Journal of Advances and Scholarly Researches in Allied Education
Vol. 15, Issue No. 9, October-2018, ISSN 2230-7540

Data Mining In Higher Education: Students Dropout Assessment "A Case Study on Aps University"

Shivendra Kumar Dwivedi[1]* Dr. Prabhat Pandey[2]

[1] Research Scholar, Department of Computer Science, APS University, Rewa

[2] OSD Additional, Directorate of Higher Education Rewa Division (M.P.), India – 486001

Abstract – In this paper, I apply different data mining approaches for the purpose of examining and predicting students' dropout through their college programs. For the subject of the study, I selected total 1295 records of various stream students graduated from APS University between years 2013 to 2015.

For this purpose, the preliminary data of 1295 students collected in prescheduled (Schedule and questionnaire) format of personal interview to find out the reasons of dropout from college. In order to classify and predict dropout students, different classifiers have been trained on my data sets including J48 Decision Tree and PART classification. These methods were tested using 10-fold cross validation. The accuracy of J48, and PART classifiers were 77.33% and 75.13% respectively. The study also includes discovering hidden relationships between student dropout status and enrolment persistence by mining a frequent cases using algorithm.

The reasons recorded for dropout of students at this university were viz; Agriculture work, Care of sibling, Poor economic condition, Lack of education facility, Ignorance of guardian, Long illness, Non friendly environment of college being to for away, no adult protection.

The information generated will be useful for better planning and implementation of educational program and infrastructure under measurable condition to find out the main reasons of dropout students in various colleges at this university.

Keywords: KDD, classification, dropout, educational data mining (EDM), decision tree, prediction, ICT (Information & Communication Technology)

---------------------------------x---------------------------------

1. INTRODUCTION

Nowadays, higher learning institutions encounter many problems, which keep them away from achieving their quality objectives. Most of these problems caused from knowledge gap. Knowledge gap is the lack of significant knowledge at the educational main processes such as advising, planning, registration, evaluation and marketing (Baepler and Murdoch, 2010). For example, many learning institutions do not have access to the necessary information to advice students. Therefore, they are not able to give suitable recommendation for them.

Data mining is a powerful technology that can be best defined as the automated process of extracting useful knowledge and information including, patterns, associations (Fayyad, et. al., 1996). The knowledge discovered by data mining techniques would enable the higher learning institutions in divers ways not limited to making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. It results in improving the effectiveness and efficiency of the processes (Baker and Yacef, 2009) (AL-Malaise, et. al., 2014) .

One of the biggest challenge that higher education faces today is predicting the academic paths of student. Many higher education systems are unable detecting student population who are likely to drop out because of lack of intelligence methods.

It remains a challenging task to accurately predict a currently enrolled student's likelihood of returning to colleges the next term or change his major (Romero and Ventura, 2010) (B. R.B. et. al., 2013)

Develop learning and teaching initiatives to improve retention and progression in education process is an important academic concern, which mainly depends on monitoring student performance and exploiting student feedback. Student marks and achievement are the main sources to study student feedback and progress, yet university and educational centers can use to predict the student performance, student dropout, and study path.

The main objective of this study is to identify those students who take dropout from college in first year. Early identification of these students is enough for the institution to accommodate its interventions and marketing strategies will greatly enhance the student persistence rate in specific majors. This paper focuses on predict reasons of dropout to use various classification techniques of various colleges from University.

2. RELATED WORK

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [4]. Mining in educational environment is called Educational Data Mining.

Al-Radaideh et al. [4], applied classification data mining techniques to improve the quality of the higher education by evaluating the main attributes of students that affect the their performance. This study was used to predict the student’s final grade in a course.

Ayesha et.al. [5], performed study on student learning behavior. For this factors like class quizzes mid and final exam assignment are studied. This study will help the tutors to reduce the ratio of drop out and improve the performance level of students.

Bharadwaj and Pal [6], used the decision tree method for classification to evaluate performance of student’s. The objective of their study is to discover knowledge that describes students’ performance in end semester examination. This study was quite useful for identifying the dropout’s student in earlier stage and students who need special attention and allow the teacher to provide appropriate advising

Kotsiantis et.al [8] conducted a comparative study of different classifiers to predict students drop-out in the middle of a course using different parameter of 350 students. It was found that the Naive Bayes and Neural Network were the best classifiers to predict about 80% of drop- outs. Based on the findings it was observed that Naïve Baves model is best suited for small data set in comparison to other methods.

Al-Radaideh, et al [9] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the NaïveBayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Pandey and Pal [10] conducted study on the student performance based by selecting 60 students from a degree college of Dr. R. M. L. Awadh University, Faizabad, India. By means of association rule they find the interestingness of student in opting class teaching language.

Divakar, R.CJain[11] applied four classification methods on student academic data i.e Decision tree (ID3), Multilayer’s perception, Decision table & Naïve Bayes classification method.

3. DATA MINING TECHNIQUE

In this section, the most common data mining techniques are discussed to understand the theory without going into details. According to Chen et al. (2005) data mining brings various techniques together to discover pattern and to construct models from database. Ngai et al. (2009) indicates data mining model: Association, Classification, Clustering, and Forecasting. Huang et al. (2012) concludes that data mining technique is used to gain useful information or interesting knowledge. Perceived usefulness and perceived ease of use are the factors that affect an individual intention to use data mining tools.

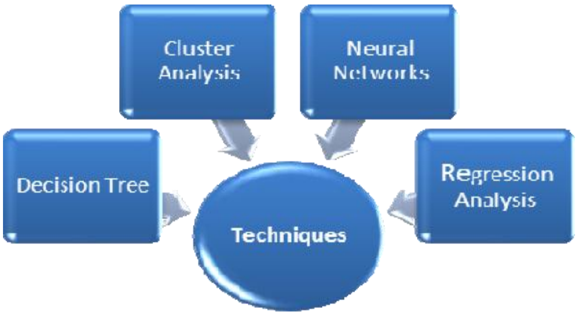


Fig. 1: Data Mining Techniques

3.1 Neural Networks

It is techniques which can be used for classification of large complex data. It can be used to study course selection by students, student course satisfaction, and specialization selection. Rokach (2010) state the neural network represents each cluster by a neuron or "prototype". The input data is also represents by neurons which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptability during learning.

3.2 Decisions Tree

Decision tree is a data mining technique that can be used for classification and prediction of large data. Decision tree is used for profiling customers. Decision tree is also called rule induction technique (Luan, 2003). According to Rokach (2010) the data, in decision tree, is represented by "a hierarchical tree where each leaf refers to a concept and contains a probability description of the concept."

It consists of nodes and branches, nodes are connected by branches, time flows from left to right, each ranch represents a decision or a possible event." In addition, decision tree make classification easy and understandable and also result-oriented. Many industries use to classify and predict customers' behavior; acquition, retention, and growth. Similarly, universities can use this technique to classify students' performance, and dropouts.

3.3 Regression Analysis

Bryman and Duncan (2005) stated that regression is "a powerful tool for summering the nature of the relationship between variables and for making predictions of likely values of the dependent variables." It is used as a continuous variable (Nisbet, 2009). Regression analysis can be applied in data mining to predict students' these techniques can be done by using SPSS software.

3.4 Cluster Analysis

Cluster analysis is an unsupervised learning technique (Tsai et al., 2011). Cluster analysis refers to identifying groups of cluster with similar characteristics (Ahn and Sohn, 2009), splitting the full data set into a set of clusters (Baker, 2010) where categories are not known in advance. Han and Kamber (2006) indicate that cluster analysis can be used to generate labels.

In brief, classification and regression models are supervised model and cluster analysis is an unsupervised model (Rokach and Maimon, 2010). Both models are used to predict and classify the large data into usable information.

4. DATA MINING TASK

There are many tasks of data mining practicing by various industries. Prediction, classification, association and clustering are the most important tasks of data mining.

4.1 Prediction

Sen et al. (2012) built models to predict secondary education placements-test using sensitive analysis on predicting models (Decision Tree Algorithm, Support Vector Machine, Neural Network and Logistic Regression). They have identified the following important predictors of placement-test: previous test experience, student has dropout at various reasons. They indicate that decision tree analysis is the best predictor, followed by support vector machine, and neural network. Logistic regression is the least predictor.

4.2 Association

Larose (2005) states that association task finds which attributes "go together". According to Gopalan and Sivaselvan (2009) association rule is "the process of discovering interesting association or relationship among data items." It summarises the entire data. Priori and GRI algorithm are used in association rules. Romero et al. (2008) indicates that association rule can be applied to discover relationship between the characteristics of the students and helped to find relationship perfectly (Aggarwal et al., 1999. Association rules are "derived from patterns in a dataset that correspondent to a particular situation" (Rajamani et al., 1999).

4.3 Clustering

Clustering means grouping similar objects. Rajshree et al. (2010) defines clustering as a process of grouping a set of physical or Abstract object into a class of similar objects. According to Larose (2005) cluster does not classify, estimate or predict the value of target variables but segment the entire data into homogeneous subgroups. Heterogeneous population is classified into number of homogenous subgroups or clusters are referred as clustering (Berry and Linoff, 2004). Furthermore, clustering task is an unsupervised classification.

4.4 Classification

Classifying data into a fixed number of groups (Soman et al., 2006) and using it for categorical variables (Nisbet, 2009) is known as classification. Classification can be classified into two types: Supervised and Unsupervised. When the objects or cases are known in advance is called supervised classification whereas unsupervised classification means the objects or cases are not known in advance. The following algorithm can be used for classification model (Gorunescu, 2011; Aggarwal et al., 1999).

- Decision/Classification tree

- K-nearest neighbour classifier

- Statistical analysis, geneticalgorithms
- Bayesian classification

Classification is called supervised learning. Data classification is two steps process. In first step, a model is built and in second step the model is used for classification there are some classification methods such as J48 decision tree and PART but decision tree is the basic and popular technique for classification

5. METHODOLOGY

Information produced by data mining techniques can be represented in many different ways. In this paper, I have used the classification data mining technique to extract the important attribute that stored in a database to analyze reasons affecting the dropout of students in various college of higher education. Two classifier algorithms J48 and PART are used for predicting the main reasons of dropout from college.

5.1 DATA MINING PROCESS

5.1.1 DATA COLLECTION & FILE CONVERSION

The data was collected from the various colleges affiliated with APS University (Both private colleges and Govt colleges) the data collected through structured questionnaire and personal interviews of students of colleges. The questionnaire has been constructed based on theoretical and empirical grounds about reasons of dropout. The Data was collected from various colleges of five districts affiliated with APS university of Madhya Pradesh. In these districts I selected 10 Government Colleges and 10 private colleges. Data was stored in MS-Excel (.xls) format then converted into CSV (Comma Delimited) *.csv file. Further CSV files are converted into Arff format WEKA enabled file for interpretation.

Table 1: Variables related to Dropout reasons

Attribute	Description	Possible values
Agriculture work	Dropout due to agriculture work	Yes and no
Care of siblings	Students dropout due to take care of brother or sister	Yes and no
Poor economic condition	Students dropout from college because economic condition is not sound	Yes and no
Lack of educational facility	Educational facility is less	Yes and no
Ignorance of guardian	Parents are not aware	Yes and no
Long illness	Students illness for a long time	Yes and no
Non friendly environment of college	College environment is not friendly	Yes and No
Migration of family	Family not staying permanently at one place	Yes and No
No proper college facilities available for girls	College does not provide proper facilities	Yes and No
Regular absence from college	Student’s continuous absence from college	Yes and No
Homeless	Having no Accommodation	Yes and No
Without adult protection	Have no Guardianship	Yes and No
College being too far away	Long Distance of college	Yes and No

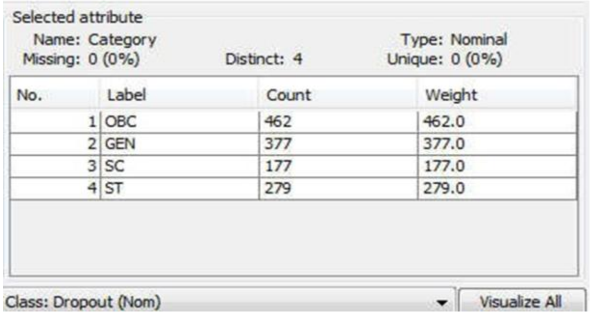
5.2 DATA MINING TOOL (WEKA)

WEKA stands for Waikato Environment for Knowledge Analysis is a popular suite of machine learning software, developed at the University Of Waikato, New Zealand. The implementation of the dataset is done using a data mining tool WEKA. WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. From the above data the student file is converted to ARFF (Attribute Relation File Format) for data analysis WEKA explorer.

5.3 DATA SELECTION AND TRANSFORMATION

After collection of data, the dataset was prepared to apply the data mining techniques. Before application of prescribed model, data preprocessing was applied to measure the quality and suitability of data. In this step only those attributes were selected which were needed for data mining. For this, removing missing values; smoothing noisy data, selection of relevant attribute from database or removing irrelevant attributes, identifying or remove outlier values from data set, and resolving inconsistencies of data.

5.3.1 Data preprocessing Category wise data



Selected attribute
Name: Category Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	OBC	462	462.0
2	GEN	377	377.0
3	SC	177	177.0
4	ST	279	279.0

Class: Dropout (Nom)

Table 2 (category wise dropout students)

Finally data set was preprocessed in WEKA tool. In this table, the selected attribute is Category. There are four categories namely OBC, GEN, SC and ST. On the basis of this, the result is as under the dropout count for OBC students is 462, for Gen 377, for SC 177 and for ST it is 279. So the OBC category students have highest dropout rate and SC category students have lowest dropout ratio.

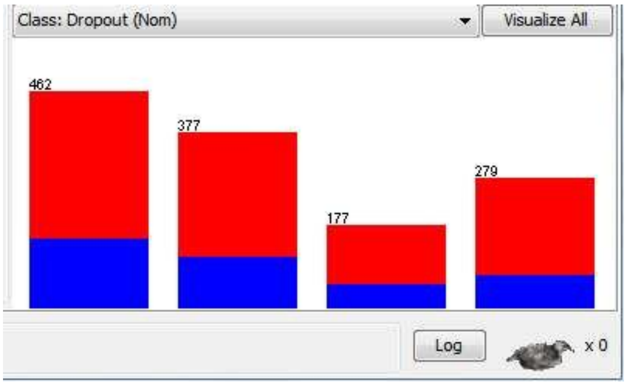


Figure 2 Graph represent category wise dropout

This graph represents the dropout rate among different students category wise.

5.4 DATA ANALYSIS TECHNIQUE

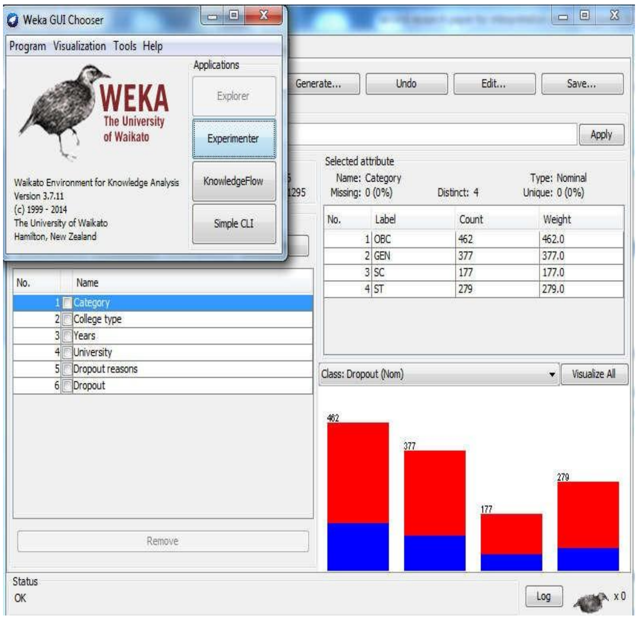
In this study data analysis techniques have to be employed using data mining methods.

5.4.1 Data mining methods:

After collection and scrutiny and transformation of data using appropriate measures, Data mining classification and decision tree approach ware applied to predict student dropout causes in early stage of their study either before or after completion of first year of their study.

5. IMPLEMENTATION OF CLASSIFICATION MODEL

In this study the data mining software WEKA tool (3.7.5) was used for data analysis of the dataset. It is independent and portable platform because it is implemented in java programming language. From above data, APS University. arff file is created.



Waikato Environment for Knowledge Analysis
Version 3.7.11
(c) 1999 - 2014
The University of Waikato
Hamilton, New Zealand

No.	Name
1	Category
2	College type
3	Years
4	University
5	Dropout reasons
6	Dropout

Selected attribute
Name: Category Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	OBC	462	462.0
2	GEN	377	377.0
3	SC	177	177.0
4	ST	279	279.0

Class: Dropout (Nom)

Status
OK

Figure 3 (weka explorer)

This file is then loaded into WEKA Explorer to generate a classification model, using decision tree method. There are several techniques in WEKA 3.7 tool for example PART and J48 are used to generate decision tree and 10 fold cross validation is selected in this study.

6. RESULT AND DISCUSSION

Data set collected from 1295 students was analyzed using frequency distribution and it was found that dropout rate is 19.96% which is mainly due to lack of educational facility and college being too far away. Particularly first year student of different colleges from University were interested to change the institution in better prospective. Further, Data set was classified using PART and J48 decision tree algorithm. A confusion matrix was constructed for PART and J48 to study the actual and predicted class. The dropout data set was classified in two groups such as Private College (Yes or No) Govt College (Yes or No) and based on two groups, 2x2 confusion matrix for PART and J48 decision tree algorithm was constructed (Table 5). The correctly classified instance is the sum of diagonal in matrix whereas all other figure is incorrectly classified. On the other hand the accuracy percentage for dropout for PART and J48 decision tree algorithm using 10-fold cross validation present in table the highest percentage 77.37 for J48 followed by PART (75.13). it indicate that J48 decision tree algorithm is the best classifier as compare to PART decision tree algorithm for present dataset to predict the student's dropout status.

Confusion matrix of PART (table 3 (a))

	Predicted Class			
Actual Class		**A**	**B**	**Total**
	A (Govt) (Y)	846	87	933
	B (Private)(N)	235	127	362
	Total			1295

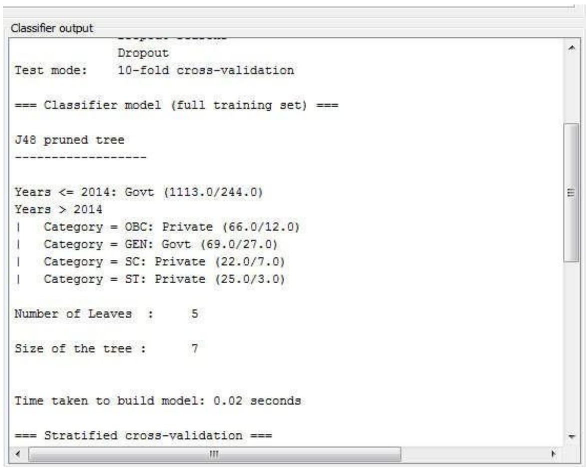
Confusion matrix of J48

	Predicted Class			
Actual Class		**A**	**B**	**Total**
	A (Govt) Y	911	22	933
	B (Private) N	271	91	362
	Total			1295

Accuracy Percentage for Dropout

Algorithm	**Correctly Classified Instances**	**Incorrectly Classified Instances**
PART	75.1351%	24.8649%
J48	77.3745%	22.6255%

Table 3(c)



```
                Dropout
Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Years <= 2014: Govt (1113.0/244.0)
Years > 2014
|   Category = OBC: Private (66.0/12.0)
|   Category = GEN: Govt (69.0/27.0)
|   Category = SC: Private (22.0/7.0)
|   Category = ST: Private (25.0/3.0)

Number of Leaves  :     5

Size of the tree :      7


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
```

Figure 4 A decision tree generated by J48 algorithm

6.1 Decision tree

Figure 5 depicts the decision tree that resulted from applying the decision-tree classification algorithm on dropout student. As it is seen from the figure, the attribute of dropout from college “Type of college” has a great influence on whether the student will drop out from college or not.

The model presented in figure 5 has an accuracy of 77 % as shown in figure 5 to interpret the rules in the decision tree.

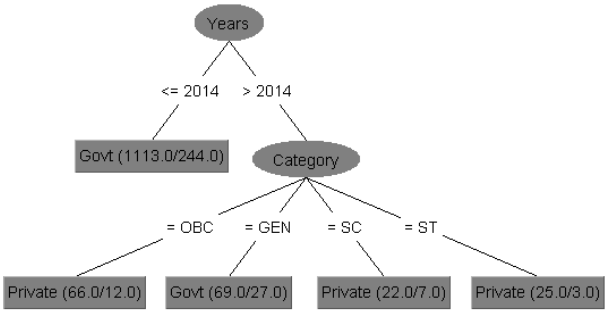


Figure 5 Decision Tree classifier output Classifier rules

Years <= 2014 AND Category = GEN : Govt(308.0/9.0)
Years <= 2014 AND Years > 2013: Govt (512.0/121.0)
Years > 2014 AND Category = ST: Private (25.0/3.0)
Years > 2014 AND Category = OBC AND Dropout reasons = Lack of educational facility: Private (11.0)
Years <= 2014 AND Category = ST AND Dropout reasons = Non friendly environment of College AND Dropout = Y: Govt (8.0/1.0)
Years > 2014 AND Category = OBC AND Dropout reasons = College being too far away : Private (8.0/1.0)
Years <= 2014 AND Category = ST AND Dropout reasons = Long illness: Govt (19.0/5.0)
Years > 2014 AND Category = OBC AND Dropout reasons = College being too far away : Private (8.0/1.0) Dropout reasons = Lack of educational facility: Govt (45.0/17.0) Dropout reasons = Agriculture work: Private (11.0/4.0)
Dropout reasons = Regular absence from college AND Dropout = Y: Private (4.0/1.0)
Dropout = Y AND Dropout reasons = No proper college facility available for girls: Govt (5.0/2.0)

7. REASONS OF DROPOUT

The data collected from 1295 students was analyzed to study the frequency distribution against each factor of those students who have completely decided to drop out during the first year of study. The result of each causing factor was listed in table 4. The highest dropout reasons were college being too far away (15.44%) and lack of educational facility (15.44 %) and those students who are dropping out because of other reasons come under the range of 1.39 % to 15.52%

Table 4

Reason	Yes (student who have dropout)	
	Number	Percentage
Agriculture work	26	2.01%
Care of siblings	114	8.80%
Poor economic condition	22	1.70%
Lack of educational facility	200	15.44%
Ignorance of guardian	62	4.79%
Long illness	188	14.52%
Non friendly environment of college	64	4.94%
Migration of family	114	8.80%
No proper college facilities available for girls	73	5.64%
Regular absence from college	18	1.39%
Homeless	18	1.39%
Without adult protection	175	13.51%
College being too far away	200	15.44%

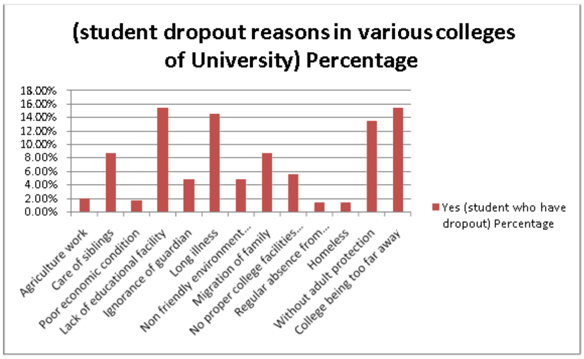


Figure 6: student dropout reasons at various colleges of University

The graphical representation (Figure 6) also displays trends against each factor. The highest reasons of dropout are college being too far away and Lack of educational facility in their colleges.

8. CONCLUSION

This study shows preliminary results for predicting student's dropout from large dataset of student's records and demonstrates the integration of different data mining technique for the purpose of EDM. During analysis of dropout students, different attributes have been checked, and some of them are found effective for the prediction. The attribute college type (Govt and private college) was the strongest attribute for the prediction and other attributes like category, year, dropout reason, and dropout are also used for prediction.

Result indicates that J48 Decision tree algorithm is best classifier as compare to PART. This study will also work in identifying those students which need special attention to minimize the dropout rate. This information is useful for developing the plan and their implementation to minimize the dropout and improving the college enrolment rate at various colleges of university.

REFERENCES

1. P. Baepler and C. J. Murdoch (2010). "Academic Analytics and Data Mining in Higher Education," International Journal for the Scholarship of Teaching and Learning, vol. 4, no. 2, pp. 1-9.

2. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (1996). "Advances in knowledge discovery and data mining,".

3. R. S. J. D. Baker and K. Yacef (2009). "The State of Educational Data Mining in 2009 : A Review and Future Visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-16.

4. A. AL-Malaise, A. Malibari and M. Alkhozae (2014). "STUDENTS' PERFORMANCE PREDICTIONSYSTEM USING MULTI AGENT DATA MINING TECHNIQUE," International Journal of Data Mining & Knowledge Management Process (IJDKP) , vol. 4.

5. C. Romero and S. Ventura (2010). "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6.

6. B. R.B., T. S.S and S. A.K. (2013). "Importance of Data Mining in Higher Education System," Journal Of Humanities And Social Science (IOSR-JHSS), vol. 6, no. 6, pp. 18-21.

7. J. Luan (2002). "Data Mining and Knowledge Management in Higher Education -Potential Applications.," in Process dings of AIR Forum, Torento , Canada.

8. Galit et. al. (2007). "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education.

9. A. Kumar and Vijayalakshmi (2011). "Implication Of Classification Techniques In Predicting Student’s Recital," International Journal of Data Mining & Knowledge Management Process, vol. 1, no. 5, pp. 41-51.

10. S. Kotsiantis (2009). "Educational data mining: a case study for predicting dropout-prone students," International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 1, no. 2, p. 101.

11. Boero, G., Laureti, T., & Naylor, R. (2005). An econometric analysis of student withdrawal and progression in post-reform Italian universities. Centro Ricerche Economiche Nord Sud - *CRENoS Working Paper 2005/04*.

12. D. G. W, P. Mykola and V. J. M. (2009). "Predicting Students Drop Out: A Case Study," International Working Group on Educational Data Mining.

13. L. Rokach (2008). Data mining with decision trees: theory and applications, vol. 69, World scientific.

14. J. F. Superby, J. P. Vandamme, and N. Meskens (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods, *Proceedings of 8th International Conference on Intelligent Tutoring Systems*, 2006, pp. 37-44.

15. S. J. Russell and P. Norvig (2009). Artificial Intelligence: A Modern Approach (AIMA), 3rd ed., Prentice Hall.

Corresponding Author

Shivendra Kumar Dwivedi*

Research Scholar, Department of Computer Science, APS University, Rewa

E-Mail – shivendra.mphil@gmail.com