Prediction of the Academic Achievement of Pupils Using Data Mining Techniques

Mokhalad Eesee Khudhur
Salah Al-Din Education Directorate.
E-mail: mokhalad2018@gmail.com

Mohammed Shihab Ahmed
College of Arts, Tikrit University, Iraq.

Saif Muhannad Maher
Cisco Networking Academy, Tikrit University, Iraq.

Received August 02, 2021; Accepted November 20, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19014

Abstract

Introduction: During this epidemic, a problem in fundamental education affecting all globe is occurring, and we note that education and learning were online and conducted in students. Academic performance of students must be forecast, so that the instructor may better identify the missing pupils and offer teachers a proactive opportunity to develop additional resources for the student to maximize their chances of graduation. Students' academic achievement in higher learning (EH) has been extensively studied in addressing academic inadequacies, rising drop-out rates, graduation delays, and other difficult questions. Simply said, the performance of students refers to the amount to which short and long-term educational objectives are met. Academics nonetheless judge student achievement from different viewpoints, from grades, average grade points (GPAs) to prospective jobs. The literature encompasses numerous computing attempts to improve student performance in schools and colleges, primarily through data mining and analysis learning. However, the efficiency of current smart techniques and models is still unanimous.

Method: This study employs multiple methods for machine learning to forecast student progress. With its accurate data sample prediction, five integrated classification algorithms have been created to forecast students' academic success (support vectors, decision-making trees algorithm and perceptron algorithm, logistic regression algorithm and a random forest algorithm).

Results: Students' academic achievement has been reviewed and assessed. The performance of five learning machines mentioned in Section 4 is discussed here. First, we displayed the data after pre-processing by simply displaying distributions to form the data packet and then

evaluated 5 important learning methods and described the variables in the data set. The entire series of 480 characteristics were examined.

Keywords

Prediction for Students, Academic Performance, Categorization, Data Mining Approaches.

Introduction

Academic performance of higher education students has been researched carefully for academic deficiencies, increasing dropout rates, graduate delays and other chronic difficulties. In other words, student performance relates to the extent to which education objectives are reached in the short and long term. But academics judge student performance from a variety of angles, from the end-level, average grade point (GPA). The literature includes a variety of computer initiatives to enhance student performance in schools and universities, especially through information mining and analysis. The effectiveness of contemporary intelligent approaches and models remains under discussion, though (Ajoodha, Klein, & Rosman, 2015; Dixson, 2015).

The early prediction of student performance allows students to recognize low performance and allows educators to take the required steps early on in the learning process. Productive treatments include student counseling, performance monitoring and the creation of intelligent tutoring technologies, but not restricted to these. Computer advancements in data mining and learning analysis substantially assist this effort (Ajoodha, Jadhav, & Dukhan, 2020; Ajoodha & Rosman, 2018).

A recent thorough evaluation has found that around 70% of the work has examined student success through graduation and GPA, while just 10% have assessed student performance using learning results. This disparity has forced us to carefully evaluate if the findings are utilized as proxy for students' academic progress (Asif, Merceron, Ali, & Haider, 2017).

Education based on results is an educational paradigm focused on adopting and achieving so-called learning outcomes. Student learning results are goals that measure how far students at the conclusion of a certain study process acquire the given talents, information, capabilities and values. We feel that the outcomes of pupils are more comprehensive than basic assessment standards for the assessment of academic achievement. This idea is congruent with the premise that student academic performance is crucial to learning outcomes. Furthermore, the results from studies of famous HE accrediting organizations

like as ABET and ACBSP form the foundations for the quality assessment of education courses (Burman & Som, 2019; Gerhana, Fallah, Zulfikar, Maylawati, & Ramdhani, 2019; Salal, Abdullaev, & Kumar, 2019).

This relevance requires more research on both class and program levels in order to predict the achievement of learning outcomes. The lack of broad surveys to anticipate student achievement using student results has led us to examine these objectives (Deepika & Sathyanarayana, 2019; Domínguez et al., 2013).

Many approaches are now being given for evaluating student performance. Data mining is one of the most important techniques for student performance assessment. Data mining has recently becoming increasingly used in education. It is termed data mining education. Data mining is the technique by which essential information and trends from a large educational database may be retrieved. Useful facts and trends to forecast the achievement of youngsters (Huang & Fang, 2013).

This would help teachers design an efficient teaching method. In addition, instructors may track the accomplishments of their children. Students can extend their learning activities and enhance the efficiency of the management system. The usage of data mining tools may therefore be targeted at unique demands of various entities. A systematic review of the problems is presented. The systematic review proposed supports the objectives of this study:

1. To assess and identify weaknesses in current prediction methods.
2. Study and identify variables used in the study of student performance.
3. To study the available methodologies for predicting student performance (Hodges, Moore, Lockee, Trust, & Bond, 2020; Hussain, Zhu, Zhang, Abidi, & Ali, 2019; Niemi & Kousa, 2020).

Methodology

This section discusses the strategies provided in this study for predicting student success through analysis of demographic and internet student data.

Data Description

There are 480 records and 16 functions to collect data. Three essential factors were identified: first, the demographics of students, such as nationality and sex. The second is academic information for graduates, students and students.

Thirdly, information on pupils' behaviour, including number of access resources, number of classroom hands and school satisfaction. Over 305 men and 175 women, 179 men from Kuwait, 29 from Palestine, 22 from Jordan, 22 from Iraq, 17 from Libya, 12 from the United States, 9 from Tunisia, 9 from the United States, 6 from Syria, 6 from Iran, 6 from Libya, 4 from Morocco, and 1 from Venezuela. Two academic semesters are followed by 245 records for the first half and 235 records for the second half. The gathering of data also covers features of school days. The number of days of absence varies by category. We noted that 289 children were away for fewer than 7 days, while 191 pupils were absent for more than 7 days.

Finally, the data set also contains parents' engagement in their children's academic journey. There are two categories: first, parent survey and second, parent school satisfaction. We found that the survey was carried out by 270 parents, whereas 210 parents did not reply. We discovered 292 school-friendly parents and 188 parents were unhappy.

Classification Field

There are three numerical intervals for pupils. The initial interval for children with a defective percentage is 0 to 69 percent (L). The second interval is between 70 and 89 percent for children with a low percentage (M). Finally, the interval for high percentage pupils is between 90% and 100% (H).

Methods

Predicted Models: The utilized data sets are constructed and assessed with five key classification algorithms (decision trees and perceptron and vector supporting machines, logistic regressions and random forests). Short description of the project's prediction models.

Support for Vector Machines: SVM helps to detect and classify outliers in the data set. SVMs are various supervised learning techniques. SVMs used kernel techniques to alter data to distinguish between plausible end outcomes and manipulated data. With the Lagrange multiplier, SVM offers the optimal solution for each feature with partial differentiation. The model decreases training data convergence owing to the supported vectors.

N data point training data set, $\{x_k y_k\}_{k=1}^{N}$ Data set having n-dimensional (x_{keR}^{N}) input data and a single-dimensional output vector space (y_{ker}As demonstrated below, SVM builds the categorizer:

$$\boldsymbol{y(x) = sign\left[\sum_{K=1}^{N} a_k y_k\, \Psi(x, x_k) + b\right]}$$

Decision tree algorithm: This study uses decision trees to determine the predictor of the variables of the predicted variable and shows the discrete objective value. "The chosen trees employ variable values to build a node and a boundary structure." A DT has inner nodes and leaves, and rectangles are sheets. The internal node is a data collection with two or more children. The branches include the value of these characteristics. Each book has a label of categorization.

Decision trees from a training set are established. Hierarchy is referred to as a tree and a section as a node. The whole collection of data is presented in the node part of the tree. The branches contain the node, and the leaves contain the terminal nodes. Each leaf is picked and all observations in the leaf are made. The choice is the predictive value.

Perceptron algorithm: Perceptron classification is a supervised training procedure; a classification of the perceptron within a specimen's classification field can be predicted. The perceptron grade receives different inputs and does not return the result. If the input number is higher, the massage can be changed. The data set in the perceptron procedure is utilized to report the total event values by $x1, x2, x3, x4, \ldots, xn$ if x and n.

The needed characteristics are stored in the first layer as an input. Total weights, inputs and results were multiplied now. The values of the training models are $w1, w2, w3, .., c$. The output value is pushed, and the output value shown on receipt of the value.

Logistic regression algorithm: The logistic regression describes the relation between components and is used to avoid performance by students in predicting the probability of an occurrence. This model equation additionally gives the likelihood of corresponding explaining components and log values for two categories.

$$log\left(\frac{(P(Y=1|X))}{1-P(Y=1|X)}\right) = \beta_1 + \beta_1 X_1 +, \ldots\ldots, \beta_N X_N$$

When the $Y = (0,1)$ is the binary variable, 1, when it surpasses the baseline level, 1 is a projected regression coefficient, or 0 percent $(X1, \ldots, Xn)$ and β $(\beta 0, \ldots . \beta n)$.

Random forest algorithm: Random Forest employs a mendicant approach to create trees that are more accurate than any forecast of a tree. Random forests have also been utilized for avoiding excess fitness and reducing bias mistakes and for producing accurate and relevant findings. RF can handle and accurately process data outliers and noise. In the training phase, RF creates multiple decision trees and class labels.

RF is employed in this study since it is allowed to overfit less and good classification results have been shown. RF is a theoretical mixing framework for decision-making institutions; $\{T1X, \ldots, TBX\}$. The ensembles $\{Y\ to\ 1 = T1(X), \ldots, Y\ to\ B = TB(X), where\ Y\ to\ b, b = 1, \ldots B\}$ the bth tree is supplied. All trees give a summarized y di forecast, the predicted class of most trees.

Results

Analysis

The data collection includes a total of 480 records. The best technique to predict the performance of children is to utilize machine-learning algorithms and vector machinery technologies. We get a total accuracy of 70.8 per cent, which demonstrates that support technology is viable for vector machines and Random Forest. Accuracy 69.7%, logistics regression 67.7%, perception 64.5%, and ultimate decision-making 46.8%.

Pre-processing

In machine learning the preparation of information and the selection of data attributes specified before processing and testing are frequently crucial. Selection of attributes shows that all appropriate combinations of features may be used to predict academic data collecting performance. Our preprocessing objective was to convert our numeric fields, containing a value like a grade ID, into a numerical value to retain this gap. The findings for sustained distance are likewise provided by our three classes, assuming $L = -1, M = 0$, and $H = 1$. We have chosen to divide the category values and to make our number fields more comprehensible. The five learning models are used for the evaluation of academic achievement and the identification of the best student model.

Data Visualization

The data collection includes a total of 480 records. The aim of this study was to identify numerical data, information category properties and classification labels. Our aim is to analyze the structure of the data set and assess if the appropriate contours are easily identifiable.

Table 1 Predictability comparison of five models

Classifier Vector machine support	Accuracy
Tree of Decision	70.8 percent
Perceptron	46.8 percent
Regression of logistics	64.5 percent
Random woodland	67.7 percent
Classifier Vector machine support	69.7 percent

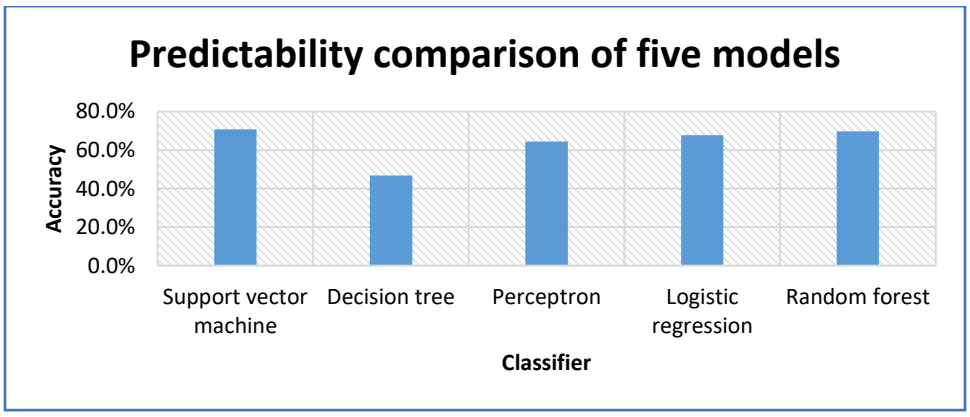

Figure 1 Predictability comparison of five models

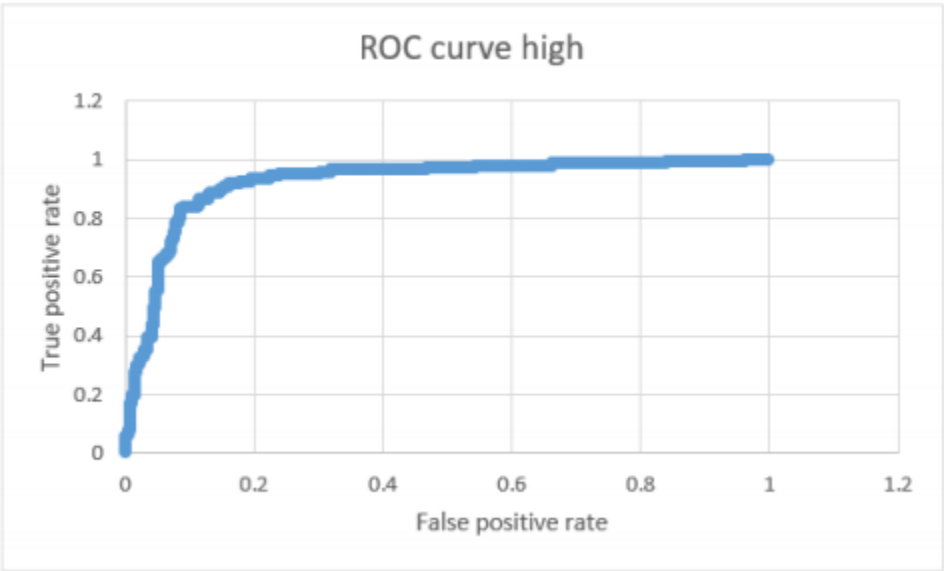

Figure 2 Receiver Characteristic curve (class high)

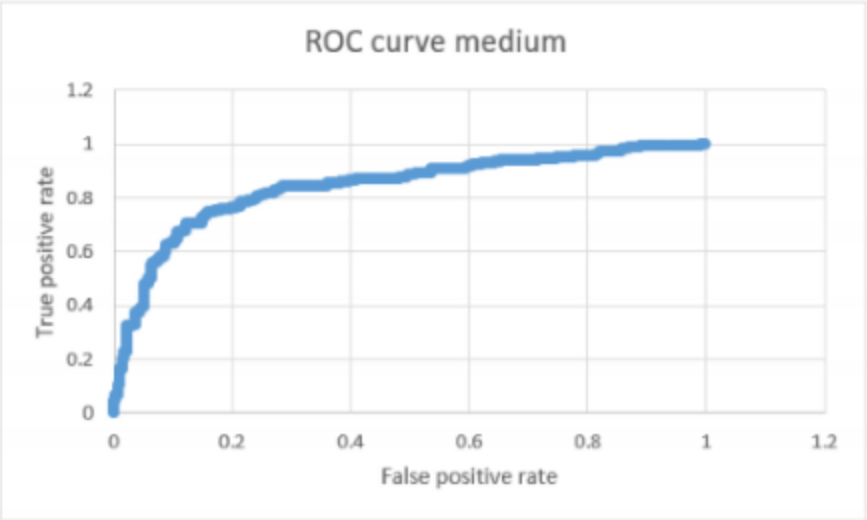

Figure 3 Receiver of operating characteristic curve (class medium)

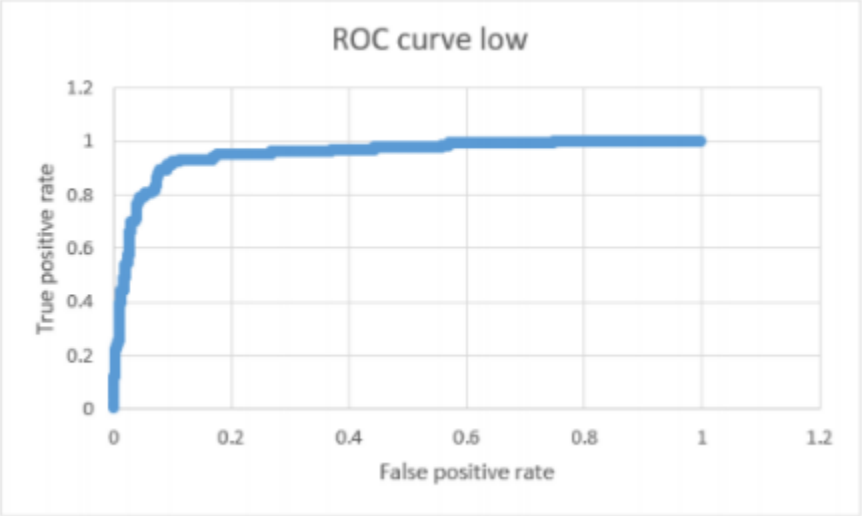

Figure 4 Receiver characteristic curve (class low)

Discussion

This section presents the findings of the latest study to forecast the performance of pupils.

The results suggest that the best approach to forecast student success is using vector machinery technology. SVM has a precision of 70.8%, 69.7% of Random Forest, 67.7% logistic regression, 64.5% accuracy and 46.7% decision timber. The entire class variable percentage was calculated. 26.46% of pupils have been decreased (less than 69%), 43.96% have smaller passes (between 70% and 89%), and 29.58% have an excellent education (90 percent to 100 percent).

Student absence days seem closely connected with the class variable. Relatively few pupils have lost more than seven days, while very few missed fewer than seven days. We've seen relatively tiny numbers from grades 5 to 9 and 10. No high school pupils and no high school pupils. The accuracy of 5 common models for student performance evaluation is shown in Figure 4. The text shows a blue vector, the reverse logistics is yellow, the tree is orange, and the woodland is blue.

The vector machine's assistance performed well in comparison with other machines. 78.75% of cases were categorized properly, whereas 21.25% of patients were misclassified. Another approach for model validity is confusion matrices in Figure 4.

The visual presentation function curve shows the performance of our best classification model at all levels-the supported vector machine. The ROC curve demonstrates our best-known vector classification performance in Figure 5, Figure 5 and Figure 7. There are three numerical grade intervals (low, medium and high). Figure 6 and Figure 7 are presented in Figure 5. The ROC curve is in the upper left corner, showing that the job categorization (SVM) is more efficient and an 86 per cent positive AUC value is predicted.

Conclusion

Companies and educational institutions utilize training management systems to design and manage courses, tests and other material.

Student success must be anticipated in order to determine the academic accomplishment of a professor and to enhance students' recognition and provide teachers an active chance to equip teachers with extra resources to boost their likelihood of graduation. It's hard for students to learn almost as in a class, and so the performance of the student is variable owing to diverse methods of teaching.

Various learning machines were used to predict LMS performance. When examined in relation to online learning systems with varying settings, each model displays distinct accuracy percentages.

The students' performance was assessed in five master training techniques, the Perceptron classification, vector support, decision making, logistical regression and random forests. With 70.8 percent accuracy, the support vector machine manages the best data. The findings show that days of absence affect student performance in academia. However, classes are not academic.

References

Ajoodha, R., Jadhav, A., & Dukhan, S. (2020). Forecasting learner attrition for student success at a south african university. *In Conference of the South African Institute of Computer Scientists and Information Technologists*, 19-28.

Ajoodha, R., Klein, R., & Rosman, B. (2015). Single-labelled music genre classification using content-based features. *In Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-Rob Mech)*, 66-71.

Ajoodha, R., & Rosman, B. (2018). Learning the influence structure between partially observed stochastic processes using iot sensor data. *In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Asif, R., Merceron, A., Ali, S.A., & Haider, N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194.

Burman, I., & Som, S. (2019). Predicting students academic performance using support vector machine. *In Amity International Conference on Artificial Intelligence (AICAI)*, 756-759.

Deepika, K., & Sathyanarayana, N. (2019). Relief-F and Budget Tree Random Forest Based Feature Selection for Student Academic Performance Prediction. *International Journal of Intelligent Engineering and Systems, 12*(1), 30-39.

Dixson, M.D. (2015). Measuring student engagement in the online course: The Online Student Engagement scale (OSE). *Online Learning*, *19*(4), n4.

Domínguez, A., Saenz-de-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J.J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*, 380-392.

Gerhana, Y.A., Fallah, I., Zulfikar, W.B., Maylawati, D.S., & Ramdhani, M.A. (2019). Comparison of naive Bayes classifier and C4. 5 algorithms in predicting student study period. In *Journal of Physics: Conference Series*, *1280*(2), 022022.

Hodges, C.B., Moore, S., Lockee, B.B., Trust, T., & Bond, M.A. (2020). The difference between emergency remote teaching and online learning. http://hdl.handle.net/10919/104648

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133-145.

Hussain, M., Zhu, W., Zhang, W., Abidi, S.M.R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review, 52*(1), 381-407.

Niemi, H.M., & Kousa, P. (2020). A case study of students' and teachers' perceptions in a Finnish high school during the COVID pandemic. *International journal of technology in education and science, 4*(4), 352-369.

Salal, Y.K., Abdullaev, S.M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *IJ of Engineering and Advanced Tech, 8*(4C), 54-59.