Predicting Students' Performance Using Machine Learning Techniques

Hussein Altabrawee [a] Osama Abdul Jaleel Ali [a] , Samir Qaisar Ajmi [a]

[a]*Al Muthanna University, Almuthanna, Alsamawa, Iraq*
hussein.a.hassan@mu.edu.iq, osama.abdul.jaleel@mu.edu.iq , samir@mu.edu.iq

ARTICLE INFO
Submission date: 23/11/2018
Acceptance date: 3 /1/2018
Publication date: 10/3/2019

Abstract

The ultimate goal of any educational institution is offering the best educational experience and knowledge to the students. Identifying the students who need extra support and taking the appropriate actions to enhance their performance plays an important role in achieving that goal. In this research, four machine learning techniques have been used to build a classifier that can predict the performance of the students in a computer science subject that is offered by Al-Muthanna University (MU), College Of Humanities. The machine learning techniques include Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression. This research pays extra attention to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students' performance. These effects introduced by using features that measure whether the student uses the internet for learning and the time spent on the social networks by the students. The models have been compared using the ROC index performance measure and the classification accuracy. In addition, different measures have been computed such as the classification error, precision, recall, and the F measure. The dataset used to build the models is collected based on a survey given to the students and the students' grade book. The ANN (fully connected feed forward multilayer ANN) model achieved the best performance that is equal to 0.807 and achieved the best classification accuracy that is equal to 77.04%. In addition, the decision tree model identified five factors as important factors which influence the performance of the students.

Keywords: Decision Tree; Naïve Bayes; ANN;Logistic Regression, Students' Performance Prediction

1. Introduction

The economic success of any country highly depends on making higher education more affordable and that considers one of the main concerns for any government. One of the factors that contributes to the educational expenses is the studying time spent by students in order to graduate. For example, the loan debt of the American students has been increased due to the failure of many students in getting graduated on time [1]. Higher education is provided for free to the students in Iraq by the government. Yet, failing of graduating on time costs the government extra expenses. To avoid these expenses, the government has to ensure that the student graduate on time. Machine learning techniques can be used to forecast the performance of the students and identifying the at risk students as early as possible so appropriate actions can be taken to enhance their performance. One of the most important steps when using these techniques is choosing the attributes or the descriptive features which used as input to the machine learning algorithm. The attributes can be

categorized into GPA and grades, demographics, psychological profile, cultural, academic progress, and educational background [2]. This research introduces two new attributes that focus on to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students' performance. Four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used to build the machine learning model. ROC index has been used to compare the accuracy of the four models. The dataset used to build the models is collected from the students at the College Of Humanities during 2015 and 2016 academic years using a survey and the student's grade book. The dataset has the information of 161 students. The activities of this research include feature engineering to create the students dataset, data collecting, data preprocessing, creating and evaluating four machine learning models, and finding the best model and analyzing the results.

2. Literature Review

Much research has been done in the area of educational data mining where a predictive model is built to forecast the performance of students to identify the at risk students. This problem can be considered a hard problem because the performance depends on many characteristics related to the students. These characteristics can be categorized into student's GPA and grades, demographics, psychological profile, culture, academic progress, and educational background [2]. The student's GPA is the most important attribute used to predict the performance. The GPA can represent the real value for the future educational and career possibilities and progression. In addition, the academic potentials can be evaluated by the student GPA. The demographics information that consists of the family background, the gender, disability, and age is also considered an important attribute [3]. This research introduces two new attributes that focus on using descriptive features related to the internet and social network usage and their effect on the performance. On the other hand, many machine learning and data mining techniques have been used to predict the students' performance such as: Artificial Neural Network (ANN); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Linear Regression; Logistic Regression; Decision Tree (DT); Random Forest (RF); Principal Component Analysis (PCA); Naïve Bayes (NB); Neuro-Fuzzy classification (NF); Decision List (DL); Bayesian Network (BN); and Discriminant Analysis (DA). Table 1 shows a summary of the research papers that relate to this study.

Paper	Features	Dataset Size	Machine Learning Algorithms	Best Algorithm
Meier et al, 2015 [4]	Grades	700	New algorithm proposed, KNN Linear regression, logistic regression, SVM	New algorithm proposed
Guleria et al, 2014 [5]	Class Performance, Attendance, Assignment, Lab Work, Sessional Performance	120	DT	DT
Xu et al, 2017 [1]	Grades, Backgrounds	1169	Linear Regression, Logistic Regression, RF, kNN, Proposed Progressive Prediction algorithm	Proposed progressive prediction algorithm

Arsad et al , 2013 [6]	Grades	896	ANN	ANN
Li et al, 2013 [7]	Grades	72	PCA	PCA
Gray et al, 2014 [8]	Aptitude, Personality, Motivation Learning strategies	914	NB, DT, Logistic Regression, SVM, ANN, KNN	SVM, KNN, NB
Buniyamin et al, 2016 [9]	Grades	391	Neuro-Fuzzy classification	Neuro-Fuzzy classification
Alharbi et al, 2016 [10]	student demographics, general performance, students' modules	1789 Testing 898 Training	Logistic regression, ANN, DL, BN, DA, DT , And Ensemble approach	No overall winners
Livieris et al, 2012 [11]	Grades	279	ANN, DT, NB, Rule-Learning,SVM	ANN SVM
Hamsa et al, 2016 [12]	Internal grades, sessional grades and admission score	168	Fuzzy Genetic Algorithm and DT	FGA model is less strict than DT
Arsad et al , 2014 [13]	Grades	896	ANN, Linear Regression	ANN, Linear Regression
Sarker et al, 2014 [14]	Personal and demographics information, student satisfaction and integration	149	ANN, Logistic Regression	logistic regression
Huang et al, 2011 [15]	GPA and Grades	239	Linear Regression, ANN, Radial Basis Function NN, SVM.	SVM

Table 1- Summary Of The Related Research Papers

3. The Proposed System

3.1 The System Components

The following diagram, figure 1, shows the main steps and components of the proposed machine learning system.

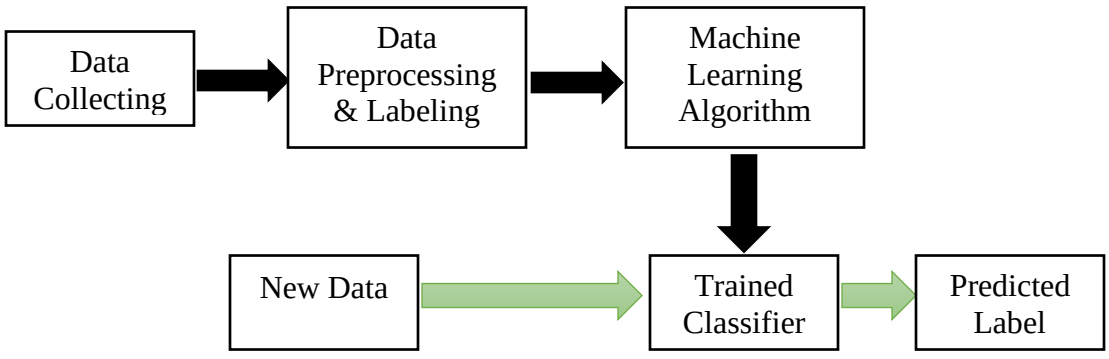

Figure 1-The Main Steps and Components of the Proposed System

The first step is collecting the data from the data sources. In our case, the data has been collected using a survey given to the students and the students' grade book. The second step is preprocessing the data in order to get a normalized dataset and then labeling the data rows. In the third step, the result of the second step, the training and testing dataset, is fed to the Machine Learning algorithm. The Machine Learning Algorithm builds a model using the training data and tests the model using the test data. Finally, the Machine Learning Algorithm produces a trained model or a trained classifier that can take as an input a new data row and predicts its label.

3.2 The Methodology

In this section, a brief review of the machine learning techniques that is used in this research is introduced.

3.2.1 Artificial Neural Networks

Artificial Neural Network represents a set of input unites and output unites that are connected to each other by weighted connections. The ANN learns by changing the weights of the connections in a way so it is able to predict the right target label for some input data instances. One of the famous learning algorithms used to train the ANN is Backpropagation Algorithm. ANN has many advantages such as its high resistance to noisy datasets and its well performance on classifying patterns that has not been trained on so it's used in situations when there is a little knowledge of the relation between the class label and the features in the dataset.

There are many real world applications of the ANNs such as image and handwritten recognition, speech recognition, laboratory medicine and pathology. There are many types of the ANNs which can be classified based on their architecture and design. One type is a fully connected multilayer feed forward ANN in which the network has an input layer, one or more hidden layers, and the output layer. In addition, its connections never cycle back to an input unit or to an output unite located in the previous layer. Also, each unit in a layer L provides input to each unit in the layer L+1.

A three layer fully connected feed forward ANN has been used in this research. The network consists of an input layer, two hidden layers, and the output layer. The input layer has twenty input unites, neurons, while the first hidden layer has six hidden unites. The second hidden layer has three hidden unites. The fourth layer is the output layer which has only one output unite. The Rectifier Linear Unit has been used as the hidden unites' activation function [16].

3.2.2 Logistic Regression

Logistic Regression represents a mathematical modeling technique which describes the relationship between several independent variables, $X_1 ... X_K$, and a dependent variable, D. The logistic model uses the logistic function as a mathematical form which has the range between 0 and 1 for any given input. The logistic model can describe a probability of an event which is always a value between 0 and 1. The following formula represents the logistic model.

$$P(D = 1|X_1, X_2, \dots, X_k) = 1/1 + e^{-(\alpha + \sum_1^k \beta_i X_i)} \quad (1)$$

Where α and β are the model's parameters that can be learned from a set of labeled instances in the training dataset. Gradient Descent Algorithm can be used to find the best values of the model's parameters during the training phase [17].

3.2.3 Naïve Bayes

Naïve Bayes classification model is considered as the simplest variation of the Bayesian network. This model assumes that every feature attribute is independent from the other attributes given the target attribute state. Each instance x in the dataset contains attribute values $a_1, a_2, \dots, a_i$. The target function f(x) equals any value from predefined finite set V=($v_1, v_2, \dots, v_j$). Naïve Bayes model uses the following equation.

$$V_{max} = \underset{v_j \in V}{Max} \; P(v_j) \prod_i P(a_i|v_j) \qquad (2)$$

Where v represents the target of the model, $P(a_i|v_j)$ and $P(v_j)$ could be found by calculating their frequencies in the training dataset[18].

3.2.4 Decision Tree

A decision tree model represents a tree structure that is similar to a flowchart. In this structure, each internal node represents a test on a dataset attribute while each tree branch represents the test outcome. In addition, each leaf node represents a target feature label and the upper first node in the tree represents the root node. Decision trees can be a binary or a non-binary trees. Decision trees are popular classification techniques because using them does not need prior knowledge of the problem domain or a complicated setting of the classification parameters. In addition, they can be converted to classification rules easily and they can be understood easily. Decision tree classification technique has been used in many real word applications such as financial analysis, medicine, molecular biology, manufacturing production, and astronomy. During building the decision tree, the algorithm uses an attribute or feature selection measure which is used in selecting the attribute or the feature that best divides the dataset instances into distinct target classes. Such measures include the Information Gain, Gain Ratio, and Gini Index. Popular decision trees algorithms include ID3, CART, and C4.5[16].

4. The Experiment

4.1 Dataset and Data Sources

The dataset used in this research is collected from the Archeology department and the Sociology department of the college of Humanities at Al-Muthanna University during the 2015 and 2016 academic years. Two data sources have been used, survey collected from the students and the students' grades data records. The dataset contains 161 student records, 76 male and 85 female. The dataset contains twenty attributes. The attributes can be divided into five categories which are personal and life style, studying style, family related, educational environment satisfaction, and student's grades. Table2 shows the attributes used in order to construct the dataset. Each student has been labeled as Weak or Good based on his/her final grade in the computer science subject. The weak student is the student who has a final grade less than sixty out of 100. On the other hand, the Good student is the student who has a final grade equal or greater than sixty.

There are 75 students with Good status and 86 students with Weak status. Identifying the weak status students is more important than identifying the good status students, therefore the weak status is considered a positive value of the target attribute. Computer Grade-Course1 attribute represents the average of the first two monthly exams in the computer science subject during the

first course. The academic year contains two semesters or courses, midterm exam, and final exam. Each of the semesters has two monthly exams. To predict the students who need support as early as possible the grade of the first semester, the average of the first two exams, has been chosen as an attribute because it could be an indicator of the final student performance. By doing that, the weak students will have an early opportunity to enhance their academic performance. In addition, the faculty members can provide the appropriate support to the students as early as possible. Similarly, English Grade-Course1 attribute represents the average of two exams in the English subject during the first semester. The English subject has been chosen as an attribute because of its relation with the computer science subject as most of the computer educational materials have been taught and presented in English.

Many factors could affect the performance such as having a job and studying and that could be very challenging. Based on the dataset, there are 65 students, out of 161, who are working and studying therefore 40.3% of the students have some kind of a job and 46.1% of them are weak students. Another factor is marriage, with marriage come more responsibilities, 21.7% of the students are married students.

Attribute	Attribute Definition
Department	Archeology=2 , Sociology=1
Gender	Male =1, Female = 2
Studying Style	Group=0,Single=1,Both=2
Using Internet For Study	Never=0,Always=1, Sometime=2
Using Extra Learning Resources	Never=0,Always=1, Sometime=2
Interest in studying computer	Ordinal from 1 to 10
Has Computer Experience	Yes=1, No=0
Studying Hours	Numeric
Family Members Education	Higher Education=1, Others=0
Family Help In Studying	Yes=1, No=0
Educational Environment Satisfaction`	Ordinal from 0 to 10
Has A Job	Yes=1, No=0
Accommodation	Dorms=1, Other=0
Residence	City Center =1, Other=0
Married	Yes=1, No=0
Sport Participation	Yes=1, No=0
Time Spent On Social Media (Hours)	Numeric
Computer Grade-Course1	Numeric
English Grade-Course1	Numeric
Final Computer Outcome	Good , Weak

Table 2- The Dataset Attributes

Although, both of the departments belong to the college of Humanities, a difference in the students' performance has been found. There are 82 students in the Sociology department. Thirty two of them are weak students therefore the percentage of the weak students in the Sociology department equals 39%. On the other hand, there are 79 students in the Archeology department.

Thirty nine of them are weak students therefore the percentage of the weak students in the Archeology department equals 49%. The college policy of placing the students in the departments could be the cause of the difference in performance. The college places the student in a department based on his/her high school GPA therefore all the student who are in the same department have a slightly similar performance. Based on that, the department attribute could be an indicator of the student performance.

The student's life style could be another factor that contributes to the performance and one of the life style activities is participating in sport. A study done by Fernando et al. [19] found that there is a positive correlation between formal sport activities and high academic performance. Another study that is done by the Centers for Disease Control and Prevention-USA [20] found that there is either positive relationship, 50.5% of the studies summarized, or not a demonstrated relationship, 48% of the studies summarized, between the physical activities and the academic performance. Only 4 associations of 251 examined showed negative relationship.

4.2. Data Preprocessing and Machine Learning Software

Each attribute value in the dataset has been normalized by subtracting the attribute mean from it and dividing the result by (the attribute maximum value – the attribute minimum value). RapidMiner Studio machine learning software has been used in order to train and test the models.

4.3 Validation Method and Accuracy and Performance Measures

In this research, three folds cross validation method has been used. In this method, the dataset is divided into three equal size sets. The learning and testing are executed three times. At each fold or execution, the machine learning algorithm selects one set to be the test set and the remaining two sets as the training sets. The accuracy and the performance measures is aggregated over all the folds in order to calculate the final performance and the final accuracy of the model. The ROC index, the area under the curve, performance measure has been used to evaluate the performance of the classification models. This measure is a well-known measure that is relying on the ROC curve and it is calculated by using the prediction scores. Equation 3 is used to calculate the ROC index [21]. In addition to the ROC index, many important measures have been used such as the accuracy, the classification error, and the F Measure. Equation 4 is used to calculate the F Measure. The F Measure is a useful alternative to the misclassification rate measure. [21]

$$ROC\ index = \sum_{i=2}^{|T|} \big(FPR(T[i]) - FPR(T[i-1]) \big) \times \big(TPR(T[i]) + TPR(T[i-1]) \big)/2 \qquad (3)$$

Where $|T|$ represents the number of thresholds that are used, $FPR(T[i])$ represents the false positive rate at the threshold i, and $TPR(T[i])$ represents the true positive rate at the threshold i. A larger ROC index indicates a better classification model. A model with ROC index above 0.7 considered a strong model while a model with ROC index below 0.6 considered a weak model. [21].

$$F\ Measure = 2 * \frac{(\ Precision * Recall\ )}{(\ Precision + Recall\ )} \qquad (4)$$

$$Precision = \frac{TP}{(TP + FP)} \qquad (5)$$

$$Recall = \frac{TP}{(TP + FN)} \qquad (6)$$

TP, True Positives, is the number of data rows in the test set which had a positive target and that were predicted to have a positive target. TN, True Negatives, is the number of data rows in the test set that had a negative target and that were predicted to have a negative target. FP, False Positives, is the number of data rows in the test set which had a negative target but that were predicted to have a positive target. FN, False Negative, is the number of data rows in the test set that had a positive target but that were predicted to have a negative target [21].

4.4 Models Implementation

All the models have been implemented by the RapidMiner Studio software. A Cross Validation operator has been used in order to execute the three folds validation operations during the training and the testing phases. The operator sampling property set to linear sampling. In order to find the best set of the models' parameters, Optimize Parameters (Grid) operator has been used.

The ANN operator has been configured to use the Rectifier activation function and the number of hidden layers sizes set to be 6 and 3 consecutively. The ANN model used 100 epochs in the training phase. All the other parameters has been set to the default values. The Optimize Parameters operator has been set to find the best value of the learning rate and the L2 regularization. For the learning rate and the L2 regularization, the configuration set to use 100 steps on a linear scale from 0 to 1.

For building the DT model, the Optimize Parameters operator has been set to find the best value of the splitting criterion, and the minimal size for split properties. Also, apply pruning property has be set by the optimization operator. All the other parameters has been set to the default values.

The Logistic Regression operator has been set to use regularization and the optimization operator set to find the best value for the solver method and the lambda. The lambda search property set to use 60 steps on a linear scale starts from 0 to 1.797. All the other parameters has been set to the default values.

For building the Naïve Bayes model, the optimization operator has been set to find the best values for the Laplace correction, the estimation mode, using the application grid, the bandwidth selection, the number of kernels, and the size of application grid. The number of kernels search property set to use 10 steps on a linear scale starts from 1 to 20. The application grid size search property set to use 10 steps on a linear scale starts from 1 to 40.

4.5 The Results

Four classification models have been created and tested using four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Logistic Regression, and Decision Tree. Table3 shows the accuracy and the performance measures for each model as well as the confusion matrices.

Model	TP	FP	TN	FN	Precision	Recall	F- Measure	Accuracy	Classification Error	ROC index
ANN	67	18	57	19	79.17	77.92	78.47	77.04	22.96	0.807
DT	67	19	56	19	77.96	77.83	77.88	76.93	23.61	0.762
Logistic Regression	62	17	58	24	79.23	71.91	74.87	74.53	25.47	0.767
Naïve Bayes	55	23	52	31	70.51	64.27	67.21	66.52	33.48	0.697

Table 3- The accuracy and the performance measures for the Models

As shown in Table3, Naïve Bayes model has the lowest ROC index that is equal to 0.697 and the lowest accuracy that is equal to 66.52 with the highest error of 33.48. The most accurate model is the model built using Artificial Neural Network classification technique which has an accuracy of 77.04. In addition, it has the best performance based on the ROC index which equals to 0.807 and the lowest classification error that is equal to 22.96. The following figures show the ROC index of each model.

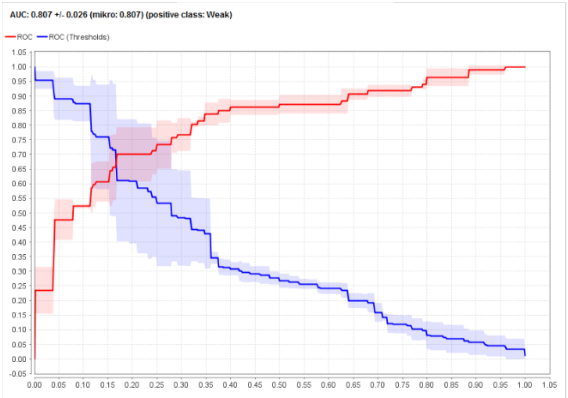

Figure 2- ANN ROC Index

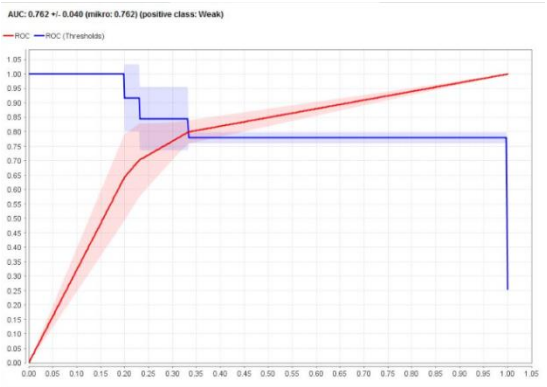

Figure 3- DT ROC Index

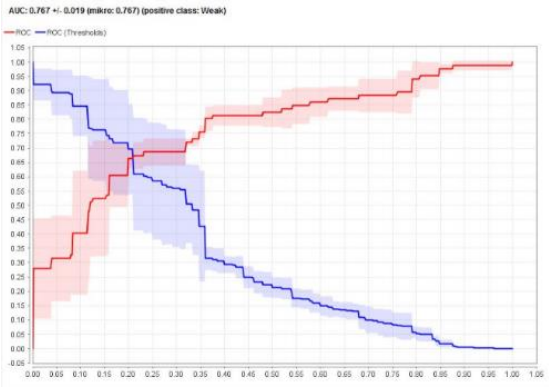

Figure 4- LR ROC Index

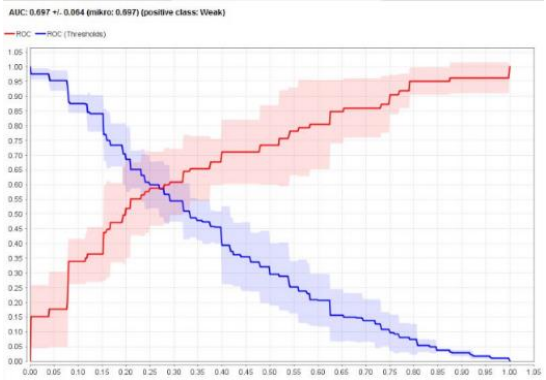

Figure 5- NB ROC Index

Decision tree model showed that not all the attributes have an impact on classifying the status of students into Good or Weak. There are five main attributes that influence the classification decision. They are Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency. Therefore, in order to enhance the weak students' performance, the faculty and the administrative members should pay an extra attention to those attributes and make decisions based on them. Many actions could be taken as the following: Providing extra sessions and lab work could enhance the student's grades. In addition, making the subject's topics much more interesting to the student or providing the students with a better educational environment also could help to enhance their performance.

5. The Conclusion

To solve the problem of identifying the students who have a poor academic performance in the computer science subject offered by Al-Muthanna University, College Of Humanities, four classification models have been built to predict the performance of the students. Four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used. The models have been compared to one another using the ROC index performance measure and the classification accuracy. ANN model has the highest ROC index that equals to 0.807 and accuracy of 77.04. In addition, the decision tree model showed that not all the attributes involve in the classification process. Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency are the attribute used by the decision tree model.

CONFLICT OF INTERESTS.

There are non-conflicts of interest.

References

[1] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 5, pp. 742–753, 2017.

[2] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in *ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology*, 2015, no. March, pp. 0–2.

[3] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in *Procedia Computer Science*, 2015.

[4] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Predicting grades," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 959–972, 2016.

[5] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," *Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014*, pp. 126–129, 2015.

[6] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "A neural network students' performance prediction model (NNSPPM)," *2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013*, no. July 2006, pp. 26–27, 2013.

[7] K. F. Li, D. Rusk, and F. Song, "Predicting student academic performance," *Proc. - 2013 7th Int. Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2013*, pp. 27–33, 2013.

[8] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in *Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014*, 2014.

[9] N. Buniyamin, U. Bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," *2015 IEEE 7th Int. Conf. Eng. Educ. ICEED 2015*, pp. 49–53, 2016.

[10] Z. . Alharbi, J. . Cornford, L. . Dolder, and B. . De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," *Proc. 2016 SAI Comput. Conf. SAI 2016*, pp. 523–531, 2016.

[11] I. E. Livieris, K. Drakopoulou, and P. Pintelas, "Predicting students' performance using artificial neural networks," *Proc. 8th Pan-Hellenic Conf. "Information Commun. Technol. Educ.*, pp. 28–30, 2012.

[12] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technol.*, 2016.

[13] P. Mohd Arsad, N. Buniyamin, and J. L. Ab Manan, "Neural Network and Linear Regression methods for prediction of students' academic achievement," *IEEE Glob. Eng. Educ. Conf. EDUCON*, no. April, pp. 916–921, 2014.

[14] F. Sarker, T. Tiropanis, and H. C. Davis, "Linked data, data mining and external open data for better prediction of at-risk students," in *Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014*, 2014.

[15] S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical

modeling techniques," *Proc. - Front. Educ. Conf. FIE*, vol. 1, pp. 3–4, 2012.
[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann publications, 2012.
[17] D. G. Kleinbaum and M. Klein, *Logistic Regression A Self-Learning Text*, 3rd ed. New York: Springer-Verlag New York, 2010.
[18] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Students ' Performance in Distance Learning Using Machine," no. Ml, 2004.
[19] F. Muñoz-Bullón, M. J. Sanchez-Bueno, and A. Vos-Saz, "The influence of sports participation on academic performance among students in higher education," *Sport Manag. Rev.*, 2017.
[20] CDC, "The Association Between School-Based Physical Activity, Including Physical Education, and Academic Performance," Atlanta, GA, 2010.
[21] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies*. 2015.

الخلاصة

الهدف الرئيسي لاي مؤسسة تعليمية هو تزويد الطلبة بافضل معرفة وتجربة تعليمية. تحديد الطلبة الذين يحتاجون الى دعم و اهتمام إضافي و اتخاذ التدابير المناسبة لتحسين مستواهم العلمي يلعب دورا اساسيا لتحقيق هدف المؤسسة التعليمية. في هذا البحث, اربع تقنيات أو طرق خاصة ب Machine Learning تم استخدامها لبناء Classifiers تقوم بالتنبأ بمستوى أو أداء الطلبة العلمي في أحد دروس الحاسوب المقام في جامعة المثنى– كلية الاداب. تتضمن التقنيات المستخدمة كل من التقنيات الاتية: Artificial Neural Network, Naïve Bayes, Decision Tree, Logistic Regression. هذا البحث يهتم بتاثير استخدام الانترنت كمصدر للتعلم و كذلك تاثير استخدام الطالب لمواقع التواصل الاجتماعي على مستوى الطالب الدراسي. هذه التاثيرات تم استخدامها ك Features لقياس فيما اذا كان الطالب يستخدم الانترنت للدراسة ام لا و كذلك لقياس الوقت الذي يقضية الطالب بتصفح مواقع التواصل الاجتماعي. تم بناء اكثر من نموذج و تمت المقارنة بينهم باستخدام مقياس الاداء ROC index و كذلك تم استخدام دقة التصنيف للمقارنة بين النماذج. تم جمع المعلومات المستخدمة في بناء النماذج من خلال استمارة استبيان تم ملئها من قبل الطلبة و كذلك من سجل درجات الطلبة. حقق نموذج ANN أعلى نسبة أداء و التي تساوي 0.807 و حقق نسبة دقة تساوي 77.04%. و بالاضافة الى ذلك و باستخدام نموذج Decision Tree , تم التعرف على اربعه عوامل مهمة تقوم بالتاثير على مستوى الطالب بصورة كبيرة.