International Journal of Scientific Engineering and Science
Volume 4, Issue 9, pp. 22-26, 2020. ISSN (Online): 2456-7361

Predictive Model to Predict the Test Scores of the Computer Skills-2 Course for Future Students in Irbid University College

Mutaz Khazal Khazaaleh[1]
[1]Department of Information Technology, Al-Balqa Applied University, Irbid, Jordan
Email address: mutaz.khazaaleh @ bau.edu.jo

Abstract— *In this paper, we are interested in predicting the computer skills-2 scores for Irbid University college students based on a relationship between the time spent studying for the computer skills-2 test and the final scores, we used it as training data to learn a model that uses the study time to predict the test scores of future students in Irbid University College who are planning to take computer skills-2 course. The predicting model use regression analysis as a type of supervised learning that use to predict the continuous outcomes, we are given a number of predictor variables (the time spent on study for computer skills-2 daily, the time spent on study before exam, the branch of secondary education (Al-Tawjihi) and the student major) and a continuous response variable (outcome/test scores), and we try to find a relationship between those variables that allows us to predict an outcome. The data for 400 students in two semesters was used as the test set for training the predictive model and we used data for 50 students to validate the predictive model. Results showed that the predicting model is useful to use to predicate students' scores. The error range between the predicative score and the actual score for the 50 students it's between +6 and -6. While the root mean squared error (RMSE) = 2.424871 and the root mean squared percentage error (RMSPE) = 4%. The time spent on study for computer skills-2 daily has the most critical influence on the student score.*

Keywords— *Machine learning, neural networks (NNs), multilayer perception (MLP), supervised learning, regression analysis, predicting model.*

I. INTRODUCTION

Occasionally some courses present great challenges and real problems for the students. These problems can be the reason for leaving studies or at least leads to negative consequences for the student's academic performance. For this reason, grasp the factors that lead to success (or failure) of students is consider one of the most an interesting problem the learning process. In our point of view, we can improve the learning process if it were possible to detect the study time that students need to get the required score in specific course before they start study this course.

Machine learning can be used effectively to learn a model that uses the study time to predict the test scores of future students or the opposite, predict the study time need to get the required score. Machine learning addresses the issue of how computers can be designed that will automatically develop over experience [1]. It is one of the fastest-growing technical fields of today, standing at the intersection of computer science and statistics and at the core of artificial intelligence and data science [1]. In order to automate complex tasks or make predictions, machine learning algorithms are widely used to detect patterns in data [2].

As shown in Figure 1 there three types of machine learning: unsupervised learning, supervised learning and reinforcement learning [3].

In supervised learning, the main objective is to learn from labelled training data a model that enables us to predict unknown or future data [3]. To predicting continuous outcomes by using supervised learning, the researchers using what is called regression analysis.

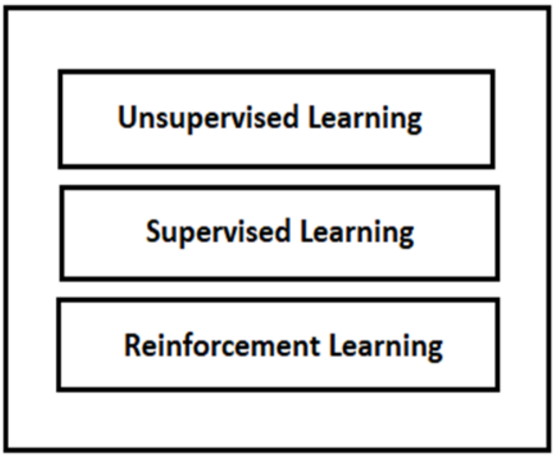

Fig. 1. The three types of machine learning.

In 1886, Francis Galton was devised the term regression in his article "Regression towards Mediocrity in Hereditary Stature" [4]. In regression analysis, a number of predictor (explanatory) factors and a continuous response (outcome) factors are given, and we are trying to find a relationship between those variables that enables us to predict the outcome. In this research we used the regression analysis to build predicting model that use to predict the continuous outcomes, we are given a number of predictor variables (the time spent studying and other variables) and a continuous response variable (outcome/test scores), and we try to find a relationship between those variables that allows us to predict an outcome. The students sample are from Al-Balqa Applied University. For this reason we need to know more about Al-Balqa Applied University.

Al-Balqa Applied University is an official Jordanian university, characterized by applied education at the level of the bachelor and intermediate diploma, especially in the fields of engineering, and overwhelming the overall disciplines of a scientific nature. Teaching in Al-Balqa Applied University began in the 1997/1998 academic year. The quarterly study method is applied according to the credit hour system. The university awards the following degrees: Master, BSC, intermediate university certificate and general professional diploma. The university is located in Al-Salt city in Jordan, which includes a group of colleges, and the university has 13 colleges distributed in different cities in Jordan. Irbid University College one of these colleges. It's located in Irbid city. Nine bachelor's degree majors are offered through Irbid University College.

The Computer Skills-2 course is a compulsory course for all of these disciplines. This course aims to develop the student's skills in programming by using visual basic 6 programming language and to provide them with the necessary and basic ones in order to create the useful applications. Topics covered include introduction, dialogue boxes, selection control structures, repetition, arrays and procedures and functions. Computer Skills-1 course is the prerequisite for Computer Skills-2 course. A student's final score in computer skills-2 course was calculated by using the relative evaluation system, the principles of which were determined by the course coordinator. Final score were evaluated out of 100 points possible. Course scores were calculated according to the contribution of the mid-term exam (40%), Assignments (10%) and the final exam (50%). As a result, the calculated course scores were converted to grade coefficients shown in Table 1.

TABLE I. Course score conversion table.

Grade	Score	Grade	Score
A	100-95	C	69-65
A-	94-90	C-	64-60
B+	89-85	D+	59-55
B	84-80	D	54-50
B-	79-75	D-	49-40
C+	74-70	F	39-35

This research aims to find predictive model using to predict the test scores of the Computer Skills-2 test for future students in Irbid University College based on the study time and other factors. This paper organized as follows: the problem introduced in the Introduction (Section I), which followed by a discussion on relevant past work in the Literature Review (Section II). The proposed framework for predictive model described in Methodology (Section II). The model analysed in the Results and Discussion (Section IV). Provide a further discussion on the highlights of the predictive model, directions for future work and conclusions in the Conclusion (Section V).

II. Literature Review

Machine learning techniques are useful, researchers use it to analyse many systems in many fields, or to invent strategies for controlling these systems, and it might be consider the cornerstone to improvement these systems output. Many scholars (such as [5-18]) studied the regression analysis and developed models to achieve predicative models. Their studies explained predicative models usage to predicate and improvement the systems output.

Many approaches for predicative models in teaching and learning systems addressed. There are several techniques for predicative models usage with teaching and learning systems to enhance and improve these systems. Me and my colleagues in 2011 we presented new e-learning quality matrix to ELQ (E-Learning Quality) assessment at AL-Balqa applied university [19]. In our study (2011) we deal with the quality of e-learning at AL-Balqa Applied University /Jordan. It was discuss the already developed measures on the basis of statistical analysis for data gathered from e-learning elements which evaluates the quality of e-learning applications and systems. The study organized around the seven categories of e-learning quality dimensions and also we identified the difficulties that prevent achieving a high level of quality in e-learning at AL-Balqa Applied University [19].

Baker and his colleagues (2010) generated a model by using a combination of feature engineering and linear regression to predict student performance on a paper post-test of PFL (preparation for future learning) [20]. To validate the model they compared the model with two models: Bayesian Knowledge Tracing that use to predict post-test performance for students and a model trained to detect transfer. To see how well each model can predict preparation for future learning. They found that by time the student completed 20% of the tutor software, the PFL detector achieves a large proportion of its predictive power, suggesting that the PFL detector can be used to drive intervention early enough to influence overall learning.

Another study to predict academic performance, career potential, creativity, and job performance had done by Kuncel et al. (2004) [21]. The study used meta-analysis (MAT) to addresses the question: could be predict performance in both educational and work domains by using general cognitive ability measure developed for predicting academic performance. The results indicate that the abilities assessed by the MAT are associated with other instruments of cognitive ability and that these abilities are generally accurate predictors of academic and vocational qualifications, as well as career potential and innovation tests.

In another trend, Kotsiantis et al. used machine learning techniques to deal with student dropout occurs quite often in universities providing distance education [5]. The data set provided by the 'Informatics' course of the Hellenic Open University. A prototype web based support tool has been constructed by implementing the Naive Bayes algorithm, one of machine learning algorithms. It can automatically recognize students with high probability of dropout and can be successfully used.

In 2015, Wehman and his colleagues developed a logistic regression model for predicting successful employment outcomes [22]. This model was developed based on a study conducted on second National Longitudinal Transition Study to determine variables associated with post-high school competitive employment. The sample of the study is 2900 of special education students who exited high school in 2002/2003 school year from United States.

In recent study in 2019, Verma et al. where presented age group predictive models by using machine learning for the real time prediction of the university students [23]. They used five supervised machine learning classifiers: K-nearest neighbour (KNN), Random forest (RF), Support vector machine (SVM), Bayesian network (BN) and decision tree (C5.1) on dataset from Hungarian and Indian University. The study found to support the real time prediction of the age group of the students towards four different ICT (information and communication technology) parameters. The study found when they joint classifiers (RT, SVM, and BN) that provides maximum accuracy of 91.4% in the prediction of age against 37 features.

III. Methodology

In this study, the scores of student in Computer Skills-2 course at the end of their semester, time spent on study for computer skills-2 daily, time spent on study before exam, the branch of secondary education (Al-Tawjihi) and the student major, were used to predict the scores for student in Computer Skills-2 course in future. Figure 2 shows an abstract representation of the proposed predictive model building.

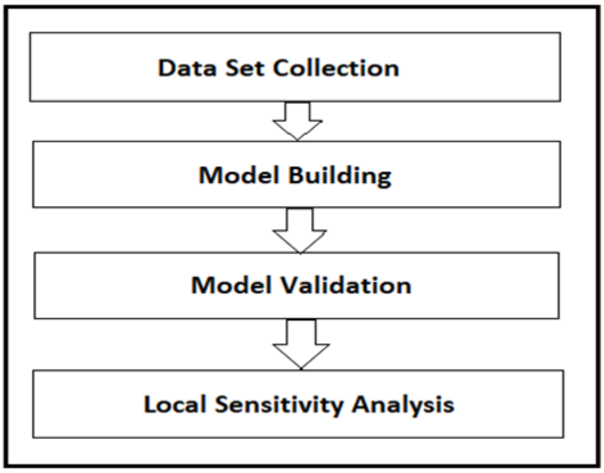

Fig. 2. Abstract representation of the proposed predictive model building.

A. Data Set Collection

The data set used in this research came from Irbid University College. This data set consists of 450 student's scores in Computer Skills-2 course. The students from nine different majors. The data collected over three consecutive semesters. Table 2 shows the numbers of students who included in the study divided by majors. The data set collected by questioner prepared especially designed for this study. We had chosen four factors that are considered the most important factors that mainly affect the score in the course of computer skills-2. These factors are: The branch of secondary education (Al-Tawjihi), student major, time spent on study for computer skills-2 daily and time spent on study for computer skills-2 before exam.

TABLE 2. The numbers of students included in the study divided by majors.

No	Student Major	Students
1	Business Management	51
2	Library and Information Management	42
3	Home Economics	78
4	Vocational Education	30
5	Nutrition and Food Processing	66
6	English Language and Literature	39
7	Arabic Language and Literature	57
8	Accounting	54
9	Raising a Child	33

B. Model Building

Without statistical assumptions, neural networks (NNs) can model the nonlinear relationship of dependent variable to independent variables completely based on data. The most popular NN structure is a multilayer perceptron (MLP); it uses a supervised learning method. In this study we used the MLP structure with back-propagation type supervised-learning algorithm. MLP is uses for modeling input–output relationships for the purposes of regression. Figure 3 shows the NN layers architecture's (input layer, a hidden layer, and an output layer) which used in this study. From input to output layer all neurons connected to each other. The number of neurons on the input layer chosen as the number of variables in the input dataset. During the iterative testing cycle it calculated the number of hidden layers and the number of neurons in each hidden layer.

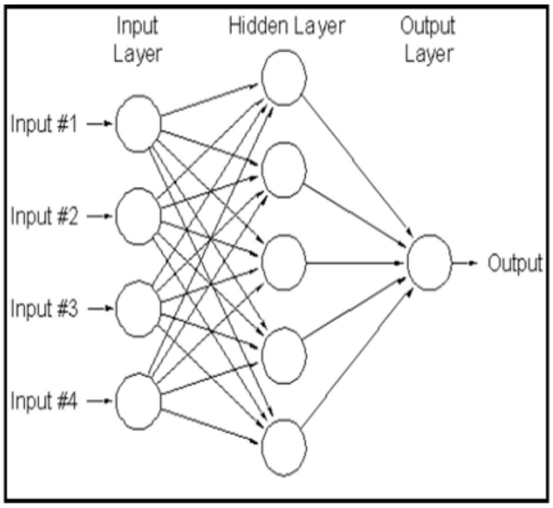

Fig. 3. Neural network layers architecture.

The input dataset was represented the data for 400 students in two semesters (first semester 2018/2019 and second semester 2018/2019) that used as the test set for training the predictive model.

C. Model Validation

Both the root mean squared error (RMSE) and the root mean squared percentage error (RMSPE) as shown in Eq. 1 and 2 were used to validate the predictive model. We used these as measures of accuracy for comparing prediction errors of the predictive model scores for 50 students in the third semester 2018/2019 against the real students score at this semester.

$$RMSE = \sqrt{\frac{\sum_{m=1}^{n}(A-P)^2}{n}}, \qquad (1)$$

$$RMSPE = \sqrt{\frac{\sum_{m=1}^{n}(\frac{A-P}{A})^2}{n}} \; X\ 100\% , \qquad (2)$$

Where, n: the number of scores, A: the actual score and P: the predicted score.

D. Local Sensitivity Analysis (LSA)

In LSA, only one factor parameter can be varied at a time for all students sample (50 students), and then we get the predictive model scores for the 50 students. The RMSPE criterion was used to evaluate the significance of the changed factor parameter (we compared the new RMSPE with the old one).

IV. Result and Discussion

The main aim of this study was to predict well in advance the students' scores in Computer Skills-2 Course in order to reveal whether a student tends to have a wanted score based on (the student major, the gender, time spent studying for the computer skills-2 test and the branch of secondary education of student) so that extra efforts can be made to improve the student's academic performance and, in turn, improve his or her score. As mentioned before, data mining prediction tool was applied to build the predicative model in this study. One hidden-layered NN model with tangent sigmoid activation function was constructed in the NN prediction method. The input layer had four neurons, while the output layer had one neuron. The hidden-layer neuron number was determined according to the neuron number of the input and output layers. Five neurons were used in the hidden layers of the NN topology. For the output layer, the linear activation function was selected. Also used to show the best performance in the experiments was the scaled conjugate gradient algorithm.

The prediction results of the predictive model scores for 50 students are presented in Table 3.

As the results indicate in Table 3, the error range between the predicative score and the actual score for the 50 students it's between +6 and -6. While the root mean squared error (RMSE) = 2.424871 and the root mean squared percentage error (RMSPE) = 4%. To show qualitatively evaluated for the predicative model results, we plotted both the predicted and actual scores of students as shown in Figure 4.

Figure 4 shows the predictive model predictions for 50 students. Except the students 4, 12, 37, and 45, predictions scores were close to the actual scores.

Table 4 shows results revealed by LSA. Table 4 summarizes the most significant factor parameters for the predicative model depend on the difference between RMSPE. The time spent on study for computer skills-2 daily (F1) parameter has the most critical influence on the predicate score, followed by the time spent on study for computer skills-2 before exam (F2) the branch of secondary education (Al-Tawjihi) (F3) and the student major (F4).

TABLE 3. The actual score (A), the predicative score (P) and the error (E=A – P) for 50 students.

No	AS	PS	E	No	AS	PS	E
1	85	83	2	26	85	84	1
2	78	81	-3	27	52	51	1
3	66	66	0	28	63	64	-1
4	57	63	-6	29	70	70	0
5	50	48	2	30	75	75	0
6	67	67	0	31	75	73	2
7	84	82	2	32	51	52	-1
8	48	50	-2	33	39	38	1
9	90	93	-3	34	90	89	1
10	88	86	2	35	57	55	2
11	57	57	0	36	58	57	1
12	50	55	-5	37	71	66	5
13	67	63	4	38	49	52	-3
14	84	84	0	39	58	59	-1
15	48	49	-1	40	92	93	-1
16	80	77	3	41	94	97	-3
17	38	40	-2	42	88	84	4
18	92	91	1	43	81	80	1
19	71	72	-1	44	63	60	3
20	70	67	3	45	88	82	6
21	57	53	4	46	72	72	0
22	50	50	0	47	76	77	-1
23	67	65	2	48	77	78	-1
24	84	87	-3	49	59	59	0
25	45	45	0	50	72	73	-1

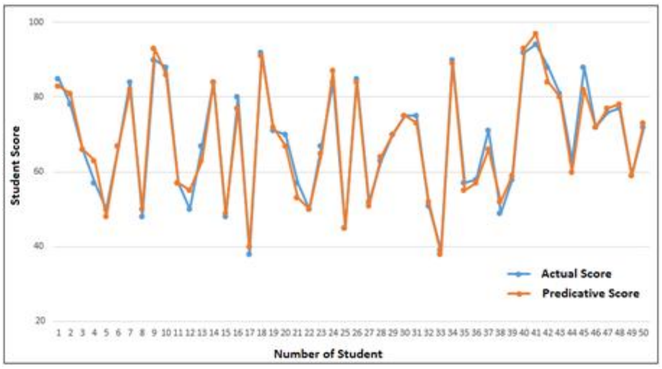

Fig. 4. Prediction results of the predictive model.

TABLE 4. The local sensitivity analysis results for the predicative model.

Factor No.	Old RMSPE	New RMSPE	The Difference
F1	4%	6.1%	2.1%
F2	4%	5.5%	1.5%
F3	4%	5.2%	1.2%
F4	4%	4.8%	0.8%

V. Conclusion

An interesting and challenging problem is to understand the factors that lead to students ' success (or failure) at university courses. Analysis of these factors may be used to improve the structure and content of the university courses and exams. Data mining can be used to effectively model and analyze the university courses and exams. In this research we used the regression analysis to build predicting model that use to predict the computer skills-2 course scores for students depending to four factors, and we tried to find the most important factor in the predicting model.

Results shows that the predicting model is useful to use to predicate students' scores. The error range between the predicative score and the actual score for the 50 students it's between +6 and -6. While the root mean squared error (RMSE) = 2.424871 and the root mean squared percentage error (RMSPE) = 4%.

The time spent on study for computer skills-2 daily has the most critical influence on the student score.

In the future, we will be adding more factors as predictive model inputs to get more accurate results.

REFERENCES

[1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[2] Sorkin, D. E. (2001). Technical and Legal Approaches to Unsolicited Electronic Mail, 35 USFL Rev. 325 (2001).

[3] Raschka, S. (2015). Python machine learning. Packt Publishing Ltd.

[4] Galton, F. (1886). Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[5] Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In International conference on knowledge-based and intelligent information and engineering systems (pp. 267-274). Springer, Berlin, Heidelberg.

[6] Salasznyk, R. M., Klees, R. F., Ward, D. F., Hughlock, M. K., Westcott, A. M., Xiang, Z., ... & Plopper, G. E. (2004). Using machine learning to build predictive models analyzing the osteogenic differentiation of human mesenchymal stem cells. In Molecular Biology of the Cell (Vol. 15, pp. 470A-470A).

[7] Agakov, F., Bonilla, E., Cavazos, J., Franke, B., Fursin, G., O'Boyle, M. F., ... & Williams, C. K. (2006). Using machine learning to focus iterative optimization. In Proceedings of the international symposium on code generation and optimization (pp. 295-305). IEEE Computer Society.

[8] Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care, S106-S113.

[9] Maneerat, N., & Muenchaisri, P. (2011). Bad-smell prediction from software design model using machine learning techniques. In 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 331-336). IEEE.

[10] Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. In AMIA annual symposium proceedings (Vol. 2012, p. 606). American Medical Informatics Association.

[11] Kanewala, U., & Bieman, J. M. (2013). Using machine learning techniques to detect metamorphic relations for programs without test oracles. In 2013 IEEE 24th International Symposium on Software Reliability Engineering (ISSRE) (pp. 1-10). IEEE.

[12] Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems, 60, 126-140.

[13] Karstoft, K. I., Galatzer-Levy, I. R., Statnikov, A., Li, Z., & Shalev, A. Y. (2015). Bridging a translational gap: using machine learning to improve the prediction of PTSD. BMC psychiatry, 15(1), 30.

[14] Chong, S. L., Liu, N., Barbier, S., & Ong, M. E. H. (2015). Predictive modeling in pediatric traumatic brain injury using machine learning. BMC medical research methodology, 15(1), 22.

[15] Parish, E. J., & Duraisamy, K. (2016). A paradigm for data-driven predictive modeling using field inversion and machine learning. Journal of Computational Physics, 305, 758-774.

[16] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[17] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D., & Doyle, A. G. (2018). Predicting reaction performance in C–N cross-coupling using machine learning. Science, 360(6385), 186-190.

[18] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 44(3), 278-297.

[19] Khazaaleh, M., Al-Omari, H., & Haziemeh, F. (2011). New e-Learning quality matrix to ELQ assessment at AL-Balqa applied university. J. Theor. Appl. Inform. Tech, 32, 169-78.

[20] Baker, R. S., Gowda, S., & Corbett, A. (2010). Automatically detecting a student's preparation for future learning: Help use is key. In Educational Data Mining 2011.

[21] Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all?. Journal of personality and social psychology, 86(1), 148.

[22] Wehman, P., Sima, A. P., Ketchum, J., West, M. D., Chan, F., & Luecking, R. (2015). Predictors of successful transition from school to employment for youth with disabilities. Journal of occupational rehabilitation, 25(2), 323-334.

[23] Verma, C., Illés, Z., & Stoffová, V. (2019). Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-7). IEEE.