WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predicting Student's Performance Using Machine Learning Algorithm

Mr. Swapnil Patil, Mr. Uday Chaudhari, Ms. Swati Kangane, Ms. Rupali Shelar, Ms.Sweety Mahajan

UG Student, Information Technology (Bachelor of Engineering). Sir Visvesvaraya Institute of Technology, Nashik Maharashtra, India

A B S T R A C T

Although the educational level of the Portuguese population has improved in the last decades, the statistics keep Portugal at Europe's tail end due to its high student failure rates. In particular, lack of success in the core classes of Mathematics and the Portuguese language is extremely serious. On the other hand, the fields of Machine Learning, which aim at extracting high-level knowledge from raw data, offer interesting automated tools that can aid the education domain. The present work intends to approach student achievement in secondary education using machine learning techniques. Recent real-world data (e.g. student grades, demographic, social and school related features) was collected by using school reports and questionnaires. The two core classes (i.e. Mathematics and Portuguese) were modelled under binary/five-level classification and regression tasks. As a direct outcome of this research, more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.

Keywords: Classification, Data Mining, Supervised Learning, Education , Traditional Methods, Grades.

1 Introduction

Extensive efforts have been made in order to predict student performance for different aims, like: detecting at risk students, assurance of student retention, course and resource allocations, and many others. This research aims to predict student performance to engage distinct students in researches and innovative projects that could improve universities reputation and ranking nationally and internationally. However, analyzing students records for startup to medium size institutes or schools, like the British University in Dubai which have small size of students records, have never been explored in educational or learning analytics domain. Yet, that were investigated in other fields, like: health sciences and Chemists (Ingrassia&Morlini, 2005; Pasini, 2015). So, this project aims to explore the utilization possibility of small students' dataset size in educational domains.

Additionally, in most researches that were aimed to classify or predict, researchers used to spend much efforts just to extract the important indicators that could be more useful in constructing reasonable accurate predictive models. They will either use features ranking algorithms or will look at the selected features while training the dataset on different machine learning algorithms, like in (Comendador, Rabago, &Tanguilig, 2016; Mueen, Zafar, &Manzoor, 2016). Instead, and until recently, there have been no research efforts to investigate the ability of visualization or clustering techniques in identifying such indicators for small dataset, especially in the learning analytics domain (Asif, Merceron, Ali, &Haider, 2017). If such studies will be conducted, its outcomes might prove the feasibility of mitigating the hassle that is normally spent on features extraction or selection processes.In the present scenario, data mining/Machine Learning isa very important field of research and playing anindispensable responsibility in educational institutions and oneof the most important areas of exploration with the aim to findout relevant facts taken from historical data stored in hugedataset. Data mining for education i.e. Educational DataMining (EDM) is the discipline which uses data miningtechniques in the environment of education. It is a veryimportant research area which helps to predict usefulinformation from educational databases to improveeducational achievement and to have better assessment of thestudents learning process. Educational Data Mining could beconsidered as a best option of the science of learning and as abranch of data mining [1][2][3]. Educational Data Mining canbe useful while creating a model of user perception, actionand trial [4]. Data Mining or knowledge discovery has gainthe popularity in such a way that it has become the emergingrelevance because it is very helpful in examining data formdivergent approach and abridge it into functional information. Educational data mining relies on many data miningtechniques like k-nearest neighbor, neural networks, decisiontrees, support vector machines, naive bayes, and many more. For doing quick analysis on data with the help of datamining techniques, there are many open source softwares likeweka, rapid miner, orange, knime, SSDt (SQL Server dataTools) designed for data investigation and to getunderstandable structure for future use. In this

paper, we useWEKA (Waikato Environment for Knowledge Analysis)which is best suited for the analysis of data and to built amodel to get predictive outcome.

2 Literature Survey

Most of the researcher have done their study in data mining using for educational purposes to get the prophecy of the students' achievement. In [8] the performance of engineering students can be judged with the help of Decision Tree (DT) algorithm. Around 340 students data was collected for the prophecy of their achievement in the first year exams. The build model was able to generate only 60% accuracy in the training set. In [9] WEKA was used for the prognosis of marks of final year students and these were based on two different dataset's parameters. There was one common information in each dataset i.e. variety of students could be taken from one college course in last four semesters. In [10] the author analyzed with his own reviews of past research work done on performance prediction of students' its analysis and assessment by applying dissimilar techniques of data mining. In [11] the authors measuring student performance using Decision Tree classification techniques and used artificial neural network to build classifier models. The produced outcome was based on various traits to foresee the outcome of the students. Analyzing the weakness and strength of student which may be helpful to improve the performance in future. This study shows the efficacy of applying the methods/procedures of data mining in course rating data and the data could be mined for education at higher level. In [12] the authors represent a study that will be beneficial to the students and the teachers for the betterment to uplift the result of the students who are having more chances of non success. There are many parameters like Attendance, Seminar and assignment marks were collected from very important resource i.e. previous database of students, to evaluate their prophecy at the semester end. The authors used Na¨ıve Bayes classification algorithm that shows a highest accuracy compared to other classification algorithms. The researchers in [13] worked on a relative research to examine various decision tree algorithms and their influence on the data set choose for education to stereotype the education related prophecy of stake holders i.e. students. It mainly cynosure on choosing the top prioritized algorithm of decision tree and explain the detailed meaning of each one of them and the result shown that the regression as well as classification methods are best because they are more compatible to produce better result with the dataset that is already tested. Researchers in [14] have concluded with an idea for the better use of data mining techniques in the prediction of student's prophecy and also it provided the strong interpretation that algorithms for prediction of data mining, Decision Tree and Neural network are the two prime methods which are highly advisable by the researchers for the prediction of student's prophecy. Authors in [15] applied Data Mining techniques to find and evaluate future results and factors which affect them. Author in [16] discussed k-Nearest Neighbor (k-NN) algorithm which plays an effective role in the accuracy of the classifier.

3 Machine Learning Algorithms

1. Decision Trees: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility. Take a look at the image to get a sense of how it looks like.From a business decision point of view, a decision tree is the minimum number of yes/no questions that one has to ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.
2. Naive Bayes Classification: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The featured image is the equation—with P(A|B) is posterior probability, P(B|A) is likelihood, P(A) is class prior probability, and P(B) is predictor prior probability.
   Some of real world examples are:

   - To mark an email as spam or not spam
   - Classify a news article about technology, politics, or sports
   - Check a piece of text expressing positive emotions, or negative emotions?
   - Used for face recognition software.
3. Ordinary Least Squares Regression: If you know statistics, you probably have heard of linear regression before. Least squares is a method for performing linear regression. You can think of linear regression as the task of fitting a straight line through a set of points. There are multiple possible strategies to do this, and "ordinary least squares" strategy go like this—You can draw a line, and then for each of the data points, measure the vertical distance between the point and the line, and add these up; the fitted line would be the one where this sum of distances is as small as possible.
4. Logistic Regression: Logistic regression is a powerful statistical way of modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.In general, regressions can be used in real-world applications such as:
   a. Credit Scoring
   b. Measuring the success rates of marketing campaigns
   c. Predicting the revenues of a certain product
   d. Is there going to be an earthquake on a particular day.

5. Support Vector Machines: SVM is binary classification algorithm. Given a set of points of 2 types in N dimensional place, SVM generates a (N—1) dimensional hyperplane to separate those points into 2 groups. Say you have some points of 2 types in a paper which are linearly separable. SVM will find a straight line which separates those points into 2 types and situated as far as possible from all those points. In terms of

scale, some of the biggest problems that have been solved using SVMs (with suitably modified implementations) are display advertising, human splice site recognition, image-based gender detection, large-scale image classification.

6. Clustering Algorithms: - Clustering, like regression, describes the class of problem and the class of methods. Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonalities.
    a. The most popular clustering algorithms are:
    b. k-Means
    c. k-Medians
    d. Expectation Maximization (EM)
    e. Hierarchical Clustering
7. Artificial Neural Network Algorithms: Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types. Note that I have separated out Deep Learning from neural networks because of the massive growth and popularity in the field. Here we are concerned with the more classical methods. The most popular artificial neural network algorithms are:
    a. Perceptron
    b. Multilayer perceptions (MLP)
    c. Back-Propagation
    d. Stochastic Gradient Descent
    e. Hopfield Network
    f. Radial Basis Function Network (RBFN)

4 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. for several types of loan. The values for these attributes can have outliers that do not fit into the regular range of data.

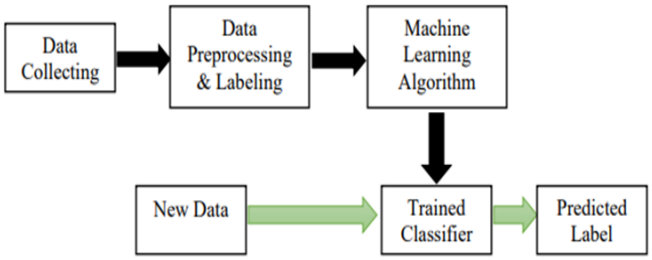

Fig. System Architecture

5 Algorithm Used

Naives Bayesian Algorithm

This method is based on Na¨ıvebayes classifier and the objective is to know what students may acquire in their end results of semester. They can be benefitted from prediction of results of students in several ways. Teachers and students take essential steps to developthe outcomes of those students whose prediction result is not fulfilled and a training set of students data is taken to construct the model of na¨ıvebayes and then it is applied on test data to find the results of students end semester. Makhtar et al. [17] examines student's performance using na¨ıvebayes classifier which is one of the methods of classification in data mining to recognize the hidden data between subjects that influenced students performance in SijilPelajaran Malaysia. The na¨ıvebayes algorithm can be employed for classification of performance of students in early stage of 2nd semester with 74% accuracy. Students choosing engineering as their discipline is developing rapidly but due to different factors and improper education in India the rates of dropout are greater. Students are not capable to shine in the subjects of engineering which are mathematical and complex hence mostly keep term or get drop out in that subject. With the use of data mining techniques the students performance can be predicted in terms of drop out and grade for a subject. Naivesbayes algorithm is used in this research and based on the rules acquired from the developed method the system can derive the major factors impacting the performance of students. Razaque et al. [19] described the method of classification which was based on the algorithm of na¨ıvebayes and use for mining academic data. It was used for students along with teachers for academic performance evaluation. It was cautionary approach for students to develop their study performance. This research was an effort to recognize students who need special attention in reducing the failure and take appropriate steps for upcoming semester exams. Divyabharathi and

Someswari [20] constructed a predictive model for academic performance of students. As there are several classification methods available this research used na¨ıvebayes classification technique. By using this model timely decisions can be taken to avoid student's academic risk. The instructor can know how poorly or how well students in class will perform. This study concentrated on validating and developing mathematical models that can be used to predict the academic performance of students in educational institutions.

6 Results

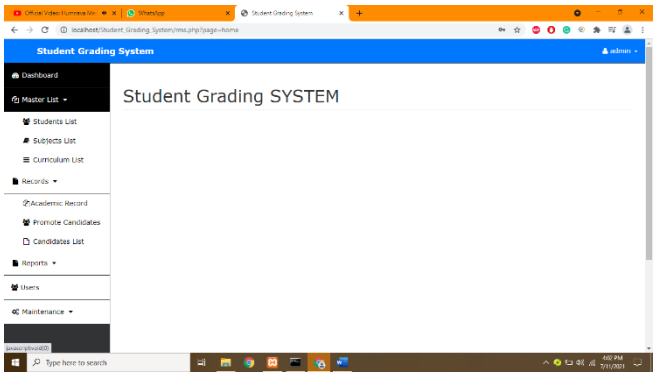

Fig. System Design

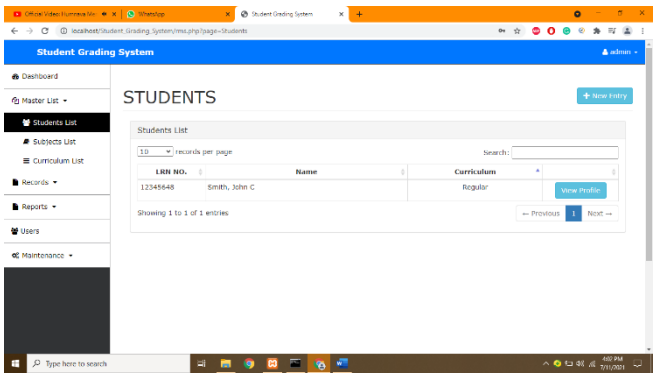

Fig. System Design

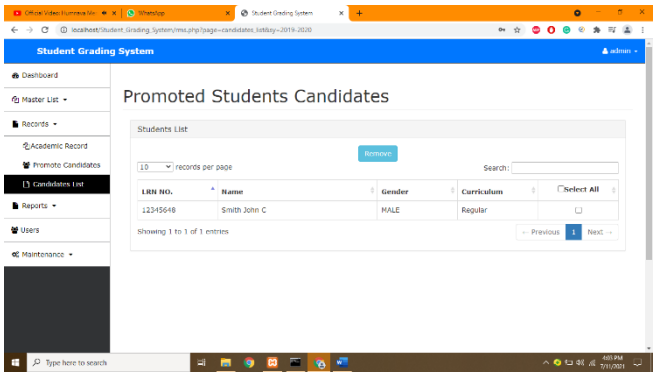

Fig. System Design

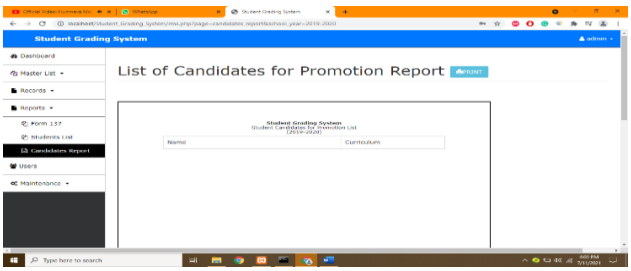

Fig. System Design

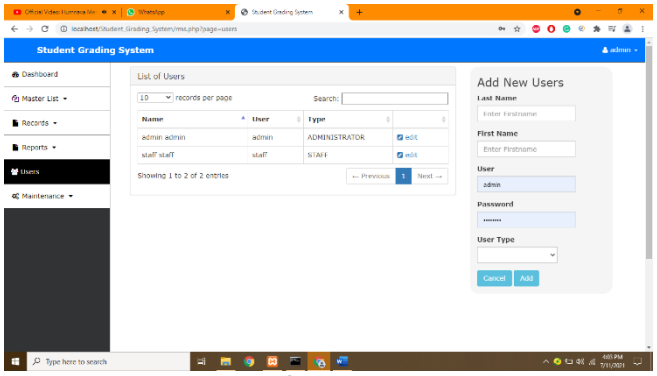

Fig. System Design

7 Conclusion

Data mining has a significant importance in educational institutions. The knowledge acquired by the usage of data mining techniques can be used to make successful and effective decisions that will improve and progress the student's performance in education. Data set contains of 163 instance and sixteen attributes. Five classifiers are used under weka and the comparisons are made based on the accuracy among these classifiers and different error measures are used to determine the best classifier. Experiments results show that Multilayer Perceptron has the best performance among other classifiers. In future work, more dataset instance will be collected and will be compared and analyzed with other data mining techniques such as association and clustering..

References

[1]M. Goyal and R. Vohra, "Applications of Data Mining in Higher Education", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[2] R. Huebner, "A survey of educational data mining research", Research in Higher Education Journal, 2012.

[3] M.S. Mythili, A.R. Mohamed Shanavas, "An Analysis of students' performance using classification algorithms", IOSR, Journal of Computer Engineering, Volume 16, Issue 1, January 2014.

[4] S. Lakshmi Prabha, A.R.Mohamed Shanavas, "Educational data mining applications", Operations Research and Applications: An International Journal (ORAJ), Vol. 1, No. 1, August 2014.

[5] C. Romero, S. Ventura and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers Education, Vol. 51, no. 1, pp. 368-384, 2008

[6] S. Ayesha, T. Mustafa, A. Sattar and M. Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, no.1, pp.24-29., 2010

[7] Weka: Data Mining Software in Java, University of Waikato,[Online].Available: http://www.cs.waikato.ac.nz/ml/index.html. [8] Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science IT Education Conference (In SITE) 2010.

[9] I. Milos, S. Petar, V. Mladen and A. Wejdan, Students' success prediction using Weka tool, INFOTEH-JAHORINA Vol. 15, March 2016. [10] P. Kavipriya, A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016 ISSN: 2277 128X.

[11] N. Ankita, R. Anjali, Analysis of Student Performance Using Data Mining Technique, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2017. [12] P. Shruthi, B. Chaitra, Student Performance Prediction in Education Sector Using Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, Issue 3, March 2016. [13] S.K Yadav, B. Bharadwaj, and S. Pal. Data Mining Applications: A Comparative Study for Predicting Student's Performance. International Journal of Innovative Technology Creative Engineering (ISSN: 2045- 711), Vol. 1, No.12, December 2012

[14] A.Mohamed Shahiria,, W. Husaina , N. Abdul Rashida, "A Review on Predicting Student's Performance using Data Mining Procedia Computer Science 72 ,414 – 422, ELSEVIER 2015.

[15] K. Kohli and S. Birla, " Data Mining on Student Database to Improve Future Performance", International Journal of Computer Applications, Vol.146 No.15, pp. 0975 – 8887, July 2016.

[16] Rashmi Agrawal, "Integrated Effect of k Nearest Neighbors and Distance Measures in k-NN Algorithms", International Journal of Advances in Intelligent Systems and Soft Computing, vol. 654, pp.759- 765 , Springer, 2017

[17] Rashmi Agrawal, Neha Gupta "Educational Data Mining Review: Teaching Enhancement", Privacy and Security Policies in Big Data, pp.149-165, IGI Global, 2017