International Journal of Engineering & Technology, 7 (4.39) (2018) 1069-1072

International Journal of Engineering & Technology

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

Research paper

Attribute Selection on Student Academic and Social Attributes 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*

Abstract

Subset selection is important when the underline dataset contains insignificant attributes as they don't contribute much to the final results especially, in the context of student performance prediction studies. Hence the exploration of procedures for such goal becomes relevant. Though this is the case, in general it happens to be NP-hard problem. In this paper we apply Best First Search, Greedy Search, and Ranker method (Information Gain Ratio) to select the attributes using weka tool with learning models based on decision rules, decision trees, neural networks, bayes NET and meta classifiers. The performance comparisons are made with ranked and non-ranked search methods over the synthetic and randomized datasets derived from original students' performance dataset.

Keywords—*Attribute Selection,Best First Search, Ranker Search , WEKA, accuracy, classification, randomization, synthetic data*

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

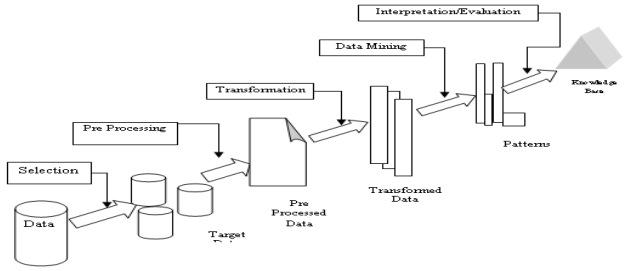

Fig.1: Knowledge discovery from databases

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 Portuguese:

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 grade G1. Table 1 shows the Metrics comparison for accuracy classifier for grade G1(with Non zero ranking applied, Table 2, Table 3, Table 4 and Table 5 shows accuracy comparison for grade by removing 5 attributes from overall dataset. Figure 1 shows the comparison of accuracy for Grade with removed of non-zero attributes from the original dataset.

Here the datasets been used by removing all non-ranking attributes, after applying the ranker search and removed the attributes with zero. This was done using ranker filter of unsupervised data in WEKA 3.8.

Table 3: Non-zero attributes the above are the accuracy for different classifiers

Sl.No.	Accuracy	Classifier
1	68.4129	bayes.Bayesnet
2	68.4129	bayes.NativeBayes
3	54.0832	bayes.NativeBayesMultinominalText
4	68.1048	rules.DecisionTable
5	65.9476	rules.Jrip
6	59.0139	rules.PART
7	69.4915	trees.J48
8	67.9507	trees.LMT
9	65.6394	trees.REPTree

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms

of accuracy for actual data with non-zero ranking attributes for Grade.

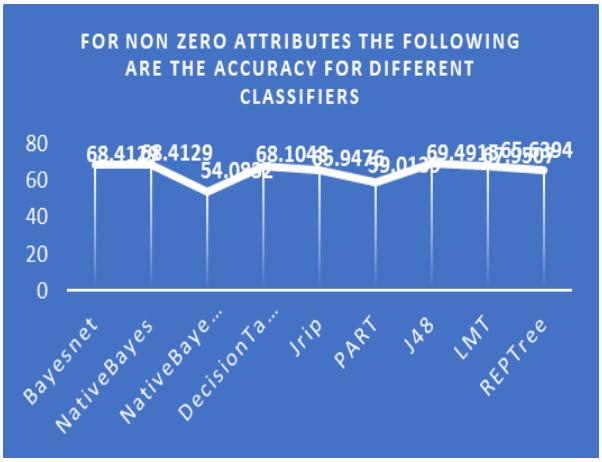

Fig .2: Non-zero attributes the above are the accuracy for different classifiers

Table 4: Metrics comparison After removing last 5 attributes (Pstatus, romantic, famsup, paid, famsize) the following are the accuracy for different classifiers

Sl.No.	Accuracy	Classifier
1	68.567	Bayesnet
2	68.8752	NativeBayes
3	54.0832	NativeBayesMultinominalText
4	65.7935	DecisionTable
5	68.7211	Jrip
6	59.6302	PART
7	68.1048	J48
8	67.3344	LMT
9	66.5639	REPTree

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms of accuracy for actual data with non-zero total attributes and with after removing last 5 attributes (Pstatus, romantic, famsup, paid, famsize) mentioned here.

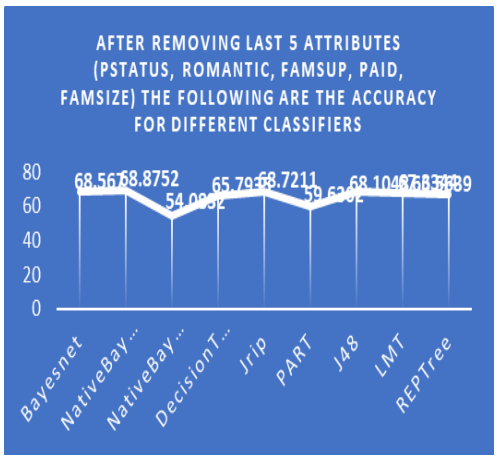

Fig. 3: Metrics comparison After removing last 5 attributes (Pstatus, romantic, famsup, paid, famsize) the following are the accuracy for different classifiers

Table 5: Metrics comparison After removing next last 5 attributes (internet, sex, nursery, activities, schoolsup) the following are the accuracy for different classifiers

Sl.No.	Accuracy	Classifier
1	67.6425	Bayesnet
2	69.7997	NativeBayes
3	54.0832	NativeBayesMultinominalText
4	66.8721	DecisionTable
5	67.0262	Jrip
6	60.5547	PART
7	68.4129	J48
8	68.567	LMT
9	65.3313	REPTree

In the below graph the variations are happening between the classifiers NativeBayes and NativeBayesMultinomialText in terms of accuracy for after removing next last 5 attributes in the dataset with non-zero. Those are next last 5 attributes (internet, sex, nursery, activities, schoolsup) in the actual dataset.

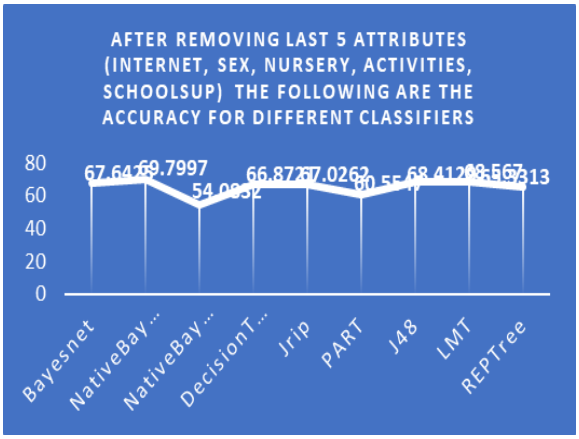

Fig. 4: Comparison of accuracy After removing next last 5 attributes (internet, sex, nursery, activities, schoolsup) the following are the accuracy for different classifiers

Table 6: Metrics comparison After removing last 5 attributes (age, Fjob, reason, guardian, address) the following are the accuracy for different classifiers

Sl.No.	Accuracy	Classifier
1	68.4129	Bayesnet
2	69.9538	NativeBayes
3	54.0832	NativeBayesMultinominalText
4	67.3344	DecisionTable
5	66.1017	Jrip
6	65.0231	PART
7	66.4099	J48
8	69.0293	LMT
9	67.4884	REPTree

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 after removing last 5 attributes in the dataset with non-zero. Those are next last 5 attributes (age, Fjob, reason, guardian, address) in the actual dataset.

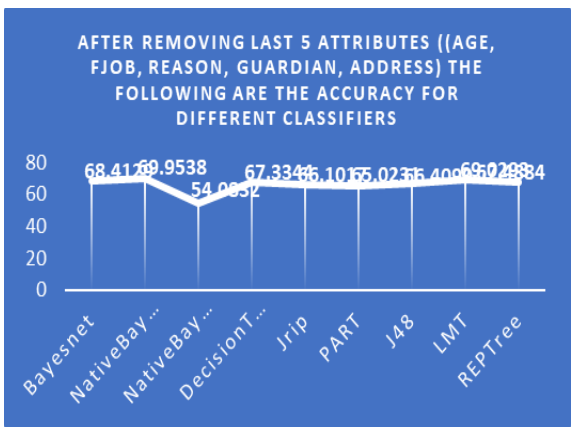

Fig. 5: Comparison of accuracy After removing last 5 attributes (age, Fjob, reason, guardian, address) the following are the accuracy for different classifiers

5. Conclusions

This work is aimed on analyzing the accuracy of the academic performance of the students using the actual dataset and applying the ranker filters which will help to find out the critical attributes needed to derive the accuracy. The study proves that required attributes needed and shows the 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.