Green University

GUBJSE: ISSN: 2409-0476

AN ANALYSIS OF STUDENTS' ACADEMIC RECORD USING DATA MINING TECHNIQUES AND IDENTIFICATION OF KEY FACTORS TO AID STUDENTS' PERFORMANCE

Md. Ashaduzzaman, Shihabuzzaman, Md. Hasanur Rahman Sagor, Md. Mizanur Rahman, and Ahmed Iqbal Pritom

Abstract**—** With the improvement of information technology, presently educational institutions generally store and compile a huge volume of students' data. This huge volume of data can be analyzed using different data mining techniques and extract hidden relation between students' result with other academic attributes. The main objective of this paper is to evaluate the impact of different academic attributes on the students' final result using data mining techniques. We used different data mining techniques to analyze students data collected from Green University of Bangladesh. We applied three well-known classification algorithms namely Decision Tree, Naïve Bayes, and SVM to develop a prediction model that can suggest probable grade by analyzing parameters like the midterm, attendance, assignment, presentation, class test, final, and CT marks. Our goal is to find out the key factors playing as a catalyst for getting good or bad CGPA. Through this research, the university authority will get the knowledge about key factors playing significant role in students' result that will help them to take proper decisions to improve students' grade that in turns will reduce students' dropout.

Index Terms**—** data mining; classification; Decision Tree; Naïve Bayes; SVM

I. INTRODUCTION

IN the era of information technology, the volume of data collection is increasing in a substantial manner every day. Hidden patterns can be found out by properly analyzing the data which can be very useful for taking future decisions. To analyze the data different tools and techniques were created which initially was used to analyze data related to finance and retail industries. Later expanded to medical, intrusion detection data, even to educational institute data [1][2].

Educational Data Mining (EDM), a relatively new feature of data mining, is a concept where educational data is analyzed using different data mining technique to find out any hidden knowledge [3]. In the education sector, especially in higher education, dropout is a major concern and the poor academic result is one of the main causes that accelerates this process. An increasing number of dropout affects the students' career at the same time harms the reputation of the institute.

Educational data mining can be used to analyze students' different academic attributes to extract any hidden pattern for students' poor academic performance. This knowledge can help educational institutes to improve their teaching or other approaches to the student which on the one side will improve students' academic performance as well as carrier and on the other side will benefit all the other stakeholders of the institute.

In this paper, using machine learning algorithms and data mining techniques, we analyze students' data and create decision trees or associative rules to better understand students' academic performance and provide the academic institution with the scope to take better decisions which can improve students' future academic performance. The main focus of this paper is to analyze the impact of different attributes on the students' result. This paper uses three machine learning algorithms namely Naïve Bayes, J48 Decision Tree and SVM on students' data to present a comparison on performance study. Also, using the Decision Tree, this paper proposes some associative rules to find the hidden relation between students' different academic attributes to their final results. Based on the total number obtained in a given subject, we categories the students' final result in four different categories and identifies the key factors that may lead result to different categories.

II. RELATED WORKS

EDM is an emerging field of research where many researchers have worked to find hidden patterns in academic data. In [4], the author presented a comparative study of Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree classification

This paper was received on 22 April 2019, accepted on 19 October 2019. This work was financially supported by Green University of Bangladesh (GUB).

Md. Ashaduzzaman is with Dept. of Computer Science of Engineering of GUB, mail: *ashaduzzaman@cse.green.edu.bd*

Shihabuzzaman is with Dept. of Computer Science of Engineering of GUB, mail: *shihabuzzaman@cse.green.edu.bd*

Md. Hasanur Rahman Sagor is with Dept. of Computer Science of Engineering of GUB, mail: hrsagor4@gmail.com

Md. Mizanur Rahman is with Dept. of Computer Science of Engineering of GUB, mail: mizanur.152015024@green.ac.bd

Ahmed Iqbal Pritom is with Dept. of Computer Science of Engineering of GUB, mail: *iqbal@cse.green.edu.bd*

algorithm to predict students' performance on introductory Programming Course.

In [5], researchers used students' educational, personal, academic and admission data to predict a students' performance model using NBTree. The resulted model proved some attributes have more influence than others in students' performance [5].

In [10], authors. used Naïve Bayes and Tree C4.5 data mining algorithms to predict students learning result. The dataset consists of 10 attributes and 279 instances. Naïve Bayes and Tree C4.5 classifiers were used to build the predicting models. For Naïve Bayes, the paper achieved 78.57% accuracy for Tree C4.5, achieved 71.43% accuracy. But the paper did not describe any association rule from the knowledge extraction.

Furthermore, researchers in [11-13] implemented various data mining algorithms to analyze semester mark and compare the performance of different classifiers form confusion matrix. Others, in [14,15] explore the students' dataset from the different academic database and comparing different models to improve prediction.

Most of the work done by researchers in EDM mainly focused on predicting students' learning performance from an academic database and to compare the performance of different data mining algorithms. The focus of this paper is to explore more predicting models and at the same time extract some important association rule for improving students' performance from the hidden knowledge base.

III. METHODOLOGY & DATASET

The knowledge discovery process in data mining at first requires a collection of data. Data can be collected in various ways and from different sources which in our case is collected from Green University of Bangladesh. Then, usually, data needs to be preprocessed so that data mining tools can be applied. The collected data was not in a suitable format to analyze so we used python to preprocess the data so that data mining tools can be applied. After that data mining techniques are applied for classification and hidden knowledge extraction. Finally, extracted information can be analyzed to formulate associative rules. Figure:1 describes the system architecture used in this paper.

A. Dataset

The data set used in this paper is collected from Green University of Bangladesh (GUB). The data set is a single table consists of the information of 687 students having 7 attributes and 38681 instances.

B. Data selection & Transformation

In this step, we identified those necessary fields from the data set for our knowledge discovery process. The dataset we collected form GUB has "StudentIdVisible", "CourseId", "component_id", "component_out_of", "SemesterCode", "Marks" and "Grades" as attributes. From here we discarded both "component_out_of" and "SemesterCode" from our final preprocessed data set. We used python to convert our data in a format where we can implement our intended data mining algorithms. In our collected data the "component_id" includes class tests, midterm, final exam, individual presentation and attendance. Using python, we converted the raw data set to a new data set of having 8 attributes and 4782 instances. All 8 attributes have numeric values and there is no value missing in the data set. There is also a class attribute named "Grade". The attributes taken for prediction is shown in table 1:

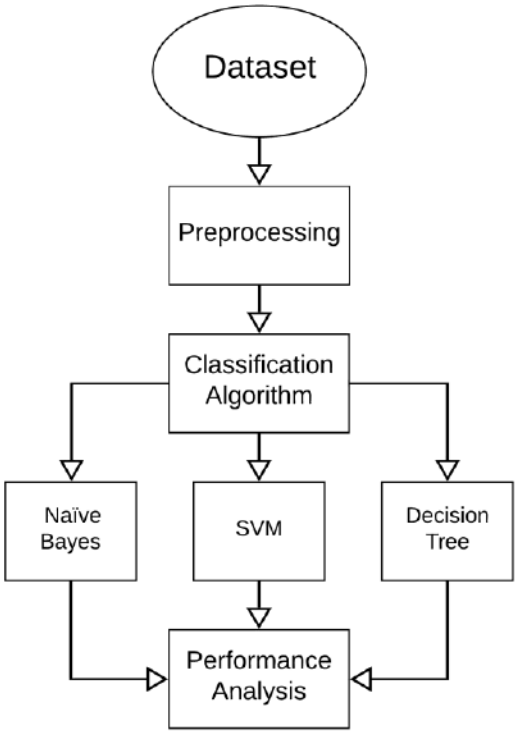

Figure 1: System Architecture

Table 1: STUDENTS DATASET

Attribute Name	Description	Possible Values
Attendance	Marks Obtained for attending classes	0 to 5
CT1	1st class test mark	0 to 15
CT2	2nd class test mark	0 to 15
CT3	3rd class test mark	0 to 15
Individual Presentation	Individual presentation mark	0 to 5
Group Assignment	Mark obtained from Group	0 to 5

	Assignment	
Midterm	Midterm examination mark	-1 to 30 {-1=Absent}
Final	Final examination mark	-2 to 40 {-1=Absent, -2=Fail}
Grade	Final grade	<A+, A, A-, B+, B, B-, C+, C, D, F>

Then in the new dataset, we categorize all the numeric values in four categories; Excellent, Good, Medium and Bad based on the following rules:

- Excellent: Mark>79%
- Good: 80%>Mark>59%
- Medium: 60%>Mark>39%
- Bad: Mark<40%

In the dataset, we also found some instances with either -1 or -2 value or both. As, -2 denotes fail, we programmed those values to be counted as Bad. But for entries with -1 value which mean absent and may or may not be sitting for makeup exams, we discarded those instances because for the uncertainty of final grade. We also make the same four categories for class attribute “Grade” following the previously mentioned rule where ‘A+’ falls under Excellent category; ‘A’, ’A-‘, ’B+’ and ‘B’ Good; ‘B-‘, ‘C+’, ‘C’ and ‘D’ Bad and ‘F’ grade falls under the category Bad. Finally, we have the dataset with 8 attributes and 1 class attributes each having four categories and the dataset has 4590 instances. After preprocessing of the data, three experimental methods are implemented on the data set and outcomes are compared.

C. *Experimental Methods*

In this paper, three different classification algorithms are used based on the attributes of the dataset. They are Naïve Bayes, Decision Tree and Support Vector Machine.

Naïve Bayes is a Bayes’ Theory based classifier which simply uses machine learning probabilistic classification approach [6,7]. Variables are evaluated independently of each other and can identify important classification parameter using small training data. It finds the probability of any tuple and based on that classify that in a particular class.

Unlike Naïve Bayes, Support Vector Machine (SVM) is a non-linear, machine learning based supervised learning model. SVM is a discriminative classifier where it rehabilitates data to a new hyperplane [8]. The hyperplane is considered as the boundaries which classify the data. Data placed the different side of the hyperplane is considered to be in a different class. The number of hyperplanes depends on the number of features in the input. SVM takes input and for each data, it predicts in which class the data belongs to. So, it is extremely slow but at the same time extremely accurate [9].

Decision Tree is unlike other described models in this paper. Where SVM and Naïve Bayes evaluate every attribute independently, Decision tree maintains a hierarchical breakdown of data. Inducer and visualizer are the two main parts of a decision tree where the role of inducer is to identify the most important attribute for classification. Then the graphical model is represented by the visualizer. Decision tree requires relatively less computation to classify data. But the main strength of Decision tree is that it can generate understandable rules.

IV. Result

In our paper, we applied three well-known classification algorithms to evaluate our work. Those are Decision Tree, Naïve Bayes and SVM. We performed our analysis by dispatching 10-fold cross validation (10CV) and 66.67% training set & 33.33% test set percentage split on the dataset. We identified the accuracy, average TP Rate, FP Rate, Precision, Recall, and F-Measure. Finally, we provided the running time for each experiment. All this information is provided in the table-2. From the table, we can observe in column 1 that 10 fold cross-validation was performed using Decision Tree algorithm and found out accuracy is 85.8139 %, average TP Rate is 0.858, FP Rate is 0.73, Precision is 0.859, Recall is 0.858, F-Measure is 0.858, and execution time is 0.03s. Among these three algorithms, we identified that SVM provided the best accuracy that is 89.1262%. The experiments also identified that 10-fold cross validation performed better for all three algorithms. However, SVM takes more time than the other two algorithms that are 0.88s and 0.86s. In that case, Naïve Bayes performed better because it takes 0s to find its results. Therefore, if we have enough time to search we can use SVM. On the other hand, if we want a good tradeoff between time and accuracy, we can use Naïve Bayes.

V. Analysis and Discussion

In our paper, we identified a set of rules to understand the relationships among attributes such as final marks, mid-term marks, class test (CT) marks etc. and the final grade. We wanted to identify why any student doing excellent in an exam or why any student doing badly. In table 3, we provided the top three rules for each category result.

For example, when a student performed well in the Final exam, excellent in Midterm, CT-2, and Individual Presentation 95.82% of them achieved excellent Grade.

Table 2: EXPERIMENTAL RESULT

	Decision Tree		**Naïve Bayes**		**SVM**	
	10CV	Percent split	10CV	Percent split	10CV	Percent split
Accuracy	85.8139 %	84.6795 %	87.8187 %	86.5385 %	89.1262%	87.7564 %
TP Rate	0.858	0.847	0.878	0.865	0.891	0.878
FP Rate	0.073	0.08	0.067	0.073	0.056	0.063
Precision	0.859	0.848	0.88	0.867	0.891	0.877
Recall	0.858	0.847	0.878	0.865	0.891	0.878
F-Measure	0.858	0.846	0.877	0.864	0.891	0.878
Time	0.03s	0.03s	0s	0s	0.88s	0.83s

Table 3: TOP THREE RULES FOR EACH CATEGORY ASSOCIATED WITH ITS ACCURACY

Category	**Rule**	**Accuracy**
Excellent	1. Final = Good -> Mid-term = Excellent -> CT-2 = Excellent -> Individual Presentation = Excellent	95.82%
	2. Final = Excellent -> Mid-term = Excellent	91.37%
	3. Final = Excellent -> Mid-term = Good -> Individual Presentation = Excellent-> CT-2 = Excellent	89.13%
Good	1. Final = Good -> Mid-term = Good	90.64%
	2. Final = Good -> Mid-term = Medium: Good -> Group Assignment = Excellent || Good	88.13%
	3. Final = Medium -> Mid-term = Good -> Individual Presentation = Excellent	91.51%
Medium	1. Final = Medium -> Mid-term = Bad	93.65%
	2. Final = Bad -> Mid-term = Medium	92.76%
	3. Final = Bad -> Individual Presentation = Excellent ->Mid-term = Medium || Bad	88.17%
Bad	1. Final = Bad -> Individual Presentation = Bad -> Mid-term = Bad	98.45%
	2. Final = Bad -> Individual Presentation = Medium -> Mid-term = Bad	85.71%
	3. Final = Bad -> Individual Presentation = Good -> Mid-term = Bad -> CT-2 = Bad -> CT-3 = Bad	87.5 %

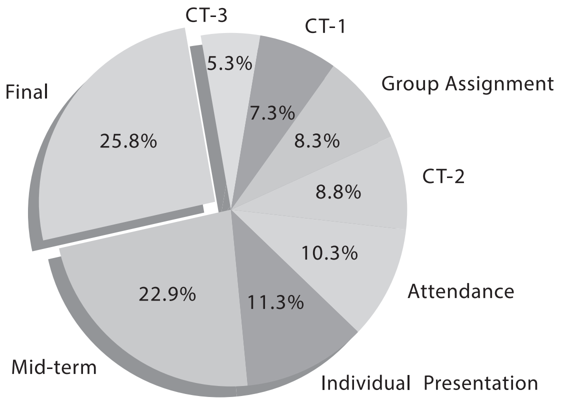

Figure 2: Pie Chart of Info Gain Attribute Evaluation

This rule is true for most of the students who achieved an excellent grade. Similarly, the students who did well in Final and Midterm also did well in total, which has an accuracy of 90.64%. Again, who performed badly in Final, Midterm, and Individual presentation, 98.45% of them did badly in total. From these rules, we can suggest that not only Final and Midterm exam played a crucial role to achieve any grade, but Individual Presentation and CT-2 also played a partial role. Therefore, apart from Final exam and Midterm exam, we also should give emphasis on Individual Presentation and CT-2. Furthermore, we performed Info Gain Attribute Evaluation [16] on the dataset to identify the pairwise dependency each of the marks column and Grade. We depicted the result in figure: 2. From the figure, we identified that Final, Midterm, Individual presentation, Attendance, CT-2, Group Assignment, CT-1, and CT-3 marks affect the final Grade by 25.8%, 22.9%, 11.3%, 10.3%, 8.8%, 8.3%, 7.3%, and 5.3% respectively. Therefore, Group Assignment, CT-1, and CT-3 marks have little effect on the final grade.

VI. Conclusion & Future work

In this paper, we investigated on finding factors which play significant roles in students' grade. By analyzing students' data provided by Green University of Bangladesh, we came up with combination of factors which affect a student's result the most. It may be obvious that midterm mark and semester final mark played the most significant role in students' final grade. But our analysis found out that CT-2 and individual presentation mark also play significant role in getting excellent grade. Again, the students performed badly if they had poor marks in Final, Midterm, and Individual

Presentation. Therefore, university authorities and the students should give more emphasis on these factors.

In future, we should perform rigorous analysis to increase the classification accuracy. We are also planning to find out course wise dependency such that the performance of a subject may predict the future performance of another related subject.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei. Data Mining Concepts and Techniques, 3rd ed. Waltham: Elsevier Inc, 2012

[2] C. Vialardi, J. Bravo, L. Shafti, A. Ortigosa. "Recommendation in higher education using data mining techniques. Available: eric.ed.gov/?id=ED539088, Retrieved September 7, 2014.

[3] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," in IEEE Access, vol. 5, pp. 15991-16005,2017.

[4] M. Sivasakthi, "Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory Programming Performance." Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017).

[5] T.M. Christian, M. Ayub, "Exploration of Classification Using NBTree for Predicting Students' Performance", 2014 International Conference on Data and Software Engineering (ICODSE).

[6] Kaur, Gurneet, and Er Neelam Oberai. "A Review Article On Naive Bayes Classifier with Various Smoothing Techniques." (2014).

[7] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6.2 (2013): 256-261.

[8] Abdelaal, Medhat Mohamed Ahmed, et al. "Using data mining for assessing diagnosis of breast cancer." Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on. IEEE, 2010.

[9] Abdelaal, Medhat Mohamed Ahmed, et al. "Using data mining for assessing diagnosis of breast cancer." Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on. IEEE, 2010.

[10] M. Wati, W. Indrawan, J.A. Widians, N. Puspitasari, "Data Mining For Predicting Students' Learning Result", 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT).

[11] Sundar, PV Praveen. "A Comparative Study For Predicting Students Academic Performance using Bayesian Network Classifiers." IOSR Journal of Engineering 3.2 (2013): 37-42

[12] Anuradha, C., and T. Velmurugan. "A comparative analysis on the evaluation of classification algorithms in the prediction of students performance." Indian Journal of Science and Technology 8.15 (2015).

[13] Mueen, Ahmed, Bassam Zafar, and Umar Manzoor. "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques." International Journal of Modern Education & Computer Science 8.11 (2016).

[14] Saa, Amjad Abu. "Educational Data Mining & Students' Performance Prediction." International Journal of Advanced Computer Science & Applications 1.7: 212-220.

[15] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6.2 (2013): 256-261

[16] Azhagusundari, B., and Antony Selvadoss Thanamani. "Feature selection based on information gain." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2.2 (2013): 18-21.

Md Ashaduzzaman received his B.Sc. Engineering degree in Computer Science and Engineering (CSE) from University of Dhaka, in 2015. He is currently working as a Sr. Lecturer of CSE Department in Green University of Bangladesh. His research interests include Data Mining, Machine Learning and Natural Language Processing.

Shihabuzzaman received his B.Sc. Engineering degree in Computer Science and Engineering (CSE) from Islamic University of Technology, in 2015. He is currently working as a Sr. Lecturer of CSE Department in Green University of Bangladesh. His research interests include Data Mining, Machine Learning and Wireless Sensor Network.

Md. Hasanur Rahman Sagor received his B.Sc. Engineering degree in Computer Science and Engineering (CSE) from Green University of Bangladesh, in 2018. His research interests include Data Mining, Artificial Intelligent and Machine Learning.

Md. Mizanur Rahman received his B.Sc. Engineering degree in Computer Science and Engineering (CSE) from Green University of Bangladesh, in 2018. His research interests include Data Mining, Machine Learning and Wireless Sensor Network.

Ahmed Iqbal Pritom received his B.Sc. Engineering degree in Computer Science and Engineering (CSE) from Islamic University of Technology, in 2015. He is currently working as a Sr. Lecturer of CSE Department in Green University of Bangladesh. His research interests include Human Computer Interaction, Data Mining and Wireless Sensor Network.