Educational Data Mining: Student Performance Prediction in Academic

Y. K. Salal, S. M. Abdullaev, Mukesh Kumar

Abstract: ***At present data mining techniques become very popular among the data analyst. It became an effective tool for finding the uncovered information from a big database. Due to this feature data mining are adopted by many areas like education, telecommunication, retail management etc to resolve their business problems. In this paper, for building classification models for 'student performance' dataset consisting of 649 different instances with 33 different attributes implement algorithms like NaiveBayes, Decision Tree (J48), RandomForest, RandomTree, REPTree, JRip, OneR, SimpleLogistic and ZeroR. After implementing these algorithms on student performance dataset, we evaluate and compare the implementation result for better accuracy of prediction. The result of this study is extremely significant and hence provides a greater insight for evaluating the student performance and underlines the significance of data mining in education. It also shows that how students attributes affect the student performance.***

Index Terms: ***Naive Bayes, k-nearest neighbor, Logistic regression, J4.8, RandomForest.***

I. INTRODUCTION

Educational system is one of the imperative parts for development of any country. So it should be taken very seriously from its start. Most of the developed countries have their own education system and evaluation criteria. Now a day's education is not limited to only the classroom teaching but it goes away from that like Online Education System, MOOC course, Intelligent tutorial system, Web-based education system, Project based learning, Seminar, workshops etc.

But all these systems are not successful if they are not evaluated with accuracy. So for making any education system to success, a well-defined evaluation system is maintained. Every educational institution generates lots of data related to the registered student and if that data is not analysis properly then all afford is going to be wasted and no future use of data happen. This institutional data is related to the student admission, student family data, student result etc. Every educational institution applies some assessment criteria to evaluate their students.

In modern education, we have lots of assessment tools which are used to observe the performance of the student in their study. Data Mining is one of the best computers based intelligent tool used to check progress of the students in their study. At present the scope of Data Mining is not limited to only Education but it all most covers the entire field where data are used. Some of the important application areas of data mining are retail management, marketing management, Banking sector, Telecommunication, Hospitality Management, Hospital Management, Production Management etc.

This entire organization takes the benefits of using data mining to increase their revenue and future growth. Actually, its technology are analyses the historical data of any organization with some inbuilt algorithm and then find the hidden information from that which is normally not identified. So we can say that with the help of data mining techniques, hidden information can get form the data warehouse of any organization.

The general purpose of this paper is to present comprehensive analysis of student performance dataset though building a classification model. For this project, we use WEKA data mining toolkit. It provides a rich collection of different data mining algorithms for classifying and analyzing data. To ensure the superlative accuracy for prediction, all development and testing of classifying models will follow the Data mining process

II. DATA EXPLORATION OF STUDENT PERFORMANCE DATA SET

This particular student's performance dataset are collected from two secondary school of Portuguese (Gabriel Pereira (GP) and Mousinho da Silveira (MS)). The dataset includes student's attributes like academic grades, demographic attributes, social attributes and school related attributes. School reports and questionnaires are used for collecting data from the students. Detail for this Dataset is shown in Table 1.

III. DATA PRE-PROCESSING FOR STUDENT PERFORMANCE DATA SET

For the analysis purpose, we are using WEKA tool kit which freely available for use with lots of supporting algorithms for classification, clustering and association rule mining. After installing the WEKA, we load our dataset using pre-process feature which is shown below in Figure 1.

Revised Manuscript Received on December 22, 2018.

Mr. Y. K Salal, Department of System Programming, South Ural State University (National Research University) (Chelyabinsk, Russian Federation).

Prof. S. M. Abdullaev, Doctor of Geographical Sciences, Professor, Chair System Programming, South Ural State University (National Research University) (Chelyabinsk, Russian Federation)

Mr. Mukesh Kumar, Assistant Professor, Chitkara University, Himachal Pradesh.

Table 1: Description of the student attribute taken into build the dataset

S.No	Attribute	Description	Variable	Possible Values
1	school	School name of student	binary	GP, MS
2	sex	Sex of the student	binary	M, F
3	age	Age of the student	numeric	15-22
4	address	Address of the student	binary	U, R
5	famsize	Family size of the student	binary	LE3, GT3
6	Pstatus	parent's cohabit	binary	T, A
7	Medu	Mother Qualification	numeric	0, 1, 2, 3, 4
8	Fedu	Father Qualification	numeric	0, 1, 2, 3, 4
9	Mjob	Job type of the Mother	nominal	Teacher, health, home, services, other
10	Fjob	Job type of the Father	nominal	Teacher, health, home, services, other
11	reason	Reason to desire this school	nominal	Home, reputation, course, other
12	guardian	Guardian of the student	nominal	Mother, father, other
13	traveltime	Home to school travel time	numeric	1, 2, 3, 4
14	studytime	Weekly study time	numeric	1, 2, 3, 4
15	failures	No's of past class failures	numeric	n if 1<=n<3, else 4
16	schoolsup	Extra study support	binary	Yes, no
17	famsup	family study support	binary	Yes, no
18	paid	Extra paid classes	binary	Yes, no
19	activities	Extra-curricular activities	binary	Yes, no
20	nursery	Attended nursery school	binary	Yes, no
21	higher	Wants higher study	binary	Yes, no
22	internet	Internet at home	binary	Yes, no
23	romantic	Relationship	binary	Yes, no
24	famrel	Family relation	numeric	1, 2, 3, 4, 5
25	freetime	Free time after school	numeric	1, 2, 3, 4, 5
26	goout	Going out with friends	numeric	1, 2, 3, 4, 5
27	Dalc	Workday alcohol	numeric	1, 2, 3, 4, 5
28	Walc	weekend alcohol	numeric	1, 2, 3, 4, 5
29	health	Current health status	numeric	1, 2, 3, 4, 5
30	absences	School absenteeism	numeric	0 to 93
31	G1	First term grade	numeric	0 to 20
32	G2	Second term grade	numeric	0 to 20
33	G3	Final grade	numeric	0 to 20 (Output target)

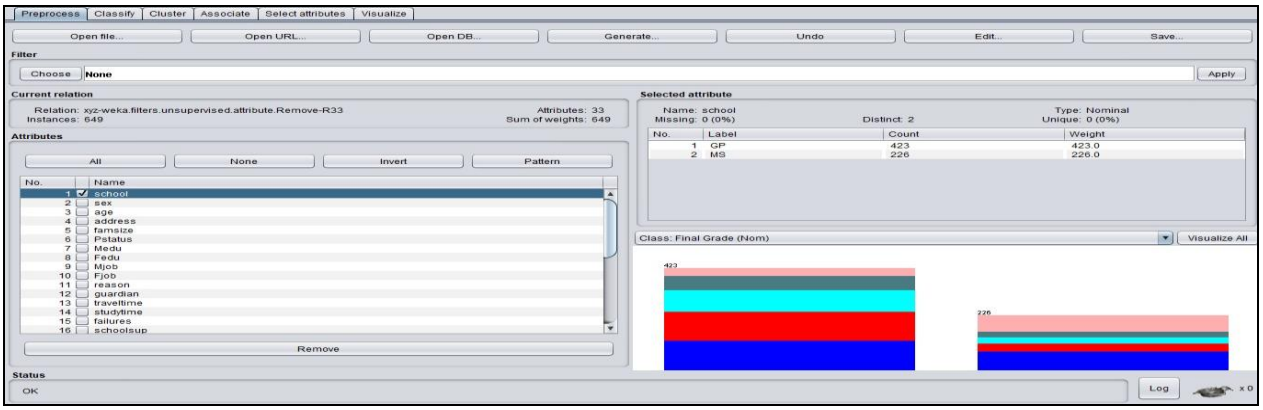

Figure 1. Student Performance Data Set on WEKA

Initially, the target output class ranges from 0 to 20 and there are 21 clusters. This is an unreasonable setting for the classification task, because it makes it extremely difficult to classify remember that the number of instances, we have is only 649. As a result, I have mapped a group of clusters to a few clusters [A, B, C, D, and F] as indicated in Table 2.

Table 2: Change the number of target class clusters

Range of initial class	New cluster number
16-20	A class
14-15	B Class
12-13	C Class
10-11	D Class
<=9	F Class

IV . CLASSIFICATION MODELS FOR STUDENT PERFORMANCE DATASET

To find best classifier's which generalize the data with more accurately, we implement different classification algorithm in WEKA in student performance dataset by choosing different parameters of the algorithms which most efficiently analysis the dataset and increase their generalized accuracy. Please note that, all the implemented classification algorithm are tested using 10-fold cross validation to calculate algorithms accuracy.

Standard classifier's Model Building using WEKA: Numerous classification algorithms were selected and implemented to the student's performance dataset. The different classification algorithms which are implemented on WEKA are NaiveBayes, Decision Tree (J48), RandomForest, RandomTree, REPTree, JRip, OneR, SimpleLogistic and ZeroR. Please note that all the classification algorithms are implemented with their default parameters. Please note that, all the implemented classification algorithm are tested using 10-fold cross validation to calculate algorithms accuracy. To check the baseline accuracy for the student's performance dataset, we are implementing ZeroR classification algorithm with overall accuracy of 30.97 %. In Table 3, different classification algorithm with finest accuracy is underline.

Table 3: Classification algorithm with their accuracy on default parameter

Classification Algorithm taken for implementation	Classification Algorithm Accuracy with all attributes
NaiveBayes	68.2589 %
Decision Tree (J48)	67.7966 %
RandomTree	53.4669 %
REPTree	75.1926 %
JRip	70.7242 %
OneR	76.7334 %
SimpleLogistic	71.3405 %
ZeroR	30.9707 % (Baseline accuracy)

Attribute Selection using Ranker method in WEKA: In this section, we are selecting the best attribute with Ranker Search Method from Student's performance dataset with mostly effect the prediction accuracy of the classification algorithm.

We are tried with three different attribute's evaluator method namely CorrelationAttributeEval, GainRatioAttributeEvall and InfoGainAttributeEval with Ranker Search Method. After finding the result of the implementing attribute evaluator algorithm, we are selected ten different attribute which mostly affect our prediction result.

Table 4. Shows the result of the entire algorithm used on student dataset. The ten selected attributes are G2, G1, failures, higher, Medu, school, studytime, Fedu with mostly attribute evaluator algorithm found most effective for prediction.

Table 4: Attribute Selection with the help of Ranker Search Method

Attribute Evaluator	Attribute in decreasing order of their Rank
CorrelationAttributeEval	32,31,15,21,7,1,14,8,27,13,4,16,28,30,22,2,25,19,3,11,9,29,18,26,20,24,10,5,23,12,17,6
GainRatioAttributeEvall	32,31,15,21,1,7,16,14,18,8,9,4,22,11,10,12,2,19,20,23,5,17,6,25,3,24,13,26,28,29,30,27
InfoGainAttributeEval	32,31,15,21,1,9,7,14,11,10,8,16,4,22,12,2,19,18,20,23,17,5,6,29,13,3,25,26,28,30,24,27

Superlative Classifier's Model Building using WEKA: To find best classifier's which generalize the data with more accurately, we implement different classification algorithm in WEKA on student's performance dataset by choosing different parameters of the algorithms which most efficiently analysis the dataset and increase their generalized accuracy.

The different classification algorithms which are implemented on WEKA are NaiveBayes, Decision Tree (J48), RandomForest, RandomTree, REPTree, JRip, OneR, SimpleLogistic and ZeroR. Please note that, all the implemented classification algorithm are tested using 10-fold cross validation to calculate algorithms accuracy with different parameter selection.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 4: Classification algorithm with their accuracy with parameter change

Classification Algorithm taken for implementation	Classification Algorithm Accuracy with eight attributes and with different parameter selection
NaiveBayes	73.1895 %
Decision Tree	76.2712 %
RandomTree	67.9507 %
REPTree	76.7334 %
JRip	74.114 %
OneR	76.7334 %
SimpleLogistic	73.6518 %
ZeroR	30.9707% (Baseline Accuracy)

NaiveBayes Classification Algorithm: When we experimented with all 32 attributes, NaiveBayes Classification Algorithm showed 68.2589 % accuracy. With eight most significant attributes, its accuracy increased to 73.1895 %. To improve the prediction accuracy, we set useSupervisedDiscretization parameter to true and useKernelEstimator parameter to false. The accuracy of NaiveBayes Classification Algorithm increased up to 4.9306 %.

Decision Tree (J48) Classification Algorithm: When we experimented with all 32 attributes, Decision Tree (J48) Classification Algorithm showed 67.7966 % accuracy. With eight most significant attributes, its accuracy increased to 76.5794 %. To improve the prediction accuracy, we set binarySplits is set to true; minNumObj is set to six, numFolds to five. The accuracy of Decision Tree Classification Algorithm increased up to 8.7828 %.

RandomTree Classification Algorithm: When we experimented with all 32 attributes, RandomTree Classification Algorithm showed 53.4669 % accuracy. But with eight most significant attributes, its accuracy increased to 67.9507 %. To improve the prediction accuracy, we set batchSize is set to 200; and seed value is set to 2. The accuracy of RandomTree Classification Algorithm increased up to 14.4838 %.

REPTree Classification Algorithm: When we experimented with all 32 attributes, REPTree Classification Algorithm showed 75.1926 % accuracy. But with eight most significant attributes, its accuracy increased to 76.7334 %. To improve the prediction accuracy, we set batchSize is set to 300; and seed value is set to 2, numFolds value is set to 5, numDecimalPlaces value is set to 4, and minNum is set to 3.0. The accuracy of REPTree Classification Algorithm increased up to 1.5408 %.

JRip Classification Algorithm: When we experimented with all 32 attributes, JRip Classification Algorithm showed 70.7242 % accuracy. But with eight most significant attributes, its accuracy increased to 74.114 %. To improve the prediction accuracy, we set folds is set to 5; seed value is set to 5, numDecimalPlaces value is set to 4. The accuracy of JRip Classification Algorithm increased up to 3.3898 %.

OneR Classification Algorithm: When we experimented with all 32 attributes, OneR Classification Algorithm showed 76.7334 % accuracy. But with eight most significant attributes, its accuracy increased to 76.7334 %. From the above implementation, we can said that the overall accuracy of OneR Classification Algorithm is same with 32 attributes as well as most effective attributes which are selected with the ranker method.

SimpleLogistic Classification Algorithm: When we experimented with all 32 attributes, SimpleLogistic Classification Algorithm showed 71.3405 % accuracy. But with eight most significant attributes, its accuracy increased to 73.6518 %. The accuracy of SimpleLogistic Classification Algorithm increased up to 2.3113 %.

ZeroR Classification Algorithm: When we experimented with all 32 attributes, ZeroR Classification Algorithm showed 30.9707 % accuracy. But with eight most significant attributes, its accuracy increased to 30.9707%. From the above implementation, we can said that the overall accuracy of ZeroR Classification Algorithm is same with 32 attributes as well as most effective attributes which are selected with the ranker method.

IV. MODEL SELECTION FOR STUDENT PERFORMANCE DATASET

From the above representation of all implemented classification algorithms in figure 4, we said that all the classification algorithms are performed well with little margin in their overall prediction accuracy.

In figure 4, classification algorithm like OneR, REPTree and Decision Tree (J48) have more than 76.00 % accuracy for predicting student result and they perform equally well.

Other classification algorithms such as SimpleLogistic, JRip and NaiveBayes have more than 73.00% accuracy. But RandomTree classifications algorithms are not gave a better prediction result and have the overall accuracy less than 70%.

In our implementation result, REPTree and OneR classification algorithms performed well and beats other classification algorithms in terms of overall classifier's accuracy which is equal to 76.7334 %.

Figure 5 represent the decision tree which was generated with WEKA tool after From figure 5, it is very much clear that G2 is the most significant factors for the final grade prediction as it appear at the top of the decision tree. G1, school and studytime are the second and third most significant attributes for prediction.

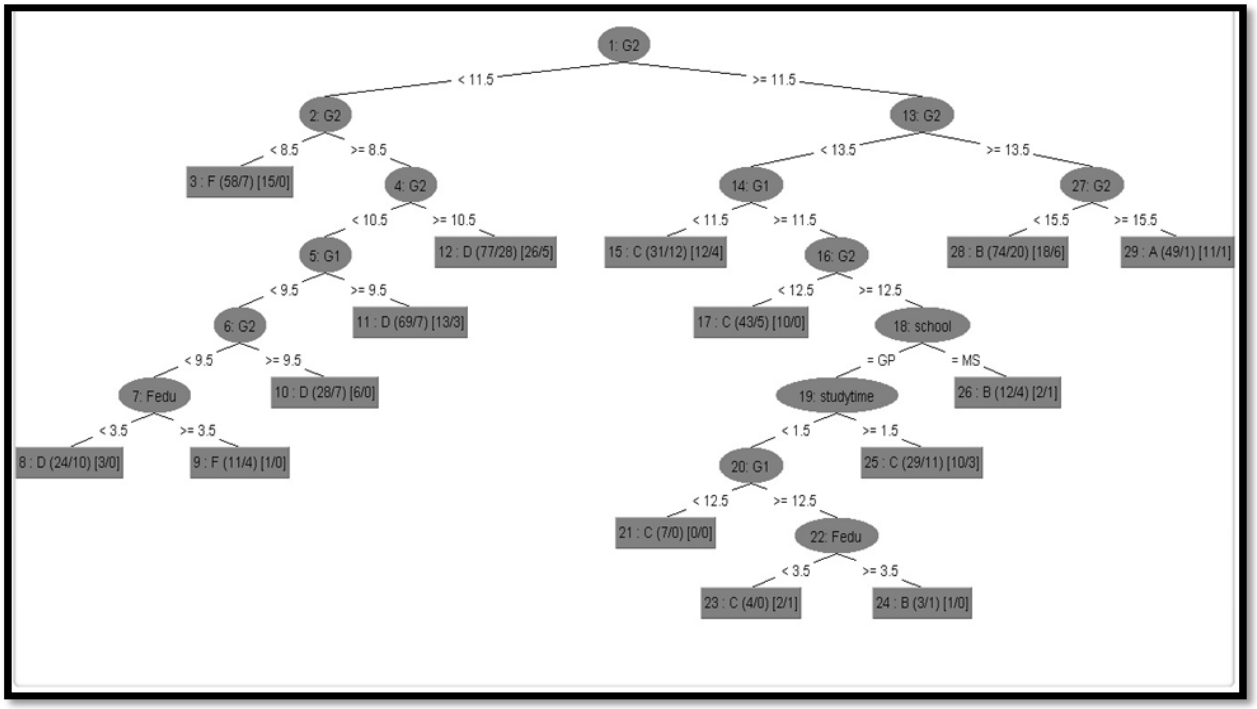

Figure 5: Tree created by REPTree with eight most significant attributes

VI. EVALUATION AND CONCLUSION

At the end, we concluded that student's performance dataset are valuable to predict the attributes that will affect their academic routine. Furthermore, this process is important to improve the educational quality which is vital to attract students to stay in the school. Generally, researchers in educational data mining are used CGPA and internal performance marks to predict academic performance. But in our implementation results we found that schools as well as studytime also affect the student final grade. We found that classification algorithm like OneR, REPTree and Decision Tree (J48) have more than 76.00 % accuracy for predicting student result and they perform equally well. As the education and evaluation system is very widespread, we assume to a large extent that further remarkable insight can still be mined from this student's performance dataset which is freely available in UCI repository. We will depart this as a part of future work.

REFRENCES

1. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. [Web Link6]
2. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. [Web Link6]
3. S. Harvey. Mining Information from US Census Bureau Data.
4. Mukesh Kumar, A.J. Singh, Disha Handa,"Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques", International Journal of Education and Management Engineering(IJEME), Vol.7, No.6, pp.40-49, 2017.DOI: 10.5815/ijeme.2017.06.05
5. Turban E.; Sharda R.; Aronson J.; and King D., 2007. Business Intelligence, A Managerial Approach. Prentice-Hall.
6. Witten I. and Frank E., 2005. Data Mining: Practi- cal Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA.
7. Luan J., 2002. Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 113, 17–36.
8. Mukesh Kumar, A.J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.8, pp.25-31, 2017.DOI: 10.5815/ijmecs.2017.08.04
9. Ma Y.; Liu B.; Wong C.; Yu P.; and Lee S., 2000. Targeting the right students using data mining. In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.
10. M. Kumar, S. Shambhu, P. Aggarwal, "Recognition of Slow Learners Using Classification Data Mining Techniques", Imperial Journal of Interdisciplinary Research, vol. 2, no. 12, 2016.
11. Mashael A. Al-Barrak And Mona S. Al-Razgan, predicting students‟ performance through classification: a case study, Journal of Theoretical and Applied Information Technology 20th May 2015. Vol.75. No.2
12. Edin Osmanbegović and Mirza Suljic, DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE, Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.
13. Raheela Asif, Agathe Merceron, Mahmood K. Pathan, Predicting Student Academic Performance at Degree Level: A Case Study, I.J. Intelligent Systems and Applications, 2015, 01, 49-61 Published Online December 2014 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2015.01.05
14. Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students‟ Performance: A Case Study, International Journal of Information and Communication Technology Research, ISSN 2223-4985, Volume 2 No. 2, February 2012.

15. Dr Pranav Patil, a study of student's academic performance using data mining techniques, international journal of research in computer applications and robotics, ISSN 2320-7345, vol.3 issue 9, pg.: 59-63 September 2015.
16. Jyoti Bansode, Mining Educational Data to Predict Student's Academic Performance, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 4 Issue: 1, 2016.
17. R. Sumitha and E.S. Vinoth kumar, Prediction of Students Outcome Using Data Mining Techniques, International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6,June 2016 ISSN: 2395-3470.
18. Karishma B. Bhegade and Swati V. Shinde, Student Performance Prediction System with Educational Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 146 – No.5, July 2016.
19. Mrinal Pandey and S. Taruna, Towards the integration of multiple classifiers pertaining to the Student's performance prediction, http://dx.doi.org/10.1016/j.pisc.2016.04.076 2213-0209/© 2016 Published by Elsevier GmbH. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/ licenses/by-nc-nd/4.0/).
20. Niranjan Lal, Shamimul Qamar , Monika Kalra, "K- Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data" Information and Communication Technology for Sustainable Development{ICT4SD), LNNS, Springer Proceeding , Volume 10, pp.61-70 2017.