International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 12
ISSN: 2454-4248
93 – 97

Student Performance Prediction Using Educational Data Mining Techniques

Harleen Kaur
Research Scholar, Department of Computer Science and Engineering, Chandigarh University, Gharuan
harleenkaur3663@gmail.com

Er. Gourav Bathla
Assistant Professor, Department of Computer Science and Engineering, Chandigarh University, Gharuan
gourav.cse@cumail.in

Abstract- Educational sector produces data in large amount that is too voluminous and complex to understand. There is a need to efficiently filter and prioritize the data so as to deliver the relevant information to get rid of information overloading. Data mining searches through the large amount of dynamically generated data to present users with the useful and understandable patterns and trends. It has the power to use the raw data effectively which has been produced by universities, to draw the hidden patterns and the relationships among the attributes that are used in predicting the student performance, their behaviour effectively. In this paper the data mining techniques have been briefly described. The literature review of educational data mining is also done. This paper, implements data mining techniques such as Naive bayes and Support vector machine to predict the student performance.

Keywords *– Educational data mining, Support vector machine, Naïve Bayes*

*****

1. Introduction

In a country's life education plays a vital role to ensure the survival of the state and the nation[1]. In today's scenario educational technologies aide the process of learning and teaching (TL) as they are being used in educational domains including the traditional form of classrooms where it's all about face to face and even the learning platforms available online. Educational actors(students, teachers and administrators) have been benefitted as they are provided with the relevant information in which they have to act upon and thereby end up in promoting the quality based innovations in this domain[2]. These days universities are run in a very powerful and dynamically viable manner. A large amount of data is gathered in the form of marks, records, documents, files, performance et cetetra all related to student performa[3].

Educational data mining act as a bridge between the two one is the education and the other is computer science. The subfields of computer science, Data mining and Machine learning are used. Data mining is used to uncover the hidden patterns in the unstructured data. It is devoted to discover the knowledge and then generate the relevant information. Due to the data mining advancements it has become possible to mine educational data to get the useful data. This relevant information serves to benefit its handlers[4].

The rapid growth in educational domain brought the fact of distilling the massive amounts data which let to the emergence of educational data mining(EDM)[5]. Educational data mining is a specific field of data mining that helps in discovering invisible patterns in the data to take decisions for students, teachers and administrators. EDM make predictions which are further characterize the learner behavior, domain content knowledge, educational functionalities, assessment outcomes, applications and the applications. The outcomes can guide the students in learning process, tutors in teaching process to enhance the educational practices and the administrators in managing process[6].

1.1Phases of Educational Data Mining

The advancements in the EDM are progressively evolving the data mining techniques to inform the educational domains. Ultimately, the objective is to gather relevant information about the learning and teaching for the pedagogical enhancements. Accordingly, EDM has been divided into four phases.

The first phase, discovers the relationships between data using data mining techniques such as classification, clustering, regression, sequential pattern mining and association rule mining.

The second phase, validates whether the selected relationships are theoratically validated or not. If the relations are validate further processing is done otherwise not.

The third phase, predictions are made according to the relationships that are theoretically validated for the future aspects in educational learning and teaching context.

The fourth phase, supports the predictions made in third phase and make policy level decisions which would help in pedagogical improvement in the educational era[6].

1.2Approaches in EDM

Prediction

 ISSN: 2454-4248
93 – 97

The aim is to infer the single label of the data that has been collected, by adjoining the other aspects of the data collection. The name itself indicates, it will predict the future aspects on the basis of the past data. In educational domain it can be used to predict the student's behavior and their educational success[7].

Classification

Classification is a supervised learning technique. It maps the different unkown samples into the given samples[8]. This technique is only used when the predicted variable is none other then categorical. It is employed on to the predefined classes in test situations, training and mapping data. The decision tree, K-nearest neighbor, support vector machine and neural network algorithms are classified under the classification technique[7].

Regression

This technique is used only when the predicted variable is numerical[8]. The regression technique is a technique to predict the relationship among dependent variables and independent variable. The independent variable is already known and the response variables are to be predicted. Regression analysis is further divided into linear regression, non linear regression and multiple regression[8].

Clustering

This technique groups the similar data together into clusters. In the same clusters each object is very similar and as compared to other clusters they are less similar. Clustering and classification differ because clustering does not follow a specific criteria, instead they find the logic between the data collected. Beforehand no one knows in how many groups or in how many different kind of groups the data is to be divided in clustering. Clustering is used for preprocessing the data[8]. It is used to map the preferencesess of the student to explore the different types of educational contents and to look for the different learning patterns. In educational data mining clustering is used for figuring out the typical behavior of the students that can be grouped together in different groups. The methods that are classified under the clustering technique are K-means, Hierarchical clustering et cetera[9].

Relationship Mining:

It is done to find out the relationships among the variables and to further use them as rules. We try to identify those variables are strongly connected to the other variable of the same kind. In educational data mining the relationship mining helps in discovering the student's behavioural patterns and difficulties they face during the learning and even the mistakes are detected in the learning process that would help the teachers[7].

Association rule mining

This technique is based on the "if then" rules. Here the relationships among the different variables is done as "If the student is willing to improve his or her performance level he or she should get the usage of the available help"[8].

Sequential pattern mining

Sequential pattern mining is the technique that helps in discovering the temporary associations among the different labels or events. In educational data mining it can be used to infer the student's request regarding the aide they require over time to improve their educational status[7].

Correlation mining

In correlation mining we identify the linear correlation between the different variables as being positive or negative. This provides aide to EDM by making the relationships among the students's attitude for a particular activity as if that is positive it means the students tried to finish it and if negative it means the student did not attempt the activity[7].

Casual data mining

Casual data mining is used to discover the relationship causes among the different variables. It identifies if the particular event is caused/originated by another. Here the predictions are made about the factors affecting the students's performance related to the activity they were given[7].

Discovery within models

The aim is to identify the structure according to the relationships without the predefinition of the idea to the fact that what is to be found. Discovery within models opposes the predicting methods as it does not allow the access to the previous correlations between the variables. In educational data mining it is used to find the factors that are responsible for the student's competencies development[7].

Distillation of data for human judgement

It provide the aide for understanding the analysis of results. It makes the complex data easily more understandable by humans to leverage their jugdement[9].

2. Applications of Educational data mining

Educational data mining is used in various fields and has enormous applications in the student modeling area as follows.

i. Performance and characteristics
ii. Undesirable behaviors
iii. Profiling and grouping
iv. Evaluation
v. Planning and scheduling[10]

 ISSN: 2454-4248
93 – 97

3. Literature survey

Alom, et al., (2018) [6] trackes the students of Australia beginning from their primary school year 1 to completion of their high school successfully and further more it too tracks their admissions into the higher educational universities or institutions. The criteria on which they are calculating the successive rates of the students is the gender of the students. In this paper data mining software Wilson calculator is used which is a practical meta-analysis effect size calculator and Orange so as to analyze data. Orange for given data sets provide the predictive modeling and the visualizations solutions.

Tavares, R., et al., (2017) [7] The prime focus of this paper is to enhance the digital educational resources in primary school for science education by implementing on the data mining techniques. There is a lot of impact on the students for the self regulated learning after adopting the learning approach. The analyses of the students behaviour is done after getting the particular help and the recommendations.

Zhang, W., et al., (2018) [8] This paper provides the guidance over the techniques of data mining and the related applications in the online educational scenario. The knowledge extraction is done with the help of educational data mining from raw data.

Asif, R., et al., (2017) [9] Studies the performance related to the education of the students. The data taken of the students is focused on two aspects. Firstly, achievement of the student is predicted at the completion of the four year study programme. Then predictions are combined with the progressions. The outcomes is the generation of two groups of students, low and high grade achieving students. By this teachers get to support the students at the low level by giving them activities and the task and to the high level students more opportunities are given.

Bakhshinategh, B.,et al., (2018) [10] In this paper different types of existing Educational data mining tasks and applications have been listed with their categorization on the basis of their purpose. A comparative study is done over the existing surveys related to educational data mining and all the task are reported in a taxonomy.

Manjula, M. (2018) [11] Implements K-means clustering algorithm and uses the weka tool for analyzing the educational data set. With weka the accuracy level improves and the results ontained are the perfect graphs. The data taken has preprocessed to clear the null values, remove the unwanted memory space and the unwanted data. Data mining two methods have been used such as classification and clustering.

4. Implementation

Software used

Python has been used in the implementation. It is the most practiced programming language as it is a open source software. It was created by Guido van Rossum in 1991. It works on different platforms such as Windows, Mac, Linux, etc. It has a very simple syntax familiar to the English language. It makes use of whitespace indentation, instead of curly brackets or keywords, to delimit blocks. An increase in indentation comes after certain statements; a decrease in indentation signifies the end of the current block. Python 3.7 has been used while implementing.

Methodology

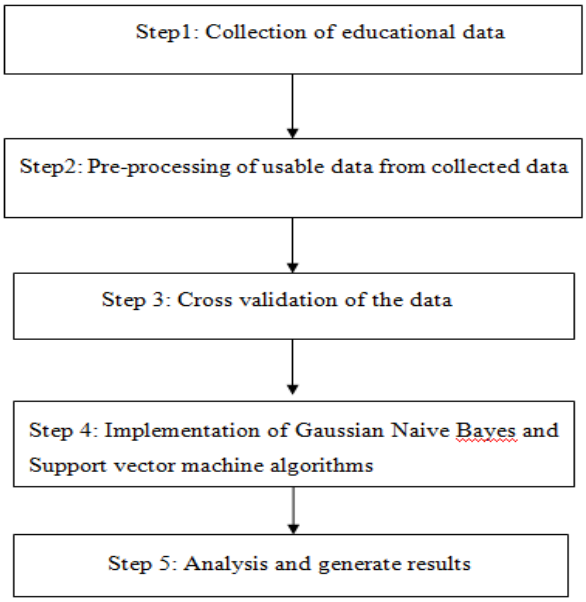


Figure 1. Process flow diagram

Step 1: Collection of educational data: Firstly, we search the dataset in educational domain from the repository. The data is collected from kaggle with these following attributes school, sex, age, address, famsize, Pstatus, Medu - mother's education, Fedu - father's education, Mjob - mother's job, Fjob - father's job, reason - reason to choose this school, guardian, traveltime - home to school travel time , studytime, failures, schoolsupport, family support, absences and activities.

Step 2: Pre-processing of usable data from collected data: The main purpose is to capture the useful and non-trivial data from the unstructured text data. In other words, it is considered as a renovation of raw material (Information) in a comprehensible format. The raw data have many variables which add the noise to distort the accuracy[14].

 ISSN: 2454-4248
93 – 97

Step 3: Cross validation of the data: It splits the dataset into two pieces, so that the model can be trained and tested on different data. It uses the data effectively.

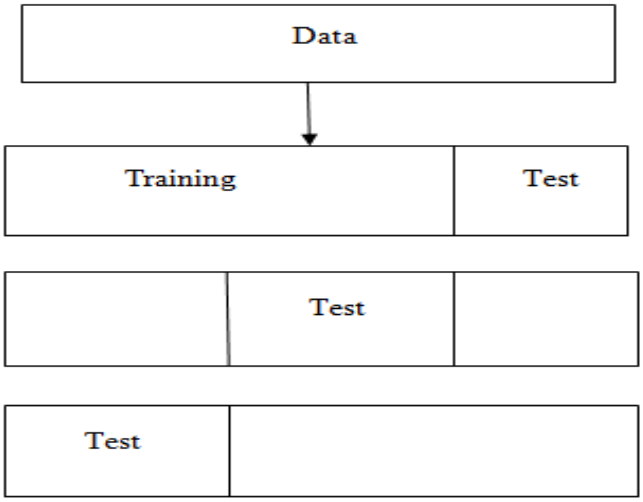


Fig 1: Cross validation

Step 4: Implementation of Gaussian Naive Bayes and Support vector machine algorithms

Gaussian Naive Bayes: The Naive Bayes Classifier technique is based on the Bayesian theorem. Bayes theorem is described as the probability of an event based on the past knowledge of the conditions that are related to the event.

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

This algorithm is used when the dimensionality of the inputs is high. The principle in Naïve bayes is that the pair of features that are being classified are independent of each other.

Support vector machine: Support Vector Machine (SVM) is a classifier that performs classifies by constructing hyperplanes in a multidimensional space that divides the cases of different attributes. Hyperplane can be written as the set of points satisfying

$$\vec{w} \cdot \vec{x} - b = 0$$

SVM implements both regression and classification methods. An SVM model plots the points in space, maps them so the points of the different groups are separated by a wide gap. New points are then mapped into that same space according to the group they belong based on which side of the gap they fall.

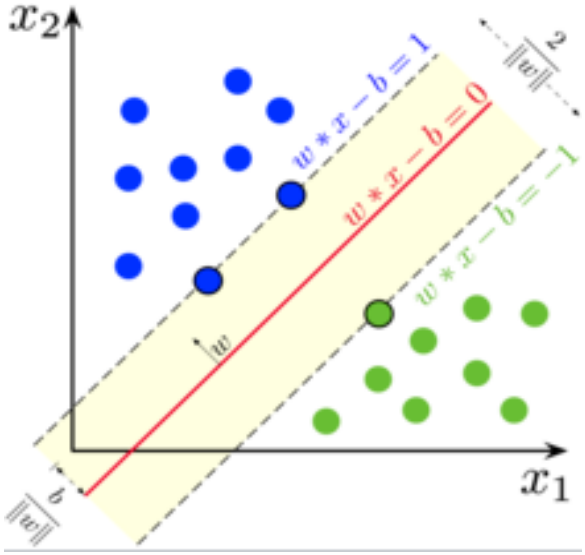
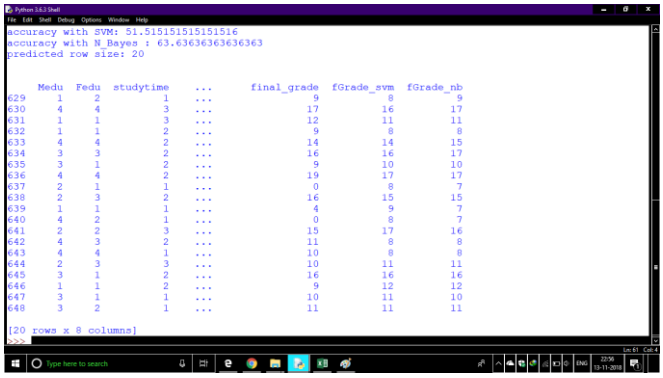
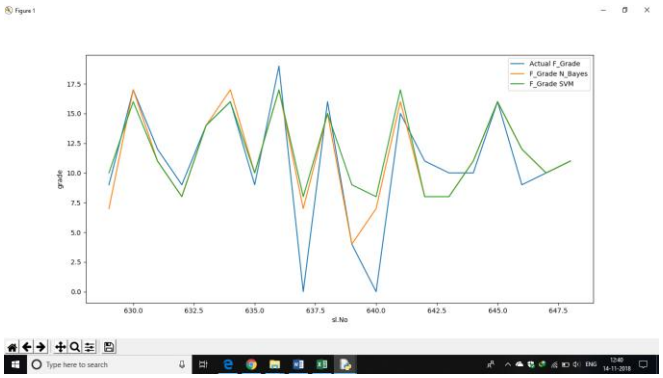


Fig 2: SVM

Step 5: Analysis and generate results: While implementing the educational data mining techniques such as support vector machine and naive bayes in python for finding out the final grade of the students according to some attributes the following are the results obtained. The accuracy for both of the data mining techniques is also calculated as shown in figure 3.



The final grade obtained by using both the educational data mining techniques are plotted with the actual grades in python. Here the Naive bayes outperforms the support vector machine algorithm in accuracy.



 ISSN: 2454-4248
93 – 97

5. Conclusion

With recent rise in the quantity of educational domain data which is being generated dynamically in a complex form needs to be understood. Solution to this problem is the usage of techniques in data mining. It has become the need of hour to implement data mining algorithms for decision making and analysis of the students in every aspect.

References

[1] Sugiyarti, E., Jasmi, K. A., Basiron, B., Huda, M., Shankar, K., & Maseleno, A. (2018). Decision support system of scholarship grantee selection using data mining. International Journal of Pure and Applied Mathematics, 119(15), 2239-2249.

[2] Rodrigues, M. W., Zárate, L. E., & Isotani, S. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics.

[3] Tegegne, A. K., & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. Information Impact: Journal of Information and Knowledge Management, 9(2), 1-15.

[4] Sheshasaayee, A., & Bee, M. N. (2018). E-learning: Mode to Improve the Quality of Educational System. In Smart Computing and Informatics (pp. 559-566). Springer, Singapore.

[5] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. IEEE Access, 5, 15991-16005.

[6] Alom, B. M., & Courtney, M. (2018). Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia.

[7] Tavares, R., Vieira, R., & Pedro, L. (2017, November). A preliminary proposal of a conceptual educational data mining framework for science education: Scientific competences development and self-regulated learning. In Computers in Education (SIIE), 2017 International Symposium on (pp. 1-6). IEEE.

[8] Zhang, W., & Qin, S. (2018, March). A brief analysis of the key technologies and applications of educational data mining on online learning platform. In Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on (pp. 83-86). IEEE.

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

[10] Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, 23(1), 537-553.

[11] Manjula, M. (2018). A Systematic Review on Educational Data Mining. International Journal of Scientific Research in Science, Engineering and Technology, 164 -170.

[12] Shingari, I., & Kumar, D. (2018). Predicting Student Performance Using Classification Data Mining Techniques, International Journal of Computer Sciences and Engineering, 43-48.

[13] Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering, 66, 541-556.

[14] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247-256.

[15] BH, H. M., & Suresh, L. (2018). Data Mining in Higher Education System and the Quality of Faculty Affecting Students Academic Performance: A Systematic Review, International Journal of Innovations & Advancement in Computer Science, 66-70.

[16] Callanan, G. A., Perri, D. F., & Tomkowicz, S. M. (2018). Data mining and automated prediction: A pedagogical primer for classroom discussion. Journal of Education for Business, 93(7), 352-360.

[17] Angeli, C., Howard, S. K., Ma, J., Yang, J., & Kirschner, P. A. (2017). Data mining in educational technology classroom research: Can it make a contribution?. Computers& Education, 113, 226-242.