DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4576

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANALYSIS OF STUDENT'S ACADEMIC PERFORMANCE USING CLASSIFICATION ALGORITHM IN WEKA

Twinkle Chawla
Research Scholar
Punjabi University, Patiala
Chandigarh,India

Gurpreet Singh
Assistant Professor, CE
Punjabi University,
Patiala,India

Abstract: As we have extensive measure of information in industry so it is important to investigate the data and extract the useful information by applying distinctive data mining techniques. Data mining is used in many fields, mining related to education is called EDM. All the institutions aimed to provide good quality education to its student. Extraction of knowledge with the help of data mining techniques helps students to know their weakness and to improve it. For better results analyse the academic performance of students and the performance will depend upon various factors like annual income of family, qualification of mother, marks of 10^{th} and 12^{th} and so on. In this study we use techniques like Random Tree, J48, Random Forest, REP Tree in WEKA. These techniques are used to build the model and to generate results in WEKA. These classification algorithms are compared based on students' social conditions, previous academic records using WEKA. The records of 175 computer engineering students are used to build the model. Random Forest with highest average accuracy 71.4% among other.

Keywords: Data mining, Educational Data mining, WEKA, J48, and REP Tree

1. INTRODUCTION

Data mining has attracted lot of attention in the research industry due to tremendous accessibility of huge measure of information and the requirement for transforming such information into valuable data and learning. Data mining, additionally called knowledge discovery in database (KDD), is the field of finding new and conceivably helpful data from immense database.

Educational data mining (EDM) is utilized to find information with respect to variables influencing understudy execution, understudies learning conduct and expectation of their execution from the educational data set. EDM is a utilization of information mining, which is a piece of the KDD procedures used to find designs from given informational index. EDM is a process used to extract useful knowledge and find the hidden patterns from a huge educational database. The derived information and the patterns will be used in predicting student performance. Remembering the ultimate objective to encounter the issues, a purposely review is proposed. The proposed productively study is to help the objectives of this examination, which are:

- To consider and see the separated in existing prediction methods.
- To study and find the variables which are used in analysing student academic performance.
- To study the existing method of predicting student performance.

The research presented in this paper was performed on the data collected from the B.Tech students of Department of Computer Engineering, Punjabi University. The data collected from students via a structured questionnaire having 30 attributes regarding social conditions and previous marks of all the students. Classification algorithms like J48, Multilayer perceptron, Naïve Bayes, REP tree, Random Forest were used to analyse the data set. All the techniques were compared with each other and find the best technique that means we want best accuracy. The main objective of this research is to find weak students which are on risk so that we can give some remedial action to improve their academic performance.

2. LITERATURE SURVEY

Arshad et al.(2011) proposed a model to anticipate understudies general CGPA. They have connected neural systems with longitudinal advance. They utilized information of 93 students of 2005 cluster, 172 of 2006 cluster and 198 of 2007 cluster who entered engineering after registration in first year and 183 of 2006 cluster, 251 of 2007 cluster and 166 of 2008 cluster who joined engineering in second year after certificate. Their outcomes demonstrate that there is a connection between general CGPA and first semester execution. There is no connection between gender and the last CGPA. [1]

Alaa M. El-Halees et al. [2012], proposed an investigation that Educational data mining utilized as a part of instructive space for finding learning to create techniques from information. They connected instructive information mining to expand performance of graduate students and to determine the performance of poor student's. for their situation consider they take helpful information from information of graduate understudies that was gathered from the school of science and innovation. The information contains fifteen years' time frame. In the wake of pre-processing, the information, classification rules were applied. In each of these they give the removed information and its incentive in instructive field. Classification naïve bayess gave 67.50% accuracy and base induction gave 70 % accuracy. [2]

Angeline DM directed an examination o the understudies execution by utilizing Apriori calculations that concentrates the arrangement of standards particular to each class and

break down the offered learning to order the researcher in light of their contribution in task, intenal assessment test, bunch activity and so forth. It distinguishes the understudies execution go like normal, beneath normal, what's more, great execution. [3]

J K Jothi and K Venkatalakshmi directed the understudies execution investigation on the graduate understudies information gathered from the Villupuram School of Engineering and Technology. The information included five year time span and connected bunching techniques on the information to beat the issue of low score of graduate understudies, and to raise understudies scholastic execution. [4]

Kumar S. Anupama, Dr. Vijayalakshmi M.N proposed C4.5 choice tree calculation can be utilized on characteristics of the understudies and forsee their execution as far as pass or fail in final exam. The anticipated outcomes and real outcomes which demonstrates, that there was a huge change in comes about as the forecast helped a considerable measure to recognize the week and good students and help them to score better marks. The ID3 choice tree calculation is better regarding effectiveness and time taken to manufacture the choice tree. [5]

R.Shanmuga Priya directed examination on enhancing the understudy execution utilizing Educational Data Mining based by choosing 50 understudies from Hindustan College of Arts and Science, Coimbatore, India. By utilizing decision tree order on 8 trait, it was discovered that the class test, course, participation, lab practical's are utilized to anticipate the understudy execution. This forecast will help to the instructor to give uncommon consideration of understudies and progress understudy certainty on their investigations. [6]

Sharabiani et al.(2014) built a model to anticipate understudies scholastic execution utilizing Bayesian Networks (BN) structure. They will likely distinguish three significant courses that understudies take in second semester and anticipate understudies evaluation in these courses to recognize powerless understudies. The information of 300 designing understudies at UIC is gathered, 70% of this information is utilized as preparing set and staying 30% as testing set. The outcome demonstrates that the exactness of their model with BN is higher than the ordinary models(Naïve Bayes, Artificial Neural System, decision tree, K-nearest neighbour). [7]

Sajadin et al directed an examination on analyse the connections between understudy behavioural and their achievement and to build up the improvement of understudy execution indicator by utilizing Smooth Support Vector Machine (SSVM) characterization and bit k-implies clustering procedures. They discover there is a solid connection between mental state of understudy and their last scholarly execution. [8]

Vaibhav P.Vasani et al [2014], proposed an examination on grouping information gathered from polytechnic foundation. This information was pre-handled to delete useless, irrelative and missing properties. Brilliant, normal, powerless diverse classes of understudies were finished using decision trees and naïve bayes calculations. They contrasted consequences of order with deference with various executions elements. They uncovered that decision tree is superior to naïve Bayesian with 95% calculation. [10]

3. DESIGN METHODOLOGY

In this section we describe the architecture of the system, tool used in the methodology, algorithms and other research methodology.

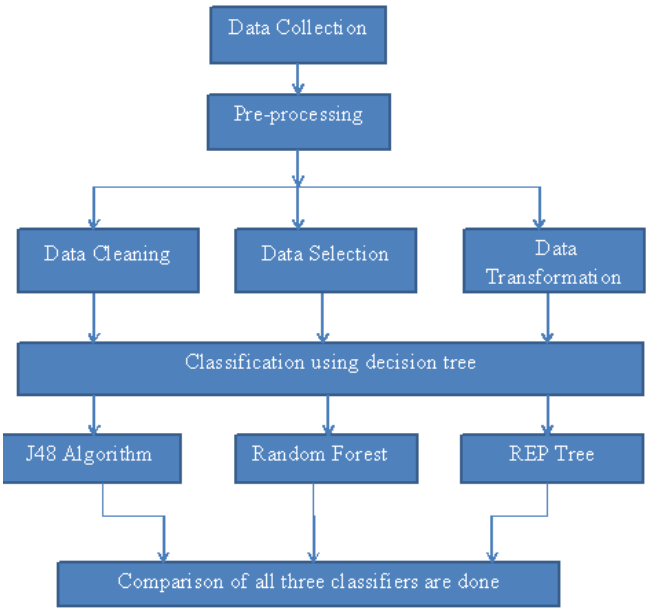

Fig1: Steps of Research Methodology

A. Data Preperation and Selection

Student related data are collected from Punjabi University via a structured questionnaire. The questionnaire includes 23 attributes that were selected from the previous studies done in the area of educational data mining as shown in Table 1. These attributes are related to student's social condition, their family annual income and their previous marks.

Table 1 Attribute Description

S.no	Description	Possible Values
1	Student's School board	CBSE, PSEB,ICSE,SLIET,HP,UP,J&K
2	Marks obtained in 10th class	EXCLT, GOOD, AVG, BAVG
3	Marks obtained in 12th class	EXCLT, GOOD, AVG, BAVG
4	Father's Qualification	N(none), M(Matric), T(12th), G(graduate), PG(post graduate)
5	Mother's Qualification	N(none),M(Matric),T(12th), G(graduate), PG(post graduate)
6	Father's Occupation	A(Agriculture),B(Businessman), J(JOB)
7	Mother's Occupation	HW(Housewife), J(JOB)
8	Student's family annual income	A(up to 5 lakh), B(up to 10 lakh), above 10 lakh
9	Performance	EXCLT, GOOD, AVG, BAVG
10	No. of backlogs	0, 1, 2, above 2
11	Loan on study	Yes, No
12	Taking extra tuitions for study	Yes, No
13	Living stay	PG(paying guest),DS(day scholar), H(Hostel)
14	No. of members in your family	3, 4, 5, above 5
15	Status of family	J(joint), I(individual)
16	Interested in higher education	Yes, No
17	No. of siblings	1, 2, 3, above 3
18	Total hours spend on studies	Below 3, Above 3, Above 5
19	Total hours spend on social sites	Below 3, Above 3, Above 5
20	Student's living location	R(rural), U(urban)
21	Student's food habit	None, drinking, smoking, other
22	Support from friends in studies	Yes, No
23	Is student suffering from any stress	Yes, No

We can group the student semester marks in following way:

- we group them into 5 classes, "BAVG" representing grades below 70, "AVG" representing grades from 70.8 to 77.9, "GOOD" representing grades from 77.94 to 85, "EXCLT" representing grades above 85.we have choose this way for the prediction of student's semester grades

B. Classification Algorithms

Classification is a data mining technique generally utilized for the predictive data mining task. This classification procedure is used to group all information into the predefined classes. This technique has different classifier to classify the data like decision tree, bayes function and so forth. Decision tree classifier represents the instance in type of a tree arranged from root to leaf hub. Every hub of the tree represents the attribute and edge descending from this hub represents value of this attribute. J48, Random Tree, REP Tree, Multilayer Perceptron algorithms are compared for eight semesters and best one is utilized to extract rules for the prediction of understudies' execution.

C. J48 Algorithm

J48 is an open source Java execution of the C4.5 calculation created by Ross Quinlan in the Weka data mining apparatus. C4.5 is a program that makes a decision tree in light of an arrangement of named input information. It utilizes greedy technique to produce decision tree. For splitting the data J48 algorithm analyses the normalized information gain. The attribute with highest normalized information gain is used as a node in decision tree and make decision. Basic steps for J48 algorithm are:

1. Check for the base cases: in case if all the events in a subset belong to the same class. Then a leaf centre is settled on in the decision tree.
2. Find the normalized information gain for all attributes from splitting on that attribute.
3. The highest normalized information gain is selected.
4. The node which represents selected attribute creates a decision node and splits on that attribute.
5. Repeat on sub list obtained by splitting on that parameter and add those nodes as its children of node.

D. REP Tree

REP tree utilizes regression logic to make numerous trees in various iterations. After this it selects the best tree from all the created trees which is considered as Agent. For pruning the tree Mean Square Error measure is utilized on the forecasts made by REP tree. It arrange all the numeric fields once toward begin of running and uses these arranged rundown to calculate right parts at every hub. REP tree is a quick choice tree student and builds a choice tree by utilization of information gain as the branching measure. REP tree utilizes general choice tree and diminished error pruning for the order of attributes. Yet, there is a little contrast in arrangement of both numeric and non-numeric attributes.

E. Random Forest

Random Forest are an assemble learning technique for arrangement, relapse and different undertakings, that work by developing a large number of choice trees at preparing time and yielding the class that is the method of the classes or mean forecast of the individual trees. Random decision forecasts correct for decision trees habit of overfitting to their training set. The random tree classifier takes input, classifies it using each tree in the forest and gives output based on the class predicted by majority of trees.

4. EXPERIMENTAL RESULTS

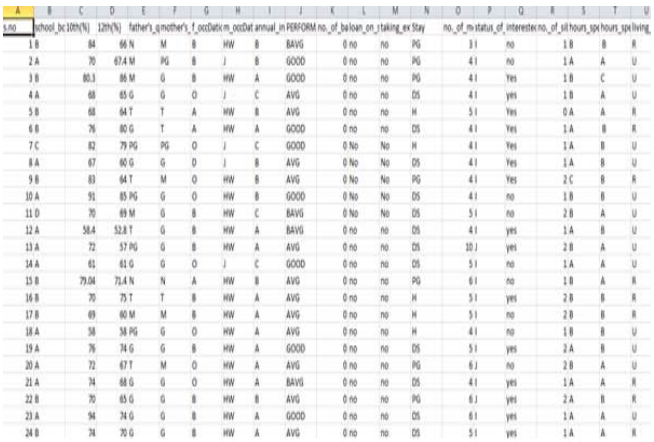

Fig2: Student Data set

fig.2 represents the students' data set collected from a database as well as a survey of approximately 175 students at Punjabi University

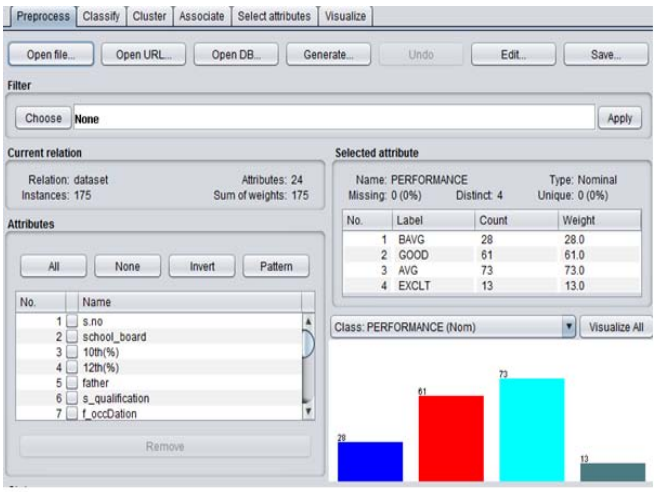

Fig3: Student Performance

Fig.3 shows the performance of students. By observation from the figure it is clearly evident that 73 students comes under AVG. The figure displays that minimum performance of students is 28 and maximum performance of students if GOOD.

The dataset during this work is tested and analyse with three classification algorithms those are J48, Random Forest and REP tree (using percentage split). And then comparison of all three classifiers are done and it is found that Random Forest has highest accuracy 71.4% .

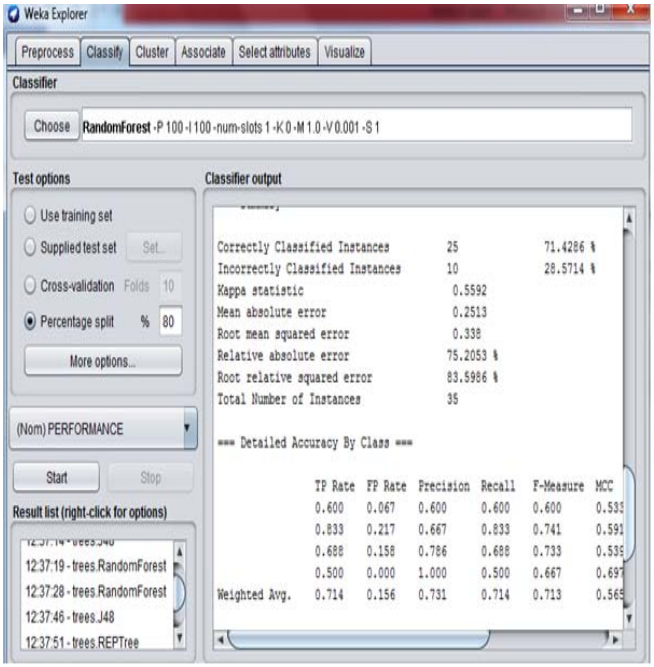

Fig 4: Shows Accuracy of Random Forest Algorithm

Table 2 Comparison of Classifier on the basis of correctly classified instances

Data mining Technique	Accuracy
Random Forest	71.4%
REP Tree	68.5%
J48	68.5%

5. CONCLUSION AND FUTURE SCOPE

The aim of this research is to find the factors that affect student's performance. There are diverse information mining classification algorithm that can be utilized to recognize patterns in the student data set. To accomplish this goal, WEKA tool is used to implement classification algorithms. The total student dataset obtained is 175. The comparison of all three classifier is done and it is found that Random Forest has highest accuracy of 71.4%. This study also concludes that factors like school board, marks of 10th and 12th have highest impact on student performance. On the other hand parental education has less impact on student performance. In future, we have to increase the accuracy and it can be done by improving the quality of data.

REFERENCES

[1] P.M. Arsad, N. Buniyamin1, J.L.A. Manan and N. Hamzah, "Proposed Academic Students' Performance Prediction Model: A Malaysian Case Study, "in 3rd International Congress on Engineering Education (ICEED), 2011, pp.90-94.

[2] Alaa M. El-Halees, Mohammed M. Abu Tair (2012), Mining Educational Data to improve students performance: A case study, International Journal of Information and Communication Technology Research.

[3] D.Magdalene Delighta Angeline, "Association rule generation for student performance analysis using Apriori Algorithm", The SIJ Transactions on Computer Science Engineering And Applications (CSEA), vol. 1, March-April 2013

[4] J.K. Jothi and K.Venkatalakshmi, "Intellectual performance analysis of students by using data mining techniques", International Journal of Innovative Research in Science, Engineering and Technology, vol 3, Special iss 3, March 2014.

[5] Kumar S. Anupama and Dr. Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science & Information Technology 02, pp. 335-343.

[6] J K. Shanmuga Priya" Improving the student's performance using Educational data mining", International Journal of Advanced Networking and Application, Vol.4, pp-1680-1685 (2013)

[7] A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov and H.Darabi, "An Enhanced Bayesian Network Model for Prediction of Students' Academic Performance in Engineering Programs, "in IEEE Global Engineering Education Conference (EDUCON), Harbiye, Istanbul, Turkey, 2014, pp. 832-837.

[8] Sajadin Sembering, M.Zarlis, "Prediction of student academic performance by an application of data mining techniques", International conference on management and Artificial Intelligence-2011.

[9] Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta (2014), Mining Students' Data for performance Prediction, Fourth International Conference on Advanced Computing & Communication Technologies.

[10] Vaibhav P. Vasani, Rajendra D. Gawali (2014), Classification and Performance Evaluation Using Data mining algorithms, International Journal of Innovative Research in Science, Engineering and Technology.

[11] V. Ramesh et.al. (2011). Performance Analysis of Data Mining Techniques for Placement Chance Prediction. International Journal of Scientific & Engineering Research, 2(8).

[12] Z.Zakaria, R.A.Kaim, A.Mohamad and N.Buniyamin. (2011). The Impact of Environment on Engineering Students' Academic Performance: A Pilot Study. 3rd International Congress on Engineering Education (pp. 113-118). Kuala lumpur: IEEE.