International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 5, Sept-Oct-2024, ISSN (Online): 2395-566X

The Impact of Behavioural Features on Predicting Academic Success: A Machine Learning Approach

[1]Nidhi Kataria Chawla, [2]Chietra Jalota
[1]B.S. Anangpuria Institute of Technology and Management (BSAITM), Faridabad

Abstract- To discover hidden patterns from educational data, researchers are developing methods by using educational data mining. Dataset and its features/attributes determine the eminence of data mining techniques. Student's academic performance model by using a new class of features i.e., behavioral features was built in this research paper. These are significant features as they are associated with the learner interactivity in e-learning system. Data was collected from an e-Learning system called Kalboard 360 using Experience API web service called (xAPI). After data preprocessing and feature selection, machine learning algorithms such as Decision Tree, Support Vector Machine and Artificial Neural Network were used to build the model. It is clearly visible from the results that there is a sturdy association between learner behaviors and its academic achievement. Results with above-mentioned classification methods using behavioral features attained up to 25% enhancement in the accuracy as compared to the results when same classification methods were applied on the data set without behavioral features.

Index Terms- Machine Learning, Educational Data mining, E-learning, Student Performance Prediction, Classification, Behavioral Features.

I. INTRODUCTION

To improve student learning experience in an academic environment, we use educational data mining (EDM). This EDM explores insight of student‟s data with the help of data mining, machine learning, and deep learning. [1, 2, 3]. A huge amount of data can be generated by alliance of learning platforms and materials with students [4, 5]. By analysing this data, the process of student learning and his success rate can be understood in a better way. By analysing data in deeper way, all significant factors like academic, demographic and communal/behavioural factors which affect student academic accomplishment could be recognized. To quantify the student academic achievement, performance of students must be assessed on the basis of academic subjects. Various methods like students‟ final grades, Grade Point Average (GPA), standardized tests are adopted by teachers to measure the student academic performance.

As per report of National Assessment of educational progress (NAEP) , various challenges are faced by the present education system like student‟s academic performance level, amplified dropout rates in university, graduation adjournment in graduation and insufficient enthusiasm by student workforce. With the passage of time, the rate of academic success of student‟s is declining day by day and is more dominant amongst the sectional students [6, 7,8]. All improved and advanced technologies like Artificial Intelligence (AI), Virtual Reality (VR), 3D printing, smart multimedia devices, Internet of Things (IoT), and Machine Learning are taken into consideration for the improvement of student‟s learning process [9]. Moreover, a study used classifiers such as J48, SVM, RF, NB, and MLP to predict students' performance levels using the data set of Kalboard 360 [26]. TheFor pattern recognition, data analysis is done by Machine Learning. After analysis, these patterns can be used to predict the student‟s outcome. To identify critical factors affecting student‟s success, Machine Learning is used by the educator in the class room. Moreover, ML also helps educators to recognize floundering students regarding their decision-making. In Machine Learning, Numerous tools such as R Software, Python Scikit-learn, TensorFlow are currently used for academic performance prediction. Several machine learning algorithms like Random Forest, Support Vector Machines (SVM), AdaBoost, Decision Tree, Naive Bayes, and K-nearest Neighbours are also used to predict academic performance of students.

In this manuscript, we built a student performance model with a new class of features, which we called behavioural features. The educational dataset is collected from Kalboard360 using learner activity tool (xAPI). Attributes/Features of the collected dataset are divided into three categories: Demographic, Academic and Behavioural features. During

© 2024 IJSRET

any educational process, learner‟s experience can be known by behavioural features as these two are related to each other.

The manuscript is systematized under the following subheadings: Section II presents literature review of educational data mining, Section III shows the brief introduction of machine learning classifiers. Data collection and preprocessing is discussed in Section IV, Proposed methodology is given in Section V, Section VI reports the results and discussion. Finally, we conclude the manuscript with conclusion and future work.

II. LITERATURE REVIEW

In the present epoch, traditional learning system has empowered by various collaborating learning tools such as e-learning systems, simulation applications, virtual reality and game-based education. The main aim of these tools to permit researchers to collect student‟s data and make its analysis for future usage. In [9], author has implemented the Decision Tree, Neural Network, and Support Vector Machine (SVM) on a student dataset of internet usage behaviours to predict academic performance. Results with accuracy (71-76%) demonstrate that there is an association between internet usage behaviour and academic performance. In [10], the authors used K-nearest neighbour and Decision tree classifiers for model projection which students‟ academic performance into good or bad category. To train this model, data was congregated from a university source and K-nearest neighbour and Decision tree classifier were implemented on that data. Decision tree classifier showed the accuracy 94.44%. To predict students‟ academic performance, the authors in [2] used SVM and Logistic regression classifiers to propose a classification model of machine learning. For this study, dataset was taken from an online educational platform to classify student academic performance as bad, average, or good. Results of this research showed that the accuracy produced by SVM was 79%, which was more than the logistic regression. Parameters of this research were accuracy, recall, precision, and f1-score to assess the system‟s performance. The authors [1] used a dataset containing 887 occurrences of 19 features of first year students and applied Random Forest classifier, Naïve Bayes and Ensemble learning classification machine learning model to predict the academic performance of students. The accuracy given by Random Forest classifier was 93% which was outclassed with other models. To measure the model performance, various parameters like recall, precision, and f1-score using confusion box metrics were considered. There are various encounters like accuracy, overfitting, underfitting, model deployment etc. that needs to be focused while doing research on machine learning in education. To identify and analyse significant features of a dataset that affect student academic performance, we use machine learning algorithms. By using these substantial features, numerous machine learning models are trained to classify and predict the category of student‟s academic performance. Model‟s performance can be measured through accuracy score and cross validation score.

III. DATA COLLECTION AND PREPROCESSING

1. Dataset Description

The dataset used in this study is an educational dataset that is collected from the learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection. [13]. The goal of X-API in this research is to track student behaviour during the educational process for evaluating the features that may have an impact on student‟s academic performance.

Table 1: xAPI Dataset features and their description

Feature Category	Name of the attribute	Description
Demographic Features	Nationality	Student's Nationality
	Gender	Male/Female
	Place of Birth	Birth Place
Academic Features	Educational Stages	Stage student belongs such as (Lower level, Middle level , and high level)
	Grade Level	Grade student belongs such as (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12)
	Section ID	Section student belongs such as (A, B, C).
	Semester	School year semester such as (First or second).
	Topic	Course topic such as (Math, English, IT, Arabic, Science, Quran)
Behavioural Features	Hand Raise	Student Behavior during interaction with Kalboard 360 e-learning system.
	Student Absent Days	
	Viewing announc4ments	

This dataset consists of 16 attributes and 480 instances. Features of this dataset are classified into three main categories: demographic, Academic Background and Behavioural features. There are 175 female and 305 male participants in the sample. The students are from a variety of countries, Kuwait is home to 179 people, Jordan to 172, Palestine to 28, Iraq to 22 people, Lebanon to 17 people, Tunis to 12 people, Saudi Arabia to 11 people, Egypt to 7

© 2024 IJSRET

people, Syria to 6 people, Libya, Iran, and the US are home to 4 people, Morocco to 1 person, and Venezuela to 22 people. Two academic semesters are used to gather the dataset. Record of 245 students is from semester 1 whereas the 2nd semester is having the data of 235 students. The summary of our dataset attributes is given in table 1.

2. Data Preprocessing and Feature Selection

To check any null, duplicates or invalid values in the dataset. data preprocessing is required. The dataset we have taken in our research is clean and is ready for encoding. The class attribute of the dataset is of multiclass category- "low (L), medium (M) and high (H)" on the basis of following conditions:

Low Level values 0-69
Middle Level values 70-89
High Level values 90-100

Feature selection is the process that focus on reducing the number of attributes that appearing in the patterns, in which it reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data, and increase the comprehensibility of the mining results [23, 24]. It is a technique that help in selecting significant features from the dataset.

IV. METHODOLOGY

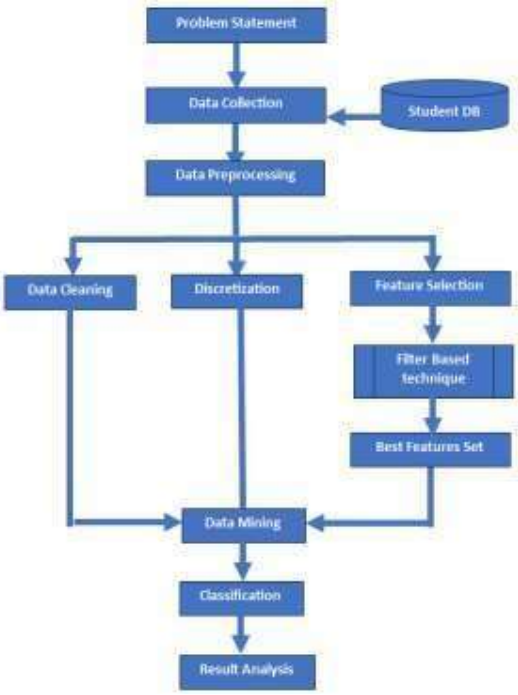

Fig.1. Flow Diagram of Methodology

To assess the features that may have an impression on academic success of students, a student‟s performance model using classification techniques is accumulated in this research paper. Main phases of the proposed methodology are shown in Fig.1. The first step of this methodology starts with data collection from Kalboard 360 E-Learning system using Experience API (XAPI). After data collection, the next step is data preprocessing step which converts the collected data into an appropriate format.

All machine learning models require only numeric values to get it train. Therefore, non-numeric values must be set to numeric before fitting it into a train model. After data preprocessing, discretization mechanism was used to transform the students‟ performance from numerical values into nominal values, which represents the class labels of the classification problem. To accomplish this step, we divide the data set into three nominal intervals (High Level, Medium Level and Low Level) based on student‟s total grade/mark such as: Low Level interval includes values from 0 to 69, Middle Level interval includes values from 70 to 89 and High-Level interval includes values from 90-100.

After discretization, feature selection process is used to select the best feature set. We used filter-based technique for feature selection. A classification model was constructed by using three different classification techniques (J48, SMO and Multilayer Perceptron)

Machine Learning Classifiers

J48: J48 is an implementation of the C4.5 decision tree algorithm in the Weka data mining tool. It is a supervised learning algorithm used for classification tasks. J48 builds decision trees by recursively splitting the dataset based on the feature that provides the highest information gain (or gain ratio). The tree structure consists of nodes where each internal node represents a decision based on a feature, and the leaf nodes represent class labels.

J48 can handle both numeric and categorical data, deal with missing values, and apply pruning to prevent overfitting by removing branches that do not contribute significantly to the decision-making process.

SMO: SMO (Sequential Minimal Optimization) is an algorithm used to train Support Vector Machines (SVMs) efficiently for classification tasks. It works by breaking down the complex optimization problem of SVMs into smaller, more manageable sub-problems that involve only two data points at a time. By solving these smaller optimization problems iteratively, SMO reduces the computational complexity and speeds up the training process.

SMO is particularly useful for large datasets, where traditional SVM training would be too slow. It supports both linear and

© 2024 IJSRET

non-linear classification by using kernel functions to handle non-linearly separable data

Multilayer Perceptron: A Multilayer Perceptron (MLP) in Weka is a type of artificial neural network used for classification and regression tasks. It consists of multiple layers of nodes: an input layer, one or more hidden layers, and an output layer. Each node (or neuron) uses a weighted sum of inputs and passes it through a nonlinear activation function. The MLP in Weka can be trained using backpropagation to minimize error, adjusting weights iteratively to optimize performance. Weka allows users to configure various parameters like learning rate, momentum, and the number of hidden layers to customize the MLP model.

V. EXPERIMENTS AND RESULTS

1. Environment

We ran the experiment on the laptop containing 4 GB of RAM, Quad Core processor. For our experiments, we used WEKA [20] to evaluate the proposed classification models and comparisons. Furthermore, we used 5-fold cross validation to divide the dataset into training and testing partitions.

2. Evaluation Measures

In this research paper, Accuracy, Precision, Recall, and F-Measure are taken for the evaluation of the classification quality: [21, 22]. Calculation of all these measures have shown in Table 2, (confusion matrix) which is based on the given below equations i.e., 1, 2, 3 and 4 respectively.

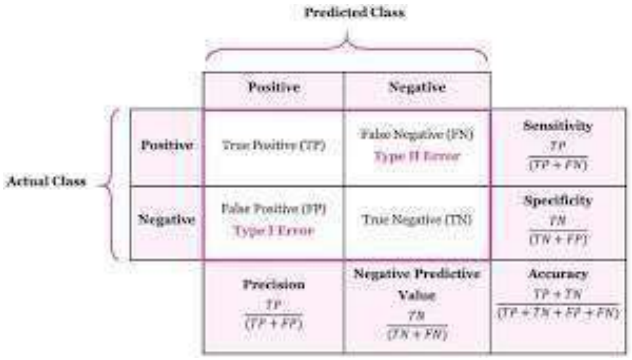

Table 2. Confusion Matrix

Accuracy, recall, precision, and F-measure are key performance metrics used to evaluate machine learning models in classification tasks Accuracy measures overall correctness, Precision evaluates positive prediction quality, Recall assesses sensitivity to positive instances, and F1 Score balances Precision and Recall [22].

$$Accuracy = \frac{TP+FN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

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

Results

After execution of different data mining classification techniques (J48, SMO and MP) on the data set, discrete results are produced. These results have shown in Table 3. Two types of results have been generated by each classifier:

- Results with student‟s behavioural features (BF)
- Results without student‟s behavioural features (WBF).

It is clearly visible from the results of Table 3, that classification results for different classification measures with behavioural features are better than the results when eliminating behavioural features. It shows that there is a sturdy association amongst learner behaviour in the classroom and student‟s academic achievement.

Table 2: Results of Classifiers With and Without Behavioral Features

Evaluation Parameters	J48		SMO		Multilayer Perceptron (MP)	
	With Behavioural Features	Without Behavioural Features	With Behavioural Features	Without Behavioural Features	With Behavioural Features	Without Behavioural Features
Accuracy	72.3	53.8	76.07	62.7	79.3	55.2
Recall	73	55.7	76.1	69.8	78.2	71.3
Precision	65.7	59.1	77	70	78.3	71.2
F-Measure	69.2	57.4	76.2	69.5	78.2	71.2

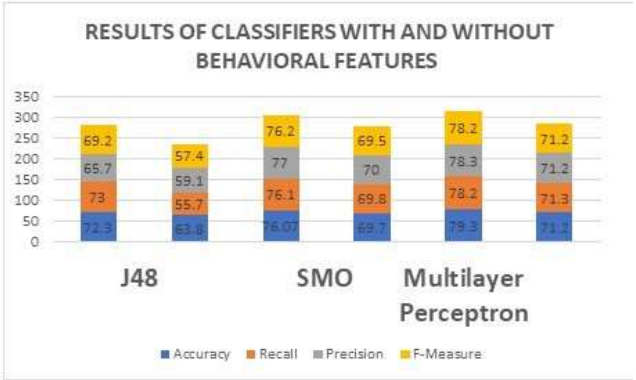

Fig. 2 Comparative Analysis of Classifiers Results

© 2024 IJSRET

It is seen from the results of Table 3 that MP model outperforms other data mining techniques. MP gives 79.8 accuracy with BF and 55.2 without behavioural features Similarly, J48 shows results 72.3 with BF and 53.8 without behavioural features. Another classifier SMO also shows more accuracy (76.07) with behavioural features as compared to accuracy (62.7) without behavioural features. Thus, it is inveterate from the results of table 3 that behavioural features play a significant role in predicting academic success of a student.

Table 3: Ranking Evaluation of Significant features using Ranker Method

Feature Name	Feature Rank
StudentAbsenceDays	54.6
VisitToResource	50.5
Raisehands	49.3
Announcement Views	39.7

Additionally, we applied ranker-based feature selection approach using Selected Attribute Eval based to check which the utmost significant features in the data set [18, 26]. Table 4 shows the Ranking Evaluation of Significant features using Ranker Method. As shown in the above table the behavioural factors got the highest feature ranks, which means learner behaviour during the educational process have an impact on student‟s academic success.

The experimental results prove that the learner behaviour has a strong impact on student‟s academic achievement. More accurate results can be achieved by increasing the training set, training time, classification methods, expanding data set with more distinctive attributes.

VI. CONCLUSION AND FUTURE WORK

A new student performance classification model which reflects the behavioural actions of the learner during learning process is introduced in this manuscript. The data was collected from an e-Learning system called Kalboard 360 using Experience API (XAPI).

The results we got from experimentation shows that learner‟s actions played a foremost role in learning process of students. The accuracy enhancement when using behavioural features are: J48 obtains 19% improvement, SMO obtains 14% improvement, and MP obtains 24% improvement.

In future, a more elaborated dataset with more distinctive attributes can be taken to get more accurate results. In addition to this, experimentation could be done by applying bagging, boosting and ensembling techniques of data mining.

REFERENCES

1. El-Halees, A. (2008) „Mining Students Data to Analyze Learning Behavior: A Case Study‟, The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18 ,2008.
2. Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students‟ Performance: A Case Study, International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.
3. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.
4. Kalboard360-E-learning system, http://cloud.kalboard360.com/User/Login#home/index/ (accessed July 31, 2015).
5. Naser, S. A., Zaqout, I.,Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2), 221-228.
6. Kevin Swingler “Data Mining Classification” http://www.cs.stir.ac.uk/courses/ITNP60/lectures/1%20Data%20Mining /3%20-%20Classification.pdf (accessed July 10, 2015).
7. Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.
8. Burcu a.m, A path model for analyzing undergraduate students‟ achievement, Journal of WEI Business and Economics-December 2013, Volume 2 Number 3.
9. Harris, A. (1999). Teaching and Learning in the Effective School. Aldershot: Ashgate.
10. Concordia Online Education, “Strategies to Improve Classroom Behavior and Academic Outcomes” http://education.cuportland.edu/blog/classroom-resources/strategies-to-improveclassroom-behavior-and-academic-outcomes/ (accessed September 19, 2015).
11. Bradley, P. Fayyad, U. and Renia C., "Scaling EM clustering to large databases". Technical Report. Microsoft Research. 1999.
12. Shannaq, B., Rafael, Y. and Alexandro, V. (2010) „Student Relationship in Higher Education Using Data Mining Techniques‟, Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.
13. Ayesha, S., Mustafa, T., Sattar, A. and Khan, I. (2010) „Data Mining Model for Higher Education System‟, European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

© 2024 IJSRET

14. Moisa, V. (2013). Adaptive Learning Management System. Journal of Mobile, Embedded and Distributed Systems, 5(2), 70-77.
15. Gunuc, S., & Kuzu, A. (2015). Student engagement scale: development, reliability and validity. Assessment & Evaluation in Higher Education, 40(4), 587-610.
16. Kuk, G. D. (2001). Assessing what really matters to student learning. Change, 33(3), 10-17.
17. Stovall, I. (2003). Engagement and Online Learning. UIS Community of Practice for E-Learning. http://otel.uis.edu/copel/EngagementandOnlineLearning.ppt
18. Ron Kohavi, George H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273-324.
19. Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2006). A comparison of classification techniques for the P300 Speller. Journal of neural engineering, 3(4), 299.
20. Arora, R., & Suman, S. (2012). Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications, 54(13), 21-25.
21. Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
22. Chen, T. Y., Kuo, F. C., & Merkel, R. (2004, September). On the statistical properties of the f-measure. In Quality Software, 2004. QSIC 2004. Proceedings. Fourth International Conference on (pp. 146-153). IEEE.
23. Asha.G.Karegowda1, A. S. Manjunath2 & M.A.Jayaram3, Comparative study of attribute selection using gain ratio and correlation based feature selection, International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277.
24. Blum, A.I., and Langley, P. “Selection of relevant features and examples in machine learning”, Artificial Intelligence, vol 97, 1997, 245- 271.
25. Jasmina Novakovic, Using Information Gain Attribute Evaluation to Classify Sonar Targets, 17th Telecommunications forum TELFOR 2009, Serbia, Belgrade, November 24-26, 2009.M. Young, The Technical Writer‟s Handbook. Mill Valley, CA: University Science, 1989
26. C. Jalota and R. Agrawal, “Analysis of Educational Data Mining using Classification,” Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 243–247, 2019, https://doi.org/10.1109/comitcon. 2019.8862214

© 2024 IJSRET