can be used to estimate the average student enrollment for each senior high school and predict future enrollments based on their previous year's GPA scores. The linear correlation coefficient value between the GPA and enrollment is 0.9238, this indicates a strong relationship between the two. The ability of the Ghana Education System (GES) to predict student enrollment for each academic year is vital for the continual effectiveness of the Free Senior High policy by enhancing its quality and ensuring that schools are fully equipped with funds and facilities to support them during the

academic year. WASSC(GPA) results are released in the fall each year. This means that with the influx of new students enrolling, GES and senior high school's administrators have approximately six months to adequately assemble needed infrastructures, equipment's and funds for the next academic year. GES can access this model to effectively predict average student's enrollment based on previous academic WASSCE scores, this will enable them to budget in advance for funds disbursement to each school. Schools administrators can also activate this model to predict number of new students enrolling for the next academic year and this will help them prepare in time if more equipment's or infrastructures are needed to accommodate them.

Example 1,
Based on enrolling, GPA chart in figure 2. Accra girls school had GPA of 3.8 before the 2017-2018 academic year, the average increase in enrollment for the 2019 academic year was 410 students. If Accra girls school average GPA increases to 5.0, based on this model, we will be 92% confident that their average student enrolment will be between 670.043 and 734.93. An average increase in enrollment of about 300 students. This means with Accra girl's school's new GPA of 5.0 score, the school expectance of more students subscribing to attend their institute for the next academic year will require extra additional funding and infrastructure to accommodate them. GES predicting this new increase number will budget ahead for this.

Example 2.
If Accra girls school drops in average GPA from 3.8 to 2.0, the model will be 92% confident that the average student's enrollment will be between 230 and 293 students. This is a sharp decline of about 148 students to enroll based on their low average GPA score. In this case fewer funds from GES to Accra girls school for the next academic year, the school will have little preparation in terms of facilities and equipment's.

REFERENCES

[1] Ghana education service, GES. (2020 February 1). Enabling an effective teaching and learning environment. Ministry of education. https://ges.gov.gh/about-us/
[2] Pearl Adiza Babah, Agyemang Frimpong, Ronald Osei Mensah, Andrews Acquah .(2020). 'Computerized School Selection and Placement System in Ghana: Challenges and The Way Forward'. European journal of education science. Vol.7 No.2 ISSN: 1857- 6036
[3] Ministry of education, MOE. (2017). changing Ghana through education. Government of Ghana. https://moe.gov.gh/free-shs-policy/#
[4] Amedahe, F.K; & Asamoah-Gyimah, E. (2005). Introduction to educational research. Cape Coast: Centre for Continuing Education of the University of Cape Coast (CCEUCC)
[5] Prakrteswar santikary, (sept 19 2019). Artificial intelligence and machine learning. ERT. https://www.ert.com/blog/artificial-intelligence-and-machine-learning-part-1-definitions-similarities-and-differences/
[6] Sanford Weisberg," Applied Linear Regression"4TH ed. John Wiley & Sons,pp.21-59,2014
[7] Montgomery, D. C. and Peck, E. A "Introduction to linear regression analysis" 5th ed. Wiley. New York, pp.12-58, 2012.
[8] Barnett, V. and Lewis, T, "Outliers in Statistical Data" Wiley and son, New York, 1994.
[9] Belsley, D. A., Kuh, E. and Welsch, R. E. "Regression Diagnostics: Identifying influential data and sources of collinearity" John Wiley & Sons, New York, 1980.
[10] Chatterjee. S and Hadi. A. S. "Influential Observations, High Leverage Points, and Outliers in Linear Regression" Statistical Science, Vol. 1, No. 3, pp. 379-393, 1986