

Prediction of Engineering Students' Academic Performance Using Artificial Neural Network and Linear Regression: A Comparison

Pauziah Mohd Arsad, Norlida Buniyamin

Faculty of Electrical Engineering
Universiti Teknologi Mara
40450 Shah Alam Malaysia
pauzial67@salam.uitm.edu.my
nbuniyamin@salam.uitm.edu.my

Jamalul-lail Ab Manan

Mimos Malaysia Berhad
Technology Park Malaysia
jamalul.lail@mimos.my

Abstract - Predicting students' performance is very important if not crucial especially in engineering courses. This is to enable strategic intervention to be carried out before the students reach the higher semesters including the final semester before graduation. This paper presents a comparison study between Artificial Neural Network (ANN) and Linear Regression (LR) in predicting the academic performance. Cumulative Grade Point Average (CGPA) was used to measure the academic achievement at semester eight. The study was conducted at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia. Students' fundamental subjects results at first semester were used as independent variables or input predictor variables while CGPA in the final semester that is at semester 8 is used as the output or the dependent variable. Performances of the models were measured using the coefficient of Correlation R and that of Mean Square Error (MSE). The outcomes of the study from both models indicate a strong correlation between fundamental results for core subjects at semester one or semester three with the final CGPA.

Keywords— Prediction; Engineering fundamentals; academic performance; ANN; LR

I. INTRODUCTION

Students' performance prediction is an important issue to most academic institutions of higher learning. This had led to many researches in prediction work that included students from different background and study areas such as business, medical and computer technology [1-3] however, most past researchers used predictor variables or the independent values from demographic profiles. Data used were collected mostly from survey forms based on students' former background education, residency region, gender and Scholastic Aptitude Test (SAT) scores.

In the past, statistical Package for Social Sciences (SPSS) has been very popular and was used extensively by researchers in this area. Recently, the use of Artificial Neural Network (ANN) is gaining popularity with researchers working in the area of students' performance prediction. [2,

4, 5]. However, these researchers developed NN models using data from demographic background as inputs to the model. The study presented in this paper, used the Grade Point (GP) of fundamental subjects scored by the students at first semester as inputs. Once the students are accepted into the Program based on merits set by the Faculty of Electrical Engineering UiTM, then every help should be offered to help students to perform in their study before graduation which is in line with the Vision and Mission of the University. [6-8].

This paper presents the results of a study that compares the academic performance of Matriculation and Diploma Students using two different methods which are Multiple Linear Regression (LR) and Neural Network (NN).

II. DATA COLLECTION AND MODEL BUILDING

Data of Matriculation and Diploma students were compiled in Excel format which included identity number, gender, Grade Points of subjects scored at semester one and semester three for Matriculation students. For the Diploma students, the Grade Points of subjects were collected from semester three as these students enter the course at semester 3 since they were given credit exemption for courses in semester one and two. The most important data is the identity number of students as that number will relate to its final CGPA at semester eight. There were 391 matriculation students from three batches, which are from intakes in July 2005, 2006 and 2007. The Diploma students totalled up to 505 from 3 intakes in July 2006, 2007 and 2008. There were seven (7) subjects attempted by students at semester one but for this study we omitted subjects like Co-curriculum and Laboratory work as input or independent variables. At semester three we again omitted Laboratory work and Tamadun Islam (Islamic Civilisation) as such subjects are not basic foundation to higher courses along the path of the program.

III. LINEAR REGRESSION MODEL AND SPSS

Statistical Package for Social Sciences (SPSS) now known as IBM SPSS (Statistical Product and Service Solutions) was first released in the original version in 1968. SPSS is most widely used for statistical analysis in social science. It is commonly used by market researchers, health researchers, survey companies, government, education researchers and marketing organizations. The original SPSS manual has been described as one of sociology's most influential books. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the data file) are features of the base software. Statistics base software package includes Descriptive Statistics (cross tabulation, Frequencies), Means, ANOVA, Correlation and prediction for numerical outcomes, linear Regression, (LR) [9]

LR is used when we want to predict the value of a variable based on the value of another variable. When there are more than one input variable, it becomes multiple linear Regression or just multiple Regression. In this research model, the dependent variable is the final CGPA8 while the independent variables are GP scored for fundamental subjects like Circuit Theory (CT), Fundamental Electronics (FE), Signals and Systems (SS), Mathematics (MAT) and Communication Theory (COMM).

Then the same students were followed through with subjects at semester three. This time around, the subjects included Digital system, Material Science, Signals and System 2 (SS2), Mathematics 2 (MAT2) and English 1.

Tables 1 and 2 below depict the model summary and the ANOVA output for Matriculation students at semester one.

TABLE 1 SUMMARY OF DEPENDENT AND PREDICTOR VARIABLES

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.736 ^a	.542	.535	.24261

a. Predictors: (Constant), CGPA1, MAT, CCT, COMM, FE, SS
 b. Dependent Variable: CGPA8

TABLE 2 ANOVA OF MODEL DATAMAXSEM1

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26.778	6	4.463	75.824	.000 ^a
	Residual	22.602	384	.059		
	Total	49.380	390			

a. Predictors: (Constant), CGPA1, MAT, CCT, COMM, FE, SS
 b. Dependent Variable: CGPA8

From Table 1, the coefficient of Correlation, **R = 0.736** and from Table 2, the **Mean Square Residual = 0.059**.

The following Figures illustrate the findings.

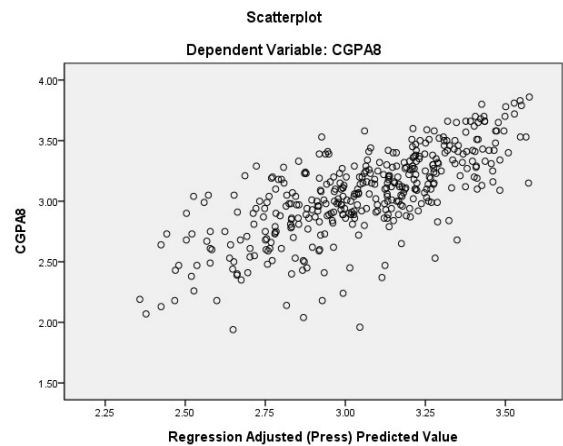


Fig. 1 Scatter plot of Actual CGPA8 and Predicted CGPA8

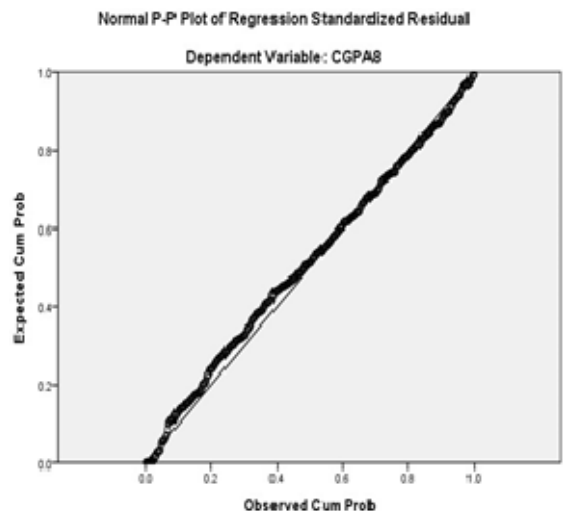


Fig. 2 Graph of Expected Residual and Observed Residual

From Fig. 1, it can be seen that at lower CGPA8, the predicted seemed to be higher than the actual value. From Fig. 2, the expected residual is above the best line and then the expected is below the best line at higher end.

Then the procedure was repeated for subjects at semester three (DataMaxSem3) with the same number of students and also tested for a different set of data of 505 students from Diploma (DataDipSem3). The following tables depict the model summary and ANOVA output respectively.

TABLE 3 SUMMARY OF DEPENDENT AND PREDICTORS OF DATAMAXSEM3

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.835 ^a	.698	.693	.19707

a. Predictors: (Constant), CGPA3, ENG1, DIGITAL, MATERIAL, MAT2, SS2
 b. Dependent Variable: CGPA8

TABLE 4 ANOVA OF MODEL DATAMAXSEM3

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.466	6	5.744	147.905	.000 ^a
	Residual	14.914	384	.039		
	Total	49.380	390			

a. Predictors: (Constant), CGPA3, ENG1, DIGITAL, MATERIAL, MAT2, SS2
 b. Dependent Variable: CGPA8

From Table 3 $R = 0.835$ and from Table 4 $MSR = 0.039$

The following figures show the Scatter plot of CGAP8 and regression Residual for MatrixSem3.

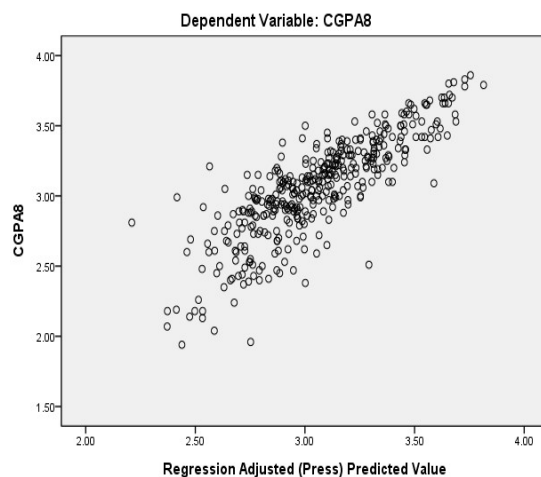


Fig. 3 Scatter plots of Actual CGPA8 and Predicted CGPA8

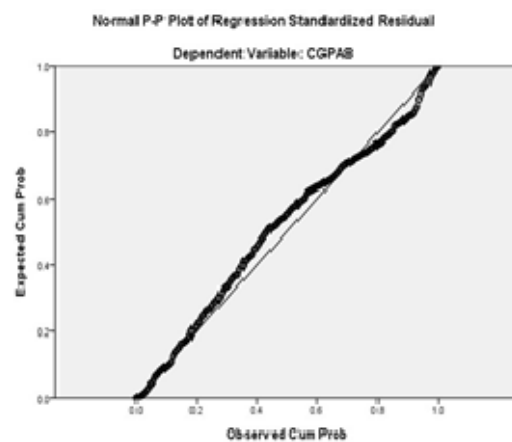


Fig. 4 Graphs Expected and Actual Residuals

TABLE 5 SUMMARY OF DEPENDENT AND PREDICTORS OF DATADIPSEM3

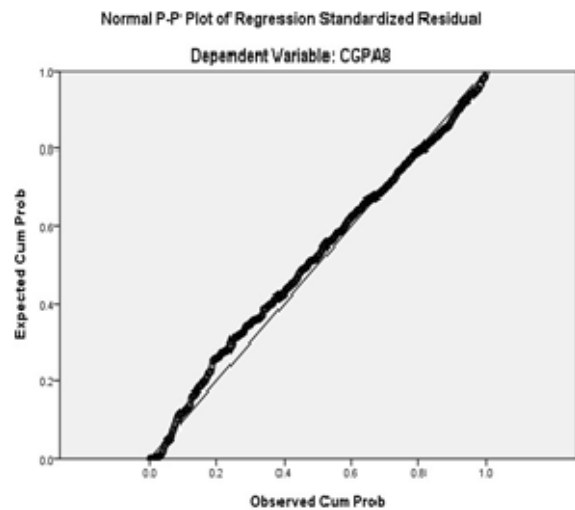
Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.740 ^a	.547	.541	.21163

a. Predictors: (Constant), CGPA3, Eng1, SS2, MATERIAL, DIGITAL, MAT2
 b. Dependent Variable: CGPA8

TABLE 6 ANOVA OF DATADIPSEM3

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26.919	6	4.487	100.177	.000 ^a
	Residual	22.304	498	.045		
	Total	49.223	504			

a. Predictors: (Constant), CGPA3, Eng1, ISS2, MATERIAL, DIGITAL, MAT2
b. Dependent Variable: CGPA8



From Table 5 $R = 0.740$ and from Table 6 $MSR = 0.045$

The following Figures depict the Scatter plot of CGPA8 and graph of Residuals.

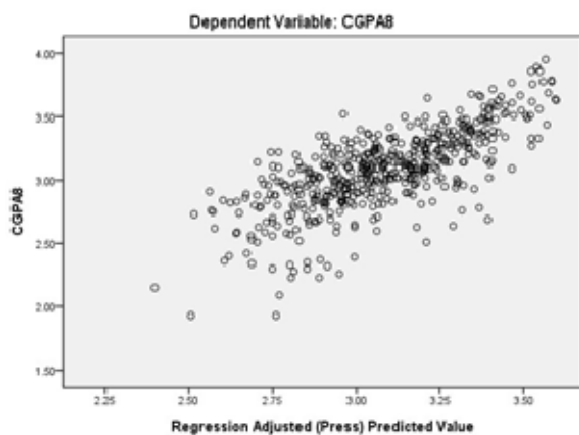


Fig. 5 Scatter plot of CGPA8 for DataDipSem3

Fig. 6 Graphs of Residuals for DataDipSem3

IV. NEURAL NETWORK MODEL

NN has been used extensively in prediction purposes. It has been used in automotive industries, medical, banking and aerospace industries. Lately NN has been applied in academic industry to predict students' performances.[8] The model developed earlier was used for Matriculation students' performance at semester one. This time the model was repeated for DataMaxSem3 and tested using another set of data comprising of 505 Diploma students with the GP scored in the same subjects attempted by the Matriculation students at semester three.

The following results only focussed on DiplomaSem3 and MatrixSem3.

TABLE 7 NN MODEL CONFIGURATIONS

Items	DataMaxSem3 (6 inputs)	DataDipSem3 (6 inputs)
Network Configuration	'logsig', 'logsig', 'purelin' [17, 35, 1]	'logsig', 'logsig', 'purelin' [17, 35, 1]
Learning Rate	0.6	0.6
Momentum Rate	0.95	0.95
Training Technique	Lavenberg Marquardt (lm)	Lavenberg Marquardt (lm)
Training Goal	1.00E-03	1.00E-03
Coefficient of Correlation	0.9774	0.9245
Training Patterns	273	355
Testing Patterns	118	150
No. of Variables	6	6
Mean Square Error (MSE)	0.0409	0.0488

From Table 7 above, it can be seen that R for DataMaxSem3 and DataDipSem3 are **0.09774** and **0.9245** respectively. The following figures depict the comparison between the Predicted and actual CGPA for the tested data both at 30% while training was set at 70%.

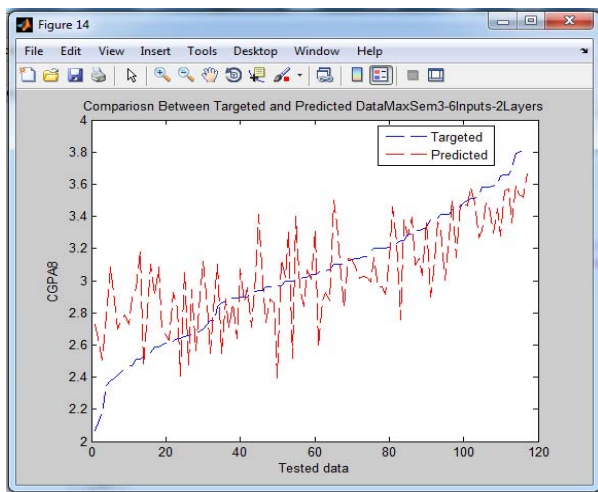


Fig. 7 Comparison between actual and predicted CGPA8 for DataMaxSem3

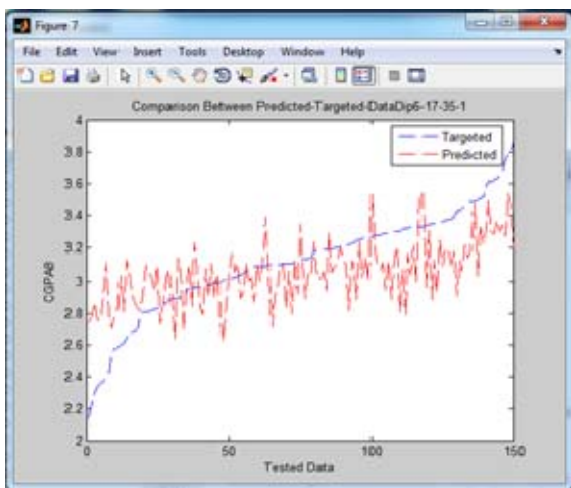


Fig. 8 Comparison between Actual and Predicted CGPA8 for DataDipSem3

From Figures 7 and 8 it can be seen that at lower CGPA, the predicted is higher than the actual or targeted while at higher CGPA the predicted is lower than the actual CGPA8.

V. RESULT AND ANALYSIS

Table 8 shows the MSE and MSR respectively for NN and LR model developed. It can be seen that both models yielded the same Residual or error in every case. However from

Table 9 the R for NN is much higher than R for LR model for all cases. The trend and pattern of the outcomes of the prediction model holds true for all three different cases.

TABLE 8 COMPARISON OF RESIDUALS

MSE/MSR	LR	NN
DataMaxSem1	0.059	0.054
DataMaxSem3	0.039	0.041
DataDipSem3	0.045	0.048

TABLE 9 COMPARISON OF COEFFICIENT OF CORRELATION

R	LR	NN
DataMaxSem1	0.76	0.9281
DataMaxSem3	0.835	0.9774
DataDipSem3	0.74	0.9245

VI. DISCUSSION AND RECOMMENDATION

As can be seen earlier, the student's performance in the fundamental subjects at semester one for the Matriculation and at semester three for Diploma students do influenced their final CGPA8 upon graduation. These fundamental subjects form the foundation or pre-requisite to all other subjects in the following semesters in the path of their study plan. Those with low CGPA at the start are advised to re-schedule to take lesser credit hours during full semester and to register for intersession or summer schools in order to improve the CGPA. The role of academic advisors (AA) and lecturers are critical this time around to motivate, convince and to inculcate positive thinking to fellow students until graduation.

As those with high CGPA at the start of the Program are to work smart and steadily maintain the CGPA. This is due to the fact that it is extremely difficult to maintain as the total credit hours that form the denominator for the final CGPA calculation. Furthermore, the subject matter is more difficult for the students as they progress further into the higher semesters of the electrical engineering courses.

VII. CONCLUSION

This paper presented LR and NN prediction models to predict students' performance based on multiple entry levels namely Matriculation and Diploma entry levels. This study is limited to Electrical Engineering Degree students at the Faculty of Electrical Engineering UiTM as the data were

obtained from Students Information Management System (SIMS) developed for academic communities of Universiti Teknologi MARA. From the findings, both prediction models indicated similar results as far as Mean square Error is concerned. The trend and pattern of the outcomes of the models hold true for all three cases. This verified that the fundamental subjects had a strong influence on the final CGPA8. Results of the study showed that strong students' abilities in engineering fundamentals will strongly influence the overall academic performance in an engineering program. Students with lower CGPA at the start can be motivated to do better by early strategic intervention of academic advisors and lecturers before reaching the final semester. Matriculation intake students could be helped as early as the first semester result without wasting further time until semester three. Diploma students can be monitored only at the end of semester three as there have only six semesters to complete the entire Program.

ACKNOWLEDGMENT

The authors gratefully acknowledge and thank University Teknologi Mara (UiTM), Shah Alam, Malaysia, for the Research Intensive Faculty (RIF) Grant (600-RMI/DANA/5/3/RIF (195/2012).

REFERENCES

- [1] O. M. N. Alfian E, "Undergraduate Students' Performance :a case of University Malaya," *Emerald Quality Assurance in Education vol 13 no 4*, vol. 13 2005.
- [2] R. D. Ibrahim Z, "Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression, 2007," in *21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur, 2007*.
- [3] W. M. C. Heather J, Nicola Price, "Predictors of academic success in a Bachelor of Nursing course," *Elsevier Sciences, Nurse Education Today*, vol. 23, pp. 246-254, 2003.
- [4] M. A. Wang T, "Using Neural Networks to Predict Student's Performance," in *Proceedings of the International Conference on Computers in Education (ICCE'02)*, 2002.
- [5] I. Z. Rusli N M, Janor RM, "Predicting Students' Academic Achievement: Comparison between Logistic Regression, Artificial Neural Network, and Neuro-Fuzzy," in *International Conference on Computers*, 2008.
- [6] Arsad P M, Buniyamin N, Rasheid N A, "Evaluation of Academic Performance of Electrical Engineering Bachelor Degree Students from Various Entry Levels: a Case Study," in *9th WSEAS International Conference on EDUCATION and EDUCATIONAL TECHNOLOGY (UDU'10)*, IWATE Japan Dec 2010 Iwate Prefectural University Japan, 2010.
- [7] Arsad P M, Buniyamin N, Manan J A, Hamzah N, "Proposed Academic Students' Performance Prediction Model: A Malaysian Case Study," in *International Congress on Engineering Education (ICEED), 7-8 December 2011, Malaysia 2011 3rd*, 2011.
- [8] Arsad P M, Buniyamin N J A Manan, "Neural Network Model to Predict Electrical Students' Academic Performance," in *International Congress on Engineering Education, ICEED 2012*, Park Royal Penang Malaysia, 2012.
- [9] SPSS, "SPSS and Linear Regression <https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php>. Accessed 16 August 2013," Accessed 16 August 2013.