Predicting Undergraduate Level Students' Performance Using Regression

Akuma, S. and Abakpa, H.
Department of Mathematics and Computer Science
Benue State University, Makurdi
sakuma@bsum.edu.ng

Abstract
Students' academic performance in the university environment changes from one academic year to another as they climb up the ladder of their academic programme. Predicting students' academic performance in higher educational institutions is challenging due to the lack of a central database of students' performance records. The other challenge is the lack of standard methods for predicting students' performance and other moderating factors like physical, economic and health that affect students' progress. In this work, we predicted students' performance based on previous academic results. A model to predict students' performance based on their Cumulative Grade Point Average (CGPA) was developed using Linear Regression Algorithm. A dataset of 70 undergraduate students studying Computer Science was analyzed and the results show that the model was able to predict the 4th year CGPA of the Students using the previous Cumulative Grade Point of the past three years with an accuracy of 87.84%, and a correlation of 0.9338. This study also identified students' second semester CGPA in the first year and their first semester CGPA in the second year as the most important CGPAs that affect the accuracy.

Keywords: Linear regression, CGPA, prediction, academic, student performance.

Introduction

Students' low academic performance at the end of a university degree has been a long-standing problem, especially among undergraduate students. Today, Universities are working in a very dynamic and powerfully viable environment. Hence, they gather large volumes of data about their students in electronic format. However, they are data rich but information poor which results in unreliable decision making. The main challenge is the effective transformation of large volumes of data into knowledge to improve the quality of managerial decisions and to predict the academic performance of students at an early stage to help University lecturers to focus on both excellent students and also identify students with low academic performance, and find ways to support them (Asif *et al* 2015). Universities have been using many data mining techniques to analyze educational reports stored in the educational institute such as enrollment data, students' performance, teachers' evaluations, gender differences, and many others. Data mining techniques may, for example, give a university the needed information to better plan several students' enrollment, students' dropout, early identification of weak students, and efficiently allocate resources with a precise approximation. Data mining is a powerful tool for academic intervention (Mitchell 2007). Through data mining, a university could, for example, predict accurately which students will or will not graduate. The university could use this information to assist weak students to improve their academic performance.

Prediction of student academic performance has long been regarded as an important research topic in many academic disciplines because it benefits both teaching and learning. It helps instructors develop a good understanding of how well or poor the students in their classes will perform, so instructors can take proactive measures to improve student learning. This paper presents a model that predicts students' final CGPA and the class of degree upon graduation, using their previous CGPAs of six semesters, without considering factors like economic, social and psychological effects on the students. We assume that from a managerial point of view, it is easier to use academic features to predict students' performance than to use economic, social and psychological factors. Thus, if a reasonable prediction can be reached with CGPA only, it makes the implementation of a Student Performance Prediction System (SPPS) in a university easier (Asif et al 2015). The other Sections of this paper includes Section 2 for related work, Section 3 for methodology, Section 4 for results and discussion and Section 5 for Conclusion and future work.

Related work

Several studies deployed and compared various data mining techniques for either classification or regression tasks in respect to predicting academic performance of different educational levels (Raheela Asif 2015; Imran et al 2019; Shahiria, 2015; Kabakchieva *et al*, 2011). Previous literature has compared different predictive models for the prediction of students' academic performance to make academic decisions. Huang *et al*. (2013) conducted research using four types of mathematical models, which are; the multiple linear regression model, the multilayer perception network model, the radial basis function network model, and the support vector machine model. Student's cumulative GPA, grades earned in four pre-requisite courses (statics, calculus I, calculus II, and physics), and scores on three dynamics mid-term exams were used as input variables, student's scores on the dynamic final exams constituted the output of the models where a total of 2907 data points were collected from 323 undergraduates in four semesters and from the four mathematical models and six different predictor variables that were used, they were able to develop 24 predictive mathematical models. It was discovered from their analysis that Average Prediction Accuracy (APA) and Percentage of Accurate Prediction (PAP) was slightly affected by these mathematical models. From this study, it was shown that the combination of predictor variables has only a slight effect on APA, but a profound effect on PAP. These models have APA of 81%−91%, and PAP of 40%−72% and the most important predictor variables that affected prediction accuracy were: dynamics mid-term exam, cumulative GPA, dynamics mid-term exam, and physics. These predictor variables, however, vary with different types of mathematical models employed (MLR, MLP, RFB, or SVM). Results show that SVM outperformed other models with an accuracy of 89% and therefore identified as the most superior model for their prediction. Asif et al. (2015) in

their study mined data of four academic cohorts comprising 347 undergraduate students using different classifiers which include Decision Tree, Rule Induction, 1-NN, Naive Bayes (NB) and Neural Network (NN) to predict students' graduation performance in 4th year at University using their pre-university marks and marks of 1st and 2nd-year courses with no socio-economic and demographic features. Their research indicates that Naive Bayes outperformed the other classifiers and it gave an accuracy of 83.65% which according to them, is better than any accuracy given by related work with socio-economic and demographic features.

Research by Abimbola *et al.* (2018) compared two neural network models (Multilayer Perceptron and Generalized Regression Neural Network) in predicting students' academic performance focusing on the academic factor (students' results) to identify the best model for predicting academic performance. Data used was obtained from the Computer Science and Engineering Department of Obafemi Awolowo University's database of graduated students. MATLAB was used to simulate the model and Mean Square Error; Receiver Operating Characteristics and Accuracy as the Performance were the criteria used to evaluate the result. From the results obtained, Multilayer Perceptron had an accuracy of 75% but Generalized Regression Neural Network outperformed Multilayer Perceptron with an accuracy of 95%. Ermiyas et al. (2017) also used Neural Network, Naive Bayesian and Support Vector Regression to predict student graduation CGPA, they performed 3 experiments which was shared into scenarios; Scenario 1: Students' university course scores from the first 2 years (i.e., scores of 23 courses) were used for predicting final CGPA, Scenario 2: Students' university course scores from the first 3 years (i.e., scores of 35 courses) were used for predicting final CGPA. Scenario 3: The students' Semester GPA at the end of each semester from the first 3 years were used for predicting the final CGPA. From their experimental result, SVR has the shortest time to build the model, while LR has the second shortest time and NN took the longest time for the first scenario. Regarding the second and third scenarios, LR has the shortest time to build the models, while SVR has the second shortest time and NN took the longest time of all the three prediction methods. Based on the correlation coefficient(R) and RMSE, the result

shows that LR was the best, SVR was the second-best while NN was the least accurate of the three. This indicates that the LR method outperforms the other two prediction methods. Abu-Naser *et al.* (2015) also used Artificial Neural Network (ANN) model, to predict student performance at the faculty of Engineering and Information Technology in the Al-Azhar University of Gaza, where a total of 150 sophomore students' records was collected. 10 factors which are High school score, Results in math l, math ll, Electrical circuit, Electronic l in the student freshman year, CGPA of the freshman year, Type of High school, Location of High school and student's gender which was obtained from student registration records, constituted the input variables for their model and student's CGPA on graduation as the output variable. They used feed-forward Backpropagation as a neural network; their model was able to predict accurately 11 out of 13 for the excellent data (which represents students CGPA in the range of 90% to 100%), 10 out of 12 of the very good Data (which represents students with CGPA in range of 80% to less than 90%) and 9 out of 11 of the good data (which represents students with CGPA in range of 70% to less than 80%), and 8 out of 9 of the poor data (which represents students with CGPA in the range of 65% and less than 70%) which was used to test the Network's topology. Artificial Neural Network gave an overall accuracy of 84.6%.

A similar study carried out by Isljamovic and Suknovic (2014) employed different Artificial Neural Network Algorithms to predict student's graduation GPA; data used for the study was collected from 1787 graduated students of the Faculty of Organizational Science, University of Belgrade. The input data (predictors) for the study consist of 15 variables which include students' characteristics (students' gender), high school information (high school GPA and high school type) admission data (entrance examination points) and the first-year examination grades (individual grades at 11 examinations of the first year of elementary studies), while the GPA at the end of their studies as output variable. In their quest to find the best quality model, they used six different methods in building Neural Network Models: Quick, Dynamic, Multiple, Prune, RBFN (Radial Basis Function Network) and Exhaustive Prune. Absolute Average Error, standard Deviation and Linear Correlation were used to measure the performance of the model and

comparative analysis of the result, on the test sample, showed that the ANN model gives acceptable results. An Absolute Average Error of prediction in all the networks was from 0.231 to 0.259 and the linear correlation coefficient was over 87%. Chaudhari *et al.* (2017) conducted research using a hybrid procedure based on a Decision tree and data clustering to predict students' GPA (performance in their next semesters). The data used for this research was collected from SSBT College of Engineering and Technology, Jalgaon. Student's performance indicators such as class quizzes, mid and final exam assignment lab work were investigated. Results of previous students and behaviours were also obtained. They used the K-Mean Clustering, Naïve Bayes, and C4.5 algorithm for prediction. From the research conducted, Naïve Bayes proved to be the best algorithm with an accuracy of 96% followed by the C-Mean algorithm with 95% accuracy and then 94% accuracy for the K-Mean algorithm. Amjad Abu Saa (2016) carried out a study using multiple decision trees (C4.5, CHi-Squared Automatic Interaction CHIAD, CART) and Naïve Bayes Classification techniques on a group of students enrolled in different colleges in Ajman University of Science and Technology (AUST) the United Arab Emirates. He used multiple performance indicators such as personal, social, and academic questions to classify students' performance (GPA) into "Excellent", "Very Good", "Good" or "Pass". Amjad Abu Saa (2016) found that after running the CART decision tree with 10-folds Cross-Validation, an accuracy of 40% was obtained which happens to be the best followed by ID3 with 33.3% accuracy, 35.19% by the C4.5 decision tree and 34.07 by CHIAD. His observation led to the conclusion that the discretization of the class attribute was not suitable enough to capture the differences in other attributes that is, the class attribute was not independent so he used the Naïve Bayes method. Naïve Bayes shows that high school performance, mother's occupation and school discount are important features in determining student's performance with an accuracy of 36.4%.

Research conducted by Dorina Kabakchieva (2012) employed four classification algorithms: a rule learner (OneR), a common decision tree algorithm C4.5 (J48), a neural network (Multilayer Perceptron), and Nearest Neighbour algorithm (IBk) to build a model that classified students into two classes- Weak and Strong, based on their university performance and pre-university data. These algorithms were applied on 10067 instances and 14 attributes using WEKA classifiers and the model was evaluated using: The percentage of correctly/incorrectly classified instances, Kappa Statistic, True Positive (TP) and False Positive (FP) Rates, Precision, Recall, F-measure and ROC Area. From the result, the highest classification accuracy (% of correctly classified instances was achieved by Neural Network Algorithm of 73.59% and this same model was identified as the only model that predicts the "Strong" class with higher accuracy (TP Rate=77%) than the "Weak" class (TP Rate=70%). Neural Network was also used in the study by Zaidah et al. (2007) where students' demographic profile and CGPA for the first semester of the undergraduate study were used as the predictor variables in predicting the final cumulative grade point average (CGPA) of students upon graduation. They used three predictive models which are; Artificial Neural Network, Decision Tree and Linear Regression. Data of 206 students were used and a correlation of 0.87654 was obtained.

The research follows Zaidah *et al.* (2007) and Isljamovic and Suknovic (2014) work which reported a high correlation for Linear regression among other classifiers based on a given number of features. This work, however, uses regression analysis to predict the result and class of degree of final year students, using previous CGPAs as inputs.

Methodology

A variety of techniques, ranging from traditional mathematical models, and data mining have been employed to predict academic performance. In these techniques, a set of mathematical formulas were used to describe the quantitative relationships between outputs and inputs (i.e dependent and independent variables). The approach used in this work for extracting knowledge out of the students' performance dataset was Linear regression. Figure 1 describes the methodology used for this work.
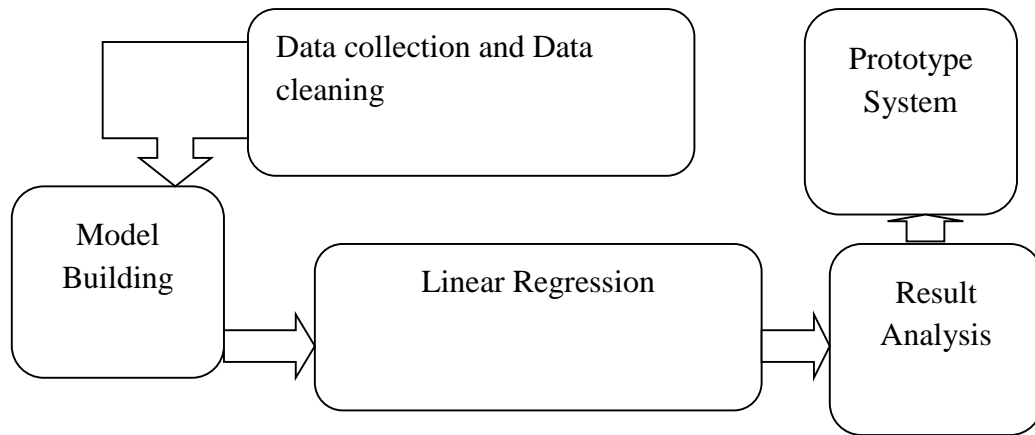
Figure 1: Proposed methodology

Data Collection and Data cleaning

The data used for this study was students' academic data of a set that graduated from the Mathematics and Computer Science Department of Benue State University, Makurdi, Nigeria. The data consist of a one-degree option (B.Sc. Computer Science), with a total of 70 graduated students who enrolled in the 2003/2004 academic session. The CGPAs of six consecutive semesters for each of these students were used as the input parameters to predict their last CGPA (CGPA8), which infer their class of degree. Data collected were stored in an Excel spreadsheet format according to the given features, but the names of the students were excluded from the data stored for confidentiality. Data cleaning was done to eliminate incomplete information and duplicates, after which the dataset was divided into two different sets: the training set and the testing set. Figure 2 shows the pictorial view of the data set.

Figure 2: A sample of the dataset Students' CGPAs

Tools used

The prediction model was created using python language. It is a language commonly used for machine learning applications. It has in-built libraries for the method selected for this study and also creates the necessary output for evaluating the results of predictions. The code written in the python language was run on Microsoft Visual Studio 2019.

Evaluation metrics

Mean Absolute Error (MAE): Mean Absolute Error is one of the evaluation metrics used for regression analysis. It measures how far away the predicted value is from the actual value. It is the average over the test sample of absolute differences between prediction and actual observation where all individual differences have equal weight. Equation 1 is the formula for calculating MAE.

$$MAE = \frac{\sum_{i=n}^{n}|y-x|}{n} \qquad (1)$$

where y is the predicted value, x is the Actual value and n is the total number of records.

Root Mean Square Error (RMSE): The second evaluation metric for regression analysis is the Root Mean Square Error (RMSE). It is used to measure the differences between the actual values and the predicted values by taking the square root of the residual (i.e the differences). Equation 2 is the formula for calculation RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=n}^{n}|y-x|}{n}} \qquad (2)$$

Results

Python and its libraries were used as a tool for analyzing the data. The Linear Regression model used for this work seeks to establish a relationship between input and output variables along with their coefficients. Linear regression is a type of predictive analysis model which examines a set of predictor variables (independent variable) in predicting an outcome and determine specific predictors of the outcome variable (dependent variable). These regression estimates explain the relationship between the dependent variable and independent variables. A prediction is accurate if the error between the predicted and actual values is within a small range. In this work, a percentage split of 70% of the data was used for training while 30% was used for the testing. The inputs of the model were the students Semester CGPA at the end of each semester for the last 3 years. Linear regression was run on the data set to predict the CGPA of the 4th year and class of degree. The model produced an accuracy of 87.84%. The evaluation of the model shows the Mean Absolute Error (MAE) to be 0.1505, the Root Mean Square Error (RMSE) to be 0.2199 and a high positive correlation of 0.934. This implies that there is a relationship between the dependent variable and the independent variable as shown in Table 1. Figure 3 shows a graph of the predicted values against the actual values.

Table 1: Summary of the prediction result

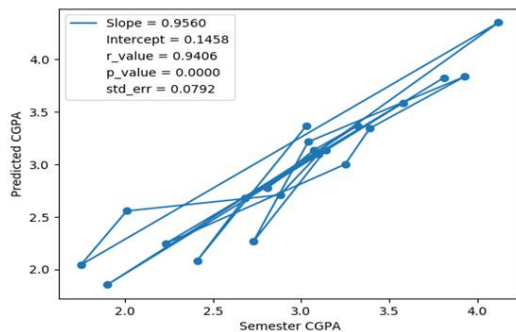| Prediction method	MAE	RMSE	CORRELATION COEFFICIENT	ACCURACY
Linear Regression	0.1506	0.2199	0.9338	87.84

Figure 3: Graph of the predicted values against the actual values

To make the system automatic, a prototype system was developed to automatically generate predicted values of the CGPA when previous students' CGPAs are inputted. Figure 4 shows the flowchart of the system.
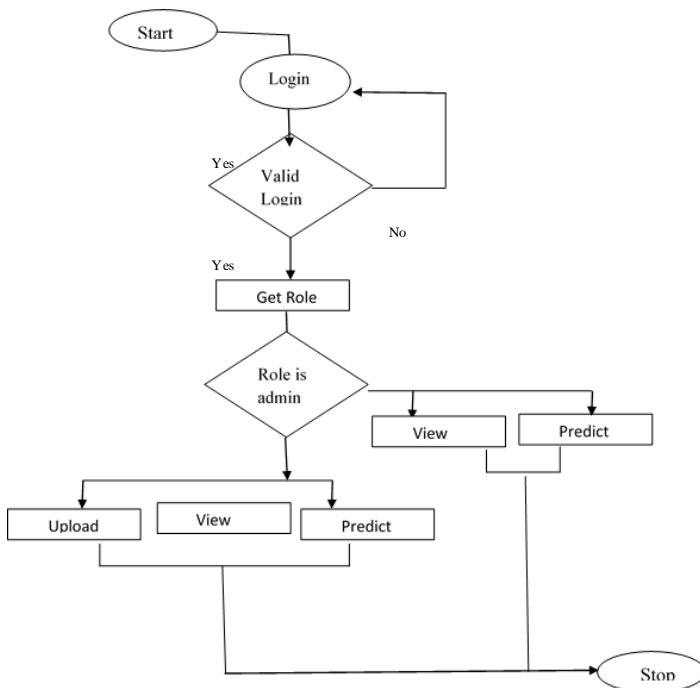
Figure 4: Flowchart of the prototype system

The Login page serves as an introductory page to the user. It gives access to the user. The username and password are security checks to grant access to the system. After a successful login as an Admin, the user will be directed to the Admin Page where he/she can upload students, upload scores, upload courses, upload CGPA. The Result page is shown in Figure 5.

STUDENTS PERFORMANCE PREDICTOR

Mat No	Actual Final CGPA	Predicted CGPA	Predicted Class of Degree
BSU/SC/03/1453	2.41	2.084314863402314	Third Class
BSU/SC/03/1456	2.81	2.7800583657354	Second Class Lower
BSU/SC/03/1471	2.01	2.557517441462371	Second Class Lower
BSU/SC/03/1475	3.81	3.8262557704367963	Second Class Upper
BSU/SC/03/1476	3.10	3.1011973013167995	Second Class Lower
BSU/SC/03/1477	2.23	2.2458085650607145	Third Class
BSU/SC/03/1479	1.75	2.0439822984660867	Third Class
BSU/SC/03/1480	3.25	3.0019402640739274	Second Class Lower
BSU/SC/03/1482	3.93	3.838068371936901	Second Class Upper
BSU/SC/03/1484	3.32	3.3711577822990897	Second Class Lower
BSU/SC/03/1493	3.39	3.344533205769255	Second Class Lower
BSU/SC/03/1494	2.68	2.6789915038452605	Second Class Lower
BSU/SC/03/1499	2.88	2.7103330588342622	Second Class Lower
BSU/SC/03/1502	4.12	4.351196562949678	Second Class Upper
BSU/SC/03/1506	3.04	3.2151525637930733	Second Class Lower
BSU/SC/03/1507	3.58	3.587412316262295	Second Class Upper
BSU/SC/03/1510	3.07	3.1348604423579376	Second Class Lower
BSU/SC/03/1513	3.14	3.1392792965306517	Second Class Lower

Figure 5: The result page

BSU/SC/03/1516	3.03	3.3688381620555434	Second Class Lower
BSU/SC/03/1521	2.73	2.2681956486984447	Third Class
BSU/SC/03/1522	1.90	1.8555405906134164	Pass

Summary	
Mean Absolute Error	0.1506
Mean Squared Error	0.0484
Root Mean Squared Error	0.2199
Spearmans correlation	0.9338
Model Accuracy	87.84%

Figure 6: The result page

Discussion

This section discusses the results. In this work, students' semester CGPA at the end of each semester during their 3 years of study were considered as the input variables to predict students' 4th year (final) CGPA and the class of degree. The result shows minimal error as returned by the MAE and RMSE. The Linear regression produces a statistically significant value ($p = 0.0000$), with the slope as 0.9560 and the intercept as 0.1458. The accuracy of 87.84% also indicates that the model did not suffer overfitting or underfitting. The result is within the range obtained by previous research. Huang et al. (2013) reported 89% using SVM, Asif et al (2015) reported 83.65% using Naive Bayes. Abimbola et al. (2018) employed a neural network (MLP) and reported an accuracy of 75%. Abu-Naser et al (2015) also used a neural network and reported an accuracy of 84.6%. Another research using the neural network model by Dorina Kabakchieva (2012) reported an accuracy of 73.59%. Our report shows that Linear regression can effectively predict the performance of undergraduate students. Unlike other research that used other parameters as inputs, our research used students' CGPA of the previous six semesters as inputs. The research didn't consider the effect of environmental factors, social factors, and economic factors on students' performance. This study focused solely on predicting students' performance based on their academic results. The prototype system for predicting students' results automatically was designed using python; it presents the predicted results and the class of degree of the students.

Conclusion

Students academic results were predicted using linear regression. The result of this work has shown that with the linear regression method, prediction of students' final CGPA is possible with an accuracy of 87.84% and a correlation coefficient of 0.934. This work shows that data mining techniques such as Linear Regression can be used efficiently for modelling and predicting students' final CGPA in higher educational institutions. This research is not devoid of limitations, one of such is that the dataset used is solely from the department of Mathematics and Computer Science. Further research can include datasets from other departments and institutions to make the model more generalizable. Also, the study does not include socio-economic and psychological factors that have the potential of affecting student's performance. Future studies will include factors such as learning style, motivation and interest, teaching and learning environment in predicting student's academic performance.

References

Abimbola R. I., Olufemi D.N., Anuoluwapo O. A., Ogochukwu G.A. (2018). Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models. International Journal of Modern Education and Computer Science (IJMECS). Vol. 10, No. 6. Pp. 1-9. DOI: 10.5815/ijmecs.2018.06.01

Abu-Naser, S.; Zaqout, I.; Ghosh, M. (2015) Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. *International Journal of Hybrid Information Technology* Vol.8, No.2, pp.221-228 http://dx.doi.org/10.14257/ijhit.2015.8.2.20.

Amjad Abu Saa (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Scien ce and Applications*, volume 7, issue 5. pp. 212-220.

Asif R., Merceron A., Pathan M.K. (2015) Predicting Student Academic Performance at Degree Level: A Case Study, *International Journal of Intelligent System and Applications.* 2014;7:49. vol.7, no.1, pp.49-61,2015. DOI: 10.5815/ijisa.2015.01.05

Chaudhari, K. P.; Sharma, R. A.; Jha, S.S.; Bari, R.J. *(2017).* Student Performance prediction using Data Mining Approach. *International Journal of Advanced Research in Computer and Communication Engineering.* DOI:10.17148/IJARCCE.2017.63195

Dorina Kabakchieva (2012). Student Performance Prediction by Using Data Mining Classification Algorithms, *International Journal of Computer Science and Management Research.* Vol 1, no. 4. Pp.686-690.

Ermiyas B.B, Feidu A.G. (2017). Student Performance Prediction Model using

Machine Learning Approach: The Case of Wolkite University. *International Journal of Advanced Research in Computer Science and Software Engineering*. Vol 7, (2). Pp. 46-50. DOI: http://dx.doi.org/10.23956/ijarcsse/V7I2/01219.

Huang, N. F. (2013). Predicting Student Academic Performance in an Engineering Dynamics Course: A Comparison of Four Types of Predictive Mathematical Models, Computer and Education. Computers & Education 61(1). Pp.133–145. DOI: 10.1016/j.compedu.2012.08.015

Imran, M.; Latif, S.; Mehmood, D.; Shah, M. S. (2019) Student Academic Performance Prediction using Supervised Learning Techniques. *iJET* – Vol. 14, No. 14, 2019. https://doi.org/10.3991/ijet.v14i14.10310

Isljamovic, S. and Suknovic, M. (2014). Predicting students' academic performance using artificial neural network: a case study from faculty of organizational sciences. ICEMST 2014: International Conference on Education in Mathematics, Science & Technology. Vol. 1, pp. 68-72

.

Kabakchieva, D.; Stefanova, K.; Kisimov, V. (2011) Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.

Raheela Asif (2015) Predicting Student Academic Performance at Degree Level: A Case Study. *I.J. Intelligent Systems and Applications*, 01. Pp. 49-61. DOI: 10.5815/ijisa.2015.01.05

Shahiria, A. M; Husaina, W.; Rashida, N. A. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. *Procedia Computer Science* 72, pp. 414 – 422

Tom M. Mitchell. (1997). Machine Learning. McGraw-Hill.

Zaidah I., Daliela R. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression.2007. 21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur