Performance Analysis of various Data mining Algorithms in Educational Domain Datasets

Nandini N

Abstract— Educational data mining applications are widely accepted now a day as they will help in analyzing and predicting information's useful for enhancing educational growth. One of the major applications of this kind is the prediction of student performance in higher education. This will help the stakeholders to understand the effect of various factors in academic performance thereby enabling them to take immediate and adequate remedial actions. This research aims to understand the various attributes and their impact on the students' academic performance. A synthetic dataset is chosen to experiment with the various data mining algorithms. Further a real time data set collected from a high school is also experimented with similar algorithms.

Index Terms— Clustering, correlation, data mining algorithms, educational data mining, KStar, PART, WEKA.

1 INTRODUCTION

IN the current scenario where everyone uses various electronic devices to collect data for variety of applications including monitoring and surveillance for security and health reasons, the amount of data generated becomes quite huge. Extracting the required information from this voluminous data is quite difficult. Here comes the importance of data mining algorithms. Data mining technology finds application in all fields wherever huge data is involved.

Educational Data Mining (EDM) is the utilization of Data Mining strategies on instructive information. The target of EDM is to break down such information and to determine instructive research issues. EDM manages growing new strategies to investigate the instructive information, and utilizing Data Mining techniques to more readily comprehend understudy learning condition The EDM procedure changes over crude information originating from instructive frameworks into helpful data that might greatly affect instructive research and practice.

Educational data mining employs a great role in enhancing the various aspects in educational domain say from predicting the students' performance and instructor excellence or even for administrative enhancements and proper resource utilization [1]. There are many areas where Educational Data Mining are used like Analysis and perception of information, Providing criticism for supporting educators, Recommendations for understudies, Predicting understudy execution, Student displaying, Detecting bothersome understudy practices, Grouping understudies, Social system investigation, Developing idea maps, Constructing courseware and Planning and booking[2] .

2 CLASSIFICATION ALGORITHMS IN DATA MINING

Various classification algorithms are used to test the data set they are explained briefly in this section.

2.1 Multilayer Perceptron

A multilayer perceptron (MLP) is a neural system interfacing numerous layers in a coordinated diagram, which implies that the sign way through the nodes just goes one way. Every node, aside from the info nodes, has a nonlinear enactment work. A MLP utilizes back propagation as a managed learning system. Since there are numerous layers of neurons, MLP is a profound learning strategy. MLP is generally utilized for taking care of issues that require regulated learning just as examination into computational neuroscience and parallel appropriated preparing. Applications incorporate discourse acknowledgment, picture acknowledgment and machine interpretation [3].

2.2 Naïve Bayes 00

The Naive Bayesian classifier depends on Bayes' hypothesis with the autonomy presumptions between indicators. A Naive Bayesian model is anything but difficult to work, with no confounded iterative parameter estimation which makes it especially helpful for exceptionally enormous datasets. In spite of its effortlessness, the Naive Bayesian classifier regularly does shockingly well and is broadly utilized in light of the fact that it frequently beats increasingly refined characterization techniques [4].

2.3 KStar

This is an occurrence based classifier that is the class of a test case depends on the class of those preparation occasions like it, as dictated by some similitude work. It varies from other case based students in that it utilizes an entropy-based separation work [5].

2.4 PART

It is a different and-vanquish rule student. The calculation creating sets of rules called „decision lists" which are arranged arrangement of rules. Another information is contrasted with each standard in the rundown thusly, and the thing is allocated the class of the main coordinating guideline. PART manufactures an incomplete choice tree in every cycle and makes the "best" leaf into a standard [6].

2.5 J48

J48 is an expansion of ID3. The extra highlights of J48 are representing missing qualities, choice trees pruning, constant property estimation ranges, inference of rules, and so on. In the WEKA information mining instrument, J48 is an open source Java execution of the C4.5 calculation [7].

The rest of the paper is organized as follows. Section 2 gives the overview of the related works, Section 3 mention about the tools and data set used. Section 4 describes the experimental results obtained followed by conclusions and future scope in Section 5. References are also mentioned.

3 RELATED WORKS

Educational data mining is quite important now a day as it helps to predict the exact results will helps in planning and improvement. Many researches is happening in this area which helps pupil to improve their scores or focus more on their areas of interest or to explore more on what they are good at. Kaur et al. had experimented on a real time dataset from a high school. They used data mining prediction and classification algorithms on this real time data set and was able to analyze the student performance and predict the slow learners [1]. A similar work in Malaysian context was performed by Shahiri et al. [8]. According to the authors the attributes selection for performance analysis is very important. They had given detailed information regarding the same along with the various algorithms suitable for effective performance prediction. Educational data mining is not related to student level achievement predictions or performance analysis alone rather it can be used for instructor and administrative level performance analysis also. Such an analysis using different classifier models for evaluating the success of instructor was done in [9]. According to the authors in [10] educational data mining can do a lot for the overall improvement of the educational institutions. It can be used to assess student performance for increasing passing rates, improve the institution performance, optimizing the resource utilization and even for curriculum updates. They had given an overview of the different effective educational data mining methodologies. In [11] an organized review of the data mining algorithms in educational data mining was done. The use of clustering as a pre-processing step before the application of data mining algorithms was emphasized here. The student background and social activities also have an impact on their academic success. This was proved in [12] where the authors analyzed these attributes and the relationship with academic success. The use of co-training semi supervised learning in order to predict the undergraduate student performance was done by the authors in [13]. For this study the student characteristics and academic achievements and also the involvement in online tutorials are considered. A similar study was conducted by Polyzou and Karypis [14] wherein the focus was to more accurately predict the poor performers. They had used specific attributes which will add accuracy for the predicted result.

4 METHODOLOGY

WEKA is an assortment of AI calculations for data mining errands. The algorithms available in the WEKA can either be applied straightforwardly to a dataset or called from your own Java code. There are some tools available in WEKA for data like pre-processing, clustering, regression. Etc. [15] .The data set is downloaded from report builder the link is given below [16].
https://wpreportbuilder.com/examples/students-exam-marks-list-generate-excel-xlsx/. The student report cards were collected from a high school in a rural area of Tamilnadu. The various characteristics influencing the students' academic performance were taken as the input. It is not a single factor that affects the performance rather a combination of factors like socio-economic – cultural factors also have an impact. The data was filtered using manual techniques and saved as an artiff file. The open source tool WEKA was used for experimentation which contains many machine learning algorithms for data mining [17].Table 1 shows the statistical result of collected data set. Table 2 shows comparison of the classification algorithm for the collected data set. Table 3. Show the statistical analysis of downloaded data set. Table 4 shows the comparison of the classification algorithm for a downloaded data set.

5 RESULTS

Data set is collected from high school and downloaded from the report builder then tested using five classification algorithm those are Multilayer Perception, Naïve Bayes, KStar, PART and J48. Results are provided in table 1 and table 2.

TABLE 1
COLLECTED DATA SET FROM HIGH SCHOOL

Name of the classification algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	PRC Area
Multilayer	P	1.000	0.273	0.961	1.000	0.980	0.997
Percetron	F	0.727	0.000	1.000	0.727	0.842	0.904
Naive	P	0.932	0.182	0.971	0.932	0.951	0.997
Bayes	F	0.818	0.068	0.643	0.818	0.720	0.924
KStar	P	1.000	0.455	0.936	1.000	0.967	0.998
	F	0.545	0.000	1.000	0.545	0.706	0.962
PART	P	0.986	0.182	0.973	0.986	0.980	0.994
	F	0.818	0.014	0.900	0.818	0.857	0.771
J48	P	0.986	0.182	0.973	0.986	0.980	0.994
	F	0.818	0.014	0.900	0.818	0.857	0.771

TABLE 2
COMPARISON CLASSIFICATION FOR THE COLLECTED DATA SET

Data Mining Techniques	Accuracy
Multilayer perception	96.39%
Naive Bayes	92.25%
KStar	94.00%
PART	97.01%
J48	97.01%

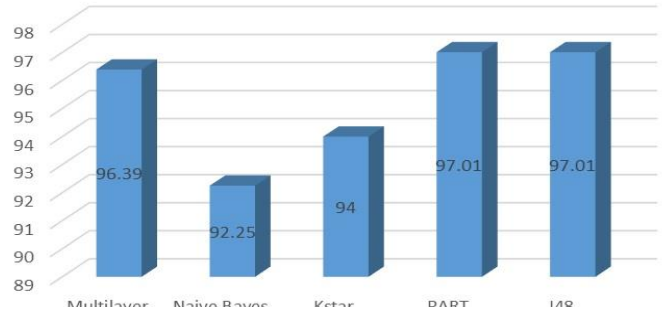

Fig-1.Graph representation for comparison of classification algorithm on collected data set

TABLE 3
DOWNLOADED DATA SET

Name of classification algorithm	class	TP Rate	FP Rate	Precision	Recall	F-Measure	PRC Area
Multilayer perception	P	0.00	0.031	0.000	0.00	0.000	0.190
	F	0	1.000	0.816	0	0.886	0.795
		0.969			0.969		
Naive Bayes	P	0.429	0.156	0.375	0.429	0.400	0.283
	F		0.571	0.871		0.857	0.809
		0.844			0.844		
KStar	P	0.714	0.188	0.455	0.714	0.556	0.501
	F		0.286	0.929		0.867	0.926
		0.813			0.813		
PART	P	0.00	0.031	0.000	0.00	0.000	0.191
	F	0	1.000	0.816	0	0.886	0.834
		0.969			0.969		
J48	P	0.00	0.031	0.000	0.00	0.000	0.191
							0.834

Fig-1.Graph representation for comparison of classification algorithm on collected data set

TABLE 4
COMPARISON CLASSIFICATION FOR THE DOWNLOADED DATA SET

Data Mining Technique	Accuracy
Multilayer perception	81.92
Naive Bayes	75.25
KStar	82.17
PART	71.83
J48	76.42

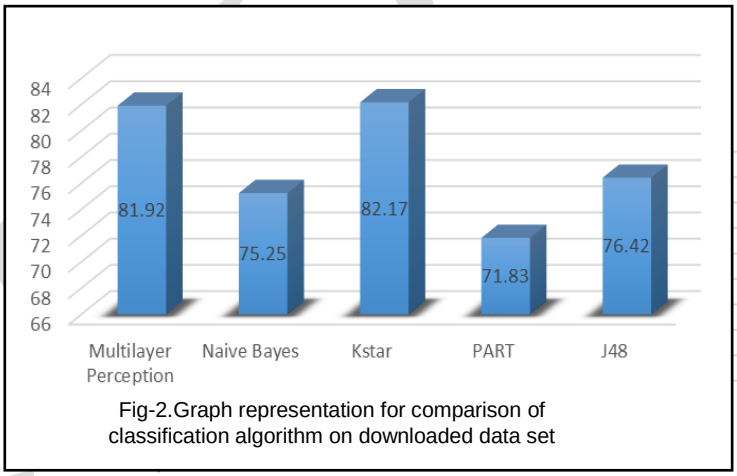

Fig-2.Graph representation for comparison of classification algorithm on downloaded data set

6 CONCLUSION

From the simulations results it was observed that the KStar is giving highest frequency in the downloaded data set and also it was observed that the PART and J48 is giving highest frequency in the collected data set. As an extension of this work the various analysis of relationship between the various attributes in finding the slow learns, the relationship between attendance and the student performance in various subjects etc will be decided. This will also be used for predicting the performance of each student well in advance so as to enable the authorities to take appropriate remedial actions to improve the overall performance.

ACKNOWLEDGMENT

I would like to thank Prof. Deepa V.Jose my research giude and Sri Barathi Vidhyalaya hr.sec School, Hosur, Tamil nadu.

REFERENCES

[1] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500-508, 2015. (Published)

[2] B. Namrata and Niteesha harma, "Educational Data Mining – Applications and Techniques," *International Journal of Latest Trends in Engineering and Technology* (IJLTET), vol. 7 issue 2 July 2016. (Published)

[3] Techopedia.com, "what is Multilayer perceptron(MLP)? – Definition from Techopedia," available at

http://www.techopedia.com/definition/20879/multilayerperceptron-mlp, 2019. (URL Link)

[4] Saedsayad.com, "Naïve Bayesian," available at https://www.saedsayad.com/naive_bayesian.htm, 2019.

[5] Wiki.pentaho.com, "KStar – Pentaho Data Mining – Pentaho Wiki," available at https://wiki.pentaho.com/display/DATAMIN/KStar, 2019.(URL Link)

[6] Dr. Vaishali S. Parsania, Dr. N. N. Jani and Navneet H Bhalodiya, "Applying Naiven bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis," *International Journal of Darshan Institute on Engineering Research and Emerging Technology*, vol.3, No. 1, pp. 60-64. (Published)

[7] Gaganjot Kaur and Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *International Journal of Computer Applications* (0975-8887) vol 98 – No.22, July 2014. (Published)

[8] A. M. Shahiri, W.Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414-422, 2015. (Published)

[9] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," *IEEE Access*, vol. 4, pp. 2379-2387, 2016. (Published)

[10] W. S. Ng, "Web Data Mining in Education," No. 6, pp. 58-77, 2016 (published)

[11] A. Dutt, M. A. Iamail, and T. Heraw Systematic Review on Educational Data Mining," IEEE Access, vol. 5, pp. 15991-16005, 2017. (Published)

[12] C. C. Ktu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," *2018 Fourth Int. Conf. Adv. Comput. Commun. Autom.*, pp. 1-5, 2019 (Published)

[13] G. Kostopoyulos, S. Karlos, and S. Kotsiantis, "Multiview Learning for Early Prognosis of Academic Performance: A Case Study," *IEEE Trans. Learn. Technol.*, vol. 12, No. 2, pp.212-224, 2019. (IEEE Transactions)

[14] A. Polyzou and G. Karypis, "Feature Extraction for Next-Term Prediction of Poor Student Performance," *IEEE Trans. Learn. Technol.*, vol. 12, No. 2, pp.237-248, 2019. (IEEE Transactions)

[15] Learning, W. *Weka – Graphical User Interference Way To Learn Machine Learning*. Analytics Vidhya. Available at https://www.analuticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/ 2019. (URL link *include year)

[16] Report Builder. *Students Exam Marks List – Excel XLSX – Report Builder*. Available at https://wpreportbuilder.com/examples/students-exam-marks-list-generate-excel-xlsx/ 2019. (URL link)

[17] "WEKA3 Machine Learning Software in Java," Available at https://www.cs.wailkato.ac.nz/ml/weka/. (URL link)