Predicting Student's Final Graduation CGPA Using Data Mining and Regression Methods: A Case Study of Kano Informatics Institute

Salim Jibrin Danbatta
Department of Computer Science
Kano State Institute for Information Technology
Kano, Nigeria
salimdambatta@gmail.com

Asaf Varol
Department of Software Engineering
Firat University
Elazig, Turkey
varol.asaf@gmail.com

***Abstract*— Data mining and regression techniques are important methods that we can use to predict students' performance to inform decision making. This study uses five regression techniques to analyse students' first-year cumulative grade point average (CGPA) and predict their final graduation CGPA. The data set used in this study is that of programming and software development students at Kano Informatics Institute. The results and the grades obtained by 163 students forms the sample data used in the study. The forecast error, mean forecast error and mean absolute forecast error are all calculated. Dickey–Fuller's stationary t-test is performed for all the regressions analysis values using the Python programming language to determine the mean and if the data is centred on the mean. We use the stationary t-test to test the null and alternative Dickey–Fuller's hypotheses to compare our P-values and critical values for all regressions analyses done. The results show that the P-values obtained for all the regressions are small and less than the critical value. However, linear regression is the model with the mean closest to zero, and, according to Dickey–Fuller's statistics, it is the model that best fits our data.**

Keywords—Educational data mining, regression, Dickey–Fuller's stationarity test, forecast error, mean forecast error, mean absolute forecast error

I. Introduction

Data mining is an interdisciplinary field of study resulting from a fusion of many different areas, such as machine learning, statistics, pattern reorganization, databases, artificial intelligence and computation capabilities. According to different studies, there are various definitions of data mining. For example, Tomar and Agarwal [1] define data mining as a process of finding meaningful information from huge data sets, while Wongchinsri and Kuratach [2] define data mining as a methodology that combines statistics, machine learning and databases to extract patterns and identify useful data from many databases.

Educational data mining (EDM) is a discipline that uses data mining techniques in the field of education. Mushtaq et al. [3] described EDM as an emerging discipline concerned with developing methods for exploring unique types of data that come from educational settings and using those methods to better understand students and the settings in which they learn. Refae and Ghaleb [4] contend that EDM techniques can be applied on the educational dataset to extract hidden knowledge for predictions concerning the enrolment of students into a particular course, alienation of traditional classroom teaching models, detection of improper values in the students' result sheets, exam malpractices and predictions on student performance.

Regression is a statistical measure used to determine the relationship between one dependent variable and a series of other mutable variables (independent variables). In regression, we can use some number of database attributes to predict another database attribute. Therefore, prediction can be achieved by designing a model that allows inferring in some aspects of the data. For example, with an effective regression algorithm, information on student dropouts or final graduation cumulative grade point average (CGPA) can be analysed to begin corrective and preventive actions targeted at probable dropout candidates.

Using data mining and regression techniques, data can be processed to discover trends and predict events in education. For example, if we can use students' previous CGPAs to predict their final graduating CGPAs, then this prediction can be used to minimise student attrition, dropout and dismissal. Hence, providing useful information to inform decision making and student counselling to boost academic performance is invaluable and cannot be overemphasised.

Kano Informatics is an information technology (IT) based institution that was established in 2011 by the Kano state government in Nigeria. This study aimed to determine the dataset of the institute and the extent to which data mining and regression techniques can be utilised to predict students' final graduation CGPAs. Our results will be analysed to see if they can be used as a basis for predicting the results of future students. Moreover, the study will recommend a course of action on the management of student academic performances.

II. Literature Review

Numerous studies have been completed on data mining regression and EDM. Gowri et al. [5] carried out a study on an EDM application that was used to estimate students' performance. The main tool used in this study is the Weka environment. The study also employs the use of k-means and

978-1-7281-6939-2/20/$31.00 ©2020 IEEE

apriori algorithms on students' databases for wider classification based on various categories. However, the parameters used in this study give more importance to psychological traits than academic features. Thus, the results of the study are tailored toward whether a student is prone to violence or not.

An EDM predictive analysis work by Fernandes et al. [6] presents a predictive analysis of the academic performance of students in the public schools of the Federal District of Brazil during 2015 and 2016. The study uses two different data sets. The first data set is historical data that was collected before the start of the academic year, while the second data set was collected two months into the academic year. Separate predictions were done using a classification model on each of the databases to predict academic outcomes of students' end-of-year performances. The study finds that grades and absences are the two most relevant attributes for the prediction. Moreover, the study concludes that neighbourhood, school and age are also potential indicators of a student's academic success or failure.

Bermudez et al. [7], in their experimental work, investigate the different attributes used in evaluating faculty performance to develop a regression model that predicts faculty performance. The study highlights how evaluating faculty performance is an immense concern for every higher education system. The study also concludes that, with the implementation of clustering and regression analysis, developing a system that predicts faculty performance could be a tool for improving teaching at every higher education institution.

The comparative study of Chertchom [8] presents a data mining tool comparison over regression methods and makes some recommendations for small and medium enterprises (SMEs). The three selected data mining tools compared in the study are WEKA, RapidMiner and IBM SPSS. The paper recommends that data mining tools for SMEs should not require programming knowledge and should have features such as customised dashboards and ad hoc reporting. Hence, the study recommends RapidMiner as the better choice for SME data mining tools since it has all of the recommended features.

Osmanbegovic and Suljic [9] use an algorithm called k-means clustering on students' real-time data, such as different subject marks in each semester, that determine the relationship between the students' learning behaviours and their academic performance. The study finds it to be a very accurate means of predicting student performance. In their research, the authors use the Weka environment as a knowledge analysis tool. Baradwaj and Pal [10] investigate the accuracy of using data sets like attendance, class tests, and seminar and assignment marks collected from the students' management systems to predict their end-of-semester performance. The ID3 decision tree technology is used on student data, such as the students' sex, grades obtained from their senior secondary school certificate examinations, entry examination scores and grades obtained by the students during graduation. The research shows that there is a significant relationship between the scores students obtained on their senior secondary school certificate examinations in specific subjects and the classes of degree they graduated with.

In [4], heuristic and artificial neural network data mining technologies are used on the traditional headcount programme used in Malaysian schools. The result appears to be a more accurate and reliable method of predicting student performance. A study conducted by Kolo et al. [11] on the different variations of student records, using the decision tree and Bayes as classification techniques, shows that data mining techniques can be applied by higher education institutions and universities to determine success and failure rates. Hence, managing students' enrolment can assist students before they are at risk of failure. It can also guide administrative officers in successful management and decision-making and ensure effective resource utilization and cost minimization.

Akinola et al. [12] conducted research at the University of Ibadan on student datasets like ordinary level results, mathematics and physics scores obtained in year one and marks obtained from the programming course in the department of computer science by year two. These datasets were put into a data mining system using an artificial neural network algorithm called multi-layer perceptron feed-forward back propagation technique. The result from the research shows that prior knowledge in physics and mathematics is central to student prosperity in computer programming and that those students at risk can be identified earlier and given necessary assistance before it is too late.

Kovačić [13] uses different classification trees, such as CHAID, Exhaustive CHAID and QUEST, to explore the factors that impact the students' study outcomes in the information system course using student enrolment data at the New Zealand Open Polytechnic. The data collected from the student forms contains both demographic and academic data. The study shows that the most important variables that separate successful from unsuccessful students are ethnicity, the course programme applied for and the course block. In the research, demographic data such as gender and age, though related to the study outcomes, were not used in the classification trees. Though the accuracy of the classification trees in the research is not very high (CHAID 59.4%, CART 60.5%). Showing that the data set used in the research does not give sufficient information needed to classify and predict the learning outcome. It still suggests that background information from pre-enrolment data, such as gender, age, disability, secondary school, work status and early enrolment, would allow both administrative and academic staff to identify students at risk of dropping out of courses before they start the programme.

III. Research Design

Data mining methods are important components of EDM. These methods are categorised into verification oriented (traditional statistics), such as hypothesis testing, analysis of variance, etc., and discovery-oriented (prediction and description), like prediction, classification, clustering, etc.

As reported by Chapman et al. [14], there are different types of data mining methodologies, such as CRISP-DM, SEMA, MY-ORGANIZATION, etc. This study is designed following the cross-industrial standard process for data mining (CRISP-DM). According to [14], the steps of the CRISP-DM include business understanding, data understanding, data preparation, modeling evaluation and deployment. Figure 1 shows the stages involved in the CRISP-DM and the possible feedbacks that may exist between the stages.

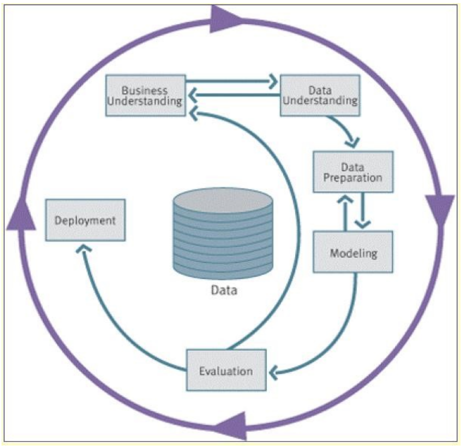
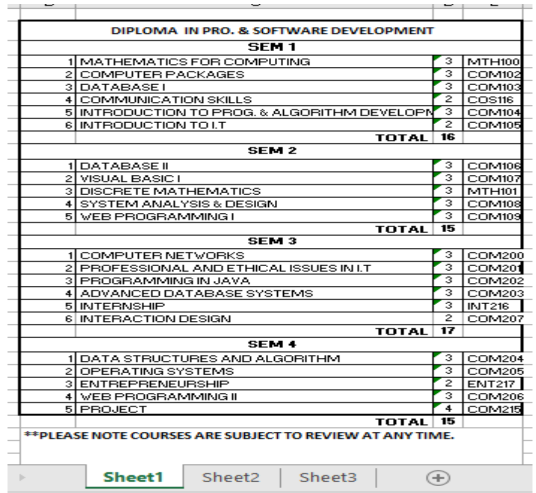


Fig. 1: Stages of CRISP-DM [14]



	DIPLOMA IN PRO. & SOFTWARE DEVELOPMENT		
	SEM 1		
1	MATHEMATICS FOR COMPUTING	3	MTH100
2	COMPUTER PACKAGES	3	COM102
3	DATABASE I	3	COM103
4	COMMUNICATION SKILLS	2	COS116
5	INTRODUCTION TO PROG. & ALGORITHM DEVELOPN	3	COM104
6	INTRODUCTION TO I.T	2	COM105
	TOTAL	16	
	SEM 2		
1	DATABASE II	3	COM106
2	VISUAL BASIC I	3	COM107
3	DISCRETE MATHEMATICS	3	MTH101
4	SYSTEM ANALYSIS & DESIGN	3	COM108
5	WEB PROGRAMMING I	3	COM109
	TOTAL	15	
	SEM 3		
1	COMPUTER NETWORKS	3	COM200
2	PROFESSIONAL AND ETHICAL ISSUES IN I.T	3	COM201
3	PROGRAMMING IN JAVA	3	COM202
4	ADVANCED DATABASE SYSTEMS	3	COM203
5	INTERNSHIP	3	INT216
6	INTERACTION DESIGN	2	COM207
	TOTAL	17	
	SEM 4		
1	DATA STRUCTURES AND ALGORITHM	3	COM204
2	OPERATING SYSTEMS	3	COM205
3	ENTREPRENEURSHIP	2	ENT217
4	WEB PROGRAMMING II	3	COM206
5	PROJECT	4	COM215
	TOTAL	15	

**PLEASE NOTE COURSES ARE SUBJECT TO REVIEW AT ANY TIME.

Sheet1 Sheet2 Sheet3

Fig. 2: Diploma programme course list

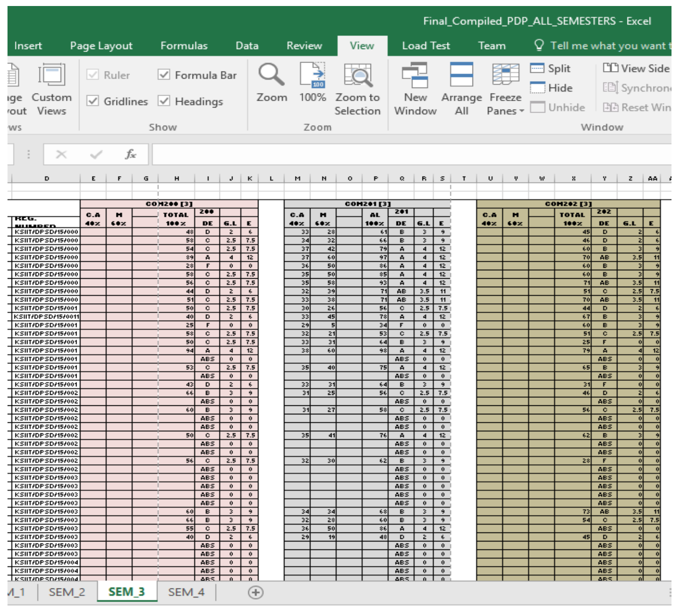


Fig. 3: Sample of students' grades

A. *Business Understanding*

The institute started as a sub-franchising informatics academy in Singapore through the Jigawa State Institute of IT for international diploma and advanced diploma programmes. This two-year programme leads to a one-year top-up degree programme at many foreign universities. The institute currently relies on high school grades plus an entrance examination to assess the competencies of the students being admitted into the institute.

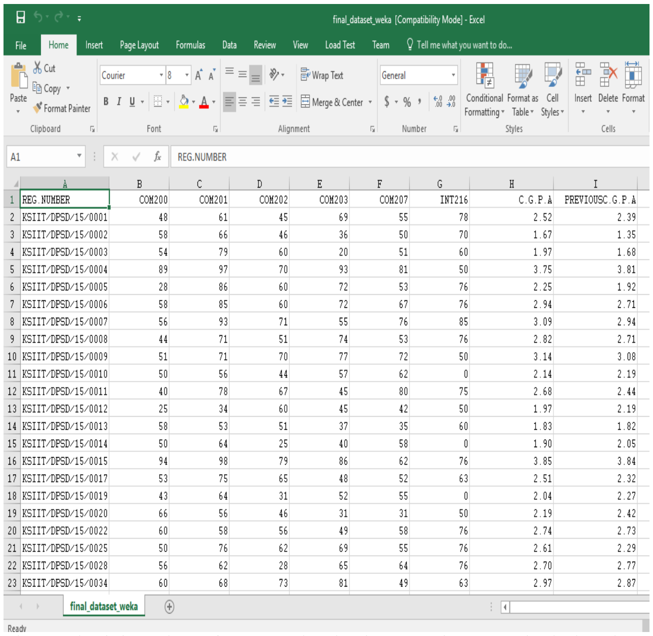
B. *Data Understanding*

The study is concerned with the historical data of the programming and software development students at the Kano Informatics Institute. The results and the grades obtained by 163 students forms the sample data used in this study. It was intended that data from all 163 students would be used but, unavoidably, only 112 student records were available and used in the research. The institute started as a parastatal under the Ministry of Science and Technology; the ministry regulated all the activities of the institute. Then, the institute started running at its temporary site, the Murtala Mohammed Library, Kano. Later, the institute was taken over by the Kano State University of Science and Technology and, eventually, the institute moved to its permanent site at Kura. It is due to these transitions that some data could not be located.

C. *Data Preparation*

The hard and soft copies of the data were collected from the examination office of the Kano Informatics Institute. The soft copy is in Microsoft Excel format and is the dataset that was used to complete this study. Figures 2 and 3 show a sample of the course list and individual grades obtained by students, respectively.

Regarding the data cleaning, many attributes exist in the data, but some of them, like the student's name, CA, PE, TPBF, etc., are considered irrelevant and therefore removed from the list of the attributes. Later on, the data are refined, re-cleansed and reselected to make them flat and normalised, as shown in Figure 4.



	A	B	C	D	E	F	G	H	I
1	REG.NUMBER	COM200	COM201	COM202	COM203	COM207	INT216	C.G.P.A	PREVIOUSC.G.P.A
2	KSIIT/DPSD/15/0001	48	61	45	69	55	78	2.52	2.39
3	KSIIT/DPSD/15/0002	58	66	46	36	50	70	1.67	1.35
4	KSIIT/DPSD/15/0003	54	79	60	20	51	60	1.97	1.68
5	KSIIT/DPSD/15/0004	89	97	70	93	81	50	3.75	3.81
6	KSIIT/DPSD/15/0005	28	86	60	72	53	76	2.25	1.92
7	KSIIT/DPSD/15/0006	58	85	60	72	67	76	2.94	2.71
8	KSIIT/DPSD/15/0007	56	93	71	55	76	85	3.09	2.94
9	KSIIT/DPSD/15/0008	44	71	51	74	53	76	2.82	2.71
10	KSIIT/DPSD/15/0009	51	71	70	77	72	50	3.14	3.08
11	KSIIT/DPSD/15/0010	50	56	44	57	62	0	2.14	2.19
12	KSIIT/DPSD/15/0011	40	78	67	45	80	75	2.68	2.44
13	KSIIT/DPSD/15/0012	25	34	60	45	42	50	1.97	2.19
14	KSIIT/DPSD/15/0013	58	53	51	37	35	60	1.83	1.82
15	KSIIT/DPSD/15/0014	50	64	25	40	58	0	1.90	2.05
16	KSIIT/DPSD/15/0015	94	98	79	86	62	76	3.85	3.84
17	KSIIT/DPSD/15/0017	53	75	65	48	52	63	2.51	2.32
18	KSIIT/DPSD/15/0019	43	64	31	52	55	0	2.04	2.27
19	KSIIT/DPSD/15/0020	66	56	46	31	31	50	2.19	2.42
20	KSIIT/DPSD/15/0022	60	58	56	49	58	76	2.74	2.73
21	KSIIT/DPSD/15/0025	50	76	62	69	55	76	2.61	2.29
22	KSIIT/DPSD/15/0028	56	62	28	65	64	76	2.70	2.77
23	KSIIT/DPSD/15/0034	60	68	73	81	49	63	2.97	2.87

Fig. 4: Final data view after removing irrelevant columns and missing data values

D. Modeling

This is the pattern discovery stage of the CRISP-DM methodology. The major activity here is using linear, exponential, logarithmic, polynomial and power regressions in SPSS to analyse the data. Since the major objective of this study is to predict students' final graduation CGPAs, prediction is the chosen data mining technique for this study. The required constant and coefficients for the prediction equation are obtained from the results of the SPSS regressions analysis.
We use the students' previous CGPAs to predict their next CGPAs. Figure 5 shows the prediction steps of the linear regression.

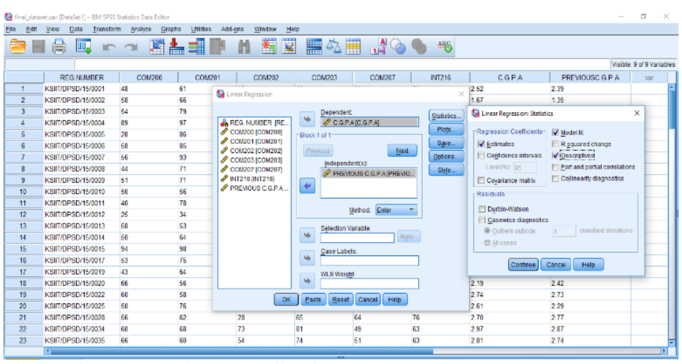


Fig. 5: Linear regression to predict the CGPA

The prediction equations obtained are used to implement a web form. For example, Figure 6 shows the web implementation of the linear regression equation obtained using JavaScript.

```
<script>
function prediction()
{
var x = document.getElementById('PCGPA').value
        var predCGPA = (0.973 * x) + 0.108
        document.getElementById('CGPA').defaultValue = predCGPA
}
</script>
```

Fig. 6: JavaScript for implementing the linear regression on a web form

A column in Microsoft Excel is created for the result of each regression analysis, and the values are plotted. The mean of the CGPA and the five regressions analyses are calculated using Microsoft Excel. The forecast error is calculated by subtracting the predicted value (regression values) from the actual value (CGPA) for every instance, according to the formula et = yt – ft provided in [15]. The mean forecast error (MFE) and the mean absolute error (MAE) for each regression are calculated using the MFE and MAE formula provided in [15], as shown in equations (1) and (2).

$$MFE = \frac{1}{n}\sum_{t=1}^{n} e_t \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \qquad (2)$$

Dickey–Fuller's stationary t-test is done on the data to assess the mean and if the data is centred on the mean. The test was done for all the regressions analyses' values using the Python programming language. Also, we use the stationary t-test to test the null and alternative Dickey–Fuller's hypotheses to compare our P-value and critical value for all completed regressions analyses. The test also obtained the standard deviation value.

IV. Results And Discussion

The SPSS linear regression analysis shows that there is a significant correlation between the previous CGPA and CGPA. For example, the coefficient obtained for the previous CGPA (independent variable) and CGPA (dependent variable) after the SPSS linear regression analysis is 0.108, and the constant obtained is 0.0973. Assuming CGPA=Y and previous CGPA=X, the values are used to generate the prediction equation in equation (3). Figures 7, 8, 9, 10 and 11 show the regression equations and scattergrams of the five regressions analyses that were done on the data.

$$Y = 0.973(X) + 0.108 \qquad (3)$$

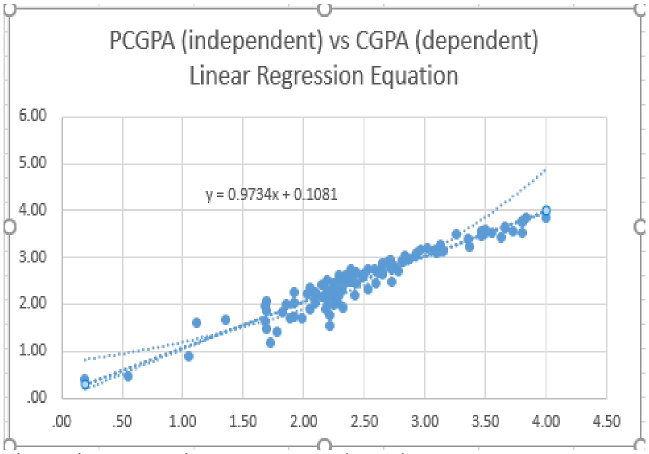


Fig. 7: Linear regression scattergram and equation

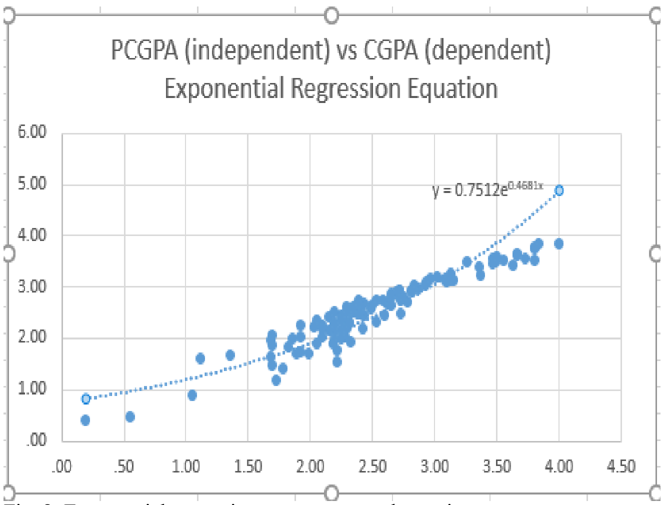


Fig. 8: Exponential regression scattergram and equation

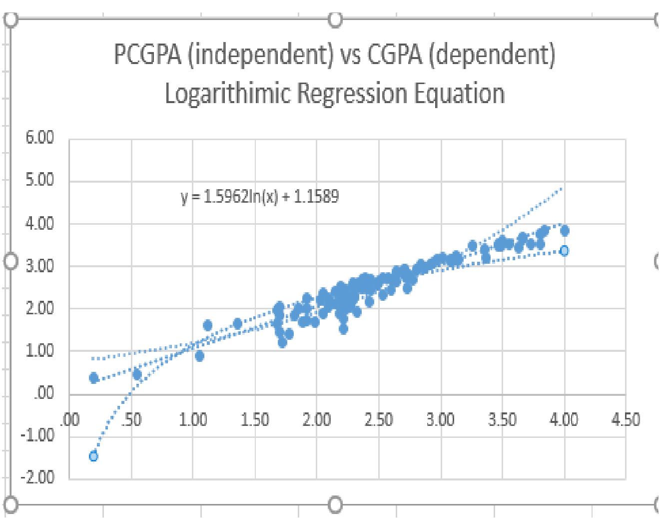


Fig. 9: Logarithmic regression scattergram and equation

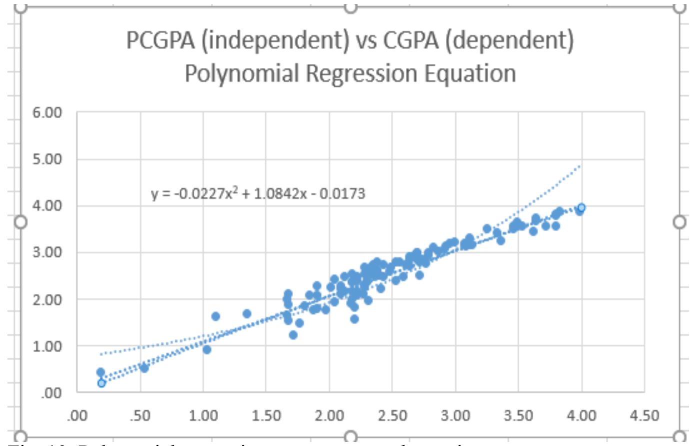


Fig. 10: Polynomial regression scattergram and equation

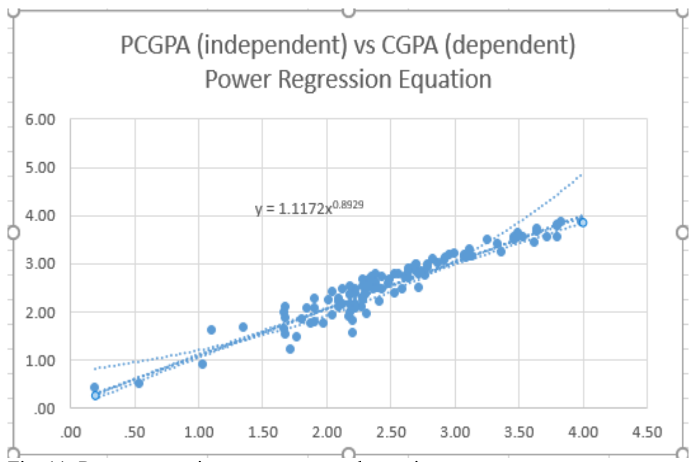


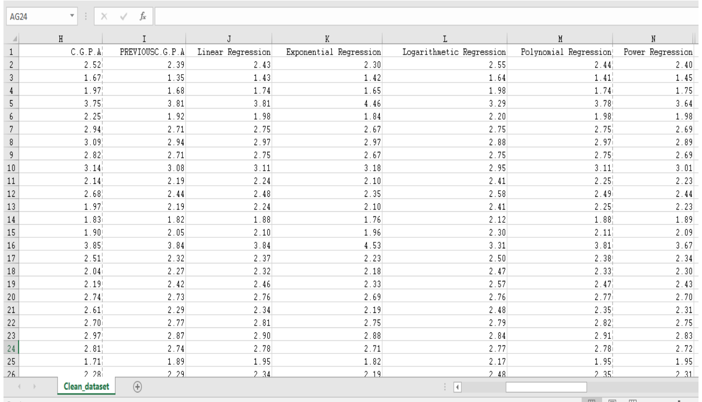
Fig. 11: Power regression scattergram and equation

The equations obtained for previous CGPA (independent variable) and CGPA (dependent variable) after the linear, exponential, logarithmic, polynomial and power regressions analysis in SPSS are summarised in Table 1.

TABLE I. REGRESSIONS EQUATIONS

Regression	**Equation**
Linear	y = 0.9734x + 0.1081
Exponential	y = 0.7512e0.4681x
Logarithmatic	y = 1.5962ln(x) + 1.1589
Polynomial	y = -0.0227x2 + 1.0842x - 0.0173
Power	y = 1.1172x0.8929

Figure 12 shows the values obtained for each regression, and Table 2 shows the MFE and MAE for the individual regression.



	H	I	J	K	L	M	N
1	C.G.P.A	PREVIOUSC.G.P.A	Linear Regression	Exponential Regression	Logarithmetic Regression	Polynomial Regression	Power Regression
2	2.52	2.39	2.43	2.30	2.55	2.44	2.40
3	1.67	1.35	1.43	1.42	1.64	1.41	1.45
4	1.97	1.68	1.74	1.65	1.98	1.74	1.75
5	3.75	3.81	3.81	4.46	3.29	3.78	3.64
6	2.25	1.92	1.98	1.84	2.20	1.98	1.98
7	2.94	2.71	2.75	2.67	2.75	2.75	2.69
8	3.09	2.94	2.97	2.97	2.88	2.97	2.89
9	2.82	2.71	2.75	2.67	2.75	2.75	2.69
10	3.14	3.08	3.11	3.18	2.95	3.11	3.01
11	2.14	2.19	2.24	2.10	2.41	2.25	2.23
12	2.68	2.44	2.48	2.35	2.58	2.49	2.44
13	1.97	2.19	2.24	2.10	2.41	2.25	2.23
14	1.83	1.82	1.88	1.76	2.12	1.88	1.89
15	1.90	2.05	2.10	1.96	2.30	2.11	2.09
16	3.85	3.84	3.84	4.53	3.31	3.81	3.67
17	2.51	2.32	2.37	2.23	2.50	2.38	2.34
18	2.04	2.27	2.32	2.18	2.47	2.33	2.30
19	2.19	2.42	2.46	2.33	2.57	2.47	2.43
20	2.74	2.73	2.76	2.69	2.76	2.77	2.70
21	2.61	2.29	2.34	2.19	2.48	2.35	2.31
22	2.70	2.77	2.81	2.75	2.79	2.82	2.75
23	2.97	2.87	2.90	2.88	2.84	2.91	2.83
24	2.81	2.74	2.78	2.71	2.77	2.78	2.72
25	1.71	1.89	1.95	1.82	2.17	1.95	1.95
26	2.28	2.29	2.34	2.19	2.48	2.35	2.31

Fig. 12: Regression values

TABLE II. MEAN FORECAST ERROR AND MEAN ABSOLUTE ERROR

Regression	MFE	MAE
Linear	-0.0000654761904762	0.1452499034889350
Exponential	0.000559566460358	0.221068848854862
Logarithmatic	-0.0000375855475638	0.0450077920599822
Polynomial	0.0003593693480253	0.0523327599390464
Power	0.048902086813533	0.178318963140582

The P-values obtained for all the regressions show that the P-values are small and less than the critical value. With this, we can reject the Dickey–Fuller's null hypothesis, which states that the data is random. Furthermore, the mean is centred around zero. The MFE and MAE obtained for all regressions analyses are not significant since the P-value is small and less than the critical value in all cases. Therefore, the models are sufficient for our data. However, with linear regression having a mean close to zero, according to Dickey–Fuller's statistics, it is the fittest model for our data. Tables 3, 4, 5, 6 and 7 present the Dickey–Fuller's stationary test results that were done using Python.

TABLE III. LINEAR REGRESSION DICKEY–FULLER'S STATIONARY TEST RESULT

MFE for Linear Regression	MAE for Linear Regression
t statistics -9.631069e+00 p value 1.611705e-16 lags used 0.000000e+00 number of observations 1.110000e+02 critical value(1%) -3.490683e+00 dtype: float64 Out[6]: count 112.000000 mean -0.000065 std 0.202129 min -0.727750 25% -0.108558 50% 0.041508 75% 0.145629 max 0.412767 Name: lr1, dtype: float64	t statistics -9.665359e+00 p value 1.319685e-16 lags used 0.000000e+00 number of observations 1.110000e+02 critical value(1%) -3.490683e+00 dtype: float64 Out[8]: count 112.000000 mean 0.145250 std 0.172349 min 0.001796 25% 0.063102 50% 0.092230 75% 0.096448 max 0.885720 Name: lr2, dtype: float64

TABLE IV. EXPONENTIAL REGRESSION DICKEY–FULLER'S STATIONARY TEST RESULT

MFE for Exponential Regression	MAE for Exponential Regression
t statistics -1.020463e+01 p value 5.847945e-18 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[9]: count 112.000000 mean 0.000560 std 0.321842 min -1.052387 25% -0.155583 50% 0.089744 75% 0.230473 max 0.432826 Name: er1, dtype: float64	t statistics -8.962246e+00 p value 8.165012e-15 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[10]: count 112.000000 mean 0.221069 std 0.184174 min 0.000399 25% 0.132014 50% 0.193761 75% 0.280250 max 1.655972 Name: er2, dtype: float64

TABLE V. LOGARITHMIC REGRESSION DICKEY–FULLER'S STATIONARY TEST RESULT

MFE for Logarithmic Regression	MAE for Logarithmic Regression
t statistics -9.330008e+00 p value 9.383905e-16 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[11]: count 112.000000 mean -0.000038 std 0.341830 min -0.893192 25% -0.190189 50% 0.044801 75% 0.180789 max 1.858257 Name: logr1, dtype: float64	t statistics -9.030881e+00 p value 5.449265e-15 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[12]: count 112.000000 mean 0.045008 std 0.017359 min 0.000524 25% 0.036760 50% 0.053744 75% 0.057982 max 0.060780 Name: logr2, dtype: float64

TABLE VI. POLYNOMIAL REGRESSION DICKEY–FULLER'S STATIONARY TEST RESULT

MFE for Polynomial Regression	MAE for Polynomial Regression
t statistics -9.625166e+00 p value 1.668167e-16 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[13]: count 112.000000 mean 0.000359 std 0.201375 min -0.736346 25% -0.114993 50% 0.039474 75% 0.151802 max 0.442972 Name: polr1, dtype: float64	t statistics -1.017576e+01 p value 6.899551e-18 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[15]: count 112.000000 mean 0.052333 std 0.043824 min 0.001397 25% 0.021065 50% 0.035169 75% 0.071805 max 0.193240 Name: polr2, dtype: float64

TABLE VII. POWER REGRESSION DICKEY–FULLER'S STATIONARY TEST RESULT

MFE for Power Regression	MAE for Power Regression
t statistics -9.406956e+00 p value 5.975783e-16 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 dtype: float64 Out[16]: count 112.000000 mean 0.048902 std 0.207610 min -0.709675 25% -0.049906 50% 0.107778 75% 0.193780 max 0.389498 Name: powr1, dtype: float64	t statistics -9.945184e+00 p value 2.600928e-17 lags used 0.000000e+00 number of observations 1.110000e+02 critical value (1%) -3.490683e+00 dtype: float64 Out[18]: count 112.000000 mean 0.178319 std 0.115891 min 0.003597 25% 0.094932 50% 0.162764 75% 0.242676 max 0.709675 Name: powr2, dtype: float64

V. CONCLUSION

In conclusion, this work has, to some extent, achieved its aims and objectives. The purpose was to examine students' historical data to find out if variables in the data can help in developing a model that can be used to predict students' final performance so that corrective and preventive measures are taken earlier. Fortunately, some attributes in the student dataset were found to serve that purpose. This shows that the use of data mining techniques on academic data can be counted as one of the critical success factors in educational management and planning, curriculum development and other forms of decision making in an academic setting.

REFERENCES

[1] D. Tomar and S. Agarwal, 'A survey on data mining approaches for healthcare', Int. J. Bio-Science Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.

[2] P. Wongchinsri and W. Kuratach, 'A survey - Data mining frameworks in credit card processing', 2016 13th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2016, 2016.

[3] H. Mushtaq et al., 'Educational data classification framework for community pedagogical content management using data mining,' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 329–338, 2019.

[4] G. A. El-refae, 'Predicting Students' Academic Performance Using Artificial Neural Networks : A Case Study', J. Comput. Sci., vol. 8, no. 5, pp. 97–100, 2010.

[5] G. S. Gowri, R. Thulasiram and M. A. Baburao, 'Educational Data Mining Application for Estimating Students Performance in Weka Environment', IOP Conf. Ser. Mater. Sci. Eng., vol. 263, no. 3, 2017.

[6] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho and G. Van Erven, 'Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil', J. Bus. Res., vol. 94, no. August 2019 pp. 335–343, 2019.

[7] R. S. Bermudez, J. O. Manalang, B. D. Gerardo and B. T. Tanguilig, 'Predicting faculty performance using regression model in data mining', Proc. - 2011 9th Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2011, pp. 68–72, 2011.

[8] P. Chertchom, 'A comparison study between data mining tools over regression methods: Recommendation for SMEs', Proc. 2018 5th Int. Conf. Bus. Ind. Res. Smart Technol. Next Gener. Information, Eng. Bus. Soc. Sci. ICBIR 2018, pp. 46–50, 2018.

[9] E. Osmanbegovic and M. Suljic, 'Data mining approach for predicting student performance', J. Econ. Bus., vol. X, no. 1, pp. 3–12, 2012.

[10] B. Kumar Baradwaj, R. Scholor, S. Pal and S. Lecturer, 'Mining Educational Data to Analyze Students & Performance', 2011.

[11] K. David Kolo, S. A. Adepoju and J. Kolo Alhassan, 'A Decision Tree Approach for Predicting Students Academic Performance', Int. J. Educ. Manag. Eng., vol. 5, no. 5, pp. 12–19, 2016.

[12] O. S. Akinola, B. O. Akinkunmi, and T. S. Alo, 'A Data Mining Model for Predicting Computer Programming Proficiency of Computer Science Undergraduate Students', African J. Comput. ICT, vol. 5, no. 1, pp. 43–52, 2012.

[13] Z. J. Kovačić, 'Early Prediction of Student Success', Informing Science & IT Education Conference (InSITE), 2010, pp. 648–665.

[14] P. Chapman et al., 'Step-by-step data mining guide', DaimlerChrysler, 2000.

[15] R. K. Agrawal and R. Adhikari, 'An Introductory Study on Time Series Modeling and Forecasting', arXiv Prepr. arXiv1302.6613, vol. 1302.6613, pp. 1–68, 2013.