**University of Ibadan**
**Journal of Science and Logics in ICT Research**

# Predicting Students' Graduating Cumulative Grade Point Average Using Difference Level, Classification and Regression Tree and Linear Regression Algorithm

**Azeez, T.O.[1], Awe, A. C.[1] and Omosebi, P. A.[2]**

**[1]Department of Computer Science,**
**College of Natural Sciences,**
**Joseph Ayo Babalola University,**
**Ikeji-Arakeji, Osun State, Nigeria.**

**[2]Department of Computer Science,**
**Faculty of Science,**
**University of Lagos,**
**Akoka, Lagos, Nigeria.**

**Abstract**

Predictive modeling using data mining methods for early identification of students' performance can be very beneficial to forecasting students' graduating class of degree. It is an innovative methodology that can be utilized by universities. The goal of the study is to have an early detection of graduating students' cumulative grade point average (CGPA) before they eventually graduate. Classification and regression tree (CART) and linear regression were the algorithms used to carry out the prediction model. Also, a novel algorithm: Difference Level (DL) was designed and incorporated in the system. The system works by taking the differences of each level grade point average and adding the resultant values together; then subtracted from their final year first semester result to give a predicted graduating cumulative grade point average. Data analysis was performed on datasets of a specific graduated class. The dataset was obtained from Joseph Ayo Babalola University (JABU) Exams and Records unit. This study found that students with a risk of graduating with a low CGPA can actually be predicted at the end of the final year first semester. The aim of this study is to help students in improving their ability in getting a better graduating CGPA.

***Keywords: CART, Difference level (DL), Linear regression, Prediction, Student graduating CGPA***

## I. INTRODUCTION

Grading in education is the process of applying standardized measurements of varying levels of achievement in a course. In some countries, all grades from all current classes are averaged to create a grade point average (GPA) for the marking period. Also, in most universities GPAs are calculated for undergraduate and graduate students [1].

The GPA can be used by potential employers or educational institutions to assess and compare applicants. A cumulative grade point average (CGPA) is a calculation of the average of all of a student's total earned points divided by the possible number of points. This grading system calculates for all of student's complete education career. Grade point averages can be unweighted (where all classes with the same number of credits have equal influence on the GPA) or weighted (where some classes are given more influence than others) [2].

A prediction (Latin *præ-*,"before," and *dicere*, "to say"), or forecast, is a statement about a future event. A prediction is often, but not always, based upon experience or knowledge [3]. The cumulative grade point average CGPA is crucial in the academic life of students, it is an interesting and challenging problem to create profiles for those students who are likely to graduate with low CGPA.

Identifying this group of students accurately will enable the university staff to mitigate further decline and improve their performances by providing them with special academic guidance and tutoring [4]. The problem faced by students in knowing their final CGPA before they actually graduate is a dominant issue in academic environment which makes it sometimes difficult for students to perform better. The goal of this

**Azeez, T. O. Awe, A. C. and Omosebi, P.A. (2018). Predicting Students' Graduating Cumulative Grade Point Average Using Difference Level, Classification and Regression Tree and Linear Regression Algorithm, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 2, No. 1, pp. 40 - 47**

work is to design a system that predicts students' graduating CGPA using the CART and linear regression algorithms, and also a newly developed difference level (DL) algorithm to assist in predicting student final CGPA, and to detect CGPA at the margin point to the next class of honours degree with parameter's in the range of 0.1 - 0.5. To achieve this goal, the algorithms aforementioned were implemented.

The significance of this study can be summarized in two ways. First, it helps in the early prediction of students graduating CGPA, in order to improve students' academic performance, which assists the department at large in having more graduating students with better class of degree. Secondly, it helps in detecting students who are close to the various class boundaries to cross into the next class of degree.

This paper is organized with a review of the literature, a research methodology where the Linear regression, CART, and DL

Algorithms were explained, followed by the implementation of our data collection and mining process following these algorithms. The results were presented and discussed. Lastly, conclusions from the research and direction of future developments were briefly highlighted

## II. LITERATURE REVIEW

In 2015, the research work of Muluken [5] investigated the potential applicability of data mining technology to predict student success and failure rates. Classification and prediction data mining functionalities are used to extract hidden patterns from students' data. The classification rule generation process was based on the decision tree and Bayes as a classification technique. The generated rules were studied and evaluated. The research result offered helpful and constructive recommendations to the academic planners in universities; in terms of learning to enhance their decision making process. Similarly, Quardri and Kalyankar in 2010 [6] stated that students' progressive academic performance is measured by their cumulative grade point average (CGPA) upon graduating. The work applied decision tree technique in predicting the drop out feature of students.

Priyanka and Ajit [7] predicted using classification techniques for the students' enrolment process in higher educational institutions, constructed a student performance model using classification technique with two decision tree algorithm (ID3 and J48 decision tree algorithm). The study helped in selecting the course for admission according to students' skills and academic line.

Furthermore, Zuhrieh and Shubair [8] presented a paper on students graduating CGPA prediction using a CRISP-DM Process Model. Correspondingly they proposed an intelligent system to predict students' graduation accumulated grad-point average AGPA in Al Ain University of Science and Technology (AAU). The prediction process was done by employing neuro-fuzzy inference systems. The dataset used to determine the model quality and validity consists of 200 student records from two colleges, Law and Business Administration. This method mines the relationship between the graduation GPA of AAU students and their scores, in order to identify those students who needed special attention to enhance and improve their low GPA. The results showed a high level of accuracy of 97%.

Siwalai and Supaporn [9] presented the development of a system for predicting students' graduation CGPA using data mining technique, to help in enhancing the quality of education system by envisaging graduation status. The result indicated that the intelligence system can help students and advisors to plan a program of study that meets the requirements for graduation in four years. Decision tree technique was used as a classifier to construct a predictive model of the system. The decision tree obtains the accuracy and the F-measure values about 79.4% and 77.9%, respectively. Then, the predictive model is deployed to web application, which tested the prediction performance using unseen data set and obtained the accuracy of 98.03%.

Chang used data mining predictive modelling to augment the prediction of enrolment behaviors of admitted applicants at a large state university [10]. Chang also applied classification and regression trees (CART), logistic regression and artificial neural networks in predicting admissions. CART yielded 74% classification rate, neural network with 75% classification rate, and logistic regression with 64% classification rate.

Comparatively to this study, previous outputs on predicting students' CGPA have been on identifying success or failure rate, enhancing the quality of education, improving students' performance as they progress to the next level based on the previous CGPA, offering helpful and constructive recommendations to the academic planners in universities; in terms of learning. In this paper; the focus is tailored towards predicting their graduating CGPA at the end of their final year first semester in school with the aim of avoiding low graduating class of honour's degree and identify students' close to the next class boundary in the range of 0.1 - 0.5 e.g. 3.49 (second class lower) is close to 3.5 (second class upper).

## III. METHODOLOGY

The data used in this study was retrieved from the Exams and Records unit of Joseph Ayo Babalola University JABU Nigeria. The data consists of information of students' academic profile from 100 to 400 level of a graduated set in the College of Natural Sciences. The total number of the dataset was One Hundred (100) which was divided into Training and test dataset. In proportion of 70-30% i.e. 70% of the dataset was used to train the algorithms and 30% for the testing the performance of the algorithm with reference to its prediction capability. The system was designed to be used for a long duration of time and can cover a number of academic sessions. In this study, the methodology that was used is CART (Classification and a Regression Tree Algorithm) and linear regression algorithm/techniques which are some of the regression algorithms used in data mining. Also, a new model Difference Level (DL) algorithm was used in order to assist in predicting student graduating CGPA. The CART was introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems.

The representation for the classification and regression tree (CART) model is a binary tree. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. For regression predictive modeling problems, the cost function that is minimized to choose split points is the sum of squared error across all training samples that fall within the rectangle:

sum (y – prediction) ^2 ……(1)

Where y is the output for the training sample and prediction is the predicted output for the rectangle.

For classification, the Gini index function is used which provides an indication of how "pure" the leaf nodes are (how mixed the training data assigned to each node is).

G = sum (pk * (1 – pk)) ……..... (2)

Where G is the Gini index over all classes, pk are the proportion of training instances with class k in the rectangle of interest. A node that has all classes of the same type (perfect class purity) will have G=0, whereas a G that has a 50-50 split of classes for a binary classification problem (worst purity) will have a G=0.5.

**Linear regression** is a common Statistical Data Analysis_technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression.

For more than one explanatory variable, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable [11].

Linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis. In linear regression, the goal is prediction, or forecasting, or error reduction linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

Linear regression formula:

$Y= \beta0+\beta1*X,$

where, $\beta0$ and $\beta1$ represents coefficients where X and Y represents variables.

Difference Level (DL) Algorithm used a pattern which predicted the final CGPA of a graduating students. The system works by taking the differences of each level's GPA and adding it together; this resultant value is subtracted from their final first semester GPA to give the predicted graduating CGPA.

### *A. Methodology Illustration*

The study was carried out based on the following:

1. Following the review of literature, results of 2013/2014 to 2016/2017 academic session of a graduating set in the college of Natural science containing grades of each courses and cumulative grade points (CGPA) was collected.
2. CART Decision tree, linear regression and DL algorithms were applied on the results to formulate the predictive model.
3. The model was trained on the data obtained and tested on the unseen dataset obtained. Comparism of the algorithms were done which will be discussed later in this work.

The diagrammatical representation of the aforementioned are shown in Figure 1.

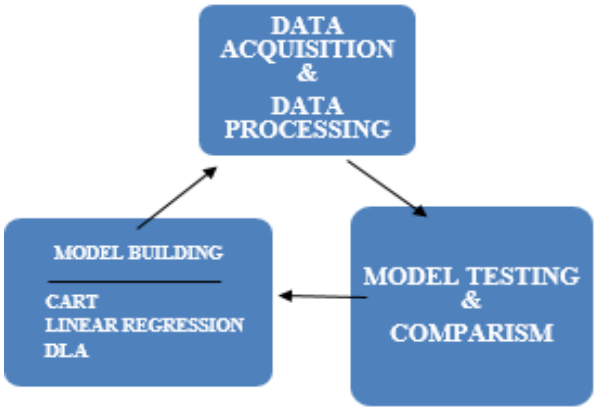



Figure 1: Methodology Illustration

*B. Difference Level (DL) algorithm and Flowchart*

Figure 3 shows the Flowchart for the difference level algorithm. The flowchart and the algorithm details is explained in the algorithm discussion below.

```
get Data
    for studentdata in Data
        get studentdata, Set diff=0
        for i in range(len(studentdata)/2)-1
        sem_year=studentdata[i:i+2]
        if semyear[0]=None
            continue
            diff+=semyear[0]-semyear[1]
        finalyear1gp=studentdata[-2]
        finalyear2gp=studentdata[-1]
        predictedgp=finalyear1gp - diff
          print(predictedgp, finalyear2gp):
```

Figure 2: Difference Level Algorithm

*C. Algorithm Discussions.*

***for studentdata in Data***:

Iterate through the dataset, picking each student data on every iteration,

The student data is a list of GPA's (both 1st and 2nd semesters for all the years' the student is expected to spend.

***get studentdata, Set diff=0:***

Initialize the difference variable = 0.

***for i in range(len(studentdata))***:

Get length of student data, compute no of iterations.

***for i in range(len(studentdata)/2)***:

Divide it by 2 (since data length is even, it will pick in two's for every iteration).

***for i in range(len(studentdata)/2)-1***:

Subtract 1 (since final year GPA will not contribute to the difference of each GPA for each level).

Iterate through number of iterations gotten.

***SemesterGP =ListName[i:i+2]***:

Fetch data in two's using index selection.

***If semesterGP = NIL, Stop***:

If student does not have result for a level(session)skip, check through to continue for such student if there are GPA's for other level otherwise continue for another student.

***diff+=SemesterGP[0]-SemGP[1]***:

Subtract each student's level's GPA and add the resultant. (difference is updated)

***finalyearGP1 and finalyearGP2***:

Means 1st semester final year which has index -2 and 2nd semester final year which has index -1.

***predictedgp=finalyear1gp - diff***:

Subtract finalyear1gp from sum of the difference of each level's gpa

***print(predictedgp, finalyear2gp)***:

Print the predicted finalyear2gp (CGPA) and the actual student's CGPA for comparison.

*D. System Description*

The system developed was linked to the web which consists of two parts; the front end and the back end. The front end is a web application that **accepts the** preferred algorithm that the user selects and which allows the user to load the comma-separated values CSV-file at which the system would work. The front end also helps in displaying the results to the user. The front end interfaces with the back end which pre-processes the files, uses the CART and linear regression algorithm to predict graduating CGPA, thereafter uses the DL algorithm to assist in the prediction and then send the results back to the front end. The interactions in the system are shown in the System Description diagram presented in Figure 3.

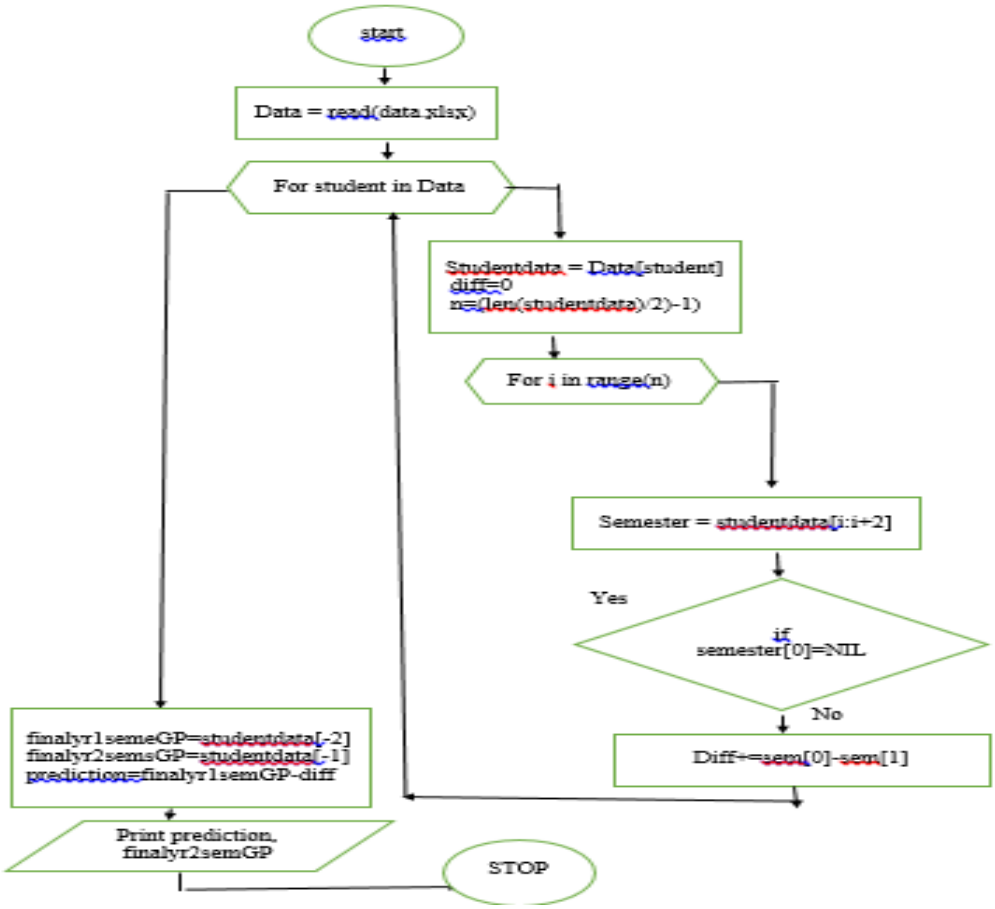



Figure 3: Flowchart

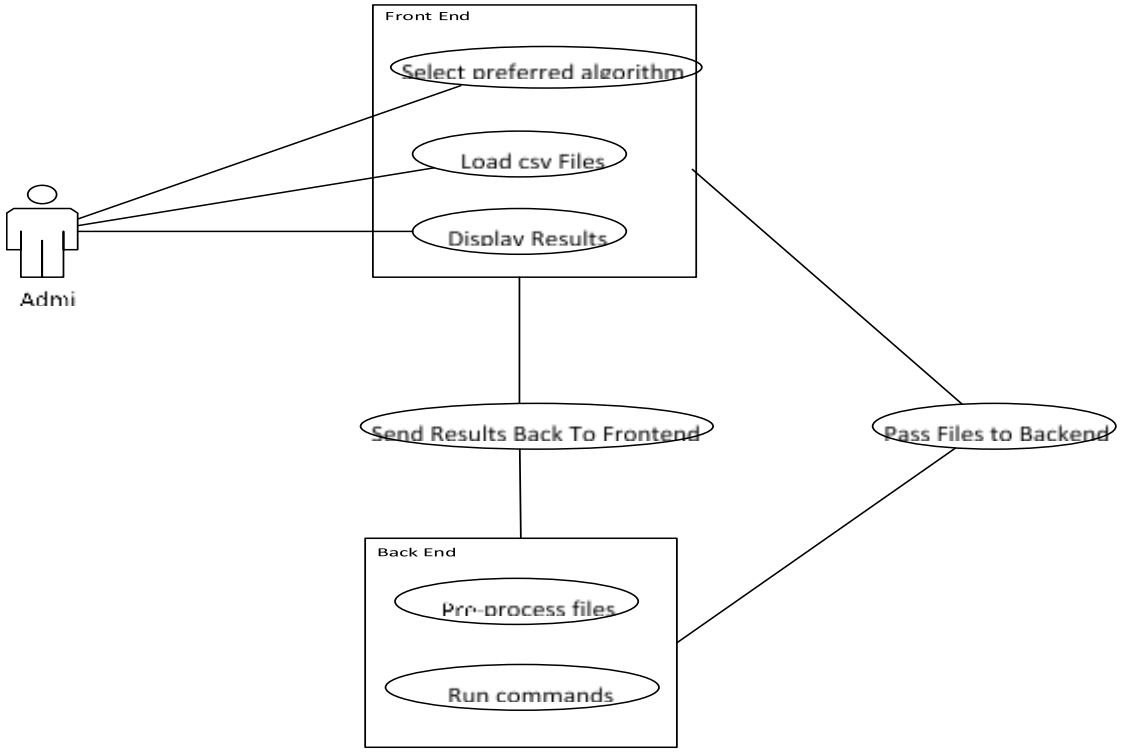



Figure 4: System Description

## IV. RESULTS AND DISCUSSIONS

The intelligent system for predicting students' graduating CGPA is a web based application that uses python programming language to implement CART, linear regression and the newly developed DL algorithms for predicting the graduating CGPA.

Figure 5 shows how the administrator can select either of the three algorithms which can be applied on students' results for prediction.

Figure 6 shows prediction results generated for some students' and also detects if the predicted students' CGPA is close to the next class of boundary i.e. next class of honours' degree.

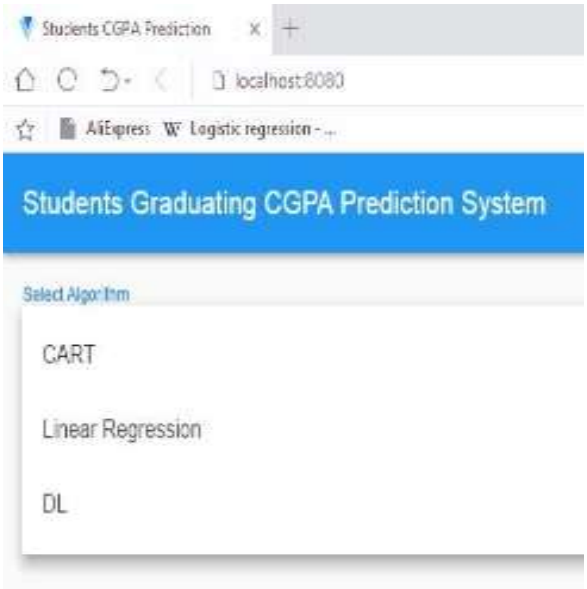



Figure 5: Algorithm selection

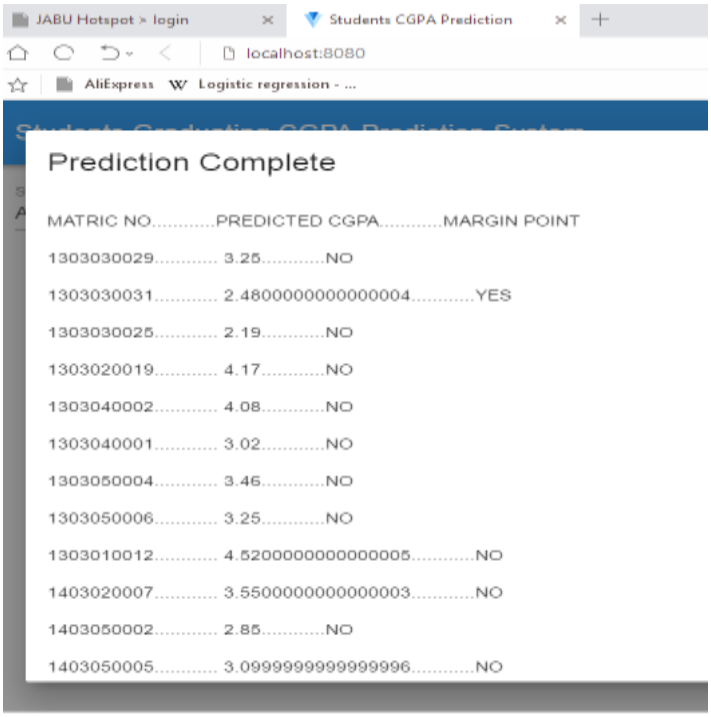



Figure 6: Prediction Results and Class Boundary CGPA's

### A. *Comparative Analysis of the Algorithms' Performance*

The performance of the prediction algorithm was measured using the metrics: precision, recall, and F-measure. Precision is the percentage of correctly predicted CGPA over all the predicted CGPA, while Recall is the percentage of the correctly predicted CGPA over all the correctly predicted CGPA and incorrect predicted CGPA. Suppose the number of correctly predicted CGPA is C, the number of all the predicted CGPA is W and the number of incorrect predicted CGPA is M, then the precision P of the approach is given by the expression given below

$$P = C/ (C + W) \quad (3)$$

and the recall, R, of the approach is

$$R = C/ (C + M) \quad (4)$$

F-measure incorporates both precision and recall. F-measure is given by

$$F = 2PR/ (P + R) \quad (5)$$

Where: precision P and recall R are equally weighted.

Table 1: F-measure of the three algorithms

| Prediction methods | C | W | M | P | R | F | Accuracy |
|---|---|---|---|---|---|---|---|
| Linear Regression | 12 | 21 | 9 | 0.36 | 0.57 | 0.44 | 44.4% |
| CART | 16 | 21 | 5 | 0.43 | 0.76 | 0.55 | 55.1% |
| Difference Level | 19 | 21 | 2 | 0.47 | 0.90 | 0.62 | 62.2% |

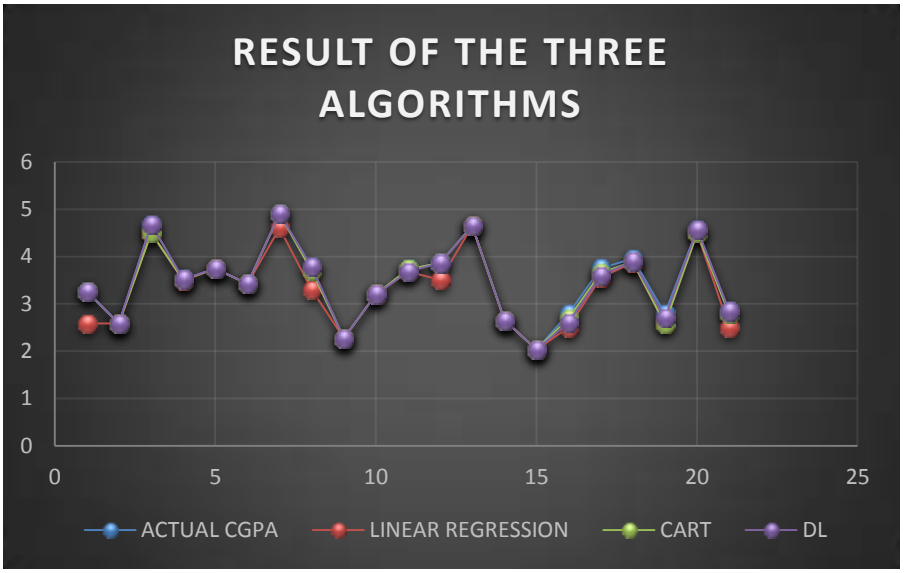
Following the results generated from all the algorithms in Table 1, it could be concluded that the best predicting algorithm was the developed DL algorithm followed by the classification and regression tree (CART) and then linear regression algorithm.

Figure 7 depicts the comparative results of the three algorithms. The graph showed the prediction result of the graduating CGPA of the three algorithms with respect to the actual students' graduating CGPA. It overlaps because the algorithm's prediction result are closer to each other. Difference Level algorithm which was developed performed a bit higher than Linear Regression and Classification and Regression Tree algorithm shown in Table 1 earlier.

## V CONCLUSION

Generally, the prediction system plays some role in the academic life of student and a department as a whole; also it could be concluded that students graduating CGPA could actually be predicted in a computerized format. The system identifies students' at the margin point to the next class of honors' degree, and also give the classified statistics of the students according to their class of honors' degree.

This system will help improve the academic performance of students, and also improve their graduating CGPA. Though, our system demonstrated an encouraging result, still there are some data quality issues to be resolved. Issues faced by the system are data completeness, reliability, and accuracy due to time limit. As thousands of data records are usually collected for such kind of prediction systems, if duplicate records, missing values, or presence of unneeded data is in the data, then all steps in the prediction process could be badly affected. However, this work can still be improved by applying other predictive algorithms with larger datasets since this will help in training the algorithm(s) to predict correctly and be more effective.




. Figure 7: Comparative Algorithms' Results

**REFERENCES**


[1] John W., (2018). Degree Classification is unfair to many students. *The Guardian*.pg. 19, ISBN 9781118391679.

[2] Warne, R. t., Nagaishi, C. Slade, Michael k.; Hermesmeyer, Paul; p., Elizabeth K.(2017) "Comparing Weighted and Unweighted Grade Point Averages in Predicting College Success of Diverse and Low-Income College Students". *NASSP Bulletin*. 98:261279. Doi:10.1177/0192 636514565171.

[3] Fox, J., (2016). Applied regression analysis and generalized linear models. *Sage Publications, London*, 791 p, Third Edition. ISBN 978-1-4522-0566-3.

[4] Ismail, S., Abdulla, S. (2015) Design and Implementation of an Intelligent System to Predict the Student Graduation AGPA *Australian Educational Computing*, Volume 30 no 02.

[5] Muluken A.Y., (2015) Application of Data Mining Techniques for Student Success and Failure Prediction (The Case of Debre_Markos University). *International Journal of Scientific & Technology Volume 04, Issue 04, April 2015 ISSN 2277-8616 IJSTR www.ijstr.org.*

[6] Quardri, M.N., Kalyankar, N.V (2010). Dropout feature of student data for academic performance using decision tree techniques, *Global Journal of Computer Science and Technology*. 10(2) ISSN 0975-4172.

[7] Priyanka S. and Ajit K. J., (2013). Prediction Using Classification Techniques for the Students' Enrolment Process In Higher Educational Institutions. *International Journal of Computer Applications (IJCA Journal)*. Volume 84(14):37-41.

[8] Zuhrieh, S. and Shubair, A., (2015) Educational Data Mining: An Intelligent System to Predict Student Graduation AGPA, *International Review on Computers and Software*.10 (6):593-601.

[9] Siwalai C. and Supaporn S., (2016). An application to improve learning effectiveness of problem facing in C++ programming. The 8th *International Conference on Science, Technology and Innovation for Sustainable Well-Being (stiswb viii)*. Yangon, Myanmar. CST-062.

[10] Chang, l. (2006). Applying data mining to predict college admissions yield: a case study. *New directions for institutional research*, (131), 53-68.

[11] Rencher, A. C., Christensen, W.F. (2012), "Chapter 10, Multivariate Regression – section 10.1, introduction", methods of multivariate analysis, *Wiley Series in Probability and Statistics*, 709 (3rd ed.) pg 19.