International Journal of Engineering & Technology, 7 (4.39) (2018) 1073-1076

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Classification of Student Academic Performance Based on Randomized and Synthetic Dataset

Mr.C.S.Sasikumar[1]*, Dr.A.Kumaravel[2]

[1]Research Scholar, Department of CSE
[2]Professor and Dean, School of Computing
[1,2] Bharath Institute of Higher Education and Research, India
**Corresponding author E mail: sasi_kumin@yahoo.com [1],drkumaravel@gmail.com[2]*

Abstract

Analyzing the students' performance has been a challenging task in the past and many data mining tools have come to derive the knowledge hidden in these data and WEKA is one such tool. In this work, a student performance standard benchmark dataset from UCI machine learning repository is analyzed using standard data mining classifiers like Bayes Net, J48, ID3,PART,LMT and REP Tree. The accuracy obtained from original data is compared with synthetic and Randomized dataset generated and this work proves that the accuracy of synthetic and Randomized data increases. This study will support educational decision makers to design the courses more effectively.

Keywords: *Accuracy, classification, data mining, randomization, synthetic data.,WEKA*

1. Introduction

Knowledge Data Discovery (KDD) refers to the process of extracting useful knowledge from data. It involves understanding of application domain, creation of target data set, data preprocessing, data reduction and projection, selection of data mining algorithms, searching of patterns, interpretation of patterns and finally consolidation of knowledge from the outcome of above processes. Clustering and Classification are the data mining methods of KDD to discover and classify the patterns from input space which could help to derive knowledge over data. Fig.1 shows the steps in Knowledge discovery from databases

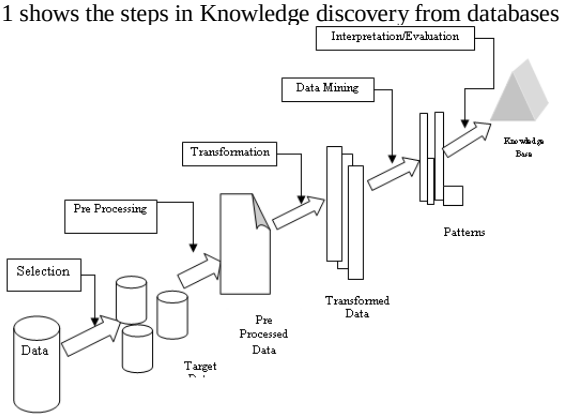

Fig. 1: Knowledge discovery from databases

As of late, Educational Data Mining has put on enormous watch in the examination domain as it has turned into a key requirement for the scholastic foundations to enhance the nature of training. For the advanced education organizations to upgrade their quality it is an unquestionable requirement for them to extricate a significant measure of concealed information. The system behind the extraction of the shrouded information is KDD process that concentrates on learning from accessible dataset and create an information base for the advantage of the organization. The paper is organized as follows: section 2 deals with related work, section 3 with methodology and dataset used. Section 4 with experimental results and section 5 with conclusions.

2. Related Work

Jai Ruby & K. David [1] presented indicators that influence the students' academic performance. The information extracted using data mining techniques can be used by educators to predict the performance of students. This has given rise to new research area named Educational Data mining (EDM)[2] that mainly concentrates on using Data Mining techniques for analyzing student performance[3]. Important uses of data mining in education are identifying preferences of students towards courses, specialization and importantly predicting the knowledge of students[4]. The behavior of students data was gathered and processed on which data mining techniques were applied to analyze hidden knowledge [5]. Higher learning institutions are now mainly involved in "knowledge creation, dissemination, and learning" [6],thereby improving the sustained competitive advantages in the academic world. In [7], using neural networks different datasets for predicting students" academic performance is compared. In [8], the authors used different feature selection methods that influence students academic performance. In [9] a study was conducted on Post Graduate students data set of size 50 to predict the performance using Decision tree. In [10] El- Halees considered a data size of 151 to extract student behaviour.

In [11] Z. J. Kovacic used CHAID and CART to predict students success. In[12]., the authors devised a method to enhance the student's performance using data mining technique. In [13] a predictive model was constructed using 772 records. Bengio Y, et.al [14] presented different classification models using Neural

networks and in [15] the authors proved that for predicting students behavior, soft computing techniques are very useful. In [16] the authors combined clustering and classification techniques and

proposed a model using which the factors influencing student's academic performance was found. In[17] many data mining algorithms were compared using student dataset and in [18], a survey on educational data mining was done for the decade 1995-2005.

The dataset used for this work was taken from UCI Machine Learning Repository and has 690 records. The dataset is a multivariate dataset for student performance with 33 number of attributes. The attributes include student grades, demographic, social and school features. The dataset is collected using reports and questionnaires. Weka tool was used for the work and estimate the accuracy of original and synthetic model. Among the different classifiers of ID3, J48, NBTree, RepTree, SimpleCart and Decision table are used. The dataset used has the following attributes.

3. Methodology Used

Table 1: Attribute name and description

No.	Attribute Name	Attribute Description
1	school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	sex	student's sex (binary: 'F' - female or 'M' - male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: 'U' - urban or 'R' - rural)
5	famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -" 5th to 9th grade, 3 -" secondary education or 4 -" higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -" 5th to 9th grade, 3 -" secondary education or 4 -" higher education)
9	Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')
10	Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')
11	reason	reason to choose school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	student's guardian (nominal: 'mother', 'father' or 'other')
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or English) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	first period grade	(numeric: from 0 to 20)
32	second period grade	(numeric: from 0 to 20)
33	final grade	(numeric: from 0 to 20, output target)

The grades obtained are related with the course subject, Math or English:

Table 2: Attribute name and description

No.	Attribute Name and Description
1	first period grade (numeric: from 0 to 20)
2	second period grade (numeric: from 0 to 20)
3	final grade (numeric: from 0 to 20, output target)

4. Experimental Results

Experiments were carried out using the Grades on the different datasets and the following results were obtained for the different grades G1, G2 and G3. Table 1 shows the Metrics comparison for accuracy classifier for grade G1(with actual data and synthetic dataset. Table 2 and 3 shows accuracy comparison for grade 2 and 3. Figure 2 shows the comparison of accuracy for Grade 1 between original and synthetic data. Figure 3 shows the same comparison for G2 and figure 4 for G3. In the similar way the dataset has been randomized using weka for each Grade and applied the classification algorithms. The table 4, 5, and 6 shows the data and classification in the respective diagrams.

Here the datasets been used as synthetic or synthesizing of data and randomizing of data to note the variation of metrics. The **Synthetic** of original data is taken the first top 100 rows and duplicated with original data which has given raise in volume of data. The synthetic happens sequentially at the bottom of the original data. The synthetic been done thrice to see the difference of metrics by incrementing 100 rows with original dataset. In the similar way, the randomization also prepared by taking the maximum synthetic of 3rd level and applied the filter **Random subset** of unsupervised data in WEKA 3.8.

Table 3: Comparison of accuracy for Grade 1 between original and synthetic data

No	Classifier-G1	Accuracy (Org)	Accuracy (Org+100)	Accuracy (Org+200)	Accuracy (Org+300)
1	BayesNet	99.2296	99.466	99.5289	99.6839
2	NativeBayes	83.5131	84.2457	83.6278	82.6133

3	NaiveBayes Multinomial Text	54.0832	46.8625	51.3545	51.4226
4	rules.PART	99.3837	100	100	100
5	trees.LMT	100	99.8665	100	99.8946

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms of accuracy for actual data and synthetic data, which was incremental duplication of original data for Grade G1.

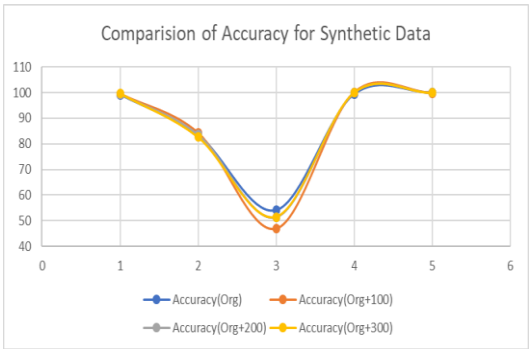

Fig. 2: Comparison of accuracy for Grade 1 between original and synthetic data

Table 4: Metrics comparison for datasets (grade G2 with synthetic data)

No	Classifier-G2	Accuracy (Org)	Accuracy (Org+100)	Accuracy (Org+200)	Accuracy (Org+300
1	BayesNet	98.6133	99.0654	99.2933	99.15
2	NativeBayes	87.3652	88.518	88.457	87.882
3	NaiveBayes MultinomialText	55.624	48.1976	48.0565	45.9431
4	trees.LMT	99.3837	99.8665	99.8822	100

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms of accuracy for actual data and synthetic data, which was incremental duplication of original data for Grade G2.

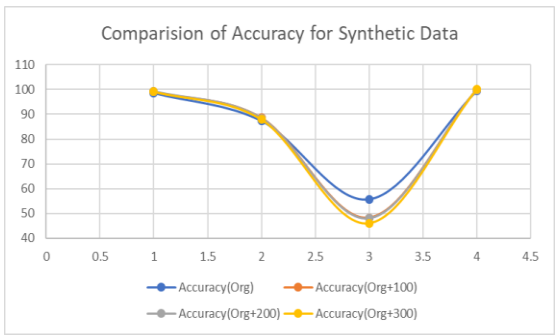

Fig. 3: Comparison of accuracy for Grade 2 between original and synthetic data

Table 5: Metrics comparison for different datasets for grade G3(with synthetic data)

No	Classifier-G3	Accuracy (Org)	Accuracy (Org+100)	Accuracy (Org+200)	Accuracy (Org+300
1	BayesNet	98.9214	99.1989	99.1755	99.68
2	NativeBayes	84.2835	85.3138	84.6879	84.19
3	NaiveBayes MultinomialText	57.0108	49.3992	43.934	49.84
4	rules.JRip	99.8459	100	99.8822	10
5	trees.LMT	99.6918	100	99.8822	99.89
6	trees.REPTree	99.8459	99.733	99.7644	99.78

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms of accuracy for actual data and synthetic data, which was incremental duplication of original data for Grade G3.

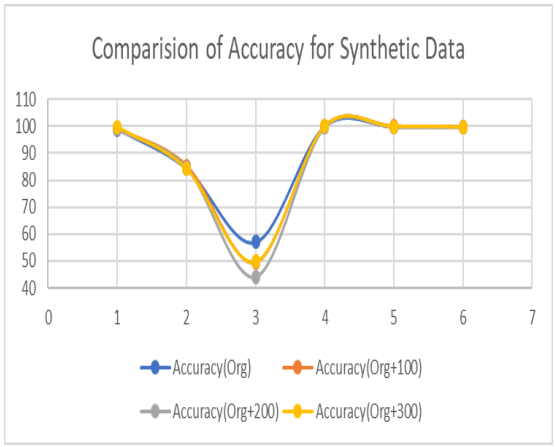

Fig. 4: Comparison of accuracy for Grade 3 between original and synthetic data

Table 6: Metrics comparison for different datasets for grade G1(with randomized data)

No	Classifier-G1	Accuracy (Randomized)
1	BayesNet	62.6976
2	NativeBayes	62.0653
3	NaiveBayesMultinomialText	51.4226
4	rules.DecisionTable	57.5342
5	rules.JRip	57.0074
6	rules.PART	71.8651
7	trees.J48	68.8093
8	trees.LMT	77.7661
9	trees.REPTree	63.5406

In the below graph the variations are happening in the classifiers of NativeBayes, NativeBayesMultinomialText, Decision Table, JRip, PART, J48, LMT and REPTree in terms of accuracy for randomized data, which was randomly kept from the data of high volume synthetic or duplicate for Grade G1.

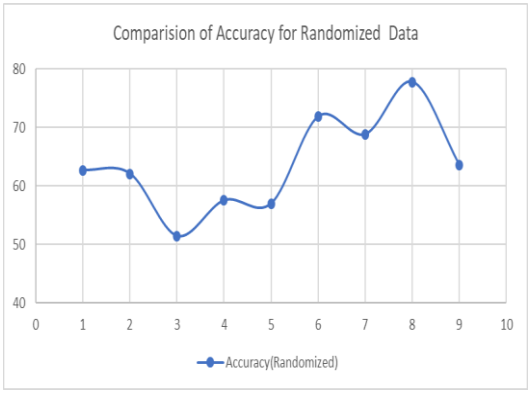

Fig. 5: Comparison of accuracy for Grade 1 randomized data

Table 7: Metrics comparison for different datasets for grade G2(with randomized data)

No	Classifier-G2	Accuracy (Randomized)
1	BayesNet	60.6955
2	NativeBayes	59.6417
3	NaiveBayesMultinomialText	45.9431
4	rules.DecisionTable	57.745
5	rules.JRip	61.2223
6	rules.PART	69.9684
7	trees.J48	71.3383
8	trees.LMT	78.9252
9	trees.REPTree	64.2782

In the below graph the variations are happening in the classifiers of NativeBayes, NativeBayesMultinomialText, Decision Table, JRip, PART, J48, LMT and REPTree in terms of accuracy for

randomized data, which was randomly kept from the data of high volume synthetic or duplicate for Grade G2.

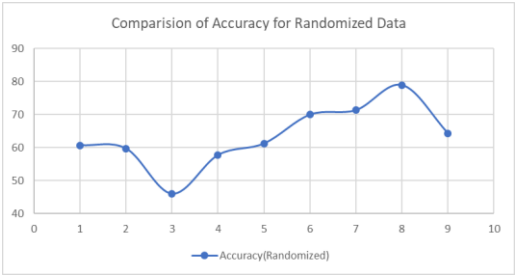

Fig. 6: Comparison of accuracy for Grade 2 randomized data

Table 8: Metrics comparison for different datasets for grade G3 (with randomized data)

No	Classifier-G3	Accuracy (Randomized)
1	BayesNet	60.4847
2	NativeBayes	59.8525
3	NaiveBayesMultinomialText	49.8419
4	rules.DecisionTable	59.7471
5	rules.JRip	64.2782
6	rules.PART	68.4932
7	trees.J48	67.8609
8	trees.LMT	78.3983
9	trees.REPTree	62.9083

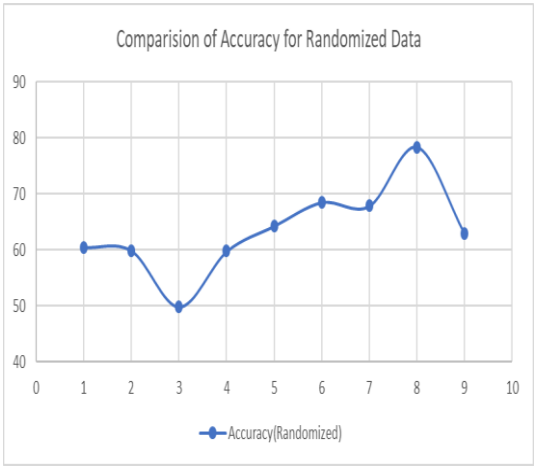

Fig. 7: Comparison of accuracy for Grade 3 randomized data

In figure 7 the variations are happening in the classifiers of NativeBayes, NativeBayesMultinomialText, Decision Table, JRip, PART, J48, LMT and REPTree in terms of accuracy for randomized data, which was randomly kept from the data of high volume synthetic or duplicate for Grade G3.

5. Conclusions

This work is aimed on analyzing the accuracy of the academic performance of the students using the actual and synthetic dataset of varying size for the different grades. The study proves that with increase in duplication of data in synthetic dataset the accuracy varies. The future work will be to use machine learning algorithms for classification purpose.

References

[1] Jai Ruby & K. David, "A study model on the impact of various indicators in the performance of students in higher education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May-2014, pp.750-755.

[2] Baker R.S.J.D., & Yacef K, 2009," "The state of educational data mining in 2009: A review and future vision", Journal of Educational Data Mining, I, 3-17.

[3] Monika Goyal & Rajan Vohra, "Applications of Data Mining in Higher Education" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp.130-120.

[4] Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383

[5] Alaa M El-Hales, "Mining Educational Data to Analyze Learning Behaviour A case study", 2009

[6] Rowley, J., "Is higher education ready for knowledge management?", International Journal of Educational Management, 2000, vol. 14(7), pp. 325–333.

[7] Lubega, J. T., Omona, W., & Weide, T. V. D., "Knowledge management technologies and higher education processes_: approach to integration for performance improvement", International Journal of Computing and ICT Research, 2011, vol. 5(Special Issue), pp. 55–68.

[8] J. Shana, T. Venkatachalam, "Identifying Key Performance Indicators and Predicting the Result from Student Data." International Journal of Computer Applications (0975 – 8887) Vol.25-No.9 July 2011.

[9] Brijesh Kumar Baradwaj, Saurabh Pal," Mining Educational Data to Analyze Students,,Performance" IJACSA, Vol.2, No.6, 2011

[10] El-Hales-A.(2008),"Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia,Dec 15-18.

[11] Kovacic Z. J., "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference 2010.

[12] Shanmuga Priya. K, & Senthil Kumar A.V., "Improving the Student"s Performance Using Educational Data Mining, 2013.

[13] Ramaswami M., & Bhaskaran R., "CHAID Based Performance Prediction Model in Educational Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.

[14] Bengio Y., Buhmann J. M., Embrechts M., & Zurada J. M., "Introduction to the special issue on neural networks for data mining and knowledge discovery," IEEE Trans. Neural Networks, vol. 11, pp. 545-549, 2000.

[15] Vasile P. B., "Analysis and Predictions on Students" Behavior Using Decision Trees in Weka Environment". Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE, (2007).

[16] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, "Prediction NDUM student's academic performance using data mining techniques," presented at the International Conference on Computer and Electrical Engineering, 2009.

[17] Jai Ruby & K. David, "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study ", IJRASET International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November 2014 [18] Romero, C. and Ventura, S. (2007) „Educational data mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146.