Predictive Analysis of Academic Performance of Students

Dr. M Lakshmi
Dept of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur
lakshmim@srmist.edu.in

Narendra Mani Tripathi
Dept of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur
ng3696@srmist.edu.in

Anees Uz Zaman Choudhary
Dept of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur
an8126@srmist.edu.in

***Abstract*—Academic records and related data collection are important aspects for tracking progress of students as well as prediction of their future results. We can have a record of such data through the institution, conduct multiple surveys for various related information through students and parents and use these records for better performance analysis. In this project, we will be using various machine learning algorithms on the generated dataset by using Clustering Algorithms, Decision Tree Algorithm, Naive Bayes Algorithms, Artificial Neural Networks and Regression for a clear idea of general accuracy for each algorithm using WEKA, which is a tool having various machine learning algorithms inbuilt that can be used for solving data mining problems found in real time world, developed by University of Waikato. Using these accuracy measurements, we will select the best algorithm for processing our data and implement a model based on the algorithm, so that even greater accuracy for the prediction of the results can be achieved. Using this model, we will be able to predict and anticipate the results of students and use this knowledge for policy making, guidelines, standardization of the course and the extra help needed for better learning experience of the students and higher chances of success in academic settings.**

Keywords— student performance prediction, classification, machine learning

I. INTRODUCTION

Education is an important sector not only for human development but the growth of community and the nation. For the better implementation of educational programs and their effectiveness using various methods, various data is collected and analyzed to understand for their improvement. The work of research in the field of educational data mining (EDM) is growing in a very fast pace. EDM is a relatively new field of study that uses various tools and techniques to get and evaluate, analyze large data that is related to activities that deals with people's learning and educational prospects. EDM is used for various purposes and people for their objectives to analyze and improve educational prospects of people and evaluate their performance. For example, the various techniques that can be implemented after evaluating and analyzing the educational data can help in identifying the problems with students and improving their performance. It can also be used by people in different fields in better decision-making. These educational data can be extracted from various sources such as surveys, educational records, various academic records created and maintained by educational institutions and web-based education records. EDM includes various Data Mining techniques, each technique is different and can be used for specific educational problems. As an example, the most popular technique for predicting an educational model is classification. There are many algorithms that comes under classification, for example: Bayesian networks, Neural networks or a Decision tree. Recently there has been a growing interest in research in educational data mining (EDM). There are different types of EDM users with their different goals and they use the techniques according to their own needs. From their use of the techniques, they use the analyzed data to improve their existing and newly found problems for better performance.
This article presents the performance model of a student with a new group of traits called behavioral traits. This proposed model will use some data mining techniques to evaluate the impact of student behavior characteristics on student academic performance. In addition, we try to understand the nature of these types of functions by expanding the data collection and pre-processing steps. "The data collection process is done using a learning activity tracking tool called the Experience API (xAPI). The collected characteristics are divided into three categories: demographic characteristics, characteristics of educational background and behavioral characteristics. Behavioral traits are a new category of traits related to lighter experience during the educational process" [1]. We then use a few data mining methods to create a model based on the academic performance of students: Artificial Neural Network (ANN), Regression, Naive Bayes, K-means clustering, Decision Tree and Random Forest. We will then use various methods that are needed to improve the performance of students. With all of the models compared, we will select the most accurate model and try to better its performance to get the maximum accuracy possible by deploying it from scratch.

II. RELATED WORK

Predicting the performance student and how they can improve is plays an important role in the field of education and learning. Multiple Data Mining techniques are used to

create predictive model models for this purpose, for example Artificial Neural Network (ANN), Regression, Naive Bayes, K-means clustering, Decision Tree and Random Forest. Ranking is one of the techniques which is most popular for analyzing the academic performance of students. We use all these techniques and try to create a model and find techniques which provide the most accurate result.

The decision tree is a technique which uses flowchart-like structure with attributes in a structure arranged based on hierarchy. "Most researchers used this technique for its simplicity, with which it can be converted into a series of classification rules. Some of the best-known DT algorithms are C4.5 and CART." [1]. "Romero et al, used the DT algorithm to predict the final grades of students based on their usage data in the Moodle system. Moodle is one of the most widely used learning content management systems (LCMS)." [8] Dataset was created by collecting real-world data from seven Moodle courses with the University of Cordoba which divided the students into two groups with one being "passed" and the other "failed". The main objective of that research was to divide students with similar final grades into distinct groups based on the activities performed by the student in a web-based course.

"The neural network is another popular technique that has been used in data mining in education. A neural network is a smart, biologically inspired technology composed of connected elements called neurons that work together to create an output function. Arsad et al used the Artificial Neural Network model to predict the academic performance of undergraduate engineering students. The study uses the grade point average (GP) of core subjects assessed by students as input regardless of their demographic background, while the cumulative grade point average (CGPA) is used as the result. Neural Network (NN) trains GP engineering students to achieve desired performance" [1].

"Naive Bayes classifiers are a family of algorithms. These classifiers are based on Bayes' Theorem, which finds the possibility of a new event based on previously occurring events. Each classification is independent of one another but has a common principle" [6]. The fundamental Naive Bayes assumption is very "naïve", which is that each feature makes an (1) independent and (2) equal contribution to the outcome. "N. T. N. Hien and P. Haddawy used the Bayesian networks to forecast the CGPA of Students based on their background during the time of their admission." [13]

To summarize, various studies have been made to predict and forecast the academic performances of students with the help of Data Mining or Machine Learning. However, not many have tried to incorporate the behaviour of the student during the learning process and its impacts on the academic performance of the student. This study will try to incorporate these behaviour patterns as well as the interaction with the e-learning methods. This will help to further understand how the learning and teaching processes are to be done, to better it and improve the academic performance of the students as well as help administration for improving the overall learning experience.

The increasing use of the Internet in education has created a new context known as a web-based learning or education management system (LMS). "The LMS is a digital framework that manages and simplifies online learning. The main purpose of the LMS is to manage learners, monitor learner engagement, and track their progress throughout the system. The LMS allocates and manages learning resources such as registration, classroom, and online learning delivery. In this article, the dataset is generated from the Learning Management System (LMS) referred to as Kalboard 360. Kalboard 360 is a multi-agent LMS made for learning and teaching through the use of modern internet technology." [2]. This system lets the user to access any educational resources and other various academic related data from any device such as mobile, laptop or desktop with an internet connection. It involves both the parents and school administration in the learning, managing deadlines for work to be done by the students and monitoring progress regularly. "This makes for a really great process that properly connects and involves all parties. The data is collected using a learning activity tracking tool called the Experience API (xAPI). The xAPI is a component of the Training and Learning Architecture (TLA) that can be used to monitor learning progress and learner actions, such as reading an article or watching a training video. The Experience API helps providers of learning activities, the learner, the activity, and the objects that describe a learning experience." [2]

III. IMPLEMENTATION

A. Algorithm Selection:

The collected data is thus analyzed using various machine learning algorithms such as clustering, Decision Tree, Random Forest and Neural Networks. Accuracy and precision of each of the algorithm is measured in general terms using WEKA.

"WEKA is a tool having collection of machine learning algorithms for data mining tasks." [15]. The data processed is used to predict the range of marks the students will attainted in their respective subjects by understanding and analyzing their behaviour, characteristics and other factors affecting their academic life. The tool also validates the data by itself so that we can find out which algorithm suits best for the task in hand.

TABLE 1. ACCURACY OF ALGORITHMS

Algorithm Used	General Accuracy
Random Forest	76.67%
Decision Tree	75.83%
Regression	72.29%
Naïve Bayes	67.70%
Neural Network	79.38%

After generating the above results of each algorithm using WEKA, we carefully decided on building a proper model for analyzing our dataset. These models will be used for the actual processing of the data and generate meaningful outputs for our studies.

B. Model Generation:

Using the above general accuracies of each algorithms, we try to build our own model on the basis of most accurate

models. For building the models, we use Python as our primary programming language since is it simple, easy to use and has multiple helpful libraries to make our work hassle-free. We use an IDE called Anaconda for ease in programming the model since it is a user-friendly platform for facilitating data science studies. Jupyter Notebook is the editor that comes built-in with Anaconda which helps us in writing and executing the codes required for developing the models.

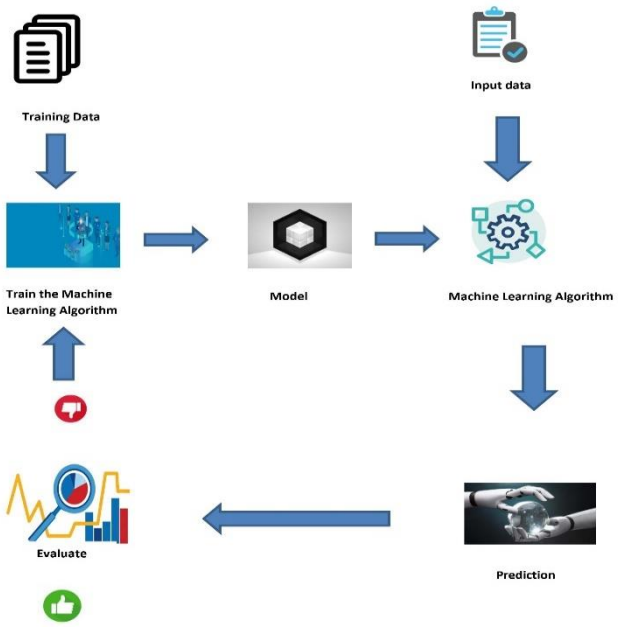

Fig 1: Machine Learning Model Lifecycle

Considering the general accuracies of each of the algorithm, we use Jupyter notebook provided in Anaconda to develop models using the respective algorithms. A model based on neural network was developed in the beginning stage considering that its general accuracy was the highest. But because of its inadequate accuracy in the model while deploying the use case, other models based on the decision tree, random forest and naive bayes and other classifiers had to be developed.

Firstly, all the required libraries to be used in the jupyter notebook was imported, then the dataset is loaded and read. Visualization and Grouping of the data are done in the next step. All the data in the columns which have categorical values are converted to numerical values. The sklearn library is used to import all the classifiers that are to be used for the machine learning. For that we use pd.get_dummies which is used for adding columns to the data. It puts 1s and 0s in those various columns. Reusing the same data for testing and training is not a good idea because we need to find out how the data will work on the data it wasn't trained on. The data is split into training and testing data using train-test-split, which is imported from the sklearn library. A function is created for the classifier in which the input and classifiers used are specified. The classifiers used for these models are Random forest (Random Forest Classifier G and Random Forest Classifier E), Extremely Randomized Trees Classifier (Extra Trees Classifier), K Nearest Neighbour Classifier, Decision Tree Classifier, Logistic Regression and Naïve Bayes (Gaussian Naive Bayes and Bernoulli Naïve Bayes). We then run these classifiers on a particular range, 20 in this particular case.

Cross validation is then used, which is way in which we can use our training data in order to get good estimated of how well our model will perform on new raw data. Cross validation allows us to compare different machine learning methods and get a sense of how well they will work in practical world. For the cross validation, a method is provided which is called cross validation score which takes the classifier, X and Y values. By taking all these cross-validation scores and calculating their mean, the accuracy of each algorithm is measured. This is done for the accuracy of each value of class i.e., L, M, H individually and then by calculating the mean of the results of respective class with each algorithm.

While converting the categorical values to numerical values manually through label encoding the accuracy of the algorithms obtained is not very high. This is because algorithms use the weight per feature or distances between samples. In label encoding, the algorithms assume the higher numbers are more important. Similarly, algorithms which use sample distance might not consider the importance of each attribute depending on the higher value of the distances. To tackle this problem, one-hot encoding is implemented in the models. One-hot encoding solves this issue of uneven values of data since it only uses 0s and 1s, and for each feature of the attribute it creates a separate column giving equal values and importance to each attribute. Thus, maximum accuracy is achieved by using one-hot encoding. And hence using these techniques multiple models based on each of the classifiers is developed, deployed and tested for our data.

IV. Results Discussion

On testing our data on all the models developed, we obtained various degrees of accuracy depending upon the nature of the classifier. Surprisingly some of the classifiers that had higher accuracy in the general accuracy in WEKA got outperformed the lesser accurate classifiers. This may be the result of particular data types and one-hot encoding implementation. Since label encoding was actually following the general given trends of the classifiers although the overall accuracy for each classifier in label encoding was quite lesser compared to its counterpart that is used in the models.

TABLE 2. ACCURACY OF MODELS

Model	L Accuracy	M Accuracy	H Accuracy	Overall Accuracy
Random Forest Classifier G	96.10%	99.36%	96.44%	97.63%
Random Forest Classifier E	96.70%	99.40%	96.57%	97.85%
Extra Trees Classifier	96.67%	99.86%	96.71%	98.08%
Decision Tree Classifier	96.90%	100%	100%	99.18%
K Neighbours Classifier	83.33%	60.41%	73.54%	70.36%
Logistic Regression	96.87%	100%	98.12%	98.62%
Gaussian Naïve Bayes	93.12%	89.16%	87.91%	89.34%
Bernoulli Naïve Bayes	90.41%	90.83%	86.87%	89.54%

From the above table, we can see the accuracy of each model for each of the outcome of the results as well as the overall accuracy of those models. Decision Tree Classifier clearly outperforms all other classifier because of its very high accuracy to predict results of M and H class. Hence it seems to be most suitable for the project that is going to be implemented.

V. Conclusion

Academic performance is of great concern to academic institutions around the world. The widespread use of LMS generates large amounts of data on teaching and learning interactions. This data contains hidden knowledge that can be used to improve student academic performance. In this article we propose to find out the effectiveness of the existing algorithms for predicting student performance based on data mining techniques with new data attributes / characteristics, called behavioral characteristics for students. These types of functions are related to the student's interactivity with the learning management system. The performance of the student's predictive model is evaluated using a series of classifiers, namely: Artificial Neural Network, Naive Bayesian, and Decision Tree. Once we figure out the best algorithm using various performance measurement matrices, we try to better those algorithms to achieve even higher efficiency and develop a better model using those algorithms. The results generated shows a close relationship between the student's behavior in the class and their performance throughout the year in that particular subject. The resource visited function seems to be the most effective behavior function for the student performance model. Ultimately, this model can help educators understand students, identify weak students, improve the learning process, and reduce rates of academic error. It can also help administrators improve learning system results. And thus, we try to help the educators as well as the learners to understand their strengths and weakness regarding their academics and help them get the support required to achieve better academic success.

References

[1] Elaf Abu Amrieh,Thair Hamtini and Ibrahim Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods", International Journal of Database Theory and Application Vol.9, No.8 (2016), pp.119-136.

[2] Hussah Talal and Saqib Saeed, "A study on adoption of data mining techniques to analyze academic perfomance", ICIC Express Letters Part B: Applications Volume 10, Number 8, August 2019, pp. 681-687.

[3] Amjad Abu Saa1, Mostafa Al-Emran2(&) , and Khaled Shaalan1, "Mining Student Information System Records to Predict Students' Academic Performance", A. E. Hassanien et al. (Eds.): AMLTA 2019, AISC 921, pp. 229–239, 2020.

[4] Liya Treesa Kunjumon1 ,Sharon Shaji2 ,Saffi Treesa Saji3 , Thasneem Naushad4 , Neena Joseph5**,** "An Intelligent System to predict Students academic performance using Data Mining", International Journal of Information Systems and Computer Sciences, Volume 8, No.2, March - April 2019

[5] Rastrollo-Guerrero, Juan & Gomez-Pulido, Juan A. & Domínguez, Arturo. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences. 10. 1042. 10.3390/app10031042.

[6] Ferda Ünal (March 28th 2020). Data Mining for Student Performance Prediction in Education [Online First], IntechOpen, DOI: 10.5772/intechopen.91449. Available from: https://www.intechopen.com/online-first/data-mining-for-student-performance-prediction-in-education

[7] González-Brambila, Silvia & Sánchez-Guerrero, Lourdes & Ardon, Irma & Figueroa-González, Josué & González-Beltrán, Beatriz. (2018). Predicting Academic Performance of Engineering Students After Approving a Mathematics Leveling Course using Decision Trees. Research in Computing Science. 147. 171-181. 10.13053/rcs-147-12-16.

[8] Romero, Cristóbal & Ventura, Sebastian & García, Enrique. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education. 51. 368-384. 10.1016/j.compedu.2007.05.016.

[9] P. M. Arsad, N. Buniyamin and J. A. Manan, "A neural network students' performance prediction model (NNSPPM)," 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, 2013, pp. 1-5, doi: 10.1109/ICSIMA.2013.6717966.

[10] A. M. Shahiri and W. Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", Proceeding Computer Science, vol. 72, (2015), pp. 414-422.

[11] Cristóbal Romero, Sebastián Ventura, Enrique García, Data mining in course management systems: Moodle case study and tutorial, Computers & Education, Volume 51, Issue 1, 2008, Pages 368-384, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2007.05.016.

[12] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting Students' Performance Using Machine Learning Techniques", *JUBPAS*, vol. 27, no. 1, pp. 194-205, Apr. 2019.

[13] N. T. N. Hien and P. Haddawy, "A decision support system for evaluating international student applications", In Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual. IEEE, (2007), pp. F2A-1.

[14] https://www.kaggle.com/c/student-academic-performance

[15] https://www.cs.waikato.ac.nz/ml/weka