IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 6, Ver. II (Nov – Dec. 2015), PP 67-73
www.iosrjournals.org

Empirical Study on Classification Algorithm For Evaluation of Students Academic Performance

Dr.S.Umamaheswari [1], K.S.Divyaa [2]
[1]*Associate Professor, School of IT and Science, Dr.G.R.Damodaran College of Science, Tamilnadu,*
[2]*Research Scholar, Master of Philosophy, School of IT and Science, Dr.G.R.Damodaran College of Science, Tamilnadu,*

Abstract: *Data mining techniques (DMT) are extensively used in educational field to find new hidden patterns from student's data. In recent years, the greatest issues that educational institutions are facing the unstable expansion of educational data and to utilize this information data to progress the quality of managerial decisions. Educational institutions are playing a prominent role in the public and also playing an essential role for enlargement and progress of nation. The idea is predicting the paths of students, thus identifying the student achievement. The data mining methods are very useful in predicting the educational database. Educational data mining is concerns with improving techniques for determining knowledge from data which comes from the educational database. However it has issue with accuracy of classification algorithms. To overcome this problem the higher accuracy of the classification J48 algorithm is used. This work takes consideration with the locality and the performance of the student in education in order to analyse the student achievement is high over schooling or in graduation.*

Keywords: *Educational Data Mining, Performance Metrices.*

I. Inroduction

Data mining (DM) is called as knowledge discovery in database (KDD), is known for its powerful role in discovering hidden information from large volumes of data. Generally, data mining is the search for hidden patterns that could be present in huge databases. Data mining is becoming gradually more important tool to make over this data into information. Educational Data Mining (EDM) develops methods and applies techniques from machine learning, statistics and data mining to analyse data collected during teaching and learning [1]. Educational Data Mining (EDM) is a growing field, concerned with developing methods for recognising the unique characters of data that come from educational surroundings, and applying those methods to better understand students, and helps in decision making. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process. Data resides from the Department of School of Information Technology and science of the Dr.G.R.Damodaran College of Science. The Data collection is done from the student database of B.Sc., (Computer Science) and B.Sc., (Information Technology) for the past 3year's i.e.2010-2012, 2011-2013, 2012-2015. The data is analysed in order to predict the improvement over schooling or in graduation.

II. Evaluation Dataset

The data is collected in two different phases. Initially, the data collected at SSLC, HSC with school name and UG level Mark/Percentage data. Data is collected from the Department of School of Information Technology and science of the Dr.G.R.Damodaran College of Science. The Data collection is done from the student database of B.Sc., (Computer Science) and B.Sc., (Information Technology) for the past 3year's i.e.2010-2012, 2011-2013, 2012-2015. The general attributes are student roll number, name, and gender, date of birth, graduation year, address, phone number, location and city. The specific attributes are school name, school location, student's mark in school, college name, department, college location and student's mark in college. The algorithms are suggested to evaluate the performance of student in school academic and college academic. The location details are such as urban school, urban home, rural school, urban college and rural home for students. The specified dataset which provides more accurate analysis as well as prediction results based on the clustering and classification algorithms. Secondarily the main parameters are considered.

Attribute	Description
Stud_Rollno	Student ID/ Roll Number
Stud_Name	Name of the Student
Gender	The gender of the student
Dob	The date of birth of the student
Enrol_year	The year of enrolment in the college
Gradu_year	The year of graduation from the college
Home_Loca	Location of the student home
Tel_no	The telephone number of the student
HSC_Perc	Percentage in the Higher Secondary Education
HSC_School	School in which the Student have studied
HSC_Loca	Location of the Higher Secondary Education
UG_Perc	Percentage in the Under Graduation
UG_Loca	Location of the UG College
S1,S2,S3…S6	Semester wise mark List
UG_Major	Major of the Degree

Figure 1: The Student Data Set Description

In Data processing the data set used in this work contains graduate students information collected from the college. The graduate student consists of 350 records and 15 attributes. Figure 1 presents the attributes and their description that exists in the data set as taken from the source database. The selected attributes and description are the selected for the analysis process.

- The data set contains some missing values in various attributes from 350 records; the records with missing values are ignored from the data set since it doesn't consider a large amount of data. The number of records is reduced.

After applying pre-processing and preparation methods, analyse the data graphically and figure out the grade of students using MAT Lab. The Data Set is shown in figure 2.

S.No	Stud_Rollno	Stud_Name	Gender	HSC_Perc	HSC_Loc	HOME_LOCA	S1	S2	S3	S4	S5	S6
1	11BIT01	AARTHI V	F	88	U	U	87	85	83	81	77	76
2	11BIT04	ABIRAMI B	F	66	U	U	70	57	62	60	61	67
4	11BIT06	ARASU RAMANAN A	M	63	U	U	61	62	63	64	54	59
5	11BIT08	ARJUN E G	M	60	R	R	70	69	62	53	55	55
6	11BIT09	ARUN MURUGAPPAN	M	60	R	R	66	62	62	57	59	66
7	11BIT10	ARUNKUMAR R	M	91	R	R	79	69	69	77	70	68
8	11BIT11	DARYL JOSEPH C	M	62	U	U	49	45	58	51	55	61
9	11BIT13	GOPINATH T K	M	65	U	U	52	42	43	50	38	60
10	11BIT14	GOWTHAM N R	M	74	U	U	61	58	60	61	57	65
11	11BIT16	HARSHINI A	F	70	R	R	71	67	63	73	68	67
12	11BIT17	INDHUMATHI R	F	89	R	R	79	67	71	79	72	72
13	11BIT18	JAIANAND S A	M	69	U	U	61	54	62	59	53	56
14	11BIT19	JENIFER ROSELIN S	F	73	U	U	83	82	90	84	86	86
15	11BIT20	KEERTHANA M	F	79	U	U	77	84	80	78	73	71
16	11BIT22	KINGSLY P	M	75	U	U	72	72	75	62	63	67
17	11BIT25	MANOJ KUMAR S	M	82	U	R	85	84	84	77	78	75
18	11BIT26	MEENA R	F	58	U	u	67	67	69	71	64	70
19	11BIT27	MITHILA D	F	83	U	u	83	74	82	79	73	69
20	11BIT28	NANDHA KUMAR S	M	59	U	U	67	69	55	62	55	66
21	11BIT29	NAVIN ARAVINTH N	M	82	U	U	59	60	61	56	59	63
22	11BIT31	PAPPU SHANKAR M	M	66	U	R	70	64	66	61	48	47

Figure:2 Sample Data

A. Train Data

The B.Sc., CS 2010-2013 batch dataset which contains 155 number of students. The training has been done on the given dataset which shows the number of students with lowest percentage as 5. The number of students with medium percentage is as 110. And the number of students with highest percentage is as 40. In this dataset, it also considers all semesters such as semester 1 up to semester 6. And it provides low, medium and highest percentages for all semesters. The random forest and J48 algorithm is used to train the specified dataset based on the tree structured format.

B. Test Data

In this section we use the B.Sc., IT 2010-2013 batch dataset which contains counts of total students. The testing has been done on the given dataset which shows the true positive, true negative, false positive and false negative values. Then it is used to predict the model accuracy values for the specified datasets.

III. Classifiction Algorithm

A. Random Forest Classification Algorithm

The Figure 3 shows the information details about the student in the department of computer science of the batch (2010-2013, 2011-2014, 2012-2015). The location of urban home students and urban college

constraints are shown the better performance of students. Total number of students is sixty and 59 student's study performance is increased and one student performance is reduced. The location of urban home students and rural college constraints produced better performance. Total number of students is 17 and 16 student's study performance is increased and one student performance is decreased. The rural home students and urban college conditions are produced good performance. Total number of students is 20 and 18 student's educational performance is increased and 2 student's educational performance is decreased. The location of rural home students and rural college conditions are produced good performance. Total number of students is 58 and 57 student's educational performance is increased and 1 student's educational performance is decreased.

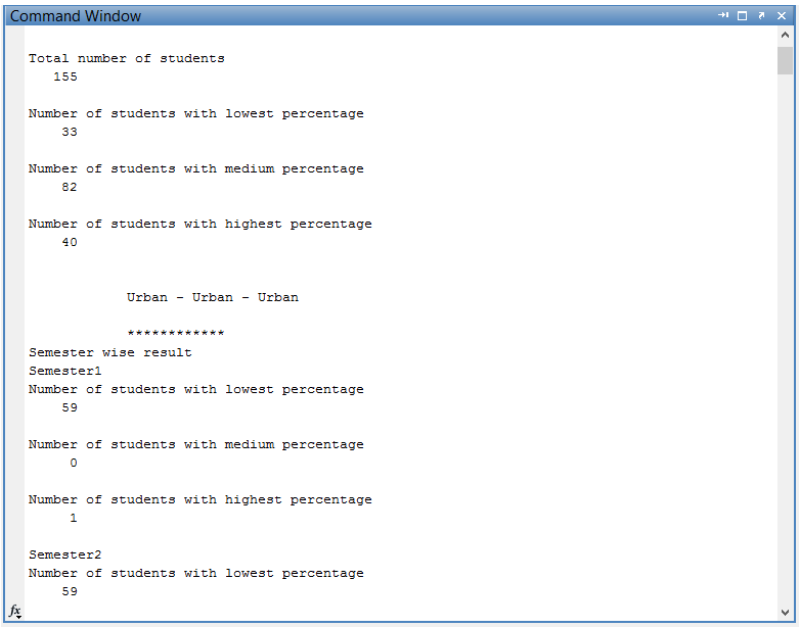

Figure 3 : Random Forest Classification Algorithm

The B.Sc., CS (2010-2013, 2011-2014, 2012-2015) batches dataset which contains 155 number of students. The training has been done on the given dataset which shows the number of students with lowest percentage as 33. The number of students with medium percentage is as 82. And the number of students with highest percentage is as 40. In this dataset, it also considers all semesters such as semester 1 up to semester 6. And it provides low, medium and highest percentages for all semesters.

B. J48 Classification Algorithm

J48 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = s1,s2,... of already classified samples. Each sample si = x1,x2,... is a vector where x1,x2,... represent attributes or features of the sample. The training data is augmented with a vector C = c1,c2,... where c1,c2,... represent the class to which each sample belongs.

At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The J48 algorithm then recurs on the smaller sub lists.

IV. Performance Evaluation

The Performance is evaluated for the existing and the proposed system. The analysis has been done for the Random Forest and J48 classification Algorithms. In the existing system and the Proposed system the accuracy, Precision, Recall, F-Measure is evaluated. From the experimental result, the scenario concludes that the j48 algorithm yields greater accuracy performances. The higher performance are in terms of precision, recall , accuracy and F-Measure metrics. From the Figure 4 describes that the existing and the proposed systems are analyzed using Random Forest and J48 Classification Algorithm.

	Random Forest Classification	J48 Classification
Accuracy	83.3333	96.2406
Precision	0.5174	0.6429
Recall	0.6691	0.9809
F- Measure	0.5836	0.7767

Figure 4 : Comparison Table

The evaluation is performed using the following performance metrices

- Precision
- Recall
- Accuracy
- F-Measure

To implement the proposed method and generate numerous results using mat lab tool in this environment. The scenario has been selected educational dataset to discover the low, medium and high performance of the students. In this section, the analysis has been done for existing and proposed research work by using algorithms. The performance metrics are such as accuracy, precision, recall and f-measure values which are evaluated by using random forest and J48 classification method. From the experimental result, the conclusion decides that the proposed method provides higher performance results in terms of accuracy, precision, recall and f-measure values.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

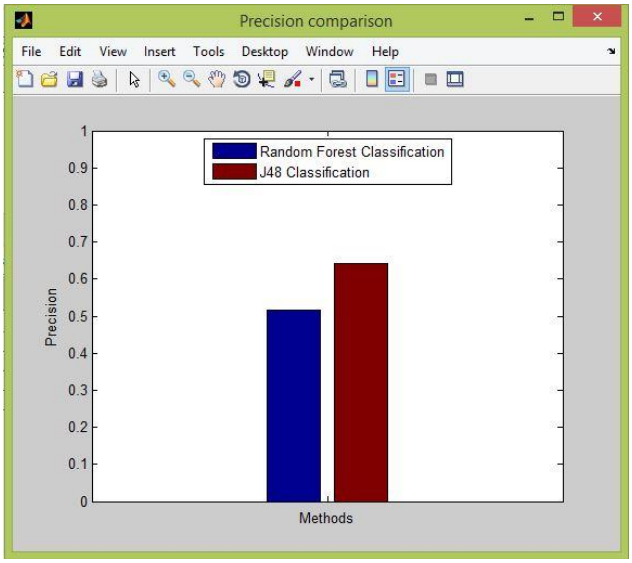

Figure 5 : Precision

From the figure 5 can observe that the comparison of existing and proposed system in terms of precision metric. In x axis we plot the methods and in y axis plot the precision values. In existing scenario, the precision values are lower by using random forest algorithm. The precision value of existing scenario is 0.55 for discover the student's performance. In proposed system, the precision value is higher by using the J48 algorithm. The precision value of proposed scenario is 0.61 for discover the student's performance. Thus it shows that effective analysis is performed by using proposed algorithm. From the result, can conclude that proposed system is superior in performance.

The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

The comparison graph is depicted as follows:

Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

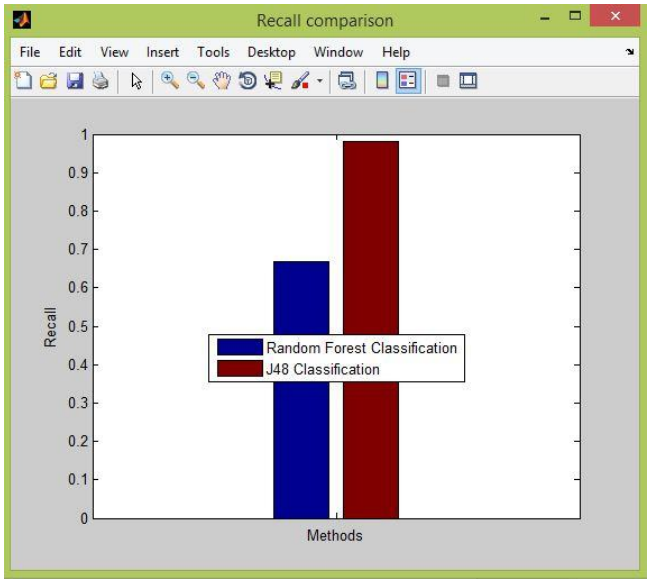

Figure 6 : Recall

From the figure 6 can observe that the comparison of existing and proposed system in terms of recall metric. In x axis we plot the methods and in y axis we plot the recall values. In existing scenario, the recall values are lower by using random forest algorithm. The recall value of existing scenario is 0.91 for discover the student's performance. In proposed system, the recall value is higher by using the J48 algorithm. The recall value of proposed scenario is 0.97 for discover the student's performance. Thus it shows that effective analysis is performed by using proposed algorithm. From the result, can conclude that proposed system is superior in performance.

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

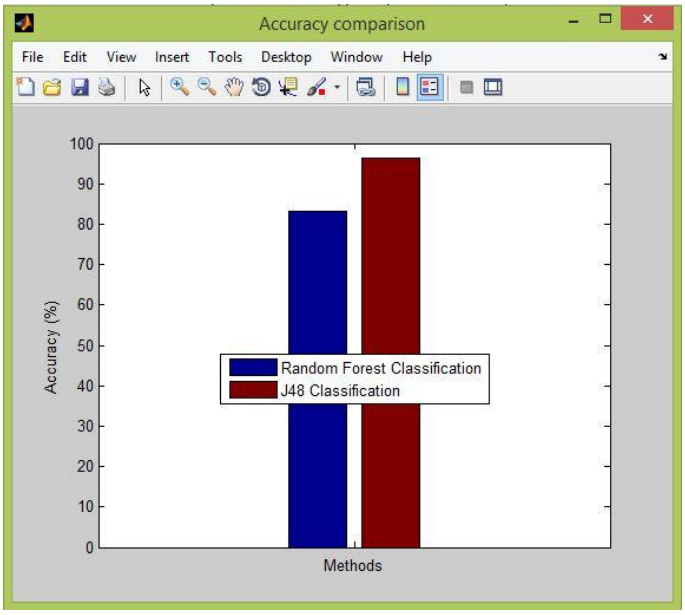

Figure 7 : Accuracy

From figure 7 can observe that the comparison of existing and proposed system in terms of accuracy metric. In x axis plot the methods and in y axis plot the accuracy values. In existing scenario, the accuracy values are lower by using random forest algorithm. The accuracy value of existing scenario is 82 % for discover the student's performance. In proposed system, the accuracy value is higher by using the J48 algorithm. The accuracy value of proposed scenario is 95% for discover the student's performance. Thus it shows that effective analysis is performed by using proposed algorithm. From the result, can conclude that proposed system is superior in performance.

F-Measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

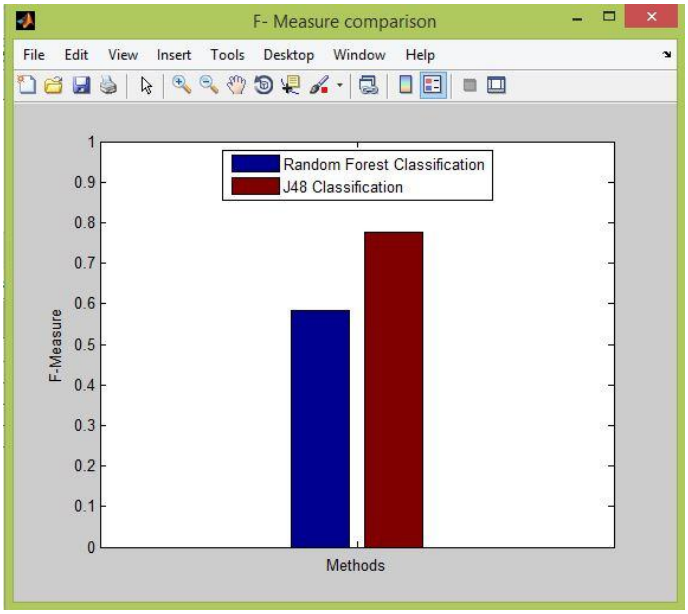

Figure 8 : F-Measure

From the figure 8 can observe that the comparison of existing and proposed system in terms of f-measure metric. In x axis we plot the methods and in y axis plot the f-measure values. In existing scenario, the f-measure values are lower by using random forest algorithm. The f-measure value of existing scenario is 0.68

for discover the student's performance. In proposed system, the f-measure value is higher by using the J48 algorithm. The f-measure value of proposed scenario is 0.74 for discover the student's performance. Thus it shows that effective analysis is performed by using proposed algorithm. From the result, can conclude that proposed system is superior in performance.

V. Conclusion

In this proposed system, J48 classification Algorithm is used to classify the student mark based on the urban and rural. To analyze the academic achievement of urban and rural areas students in order to identify superior performance is over schooling or in graduation. The analyzing is performed in class wise data. The most of the students are performing well in their graduation. The B.Sc., Computer Science student data is trained and B.Sc., Information Technology student data has been tested. It is observed from the experimental results the Random Forest and J48 Classification Algorithm are shown the higher Precision, recall, accuracy and f-measure values. The Proposed J48 Classification Algorithm is superior performance for all metrics than the other algorithm. From the result, the Proposed J48 Classification Algorithm is better for providing efficient performance.

References

[1] Amershi, S., and Conati, C., (2009) "Combining unsupervised andsupervisedcassificationtobuild user models for exploratory learning environments" Journal of Educational Data Mining. Vol.1, No.1, pp. 18-71.
[2] Baker, R. S. J. D. "Data mining for education." International encyclopedia of education 7 (2010): 112-118.
[3] Sachin,R.B., & Vijay, M. S, "A Survey and Future Vision of Data Mining in Educational Field", Paper presented at the AdvancedComputing & Communication Technologies(ACCT), Second International Conference on 7- 8 Jan. 2012.
[4] Tair, Mohammed M. Abu, and Alaa M. El-Halees. "Mining educational data to improve students' performance: a case study." International Journal of Information 2.2 (2012).
[5] Goyal, Monika, and RajanVohra. "Applications of data mining in higher education." International journal of computer science 9.2 (2012): 113.
[6] Abdul Aziz, Azwa, NurHafieza Ismail, and Fadhilah Ahmad."MINING STUDENTS'ACADEMIC PERFORMANCE." Journal of Theoretical and Applied Information Technology 53.3 (2013): 485-485.
[7] Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." arXiv preprint arXiv:1201.3418(2012).
[8] M.Kebritchi,andA.Hirumi,Examiningthe pedagogical foundations of modern educational computer games. Computers and Education, 5 (4): 1729-1743, 2008
[9] A. McFarlane, N. Roche, and P. Triggs, Mobile Learning: Research Findings. Becta, July 2007. http://partners.becta.org.uk/uploaddir/downloads/page_documents/research/mobile_learning_july07.pf (accessed February 4, 2008), 200
[10] MoMath, Mobile Learning for Mathematics: Nokia project in South Africa. Symbian Tweet, http://www.symbiantweet.com/mobile-learning-