Data Mining Approach to Analyse and Predict the Student Academic Performance

Quazi Rayad Uddin, Md. Hasnat Riaz, Md. Auhidur Rahman

Abstract— Data mining is a process of extracting interesting pattern from existing dataset and predict the future. It can be used in educational dataset. An Institute has different types of students. All students are provided same facilities and resources. But the academic performance is not same. The academic performance differs because of the way students interact with different resources and the way they utilize this resources are not same. In this case data mining can be used to find out the knowledge from educational dataset. In this research a survey was conducted on the students of Computer Science and Telecommunication Engineering of Noakhali Science and Technology University. Classify the result into three classes good, average and bad. Apply different types of classifier and find support vector machine and K-nearest neighbor classifiers work more accurately compare to others.

Index Terms— Educational data mining, Classification, Academic performance, Prediction, Learning behavior.

1 Introduction

EDUCATIONAL data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students' and the settings in which they learn [1]. By analyzing students previous data to classify students and predict their performance is an interesting field of research.

For this research a survey was conducted. We considered student background, family status, current living location, class interest, reading hours, library interaction etc. A Google forms was created and asked to fill it to the students of department of Computer Science and Telecommunication Engineering of Noakhali Science and Technology University. We classify the result into three classes and try to predict student's upcoming semester result. The main objectives of the research are to implement data mining techniques and methods on collected student's dataset. Find out patterns in the available data for predicting students' performance. Discovering the factors that have great impact on student's academic performance. Improve learning behavior of students. In this research we apply multiple classification methods and find the best one which is more suitable for our predictive model.

For all kind of implementation and visualization we use python. Because of python is more popular for research nowadays. And python is also an open source language. We also use Jupyter Notebook as editor because it is specially made for data science. Python has several powerful packages for data visualization and model execution.

The paper is organized into five sections. In introduction section a summary of the conducted research work is presented. In the second section a review of the related works are provided. Third section contains data analysis which includes representation of the collected dataset, an exploration and visualization of the data. The obtained results and the comparative analysis are given in fourth section. The paper concludes with a summary of the achievements and discussion of further work.

- *Quazi Rayad Uddin is currently pursuing bachelor degree program in Computer Science and Telecommunication Engineering in Noakhali Science and Technology University, Bangladesh. E-mail: kazirahat66@gmail.com*
- *Md. Hasnat Riaz and Md. Auhidur Rahman is currently working Noakhali Science and Technology University, Bangladesh. E-mail: Hasnat.riaz90@gmail.com, auhid.sumon@gmail.com.*

2 Related Research

V. Shanmugarajeshwari and R. Lawrance [2] conducted a research on student of Ayya Nadar Janaki Ammal College, Sivakasi, TamilNadu, India of computer Applications department (Master of Computer Applications). Initially the data size was 47 records and 12 attributes. They classify the result into pass and reappear. They used C5.o algorithm. And the accuracy of this research is 100 percent.

Asraful Alam and Mehedi Hasan [3] conducted a similar research on student of Department of Computer Science and Engineering, United International University, Bangladesh. The dataset contains 70 instances, 16 attributes (values are both real & nominal), and 1 class attribute (Good, Average and Poor). They used Decision tree classifier to classify student's programming skills. The proposed decision tree models can correctly classify 87% students.

Similarly Amin Zollanvari [4] with three others conducted a research on fourth-year students enrolled only in the electrical engineering program at Nazarbayev University. They constructed a predictive model of GPA based on a set of self-regulatory learning behaviors.

Nguyen Thai Nghe, Paul Janecek and Peter Haddawy [5] compared the accuracy of Decision Tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes.

Bo Guo with four others [6] conducted a research to develop a classification model to predict student performance using Deep Learning which automatically learns multiple levels of representation.

A.F.ElGamal [7] conducted a research to predict student performance in programming courses.

3. DATA ANALYSIS

A. Dataset

The dataset used in this study was collected through a survey using Google Forms. The initial size of the dataset is 77 records. Table 1 describes the attributes of the data and their possible values.

TABLE 1
Attributes description and possible values

Attribute	Description	Possible Value
Gender	Binary	Male, Female
adtype	Student's admission	1st time, 2nd time, readmit
Region	Belongs which division to Bangladesh	Chittagong, Dhaka, Sylhet, Rajshahi, Khulna, Barisal, Rangpur
Cloc	Current Location	Hall, Mess, With family
PositionFamly	Student's position in family	First, Middle, Last
ClAtten	Class attendance	Good(>80%), Average(60 to 80), Poor(< 60)
Ctmark	Class Test Mark	Good(>20), Average(15 to 20), Bad(<15)
Assignment	Assignment done by own	Yes, No
ClasNote	Like to take class note	Yes, No
ClassRespond	Try to respond in class	Yes, No
QuesInClass	Asked question in class	Yes, No
LibraryBooks	Brought Books form Library	Yes, No
SubBooks	Bought/Collected Books	Yes, No
DailyFB	Daily spent time in Facebook	Around 1 hour, 2 to 4 hours, More than 4 hours
ExamNote	Made notes for final exam	Yes, No
Backlog	Number of backlog(one semester)	1,2,3..
Class	TGPA(term result)	2.5 to 4

The student performance is measured by the Grade Point Average (GPA), which is a real number out of 4. For our prediction we classify the result into following three groups.

TABLE 2
Proposed class for the model

Class	Range
Good	>3.49
Average	3.0 to 3.49
Bad	< 3.00

B. Data Exploration

For understanding the dataset and relation between different attributes it must be explored in a statistical manner and visualize it using graphical plots and diagrams. It is essential because it allows us to understand the data before jumping into applying more complex data mining tasks and algorithms. Here we summarize the overall dataset and visualize class and relation among class and different attributes. We also visualize the data correlation matrix to observe correlation among attributes.

TABLE 3
Summary of the dataset

Attribute	Range
Gender	M(64), F (13)
adtype	1st(45), 2nd(26), readd(6)
Region	CHI(42), DHA(25), KHU(3), MYM(3), SYL(2), BAR(1), RAJ(1), RAN(0)
Cloc	Hall(24), Mess(39), Family (14)
PositionFamly	First(26), Middle(34), Last(17)
ClAtten	G(35), A(29), P(13)
Ctmark	G(13), A(46), B(18)
Assignment	Yes(34), No(43)
ClassRespond	Yes(40), No(37)
QuesInClass	Yes(31), No(46)
LibraryBooks	Yes(24), No(53)
SubBooks	Yes(44), No(33)
DailyFB	Around 1 hour(23), 2 to 4 hours(48), More than 4 hours(6)
ExamNote	Yes(23), No(54)
ClasNote	Yes(31), No(46)
Backlog	0(42), 1(25), 2(4), 3(5), 5(1)
Class	G(17), A(39), B(21)

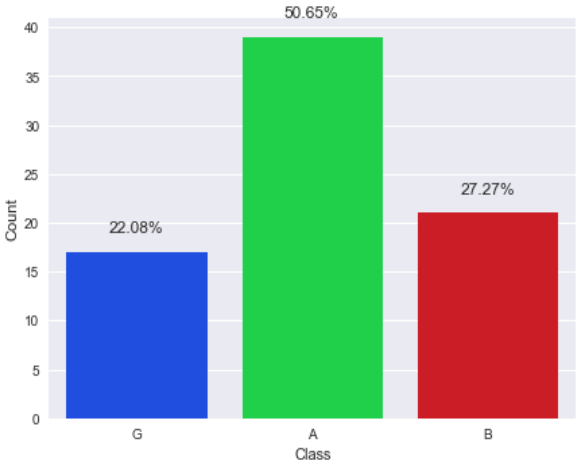

Fig 1: Histogram of Class Attribute

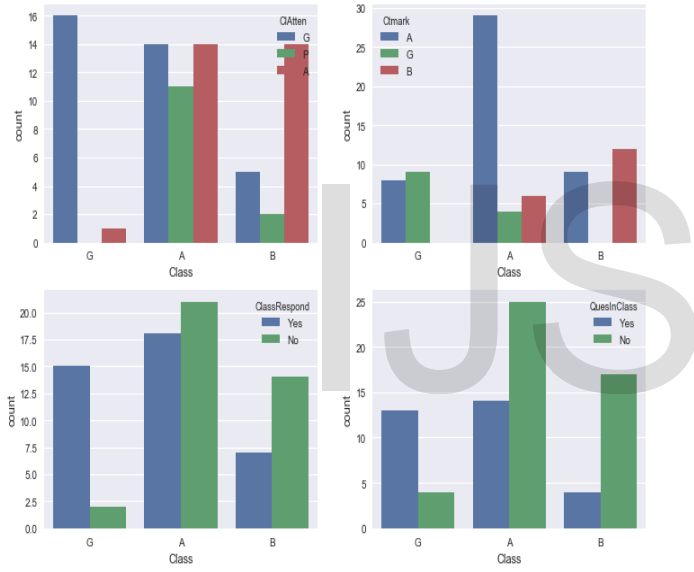

Fig 2: Class comparison with attributes (ClAtten, Ctmark, ClassRespond, QuesInClass).

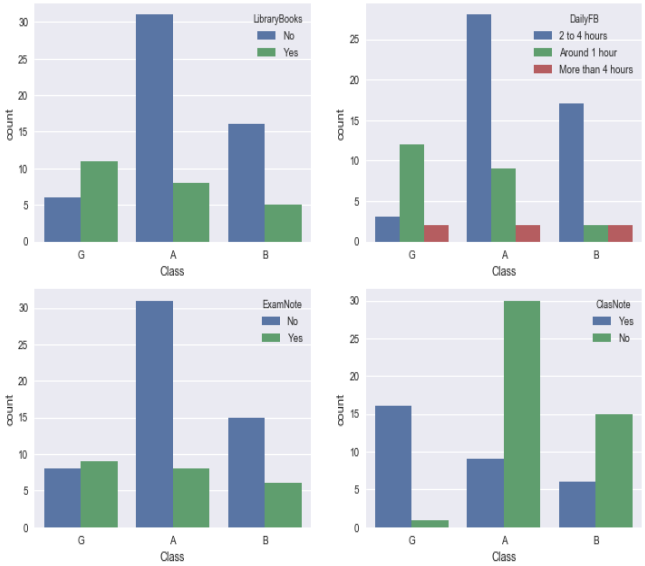

Fig 3: Class comparison with attributes (LibraryBooks, DailyFB, ExamNote, ClassNote).

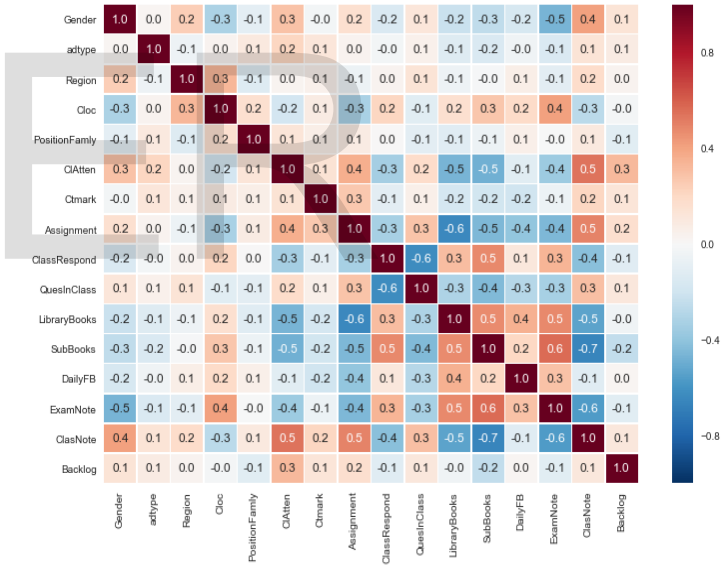

Fig 4: Heat map of correlation matrix of dataset.

4. EXPERIMANTAL RESULT

A. Selected algorithm

In this study, multiple classification techniques are used in the data mining process for predicting the students' grade at the end of the semester. This approach is used because it can provide a broader look and understanding of the final results and output, as well as, it will lead to a comparative conclusion over the outcomes of the study. There are multiple classification techniques available in data mining. For this research we use following classifiers.

1. Naive bayes classifier

2. Logistic regression
3. K-nearest neighbor
4. Support vector machine
5. Decision tree classifier

B. Results from analysis

We split our dataset into training and testing datasets. The split ratio is 8:2, 80 percent data used for train model and with remaining 20 percent we test our model. For model evaluation we use confusion matrix.

TABLE 4:
Model evaluation with NAIVE BAYES

Confusion Matrix		Actual			Class Precision (%)
		A	B	G	
Prediction	A	3	2	4	75
	B	1	2	0	50
	G	0	0	4	50
Class Recall (%)		33	67	100	Accuracy: 56.25%

TABLE 5:
Model evaluation with LOGISTIC REGRESSION

Confusion Matrix		Actual			Class Precision (%)
		A	B	G	
Prediction	A	5	2	2	100
	B	0	3	0	60
	G	0	0	4	67
Class Recall (%)		56	100	100	Accuracy: 75%

TABLE 6:
Model evaluation with K-NEIGHBOURS

Confusion Matrix		Actual			Class Precision (%)
		A	B	G	
Prediction	A	7	0	2	88
	B	1	2	0	100
	G	0	0	4	67
Class Recall (%)		78	67	100	Accuracy: 81.25%

TABLE 7:
Model evaluation with SVM

Confusion Matrix		Actual			Class Precision (%)
		A	B	G	
Prediction	A	7	1	1	88
	B	1	2	0	67
	G	0	0	4	80

TABLE 8:
Model evaluation with DECISION TREE

Confusion Matrix		Actual			Class Precision (%)
		A	B	G	
Prediction	A	6	2	1	75
	B	2	1	0	33
	G	0	0	4	80
Class Recall (%)		67	33	100	Accuracy: 68.75%

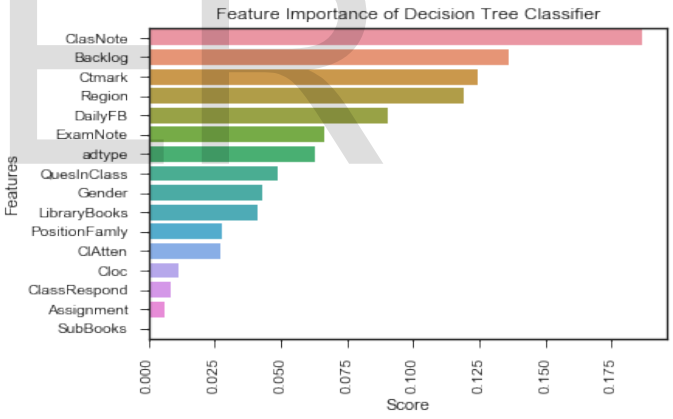

Fig 5: Feature Importance of Decision Tree.

TABLE 9:
Accuracy comparison of classifiers

	Class Accuracy			Overall Accuracy
Algo-rithms	A	B	G	
Naive Bayes	33	67	100	56
Logistic Regres-sion	56	100	100	75
K-Neigh-bors	78	67	100	**81.25**
SVM	78	67	100	**81.25**
Decision Tree	67	33	100	68.75

5. CONCLUSION

In this research paper, multiple data mining tasks are used to create qualitative predictive models which are efficiently and effectively able to predict the students' grades from a collected training dataset. First, a survey is constructed that has targeted university students and collected multiple personal, social, and academic data related to them. Second, the collected dataset is preprocessed and explored to become appropriate for the data mining tasks. Third, the implementation of data mining tasks is presented on the dataset in hand to generate classification models and testing them. Finally, interesting results were drawn from the classification models. And we find that support vector machine and K-nearest neighbor classifier work better for our model compare to other classifiers. Also decision tree classifier has been implemented. From decision tree we find the different attributes impact on student's academic performance. In conclusion, this study can motivate and help universities to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students in many ways.

5. REFERENCES

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

[2] Shanmugarajeshwari, V., & Lawrance, R. (2016, January). Analysis of students' performance evaluation using classification techniques. In Computing Technologies and Intelligent Data Engineering (ICCTIDE), International Conference on (pp. 1-7). IEEE.

[3] Pathan, A. A., Hasan, M., Ahmed, M. F., & Farid, D. M. (2014, December). Educational data mining: A mining model for developing students' programming skills. In Software, Knowledge, Information Management and Applications (SKIMA), 2014 8th International Conference on (pp. 1-5). IEEE.

[4] Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., & Hernández-Torrano, D. (2017). Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors. IEEE Access, 5, 23792-23802.

[5] Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual (pp. T2G-7). IEEE.

[6] Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In Educational Technology (ISET), 2015 International Symposium on (pp. 125-128). IEEE.

[7] El Gamal, A. F. (2013). An educational data mining model for predicting student performance in programming course. International Journal of Computer Applications, 70(17).