EDUCATIONAL DATA MINING: A COMPARATIVE STUDY ON PREDICTING STUDENT'S PERFORMANCE USING DIFFERENT ALGORITHMS

A.PAVITHRA
Ph.D Research Scholar
Department of Computer Science
Pollachi College of Arts and Science
Pollachi
amspit25@gmail.com

Dr.B.SELVANANDHINI
Assistant Professor
Department of Computer Science
Pollachi College of Arts and Science
Pollachi
selvanandhini.n@gmail.com

***Abstract*— Education is the acceptance that education develops the possible of each person. Each person is outstanding and promotes the individual academically, significantly and keenly. It is a notable for individual development and to be awareness in life. Educational Data Mining (EDM) is an forthcoming research field on the growth of educational psychology and the learning sciences. For an institutional development educational enactment records plays a main role for the good will. Though there are many features like socio- economic, non- academic, some academic things prompting the student performance in an institution. Data mining methods is the knowledge of bring out the concealed information from the vast raw datasets.**

Finally, it brings a Knowledge discovery and Decision making. There are many classification and prediction algorithms in data mining. In this paper it is absorbed on student performance and prediction of a characteristic that, were the student will be successful or failure. Also it is associated with two sets of algorithms. The datasets are original data of a private institution named ‒Sree Saraswathy Thyagaraja College‖, Pollachi which is situated in rural area. One set of investigation work done with five different algorithms like Naïve Bayes (NB), Rep tree, J48, Decision tree and Multi-Layer Perception (MLP) along with the confirmation of student destination survey. Another set of algorithms are Random Forest (RF), Support Vector Machine (SVM), Neural network (NN) and K- Nearest Neighbour (kNN) with three different factors. This work also displays the different factors which influences student performance. At last we come to a comparative analysis that which prediction algorithm shows us best correctness rate with the help of ROC values.

Keywords— *Educational Data Mining (EDM), Multi-Layer Perception (MLP), Classification, Prediction, Knowledge Discovery, Decision Making, Random Forest, Support Vector machine, Neural network*.

I.INTRODUCTION

Data mining is the concept of finding the patterns from huge raw datasets. The data mining provides many algorithms like classification algorithm, association algorithms, machine learning algorithms and rule based algorithms so on. The datasets are getting processed with the help of all these algorithms and gives us a meaningful mining result. That result is fine tuned for some decision making. Thus data mining is fully concentrates on ‒Knowledge Discovery and Decision Making‖ (KDD). Educational Data Mining is based on education data which is to be processed for bringing up meaningful results that is decision making. EDM plays a big role in student performance prediction and analysis. It is an upcoming research area in education. WEKA tool is used in this research work for better understanding. It is very user friendly and provides many visualization effects on datasets.

There are many prediction algorithms in data mining. In this study it is concentrated on student performance prediction with different prediction algorithms. Totally 50 students data of Bachelor of Computer Science is taken from the college and they are compared with the results of two different sets of algorithms. All are formatted to WEKA tool accepted ARFF format. Different fields of datasets are student gender, place, father occupation, annual income, caste and finally with their higher secondary marks. They are analysed deeply and formatted to user need type. First and foremost, the dataset is analysed to which attribute plays a major role and obtained by Info gain attribute selection. As per the first result Naïve Bayes shows the best result. At the same time on the other end RF, SVM, NN shows us the best accuracy predictions. The importance of this comparative study is to find the best possible changes we need to implement for the students betterment, if they are in failure stage. Finally, the knowledge flow is drawn and all data are visualized with visualize errors.

I. ABOUT STUDENT DATASET

The dataset consists of 50 students details and they are belong to rural area institution. On working with this dataset it has been know that their parents are mostly working as an agriculturist, belongs to annual income of 50,000 and below. The students need to concentrate little better in their studies according to the observation. The below table explains the different attributes of the dataset along their description and the different values it have.

Table1.1a Student Dataset

Parameter	Description	Values
SEX	Gender of the student	{M,F}
PLACE	Place where the students are from divided into two values	{VILLAGE, TOWN}
OCCUPATION	Occupation of the student father	{AGRICULTURE, BUSINESS, OTHERS} OTHERS: worker, government employee, police, teacher, weaving, water man, sales man
INCOME	Annual income of the student	A: ABOVE 50000 B: 50000 C: BELOW 50000
CATEGORY	Different category of the student	OBC: BC, OC UNRESERVED: MBC, BCM SC: SC ST: ST
HSLC	Student higher secondary mark	A: ABOVE 1000 B: 800 - 1000 C: 700 - 800 D: 600 - 700 E: BELOW 600
SEM	Semester percentage	A: ABOVE 70% B: 50% – 69% C: 49% - 40% D: FAILURE

1.2. PROPOSED SYSTEM FOCUSES ON

The proposed system focuses on student performance prediction with different sets of algorithms. There may be many algorithms used on the prediction on academic performance but here we have selected the results of two sets of algorithms. In both the results the authors used the original datasets. A Review on Predicting Student's Performance Using Data Mining Techniques [1] explains on student performance prediction along with the factors which influence student performance. Educational data mining: prediction of students' academic performance using machine learning algorithms [2] shows 75% accuracy with student's midterm marks as an influencing attribute.

There are many factors which influence the student performance and day by day there are new upcoming challenges which affects the performance, especially this pandemic period, learning management system is completely changed. So, it will be not stick on to any single factors that affects the performance of students. The proposed system mainly focuses on best student prediction algorithm and here we have compared the different ROC values. The algorithm which has highest average ROC value is taken for the final prediction of the student dataset attribute namely, place able/ not place able or successful/failure.

On analysis of this student dataset it has been find out that maximum students are from rural areas, they have only a minimum income of 50000 and below. Most of the students family are belongs to agricultural background and their performance on higher secondary are little poor. And also it is noted that on reviewing many papers it has been cleared that many factors are influencing student performance and we cannot stick into any single factors. This study made a comparative view of best accuracy algorithm.

II.FRAMEWORK OF PROPOSED SYSTEM

The 50 students dataset is taken from the database and purified that is, pre-processed with Infogain in weka explorer interface. But in prediction on student performance we cannot stick on to any single factors which influence the student performance, there may be many time being and students places decide the factor which affects their performance.

In article [6] explains the prediction of student performance in a different manner when compared to others. They have explained the different factors which affect student performance. One is parameter which affects student performance, secondly data mining algorithms that we used to predict and finally, data mining tool. The methodology used in this research work is explained by a hybrid algorithm. [1]

Here in this study the dataset is run on different algorithms which was already reviewed and discussed why it is best for predicting students' performance. So, we are using RF, SVM, NN and NB algorithms to the dataset for predicting and the ROC value and confusion matrix decides which algorithm is best on prediction. There are some factors need to be concentrated as deciding factors. They are explained below.

II.I.1 CORRECTLY CLASSIFIED ACCURACY

It shows the correctness ratio of investigation that is correctly classified.

II.I.2 INCORRECTLY CLASSIFIED ACCURACY

It specifies the accuracy percentage of test that is incorrectly classified.

II.I.3 MEAN ABSOLUTE ERROR

It prove the number of errors to study algorithm classification accuracy.

II.I.4 TIME

It illustrates how much time necessary to build a model.

II.I.5 ROC AREA

Receiver Operating Characteristic shows test performance director for classifications accuracy. Here in this research work ROC plays a major role.

II.I.6 CONFUSION MATRIX

The confusion matrix shows the current situation in the dataset and the number of correct/incorrect predictions of the model.

III.LITERATURE REVIEW

The proposed system focuses on student prediction performance with the help of their higher secondary marks and with their semester wise improvements. Here the literature review is made on different student a mining paper that gives us more information regarding students influencing factors. The influencing factors may vary to the students' locality and it mainly affects their psychology. The student prediction is done by many different data mining algorithms and it shows different us different results.

In various studies on EDM, e-learning systems have been successfully analysed (Laraet al., 2014). Some studies have also classified educational data (Chakraborty et al., 2016), while some have tried to predict student performance (Fernandes et al., 2019). Asif et al. (2017) focused on two aspects of the performance of undergraduate students using DM methods. The first aspect is to predict the academic achievements of students at the end of a four-year study program. The second one is to examine the development of students and combine them with predictive results. He divided the students into two parts as low achievement and high achievement groups. He have found that it is important for the educators to focus on a small number of courses indicating particularly good or poor performance in order to offer timely warnings, support underperforming students and offer advice and opportunities to high-performing students. Cruz-Jesus et al. (2020) predicted student academic performance with 16 demographics such as age, gender, class attendance, internet access, computer possession, and the number of courses taken. Random forest, logistic regression, k-nearest neighbours and support vector machines, which are among the machine learning methods, were able to predict students' performance with accuracy ranging from 50 to 81%. Fernandes et al. (2019) developed a model with the demographic characteristics of the students and the achievement grades obtained from the in-term activities. In that study, students' academic achievement was predicted with classification models based on Gradient Boosting Machine (GBM). The results showed that the best qualities for estimating achievement scores were the previous year's achievement scores and unattendance. The authors found that demographic characteristics such as neighbourhood, school and age information were also potential indicators of success or failure. In addition, he argued that this model could guide the development of new policies to prevent failure. Similarly, by using the student data requested during registration and environmental factors, Hoffait and Schyns (2017) determined the students with the potential to fail. He found that students with potential difficulties could be classified more precisely by using DM methods. Moreover, their approach makes it possible to rank the students by levels of risk. Rebai et al. (2020) proposed a machine learning-based model to identify the key factors affecting academic performance of schools and to determine the relationship between these factors. He concluded that the regression trees showed that the most important factors associated with higher performance were school size, competition, class size, parental pressure, and gender proportions. In addition, according to the random forest algorithm results, the school size and the percentage of girls had a powerful impact on the predictive accuracy of the model.

IV. ALGORITHMS USED IN PROPOSED SYSTEM

In the proposed system it is concentrated mainly on Random forest, Support Vector Machine, Neural network (NN) and K- Nearest Neighbour (kNN) and Naïve Bayes algorithms. Prediction on student dataset is more cautious because of the influencing factors. There are many factors influence the student performance but here it has been identified at the end with semester marks will plays a major role on the student performance prediction. Herewith we will discuss the different algorithms which are used in this comparative work.

A) Naïve Bayes(NB)

Naïve Bayes is one of the algorithms that works as a probabilistic classifier of all abilities enclosed in data model discretely and then groups' data problems. Running the algorithms Naïve Bayes we inspect the classifier output with some statistics set up output by using 10 cross validation to make an estimate of all instance of the dataset. In most of the research article NB is mainly used in student prediction.

B) Random Forest (RF)

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned

C) Support Vector Machine(SVM)

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

D) Neural Network (NN)

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to manual identification by human experts.

E) K Nearest Neighbor (kNN)

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

V. RUN ON DIFFERENT ALGORITHMS

Run on Naïve Bayes

In run 1 the dataset is run on Naïve Bayes algorithm which shows the following values. The average ROC

Table5.1 Run on NB

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Time taken	ROC Area
72%	**28%**	**0.163**	**0 sec**	**0.944**

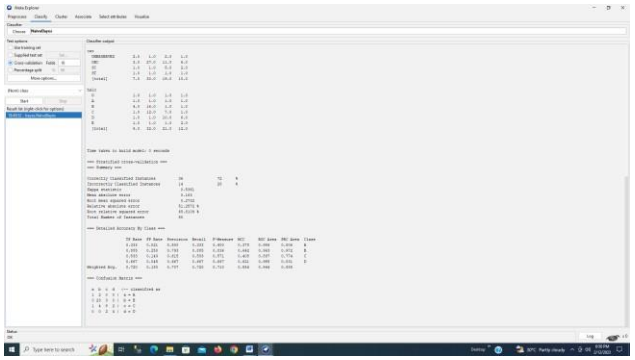

Fig 5.1. NB

Run on Support Vector Machine

In run 2 the dataset is run on Support Vector Machine algorithm which shows the following values. The average ROC value of SVM is 0.912.

Table5.2 Run on SVM

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Time taken	ROC Area
82%	**18%**	**0.266**	**0.06 sec**	**0.912**

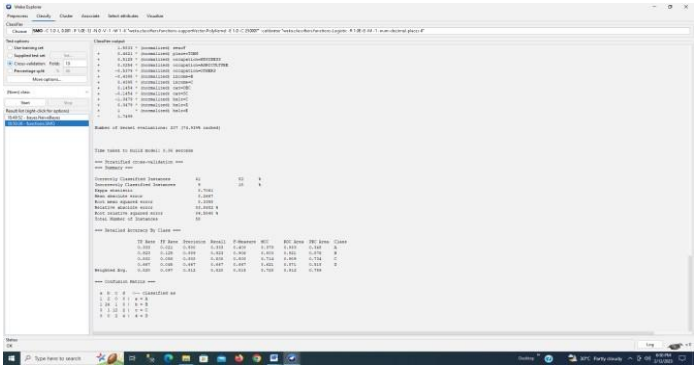

Fig 5.2. SVM

Run on Random Forest

In run 2 the dataset is run on Random Forest algorithm which shows the following values. The average ROC value of RF is 0.998.

Table5.3 Run on RF

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Time taken	ROC Area
94%	**6%**	**0.088**	**0.03 sec**	**0.998**

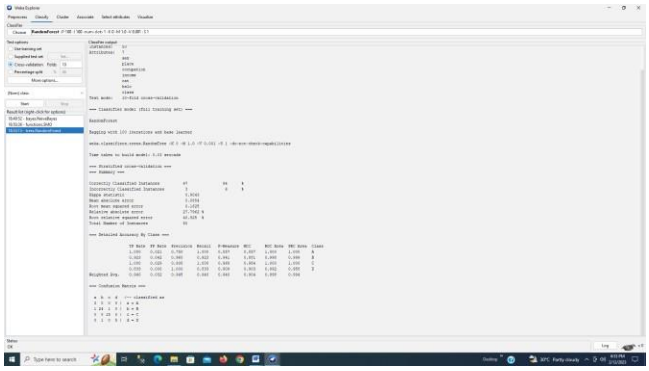

Fig 5.3. RF

Run on k- Nearest Neighbour

In run 4 the dataset is run on kNN algorithm which shows the following values. The average ROC value of kNN is 0.965.

Table5.4 Run on kNN

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Time taken	ROC Area
96%	**4%**	**0.061**	**0 sec**	**0.965**

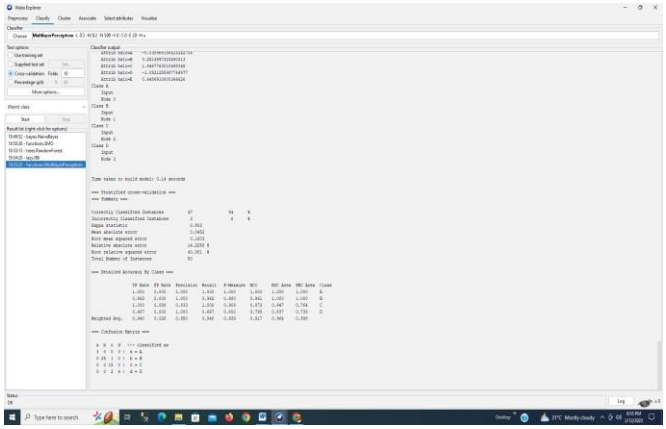

Fig 5.4. kNN

Run on Neural Network

In run 5 the dataset is run on Neural network algorithm which shows the following values. The average ROC value of NN is 0.964

Table5.5 Run on NN

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Time taken	ROC Area
94%	**6%**	**0.0452**	**0.14 sec**	**0.964**

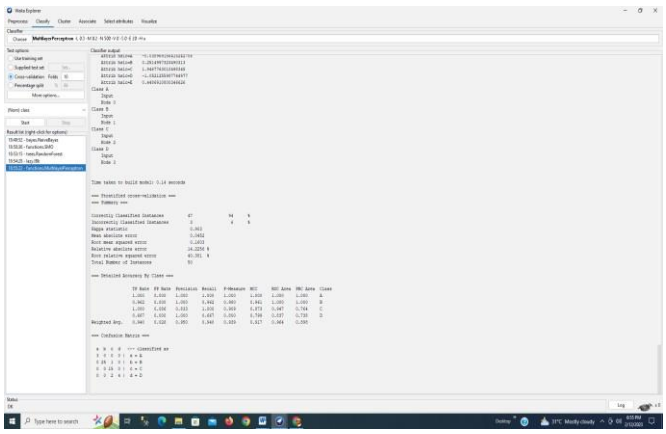

Fig 5.5. NN

VI. COMPARING THE ROC VALUES

The comparative study has been done on different algorithms. Now it needs to calculate the average ROC value of the each algorithm that has been explained above. Now it is need to know which algorithm gives best ROC value. It is clear that Random Forest gives us highest ROC value. So we can come to a conclusion that Random Forest performed good in comparing to other four algorithms. Below in a table the average value is mentioned.

Table 6.1 Average ROC values

ALGORITHM	AVERAGE ROC VALUE
RF	**0.998***
kNN	**0.965**
NN	**0.964**
NB	**0.944**
SVM	**0.912**

In the above table the average value is calculated and it has been showed that the Naïve Bayes algorithm has highest average ROC value and the research work will be carry on next with Naïve Bayes algorithm.

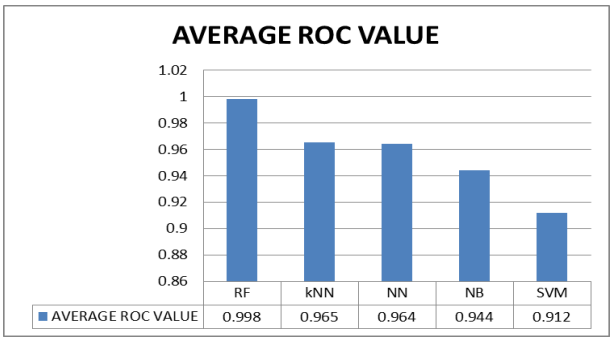

Fig 6.1. Graph for average ROC values

VII. KNOWLEDGE FLOW

The WEKA provides us the knowledge flow of the datasets. There we have different tools for different process. We just click and drag to the working area. Here they have different icons on different usages. The arff icon which shows that we have included the data in arff format. It includes the different runs in a flow. Also this makes the user to understand it very clearly.

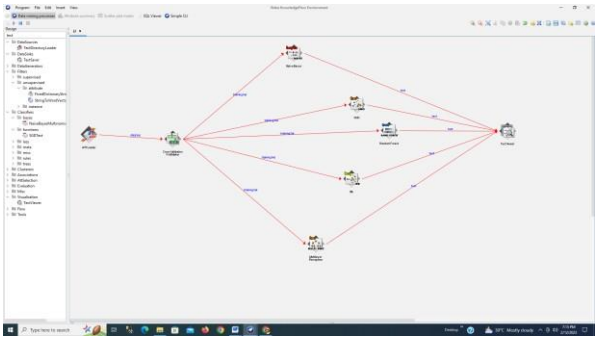

Fig.7.1 Knowledge Flow

VIII. RESULTS AND DISCUSSIONS

The dataset is run with different five algorithms which gives us the best prediction values. Each time the dataset is run we can see the results with different ROC values. Finally we come to know that Random Forest gives us the best result when compared to other four algorithms. Next kNN, NN, NB gives us the possible accuracy values. At last we come to know SVM is having least ROC values.

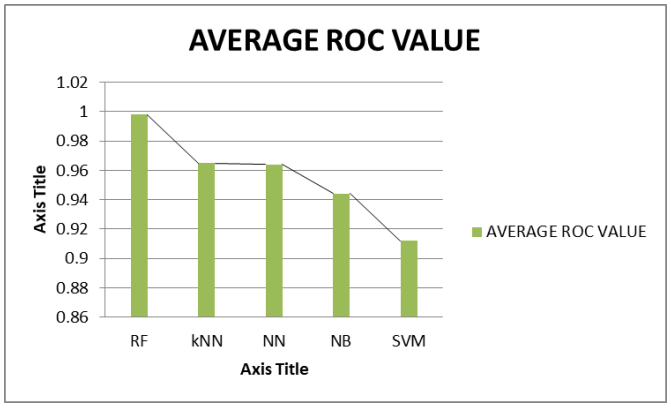

Figure 8.1 Average Roc Values Of Different Algorithms

IX. CONCLUSION AND FUTURE WORK

I. CONCLUSION

Educational Data Mining (EDM) is an upcoming field in research development. This study proposes a new model based on machine learning algorithms to predict the final exam grades of undergraduate students. The performances of the Random Forests, Support vector machines, Neural network, Naïve Bayes, and k-nearest neighbour algorithms, which are among the machine learning algorithms, were calculated and compared to predict the final exam grades of the students. This study focused on two parameters. The first parameter was the prediction of academic performance based on previous achievement grades. The second one was the comparison of performance indicators of machine learning algorithms.

This research work is primarily dedicated on examining the prediction accurateness of the academic performance of the students using different. The final conclusion comes under Random Forest algorithm as it gives best ROC value. Secondly, it helps the institution to find out the students' performance in future and to know the slow learners who needs extra care on them in academic. It is also concluded that many factors will influence the student performance and it may differ to different locality of students.

II. FUTURE WORK

Future research can be conducted by including other parameters as input variables and adding other machine-learning algorithms to the modeling process. In addition, it is necessary to harness the effectiveness of DM methods to investigate students' learning behaviors, address their problems, optimize the educational environment, and enable data-driven decision-making. Future work of this study will be go on with the different influencing factors in detail because the locality of the students will affect their mindsets and development towards academic. Also, may include the psychology of the student.

III. ACKNOWLEDGEMENTS

This research paper is truthfully involvement and guidance who belong to the list of authors. The authors pay morality to Pollachi College of Arts and Science, Pollachi for continuous encouragement, guidance and support during their work.

References

[1] Amirah Mohamed Shahiri, Wahidah Husain, Nuraini Abdul Rashid. A Review on Predicting Student's Performance Using Data Mining Techniques. Elsevier, Procedia Computer Science, Volume 72, 2015. Pages 414-422.

[2]Educational data mining: prediction of students' academic performance using machine learning algorithms Mustafa Yağcı* Yağcı *Smart Learning Environments (2022) 9:11* https://doi.org/10.1186/s40561-022-00192-z

[3] Amjad Abu Saa. Educational Data Mining & Students' Performance Prediction. International Journal of Advanced Computer Science and Applications, Volume 7, No. 5, 2016.

[4] Brijesh Kumar Bhardwaj, Saurabh Pal. Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security, Volume 9, N0. 4, April 2011.

[5] Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. Bulletin of Education and Research, 40(3), 157–164.

[11] Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE

Access, 8, 203827– 203844. https:// doi. org/ 10. 1109/ access. 2020. 30365 72

[6] Arias Ortiz, E., & Dehon, C. (2013). Roads to success in the Belgian French Community's higher education system:predictors of dropout and degree completion at the Université

[7] Libre de Bruxelles. Research in Higher Education, 54(6), 693–723. https:// doi. org/ 10. 1007/ s11162- 013- 9290-y

[8] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzingundergraduate students'performance using educational data mining. Computers and Education, 113, 177– 194. https:// doi. org/ 10. 1016/j. compe du. 2017. 05. 007

[9] Aydemir, B. (2017). Predicting academic success of vocational high school students using data mining methods graduate.[Unpublished master' hesis].Pamukkale University Institute of Science.

[11] Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. Croatian Operational Research Review, 8(2), 443–461. https:// doi. org/ 10. 17535/ crorr. 2017. 0028

[12] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. Learning analytics (pp. 61–75). Springer.

[13] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.

[14] Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158(August), 103999. https:// doi. org/ 10. 1016/j. compe du. 2020. 103999

[15] Burgos, C., Campanario, M. L., De, 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 and Electrical Engineering, 66(2018),541–556. https:// doi. org/ 10. 1016/j. compe leceng. 2017. 03. 005

[16] Capuano, N., & Toti, D. (2019). Experimentation of a smart learning system for law based on knowledge discovery and cognitive computing. Computers in Human Behavior, 92, 459–467. https:// doi. org/ 10. 1016/j. chb. 2018. 03. 034

[17] Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: A case study in higher education using learning analytics approach. Interactive Learning Environments, 24(1), 49–67. https:// doi. org/ 10.1080/ 10494 820. 2013. 817441

[18] Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. In Proceedings of 2nd IEEE international conference on engineering and technology, ICETECH 2016, March (pp. 431–436). https:// doi. org/ 10. 1109/ ICETE CH. 2016. 75692 90

[19] Parneet Kaur, Manpreet Singh,Gurpreet Singh Josan. Classification and prediction based data mining algorithms to predict slow learners in education sector, Procedia Computer Science 57 (2015) 500 – 508.

[20] K.Sumathi, S.Kannan, K.Nagarajan. Data Mining: Analysis of student database using Classification Techniques, International Journal of Computer Applications (0975 – 8887) Volume 141 – No.8, May 2016.

[21] Jai Ruby, Dr. K.David. Prediction Accuracy of Academic Performance of Students using Different Datasets with High Influencing Factors, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016.

[22] M.Durairaj, C.Vijitha. Educational Data mining for Prediction of Student Performance Using Clustering Algorithms, International journal of Computer Science and Information Technologies, Volume 5 (4), 2014, 5987-5991.

[23] A.Pavithra,S.Dhanaraj. Comparative Study Of Effective Performance Of Association Rule Mining In Different Databases, in CIIT- Data Mining and Knowledge Discovery Volume 10, No.4 on May 2018