Indian Journal of Science and Technology, Vol 9(12), DOI: 10.17485/ijst/2016/v9i12/83335, March 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Customized Prediction Model to Predict Post-Graduation Course for Graduating Students Using Decision Tree Classifier

Jaimin N. Undavia[1]*, Prashant Dolia[2] and Atul Patel[3]

[1]Faculty of Computer Science and Applications, Charotar University of Science And Technology (CHARUSAT), Changa - 388421, Gujarat, India; jaiminundavia.mca@charsat.ac.in
[2]Department of Computer Science and Applications, Maharaja Krishnakumarsinhji Bhavnagar University, Bhvnagar - 364001, Gujarat, India; prashant_dolia@rediffmail.com
[3]CharotarUniversity of Science and Technology (CHARUSAT), Changa - 388421, Gujarat, India; atulpatel.mca@charsat.ac.in

Abstract

Background/Objectives: Excellence of Universities is based on students' success in their academic and it is possible if the students are instructed or counseled before getting admitted in their post graduation. So, we have developed a model for the post graduating students to utilize their intelligence in right direction. **Methods/Statistical Analysis**: If students are given admission in right course then their academic success is guaranteed by the university. To formulate the prediction, decision tree classifiers are best suitable as it has potential to generate comprehensible output. It is generating the tree and rules which will be used to formulate the predictions. Hence, this approach is of two steps approach known as training phase and testing phase. **Findings:** The model trains on the basis of the defined instances and from the defined instances the classified builds the rules. These rules are used to formulate prediction for unknown valued instances. This article depicts the customized classification model to predict the Post-Graduation degree of the students. The model is based on J48 decision tree algorithm for classification. The model is trained by the data collected through survey of different institutions with the purpose of differentiating and predicting students' choice and to generate unbiased result. We obtained certain patterns of the students preferences to select their post graduation course. On the basis of such rules which are derived from historical data, are used to predict post graduation course for unknown instance. We have used J48 classification algorithm for decision tree to predict the post graduation course based on their academic history and other identified parameters. We have identified total 14 parameters to predict the class label of 15thattribute. **Applications/ Improvements:** We have customized a model using Weka which uses the J48 algorithm to predict students' post graduation degree. We have obtained 94.03% accuracy of prediction against 4 classes as final attribute.

Keywords: Classification, Customization in Weka, Post Graduation Course Selection, Prediction Model, Weka

1. Introduction

Enforcement of people in education in adequate direction is the ultimate goal of education[1]. One of the major challenges for knowledge discovery and data mining systems stands in developing their data analysis capability to discover out of the ordinary models in data[2].

Student's success in Post-Graduation course is most crucial as it defines the career of the students. Students take admission in Post Graduate courses not based on their inclination for a particular subject but they may take the admission based on many other aspects also. This problem has grown up and became concern in many countries. This problem is known as "The One Hundred Factor

*Author for correspondence

Problem" and identified as potential field of research[1]. This problem can be addressed or resolved by the advent of Data Mining. Data Mining has many applications in the field of business now a days but its use in higher education is relatively new[2]. The application of Data Mining in the field of education is referred as Education Data Mining. It is a filed which exploits statistical, machine learning and Data Mining algorithms over the different types of educational data[3].

Education Data Mining has huge impact on the operations and efficiency of the higher education system. In most cases, the EDM follows the same steps as traditional Data Mining operations. This steps are followed with preprocessing, Data Mining & post-processing. Apart from the traditional Data Mining techniques like clustering, classification, etc., EDM supports some innovations in the discovery of models which are not seen in the traditional Data Mining.

Two main objectives can be distinguished in the Data Mining process with the perspective of management. First one is known as descriptive objective where variable and their impact are identified. The second objective is known as predictive objective in which the identified variable is mined to discover new knowledge.

From practical point of view, EDM allows to discover new knowledge from the students' data in order to evaluate/improve the educational system[3]. It helps the students and institutions as well for the betterment in overall process.

In this paper, we have discovered a new model by customizing WEKA tool to predict best suitable post graduate course for the graduating students. As pointed earlier, the descriptive objective is to identify the affecting parameters and their influence for the selection of Post-Graduation course is identified. Once these parameters with their influence are identified, the second predictive objective comes into picture and uses the parameters for the prediction of Post-Graduation course for the particular student.

2. Education Data Mining and Higher Education Systems

Higher Education Systems in Gujarat, India is the source of excessive amount of data. These data are versatile in nature and needs to be analyzed and study for the management use, decision making, prediction, etc., for the betterment in academic systems. These data are subject to various data mining techniques for decision making and prediction. At a glance, when educational data is undergone to the data mining technique for various purposes then this is called Education Data Mining.

Higher Education is provided through higher education systems like Universities, post - academic study centers and Post-Graduation departments. Higher education centre of UNESCO defines the higher educational system as "an institute with high level of learning and cooperation ability in completing knowledge, which is facilitated for educating and researching, and have got the responsibility of setting exam and granting science degree"[4].

Some related research in the fields of data mining has been found by many researchers to predict students' performance based on their past performance. Research emphasizes on the prediction of student performances on the basis of their academic history. Most of the research in the field of education data mining is with the use of classification and prediction techniques[5].

3. Decision Support System for Problem

Students' success in terms of academic and career is the major concern of any Post-Graduation institutions. Students' career is dependent on their Post-Graduation degree's grade and expertise in concerned subjects.

Post-Graduation degree selection is a major decision which needs to be taken by the students once they complete their graduation course. Here in this paper we have analyzed the academic history of the student and then based on the historical data support, we are predicting the Post-Graduation course for the student. The system takes the decision based on the training of the model and then predicts the Post-Graduation course for the student. Recent educational systems are target activities of the students that can affect their performance and maintain them with the help of ERP systems[5]. These maintained data are used for the betterment in their study and future study as well. Technological developments and new programming techniques have evolved and used for the decision making process with advent use of Artificial Intelligence.

4. Parameters for Decision Making

Researchers have worked on several data mining techniques to predict performance of employees, result or score of the students in the different branches, weather

forecasting, crop production, etc., for the betterment in the efficiency of the overall systems. Selection of parameters for such study or prediction is extremely important for the precise output of the system[6].

Sayed Zakarya Taghavinezhad, Fariba Nazari and Zahed Bigdelihave developed "Davis Technology Acceptance Model" which is used to determine factors affecting the technology acceptance and application of social network[7].

Edin Osmanbegovic and Mirza Suljic have developed a data mining approach for predicting student performance by taking 11 different student related parameters and predicted students' performance with different algorithms[8].

Chein and Chen have developed a model to improve the selection process of employee. They have improved the process by predicting performance of the new applicants. They have decided the attributes and fetched the value of those attributes from their CVs, job applications and interviews.

Behrouz Minaei-Bidgoli and others have developed a model to predict students' performance by means of classification technique. They have developed a model to predict students' performance and segregated students into four different classes based on their performance[9]. They have developed their dedicated system named LON-CAPA system and fed up the system with the data of a test result. The test result yields students parameters like, number of correct answers, first attempt to reply the answer, time taken to reply for each question, etc.

Vasile Paul Bre_felean invented a model to predict students' behavior on the basis of analysis and prediction technique of data mining. They have conducted a survey and collected student's data and those data fed up to the system for analysis[10].

From the basis of literature survey, we have kept following parameters for our proposed research to predict students' performance.

Table 1. List of Attributes

Sr. No.	Name of Attribute	Possible Value	Remarks
1	HSC_Per	Distinction (70%& above), Fclass (60%-69.99%), Sclass (50%-59.99%), Tclass (less than 50%)	Percentage of HSC
2	HSC _Stream	{SC, CM & AT}	Science, commerce & Arts
3	Graduation	{BSC, BCA, be, OC, B.Com, BBA}	Graduation Degree
4	MSG	{Basic Science, Computer, Accountancy, OS, Other}	Major subject opted for graduation
5	POG	Distinction (70% & above), Fclass (60%-69.99%), Sclass (50%-59.99%), Tclass (less than 50%)	Percentage of Graduation
6	Gender	{Male, Female}	Gender of student
7	Category	{Res, Op, Other}	Category of Student
8	ROPG	Distinction (70% & above), Fclass (60%-69.99%), Sclass (50%-59.99%), Tclass (less than 50%)	Result of Post Graudation
9	No Back	Number of backlog during their PG	No of backlogs in PG course
10	Clstat	{Government, Self-Financed}	Status of college
11	NoFB	{Job,Business,NA,NO,DairyProducts,Farmer,Government officer,Governmentservice,Service,Self-Employee}	Nature of father's business
12	NoS		No of sibling
13	Personal Choice	{Yes, No}	Current pg course is personal choice or not.
14	Reason	{JobProspects,PeerPressure,FamilyBusiness,Personal Choice,MarketDemands,Programming}	Reason for opting the PG course.
15	PG Course	Final Class Lable	{MCA/M.Sc.(IT), M.Sc, M.Com, MBA}

5. Data Collection and ARFF Creation

For the accomplishment of the research work, we have conducted a survey across the 13 universities of the state of Gujarat. We have collected data through questionnaire by online survey and personal interviews as well. We have collected 14,000 records from different colleges and they are cleaned for the accurate results.

For data collection and training the model a survey is conducted for current post graduating students. The information is collected through questionnaire and personal interview as well. The value we have collected is also referred as experiment variable[11].

The collected data were converted into Attribute Relationship File Format for the input to the WEKA. Here is the small portion of the ARFF file.

@relation student
@attribute HSC Per {Distinction, First Class, Second Class, Third Class}
@attribute HSC Stream {Science, Commerce}
@attribute Graduation {BSC, BCA, be, OC, B.Com, BBA}
@attribute MSG {Basic Science, Computer, Accountancy, OS, Other}
@attribute pog {Distinction, First Class, Second Class, Third Class}
@attribute Gender {Male, Female}
@attribute Category {Open, Reserved, Other}
@attribute ROPG {Distinction, First Class, Second Class, Third Class}
@attribute No back real
@attribute Clstat{Government, Self Financed}
@attribute NoFB{Job, Business, NA, NO, Dairy Products, Farmer, Government officer, Government service, Service, Self Employee}
@attribute NoS real
@attribute Personal Choice {Yes, No}
@attribute Reason{Job Prospects, Peer Pressure, Family Business, Personal Choice,
Market Demands, Programming}
@attribute PG Course {MCA/M.Sc.(IT), M.Sc, M.Com, MBA}
@data
First Class, Science, OC, OS, Second Class, Female, Open, Distinction, 0, Government, Business, 1, Yes, Job Prospects, MBA
Second Class, Commerce, OC, OS, Second Class, Male, Other, Distinction, 0, Government, Farmer, 0, Yes, Market Demands, MCA/M.Sc.(IT)
FirstClass,Science,OC,OS,SecondClass,Male,Open,Distinction,0,Government,Business,1,Yes,JobProspects,M.Com
FirstClass,Science,OC,OS,SecondClass,Female,Open,Distinction,0,Government,Business,1,Yes,JobProspects,MCA/M.Sc.(IT)
SecondClass,Commerce,OC,OS,SecondClass,Male,Other,SecondClass,0,Government,DairyProducts,1,Yes,MarketDemands,MCA/M.Sc.(IT)
ThirdClass,Science,OC,OS,FirstClass,Male,Open,FirstClass,1,SelfFinanced,NA,1,Yes,FamilyBusiness,MBA
SecondClass,Commerce,OC,OS,SecondClass,Male,Other,Distinction,0,Government,Farmer,0,Yes,MarketDemands,MCA/M.Sc.(IT)

6. Development of Model by Customizing Weka

Weka tool is customized to develop the specific classification problem and the model is devised.

As the classification is a two-step process, the model is also based on the following steps of classification to classify the students for a specific Post-Graduation Course.

For the accuracy prediction, we have developed model in two phases. The results, we got in simple model (with selected parameters) and advanced model (with all parameters) are different significantly. But we have shown the results in that we got through advanced model only[12].

As shown in the figure data classification is a process consisted of two steps. A model is built on predetermined set of data classes and this is called training phase. During the training phase some rules are constructed and in the

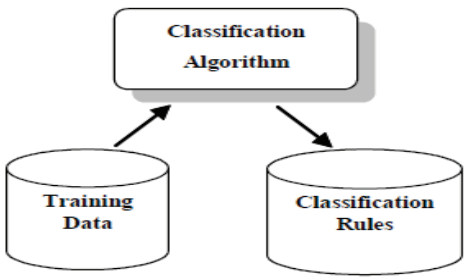

Figure 1. Classification Process.

second phase the class attributes are predicted by combining the training and classification rules[13].

First, we have created an ARFF file for training the model and then another ARFF file supplied to the model for the purpose of testing. Then the Weka is customized to specific level of the problem.

The initial screen of the model is shown below.

This screen is used to open ARFF file and after that we can select classification algorithm to be performed over the ARFF file. We are using J48 algorithm for the research, so we select J48 and first we will train the model using training option. J48 algorithm is the implementation of C4.5 algorithm and most commonly used algorithm. It is also a modified version of ID3 algorithm. Features like handling missing values, categorization of continuous attributes, pruning decision trees, rules derivation and many more are supported by J48 algorithm. J48 algorithms uses a method known as divide and conquer to construct the decision tree[14].

Once we provide the ARFF file then, tool is used to train the algorithm with class labelled records and then we can supply the ARFF file for the testing. Once we supply the ARFF file for testing, then the model generates the result.

7. Result and Conclusion

The result generated through model has following outcomes. The results are as efficient as we applied on standard data mining tool[15]. Though the result of classifier algorithms are not uniform among the classification algorithms, J48 produces classification accuracy at maximum extent. Selected data attributes have found to be influenced the classification process. The result can be extended and enhanced too by variation in attributes and increasing number of records too[16]. The model is trained by 919 instances and we obtained 892 records classified correctly and 27 records are incorrectly classified. Hence, we are getting 97.062% accuracy of the model with 0.94 kappa statistics. However, the rules we found from the model are narrated below:

```
Clstat = Government
|   HSC Stream = Science
|   |   Personal Choice = Yes
|   |   |   Category = Open: MBA (11.0/4.0)
|   |   |   Category = Reserved: MCA/M.Sc.(IT) (9.0)
```

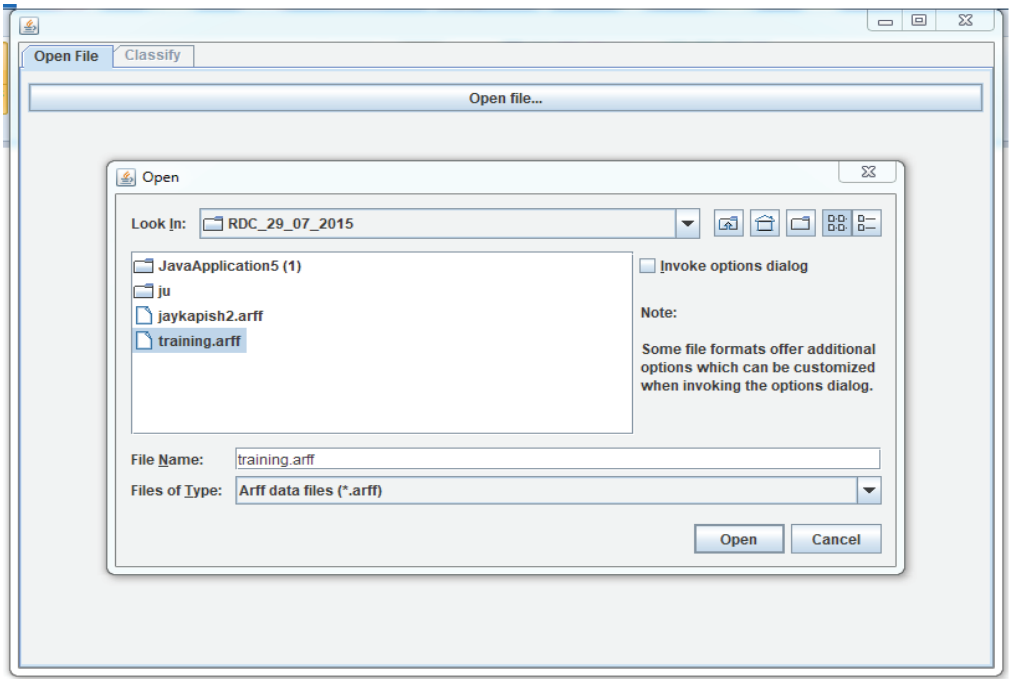

Figure 2. Customized Prediction Model.

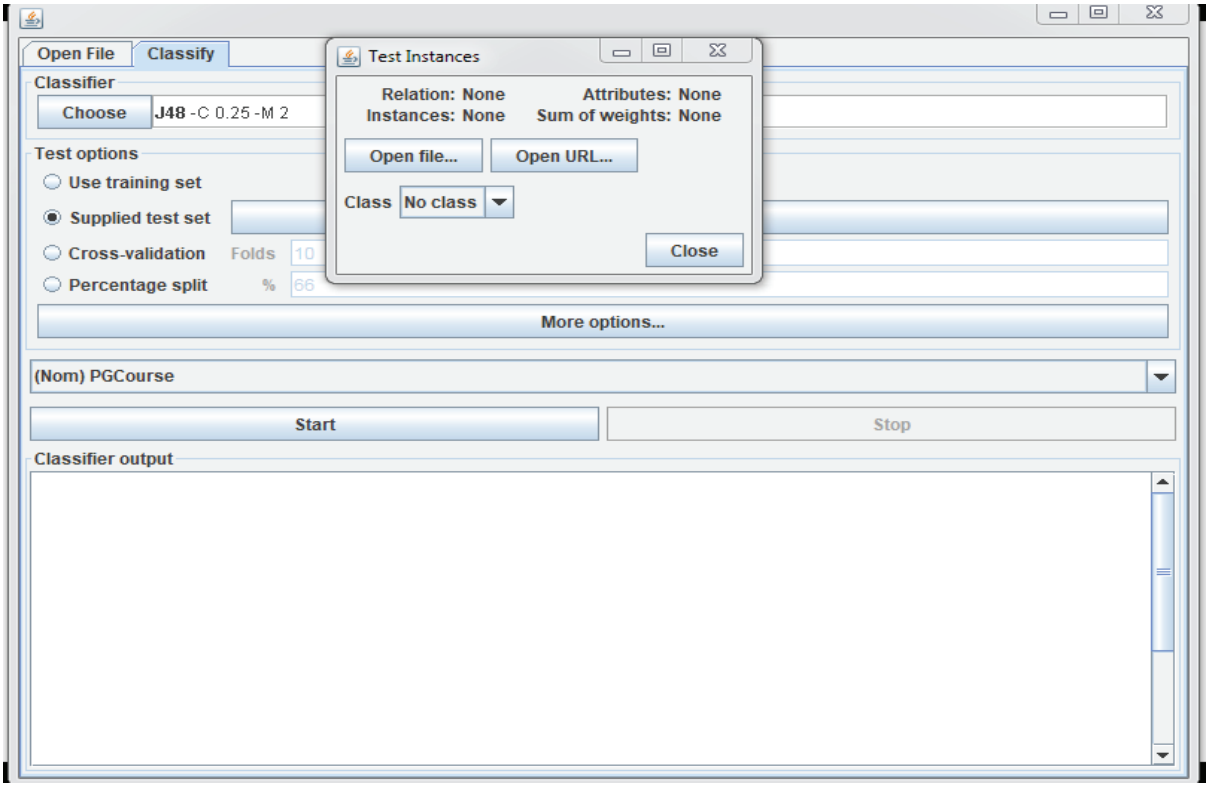

Figure 3. Testing in Prediction Model.

| | | Category = Other: MCA/M.Sc.(IT) (0.0)
| | Personal Choice = No: M.Sc (4.0)
| HSC Stream = Commerce
| | Reason = Job Prospects: M.Com (15.0)
| | Reason = Peer Pressure: M.Com (3.0/1.0)
| | Reason = Family Business
| | | Graduation = BSC: M.Com (0.0)
| | | Graduation = BCA: M.Com (0.0)
| | | Graduation = be: M.Com (0.0)
| | | Graduation = OC: M.Com (0.0)
| | | Graduation = B.Com: M.Com (155.0/6.0)
| | | Graduation = BBA: MBA (8.0)
| | Reason = Personal Choice: MCA/M.Sc.(IT) (9.0)
| | Reason = Market Demands: MCA/M.Sc.(IT) (6.0)
| | Reason = Programming: M.Com (0.0)
Clstat = Self Financed
| Graduation = BSC: MCA/M.Sc.(IT) (58.0/2.0)
| Graduation = BCA: MCA/M.Sc.(IT) (534.0/10.0)
| Graduation = be: MCA/M.Sc.(IT) (0.0)
| Graduation = OC: MBA (6.0)
| Graduation = B.Com
| | pog = Distinction: M.Com (2.0)
| | pog = First Class: M.Com (2.0/1.0)
| | pog = Second Class: MCA/M.Sc.(IT) (10.0)
| | pog = Third Class: MCA/M.Sc.(IT) (0.0)
| Graduation = BBA: MBA (87.0/3.0)

The model is then subjected to the testing phase and new ARFF file is supplied for the same. The testing file is kept low to accommodate results in this paper. As the rules we have narrated above, the locus of the model is completely based on the above rule and classifies the instances according to formed rules. We have taken 10 records for the testing against the trained model and we have got 50% accuracy with 5 instances classified correctly and 5 classified incorrectly. The detailed result instances wise is Table 2.

The table 2 shows that instance number 2, 5, 8, 9 and 10 are wrongly classified as actual class is not as per predicted class. So the predicted value can be replaced by actual class and then the instance will be classified as correctly classified instance. So, the model predicts the

Table 2. Results

inst#	Actual	Predicted	error	Prediction
1	3:M.Com	3:M.Com		1
2	4:MBA	3: M.Com	+	0.961
3	1:MCA/M.Sc.(IT)	1:MCA/M.Sc.(IT)		0.981
4	4:MBA	4:MBA		1
5	2:M.Sc.	3:M.Com	+	0.961
6	3:M.Com	3:M.Com		0.961
7	3:M.Com	3:M.Com		0.961
8	4:MBA	3:M.Com	+	0.961
9	1:MCA/M.Sc.(IT)	3:M.Com	+	0.961
10	4:MBA	1:MCA/M.Sc.(IT)	+	1

Post-Graduation degree for the students from the historical dataset and predicts correct Post-Graduation course for the student.

8. References

1. Shirsavar HR, Moosavi MM. Finding Aptitude in Post-graduates in Educational Management Course in Statistics by the Means of Evaluating the Multiple Intelligence Factors (case Study of Azad University- GARMSAR Branch). Indian Journal of Science and Technology. 2012 Jan; 5(1).
2. Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. ITI 2007 29th Int Conf on Information Technology Interfaces. 2007. p. 51–6.
3. Ventura S, Romero C. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews. 2010; 601–17.
4. Delavari N, Shirazi MRA, Beikzadeh MR. A New Model for using Data Mining Technology in Higher Educational Systems. IEEE. 2004; 319–24.
5. Ogor EN. Student academic performance monitoring and evaluation using data mining techniques. Fourth Congress of Electronics, Robotics and Automotive Mechanics. 2007; 354–9.
6. Ogor EN. Student academic performance monitoring and evaluation using data mining techniques. Fourth Congress of Electronics, Robotics and Automotive Mechanics. 2007; 354–9.
7. Taghavinezhad SZ, Nazari F, Bigdeli Z. Review of factors effecting social networks acceptance among graduate students at Islamic Azad University of Ahvaz. Indian Journal of Science and Technology. 2015 Sep; 8(21).
8. Edin Osmanbegovic MS. Data mining approach for predicting student performance. Economic Review – Journal of Economics and Business. 2012; 3–12.
9. Minaei-Bidgoli B, Kashy DA, Kortemeyer G, Punch WF. Predicting student performance: An application of data mining methods with an educational web-based system. 33'd ASEEIIEEE Frontiers in Education Conference. 2003. p. 13–8.
10. Breasfelean VP. Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. ITI 2007 29th Int Conf on Information Technology Interfaces. 2007. p. 51–6.
11. Alshareet OM. An empirical study to develop a Decision Support System (DSS) for measuring the impact of quality measurements over Agile Software Development (ASD). Indian Journal of Science and Technology. 2015 Jul; 8(15).
12. Chamoli S, Tenne G, Bhatia S. Analysing software metrics for accurate dynamic defect prediction models. Indian Journal of Science and Technology. 2015 Feb; 8(S4).
13. Padmapriya DA. Prediction of higher education admissibility using classification algorithms. International Journal of Advanced Research in Computer Science and Software Engineering. 2012; 330–6.
14. Osmanbegovic E, Suljic M. Data mining approach for predicting student performance. Economic Review. Journal of Economics and Business. 2012; 3–12.
15. Undavia JN, Dolia PM, Shah NP. Prediction of graduate students for master degree based on their past performance based on their past performance using decision tree in weka environment. International Journal of Computer Applications. 2013; 23–9.
16. Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. Indian Journal of Science and Technology. 2015 Jul; 8(15).