PREDICTION OF ACADEMIC PERFORMANCE OF UNDERGRADUATE STUDENTS IN JABALPUR

[1]Khushbu Gupta, [2]Sonam Katare
[1]Assistant Professor, [2]Student
[1,2]Department of Computer Science and Application
St. Aloysius' College (Autonomous), Jabalpur (M.P.), India

Abstract: Colleges, Institutes and universities have huge amount of students' data and this data is growing rapidly. Educational Data Mining helps enormously to extract hidden knowledge from this students' data. In the past, researchers predicted the students' performance based on their academics' attributes only, but we realized that there are many other attributes which also affects the students' academic performance. We have collected data of students' currently studying in final year as testing data and recently graduated students as training data from Jabalpur city through online survey. This survey consists of 34 questions for final year students and 33 questions for graduated students. Collected data was applied on various classification algorithms such as J48, Naïve Bayes, Random Forest, Random Tree and REPTree using WEKA. The intent of this paper is to find the classifier(s) which gives high accuracy.

***IndexTerms* - EDM, Naïve Bayes, Decision Tree, J48, REPtree, ID3, Random Forest, Random Tree.**

I. INTRODUCTION

Data mining is defined as extracting useful information from abundance of data. The information or knowledge extracted called Knowledge Discovery from Data (KDD). Its advantages have landed its application in numerous fields including e-commerce, bioinformatics and lately, within the educational research which commonly known as Educational Data Mining (EDM).

EDM is defined by The Educational Data Mining community website, www.educationaldatamining.org "as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting, and using those methods to better understand students, and the settings which they learn in"[1].

Educational data mining is a research area which is a package of mathematical and dispositional method forcomprehending the learning process of students and also derives what are the strategies that should be taken up by the teachers to enhance their teaching skills along with the improvement in the education system [2].

EDM inherits properties from different fields like Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine learning, Statics, Database Management System, Computing and Data Mining. It can be considered as interdisciplinary research field which provides knowledge of teaching and learning for more effective education system. Various data mining techniques like prediction, classification, clustering and relationship mining can be applied on educational data to study the behavior and performance of the students. EDM is working widely for research area not only in mining and evaluating the educational data available but even in analyzing the learning from educational data globally.

Prediction is the technique of EDM which helps to predict future state from present state [3]. This technique was used to predict undergraduate retention by Lehr S et al [4].

Classification is another technique in EDM, which is a two way technique (training and testing) that maps data into predefined classes. It's a supervised learning which classifies the testing data according to the outcomes of training data. This technique was used to predict students' academic performance by classifying characteristics of students like drop out, weak and good but lately deteriorated [5].

Classifying students' academic performance and students' characteristics can be useful for educational organizations in many different contexts. Classification models can be evaluated by many different parameters such as accuracy, recall, precision, speed, robustness, scalability and interpretability [6].

II. TOOLS AND TECHNIQUES USED

2.1 Weka

WEKA stands for "Waikato Environment for Knowledge Analysis"[7], developed at the University of Waikato in New Zealand. It is basically suite of machine learning software written in Java and runs on almost any platform. It is a collection of machine learning algorithms which help to mine a lot of data. It contains tools for data preprocessing, classification of data, regression, clustering, association rules and visualization of data.

2.2 J48

J48 algorithm generates the rules for the prediction of the target variable by using the tree classification algorithm; the critical distribution of the data is easily understandable. It is an extension of ID3. The J48 algorithm is WEKA's implementation of the C4.5 decision tree; it generates the rules from which particular identity of that data is generated. The extra features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges etc[8]. The algorithm uses a greedy technique to induce decision trees for 20 classification and uses reduced-error pruning method.

2.3 Naïve Bayes

Naive Bayes algorithm is based on Bayes' Theorem. In this algorithm, dataset is divided into two parts, feature matrix and the response vector. The Naive Bayes classification method assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. It is easy to build and very useful for large data sets.

2.4 Random Forest

Random forest algorithm is developed by Leo Breiman and Adele Cutler, In Random Forest algorithm combines the "Bootstrap aggregating" method and "random subspace method" to build a set of decision trees. In the construction of a single decision tree, the random forest algorithm uses two random selection methods: the first is the random selection of training samples (called features or variables), and the second is the random selection of the characteristics attributes of the sample. After all the decision trees are generated, the final classification result is decided by the equal-weight voting method.

2.5 Random Tree

Random tree algorithm introduced by Leo Breiman and Adele Cutler. It is a supervised Classification technique, it employs the idea to construct a random set of data for constructing a decision tree. The algorithm can work with both classification and regression problems. It is essentially the combination of two existing algorithms single model trees are merged with Random Forest tree. It is the group of tree predictors that is called forest. It gets the input feature vector, classifies with every tree in the forest, and generates output class which received the majority of "votes"[9]. Output is created by two ways of randomization first, the training data is sampled with replacement for each single tree like in Bagging. Second, when growing a tree instead of computing the best possible split for each node only a random subset of all attributes is considered at every node, and the best split for that subset is computed.

2.6 REPTree

REPTree(Reduced Error Pruning Tree) classification is fast decision tree learning algorithm and it builds a decision tree based on the principle of computing the information gain with entropy or reducing the error arising from variance. REP Tree applies regression tree logic and creates multiple trees in different iterations[10]. Afterwards it selects the best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. At the beginning, it only sorts values for numeric attributes. Missing values are dealt with using C4.5 algorithm by splitting the corresponding instances into pieces.

III.PROPOSED METHODOLOGY AND DATASET

An online survey cum experimental methodology is used. The collected data from survey is possible values of some attributes related to students' academic, family, behavior, health and social life. The collected students' data is from different academic organization in Jabalpur City. This study used those students' data that had enrolled in the academic batches of 2015-16 and 2016-17.

Following table shows list of attributes and their possible values:

Attributes	Description	Values
attitude	Attitude	{ Positive, Negative }
talking	Way of Talking	{ Slowdown, Breath, Quickly, Nervousness }
frnds	Friend's Type	{ Helping, Understandable, Bad }
nature	Nature	{ Extroversion, Directness, Bad }
famatm	Family Atmosphere	{ Positive, Negative }
famtyp	Family Type	{ Joint, Nuclear }
class	Class	{ Poor, Lower_Middle, Middle, Upper_Middle }
spndtime	Spent Time	{ More, Average, Little }
addicted	Addicted to	{ Drug, Social_Media, Music, Dance, Smart_Phone, Television, Reading, Smoking, Alcohol, Other }
tmspnt	Time Spent	{ More, Average, Little }
parc	Participation in Extra-curricular activity	{ Yes, No }
eact	Type	{ Sports, Cultural, NCC, NSS, Other, No }
punc	Punctuality	{ Excellent, Good, Average, Bad }
atd	Attendance	{ Excellent, Good, Average, Bad }
tcp	Theory Class Performance	{ Excellent, Good, Average, Bad }
pcp	Practical Class Performance	{ Excellent, Good, Average, Bad }
ten	SSCG (10th Grade)	{ A=90-00%, B=80-89%, C=70-79%, D=60-69%, E=less than 60% }
fyr	1stYear Grade	{ A=90-00%, B=80-89%, C=70-79%, D=60-69%, E=less than 60% }
syr	2ndYear Grade	{ A=90-00%, B=80-89%, C=70-79%, D=60-69%, E=less than 60% }
thyr	3rd Year Grade	{ A=90-00%, B=80-89%, C=70-79%, D=60-69%, E=less than 60% }

satlvl	Student's satisfaction level with their UG course	{ V_Satisfied, Satisfied, Not_V_Satisfied, Not_Satisfied }
stdy	Study Duration	{ More, Average, Little }
mart	Marital Status	{ Single, Married }
wrkg	Working	{ Yes, No }
conft	Confidence Level	{ Over, High, Average, Low }
clgbck	College Background	{ Good, Average, Poor }
hndwrtg	Hand Writing	{ Good, Average, Poor }
hndicap	Physical Handicapped	{ Yes, No }
cclass	Do Coaching Classes	{ Yes, No }
unbrstnd	Understanding Level	{ High, Medium, Low }
hlthprb	Health Problem	{ Yes, No }
Aggr	Aggregate Result	{ Excellent, Good, Average, Bad }

Following figure shows the stepwise process of predicting academic performance of a student:

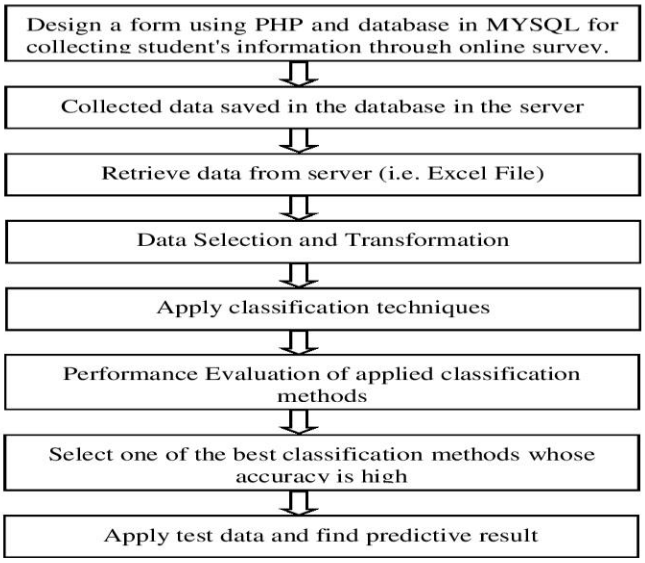
Design a form using PHP and database in MYSQL for collecting student's information through online survey.
Collected data saved in the database in the server
Retrieve data from server (i.e. Excel File)
Data Selection and Transformation
Apply classification techniques
Performance Evaluation of applied classification methods
Select one of the best classification methods whose accuracy is high
Apply test data and find predictive result

fig. stepwise process of predicting students' academic performance

IV.EXPERITMENT

We have applied training set in the following classifiers using WEKA tool then we have applied testing data for prediction of academic performance of final year student. The performance of the applied classifiers is shown in the following table:

Table. Performance of classifiers

Classifiers	Time taken to build model (seconds)	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	TP-Rate (Avg)	FP-Rate (Avg)	Precision (Avg)	Recall (Avg)	ROC area (Avg)	F-measure (Avg)
J48	0.17	64.42	35.57	0.644	0.109	0.648	0.644	0.817	0.628
NavieBayes	0	51.92	48.07	0.519	0.137	0.538	0.519	0.792	0.515
Random Forest	0.16	76.92	23.07	0.769	0.062	0.776	0.769	0.847	0.758
Random Tree	0	77.88	22.11	0.779	0.059	0.784	0.779	0.86	0.767
REPTree	0.03	46.15	53.84	0.462	0.154	0.522	0.462	0.725	0.459

V.RESULT

Random Tree and Random Forest classifier give high performance as compared to other classifiers and J48 gives high accuracy.

VI.CONCLUSION

Predicting academic performance of any student will help the academic organizations for keeping the prior knowledge of the student about their performance in final year. So, the academic organizations will give extra attention to those students whose predicted value is poor. As a result student will strive to get high grade. It will also help the academic organizations to make students achieve 90 to 100 % result. Here we have used various classification algorithms with WEKA tool for predicting students' academic performance, evaluated the outcomes of the algorithms based on their predictive accuracy and performance. As result Random Tree classifier had high accuracy and high performance.

VII.REFERENCE

[1] Mohamad, S. K. and Tasir, Z. 2013.Educational data mining: A review. Procedia - Social and Behavioral Sciences (Published by Elsevier), 320-324.

[2] Jayanthi, M. A., Surendran, A., Kumar, R. L. and Prathap, K.2016.Research Contemplate on Educational Data Mining. IEEE International Conference on Advances in Computer Applications (ICACA), 110-114.

[3] Jindal, R. and Borah, M.D. 2013.A Survey on Educational Data Mining And Research Trends. International Journal of Database Management Systems (IJDMS), 53-73.

[4] Lehr, S., Liu, H., Klinglesmith, S., Konyha, A., Robaszewska, N. and Medinilla, N. 2016. Use Educational Data Mining to Predict Undergraduate Retention. IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 428-430.

[5] Roy, S. and Garg, A. 2017. Predicting Academic Performance of Student Using Classification Techniques. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura.

[6] Buniyamin, N., Mat U. b. and Arshad, P. Md. 2015. Educational Data Mining for Prediction and Classification of Engineering Students Achievement. 2015 IEEE 7th International Conference on Engineering Education (ICEED '15), 49-53.

[7] https://en.wikipedia.org/wiki/Weka_(machine_learning)

[8] Chauhan, Y. and Vania, J. 2013. J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient. International Journal of Computer Trends and Technology (IJCIT), 305-309.

[9] Mishra, A.K. and Ratha, B.K. 2016. Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis. International Journal on Advanced Electrical and Computer Engineering(IJAECE).

[10] Kalmegh, S.K. 2015.Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News. International Journal of Innovative Science, Engineering & Technology (IJISET).