IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Predicting Student Performance Using Classification Data Mining Techniques

Isha Shingari[1*] Dinesh Kumar[2]

School of Computer and System Sciences, Jaipur National University, Jaipur

[*]*Corresponding Author: shingariisha@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 17/Jul/2018, Published: 31/July/2018

Abstract – The term education data mining deals with extracting knowledge out of academic database which can be used for providing suitable patterns to education managers, teachers, and students. Education is a progressing field and students need to put in extra efforts to keep the right move towards learning. This paper presents on approach to study the student data and implementing various data mining classification algorithms. Thus, finding out the best algorithm, that can help in evaluating the final grade of a student and finding the best fit for identification of possible results beforehand, so that appropriate interventions can be planned. For our research we collected the data from a reputed higher education institute related to a set of students pertaining to their current and previous academic records. The data were filtered, cleaned, and processed for training different data mining models to define classifications based on different criteria. This method may be considered useful in finding out the students who are at the state of high risk in a very early stage, thus allowing the educationists to provide the appropriate advice to learners in a timely manner.

Keywords– Education Data Mining, academic intervention, Data Classification, pattern identification.

I. INTRODUCTION

Education system has evolved a lot over the years. New innovation takes into consideration the customized training, which empowers the students to study more productively and thereby giving educators the ways which help each student separately if necessary, regardless of whether the class is expansive [1]. Evaluations should compress in a solitary number or letter how well a scholar could comprehend and apply the information passed on a course. Hence it is essential for the students to obtain the fundamental help to pass and do well in a class. In any case, with huge class sizes at colleges and much bigger class sizes in Massive Open Online Courses (MOOCs) it has turned out to be incomprehensible for the teacher and instructing partners to monitor the execution of every learner separately. Subsequently, in both disconnected and online instruction, it is of much significance to create robotized customized frameworks that foresee the execution of a student in a course before the course is finished and at the earliest opportunity [2].

In this research paper, our focus is to predict the final grades in the conventional classroom teaching, where the previous assessments are available. Predicting the final grades is quite challenging. First of all, even if the same curriculum is followed each year, the examinations and the assignments are altered every term. Thus, the criteria for final grade prediction can be different with the change in assignments. Secondly, the student background is very distinctive, so an overall prediction could not work for all students. There has to be an algorithm which could predict the final grade keeping in mind the individualistic approach for every student. The rest of the paper is organized as follows, section II contains the related work, Section III has the Methodology, which describes the various methods required in the research, Section IV shows the Results and discussion and finally in the Section V , the future scope of the research is proposed and conclusions are made.

II. RELATED WORK

Figure 1, explains the procedure of data mining in a nutshell. The process includes five steps [4]:

- Data pre-processing: It helps in removing the noisy data and building an understanding of the educational data set.
- Data standardization: After the irrelevant data is removed, the data has to be cleaned and prepared in order to achieve the desired results.

- Modeling: In this stage, the various data mining techniques are applied over the data.
- Model Evaluation: The models and techniques employed in the previous stage and evaluated and on the basis of which results are evaluated.
- Deployment: This process involves the evaluation of results and presenting them in graphs, diagrams etc. It is the final result and helps in figuring out a clear understanding of the research being carried out.

In our research paper we have undergone the above stated steps and produced the desired results.

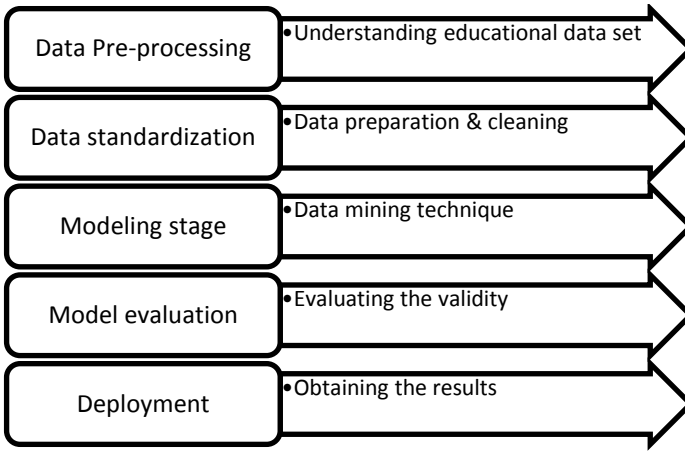

Figure 1. Data mining process

III. METHODOLOGY

The educational data mining involves the process that helps educators for planning, designing, building and maintaining the educational systems. The students make use of this data. The process of data mining then be applied in studying these patterns and making the recommendations by evaluating these patterns [3]. As shown in the Figure 2, it explains the process of education data mining. The process of data mining extracts the hidden patterns out of the educational systems and discovers the useful recommendations. These recommendations are thereby useful in predicting the valuable results.

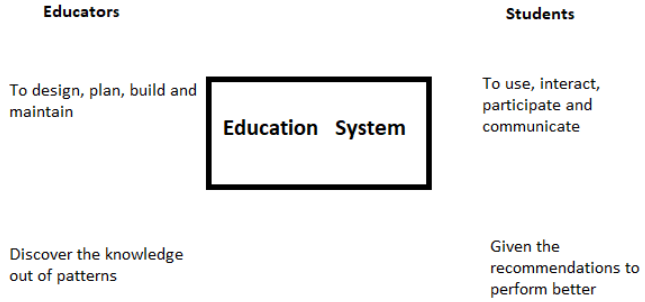

Figure 2: The process of data mining on education system.

The process of education data mining uses many techniques like K-Nearest neighbor, Naïve Bayes, Neural Networks, Decision trees and many more [5]. These techniques help in discovering clustering, classifications and association rules. The explored knowledge can then be used for making predictions regarding the student performance. Examinations play a crucial role in a learner's future. The marks incurred by him decide the result. If at all the prediction says that the scholar tends to fail in the examination, he can put in extra efforts in order to improve his results, and pass with good grades in the exam, the good grades obtained in the exams can also help in securing him a good job.

In this connection, the objectives of the present investigation were framed so as to assist the low academic achievers and they are:

a) It involves the generation of a data source of predictive variables,

b) The next thing is the identification of different factors, which affects a student's learning behavior and performance during academic career,

c) Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables and

d) It should conduct the validation of the developed model for engineering students studying in Indian Universities or Institutions [6].

IV. RESULTS & DISCUSSION

A. The Data Set

Every year, lots of students seek admission aspiring to become successful professionals and get quality education. One such course is engineering. In order to get good placements it's important to score great marks and develop the skills required for the corporate world. In this paper we have taken the data of 36 students of engineering first year.

We have taken the marks of 1st semester as the target data which we have predicted using the input variables like gender, marks scored in 10th and marks scored in 12th. We have used various data mining techniques and various models, the results have been shown in the results section.

B. *Evaluation of the Experiments*

In the figure 3, we have shown the box plot analysis of the Distribution of 12th and 1st semester marks according to gender. In Figure 4, the correlation summary is shown.

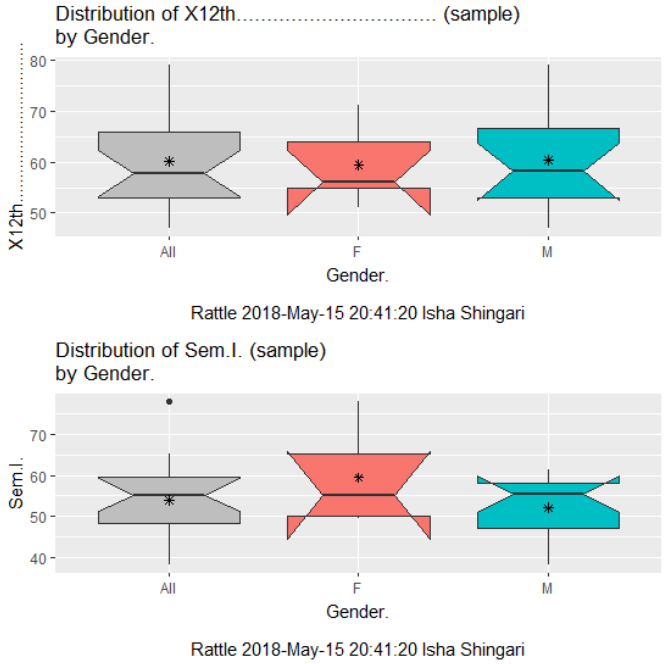

Figure 3: Distribution of 12th and 1st semester marks according to gender.

A correlation could be positive or negative. The positive correlation means both variables move in the same direction, and negative means that when one variable's value increases, the other variables' values decrease. Correlation can also be neural or zero, meaning that the variables are not related to each other. The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample [7]. It is the normalization of the covariance between the two variables to give an interpretable score as shown in Table 1.

- Correlation summary using the 'Pearson' covariance.

Table 1: Correlation summary of marks obtained in 10th and 12th

	12th
12th	1.0000000
10th	0.4137873

	10th
12th	0.4137873
10th	1.0000000

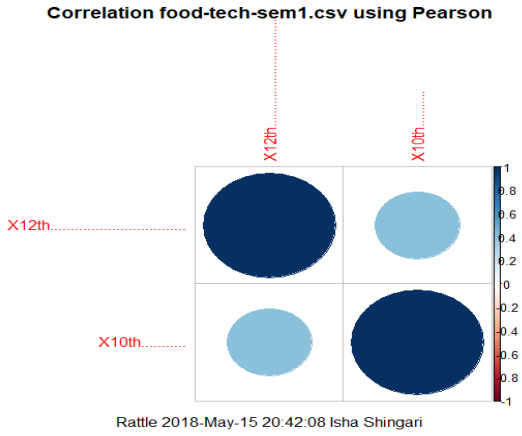

Figure 4: Correlation Data Summary

The Positive correlations are displayed in blue color, the size of the circle and the color intensity are directly proportional to the correlation coefficients. The right side of figure 4, indicates legends corresponding to the correlation coefficients [9].

- Importance of components:

Table 2: Principal Component Analysis

	PC1	PC2
Standard deviation	**1.1890**	**0.7656**
Proportion of Variance	**0.7069**	**0.2931**
Cumulative Proportion	**0.7069**	**1.0000**

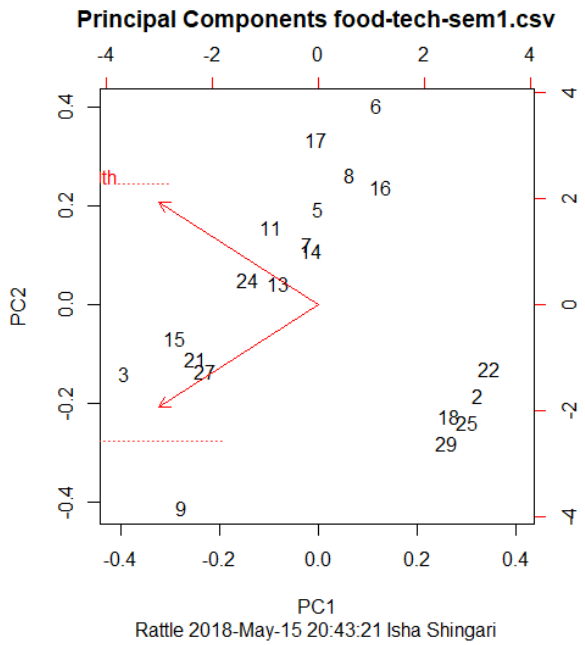

Figure 5: Principal Components of the data.

Table 3: Time taken to build the models

S.No.	Model	Time Taken to build the model(Seconds)
1	Decision Tree	0.00
2	Random Forest	0.46
3	Linear Model	0.02
4	Neural Network	0.02

According to the statistics obtained as shown in Table 2, the decision tree is the fastest model obtained, as it requires the least time to build. The figure 6 shows the decision tree plot for the data.

- Summary of the Decision Tree model for Classification (built using 'rpart'):

- Rpart is the library in R that is used to construct the decision tree. Classification indicates that the modeling technique was applied to a set with a categorical dependent variable.
- The value of n indicates the number of observations used in the model.

n= 20
node), split, n, deviance, yval
* denotes terminal node, it indicates what information is printed in the output for each node

1) root 20 1587.6380 53.97500
2) X12th.....>=56.7 11 773.7873 51.84545 *
3) X12th.....< 56.7 9 702.9956 56.57778 *

- This output prints the tree in an extended form, that is, it describes exactly each node in the tree accordingly to the print specifications described in the previous bullet. The indentation is used to indicate the tree topology, that is, it indicates the parent and child relationships
- The following data shows the number of splits for the tree

Regression tree:
rpart(formula = Sem.I. ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)], method = "anova", parms = list(split = "information"),
control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] X12th.....
Root node error: 1587.6/20 = 79.382

- This is the error rate for a single node tree, that is, if the tree was pruned to node 1. It is useful when comparing different decision tree models.

n= 20

- The complexity table provides information about all of the trees considered for the final model. It lists their complexity parameter, the number of splits, the re substitution error rate, the cross-validated error rate, and the associated standard error.

Table 4: Complexity Table

CP	nsplit	rel	error	xerror	xstd
1 0.069824	0	1.00000	1.0916	0.39537	
2 0.010000	1	0.93018	1.0916	0.39537	

Nodes are labeled with unique numbers. Those numbers are generated by the following formula: the child nodes of node X are always numbered 2x (left child) and 2 x+1(right child). The root node is 1.
Primary Split. Marks in 12th is the predictor variable used for the primary split. The same predictor variable can be used to split many nodes.
Split Point. Nodes 2 and 3 were formed by splitting node 1 on the predictor variable ,arks in 12th. The split point is 57. Node 2 consists of all rows with the value of marks greater than 57, whereas node 3 consists of all rows with marks less than 57[8].

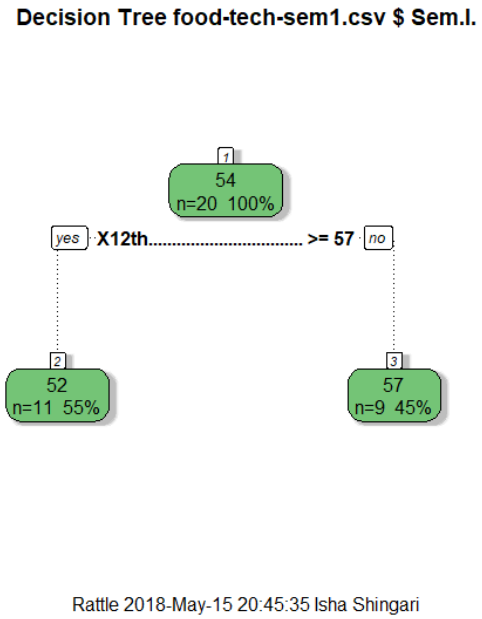

Figure 6: Decision Tree for the data.

Predicted Versus Observed

The Predicted Versus Observed plot is relevant for regression models (predicting a continuous value rather than a discrete value). It will display the predicted values against the observed values, as the name suggests! Two lines are also plotted, one being a linear fit to the actual points, and the other being the perfect fit, if the predicted values were the same as the actual observations. The Pseudo R-Squared is a measure that tries to mimic the R-Squared. It is calculated as the square of the correlation betweenthe predicted and observed values [10]. The closer to 1, the better. The figure 7 and Figure 8 shown below acknowledges the observed vs predicted marks.

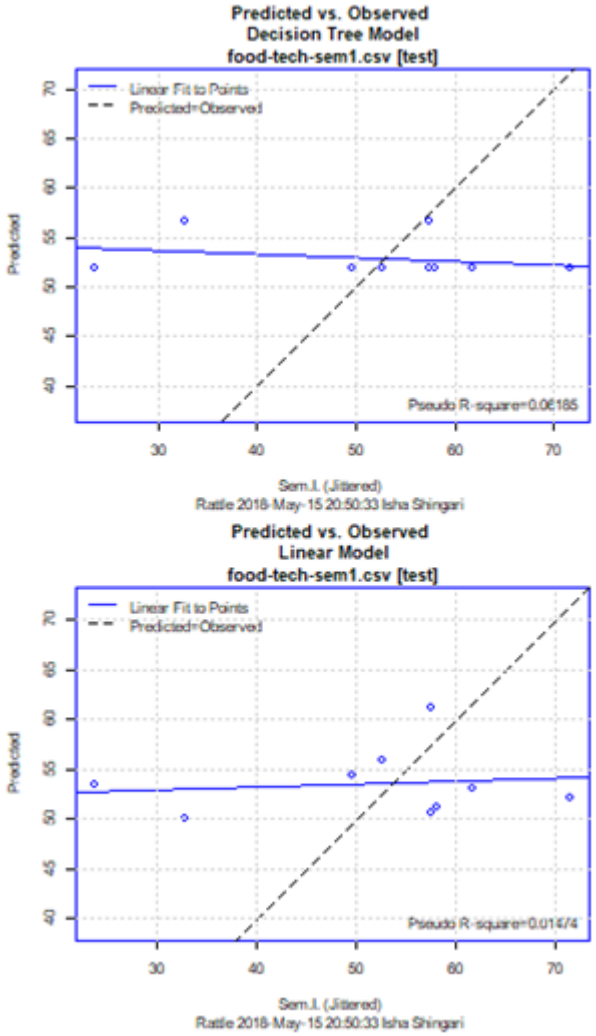

Figure 7: Predicted vs observed values decision tree model, linear model

A common and simple approach to evaluate models is to regress predicted vs. observed values (or vice versa) and compare slope and intercept parameters against the 1:1 line.

However, based on a review of the literature it seems to be no consensus on which variable (predicted or observed) should be placed in each axis. Although some researchers think that

it is identical, probably because *r*2 is the same for both regressions, the intercept and the slope of each regression differ and, in turn, may change the result of the model evaluation[11].

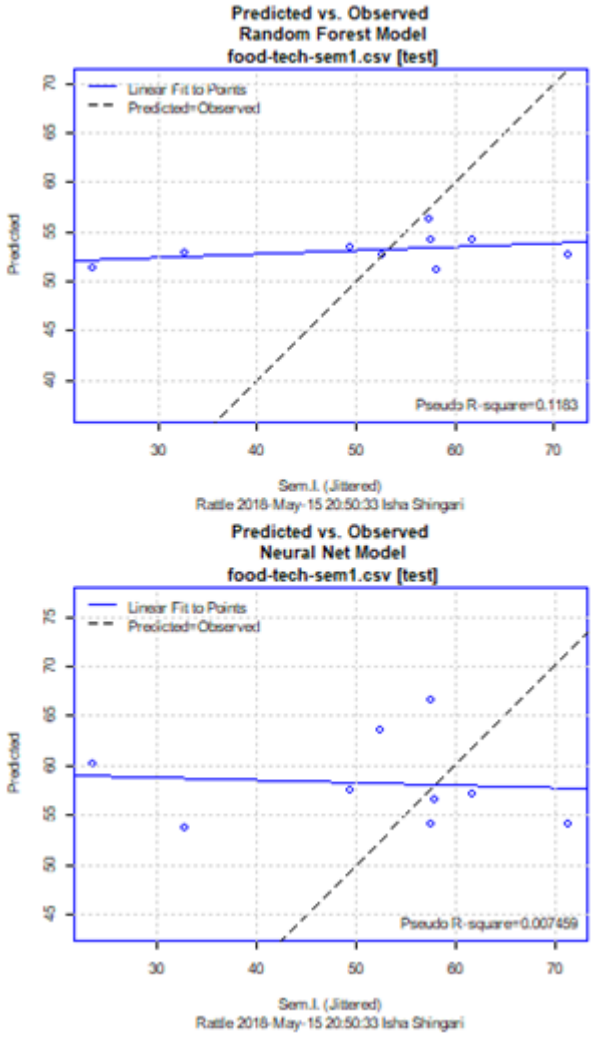

Figure 8: Predicted vs observed values Random Forest model, neural network model

V. CONCLUSION & FUTURE WORK

The Rattle package provides a GUI platform toward using R as a programming language. Rattle is open source data mining tools packed under the regime of R. In this paper, one data set was mined. If one compares the data set results, then it may be concluded that decision tree is the best suited classification algorithm. We hence found that the female candidates of the University did better than the boys. Moreover, as this paper dealt with only one examination i.e. Bachelor of Technology, there are lots of another Examinations to deal with as well as one may extract valuable patterns and information from them. The future plan is to compare entry and exit data of more courses in bachelors, masters level students. Hence it will help in predicting student performance and the factors associated and thereby helping them to fetch a better job and build a successful career.

REFERENCES

[1] C. Tekin, J. Braun, and M. van der Schaar, "etutor: Online learning for personalized education," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015.

[2] Y. Meier, J. Xu, O.Atan and M.V. D. Schaar" Personalized Grade Prediction: A Data Mining Approach" 2015 IEEE International Conference on Data Mining.

[3] K. H. Rashan, Anushka Peiris, "Data Mining Applications in the Education Sector", MSIT, Carnegie Mellon University, retrieved on 28/01/2011

[4] A. Dutt, M. A. Ismail, T. Herawan "A Systematic Review on Educational Data Mining" 2169-3536 (c) 2016 IEEE.

[5] S. K. Yadav, S. Pal " Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[6] I. Shingari , D.Kumar" A Survey On Various Aspects Of Education Data Mining In Predicting Student Performance" JASC, June , 2018

[7] A. Auysha1, A. Jayachandran" Online Ensemble Learning of Data Streams with Gradually Evolved" International Journal of Scientific Research in Computer Science, Engineering and Information Technology , Vol 2, Issue 2.

[8] Himanshi, Komal Kumar Bhatia" Prediction Model for Under-Graduating Student's Salary Using Data Mining Techniques" April, 2018

[9] A.Deepa , E. Chandra Blessie" Input Analysis for Accreditation Prediction in Higher Education Sector by Using Gradient Boosting Algorithm" June, 2018

[10] V. Kumar "Data Mining with R Learning with Case Studies" Chapman & Hall/CRC Data Mining and Knowledge Discovery Series

[11] G. P. eiro, S.Perelman, J. P. Guerschman, J. M. Paruelo "How to evaluate models: Observed vs. predicted or predicted vs. observed?" Elsevier Ecological Modelling · September 2008

Author's Profile

Isha Shingari obtained her Bachelors in Technology (honors) in Information Technology from Rajasthan Technical University and Masters in Technology(computer science engineering) from Mody Institute of Technology and Science in year 2010 and 2013 respectively. Currently, she is pursuing her PhD. From Jaipur National University. Her area of interests lie in pattern mining, data mining, machine learning.

Dr. Dinesh Kumar obtained his Bachelors and Masters degree (computer science) from Birla Institute of Technology and Allahabad Agricultural University. He obtained his PhD. From NIMS university,Rajasthan. He has published in various journals and conferences of International repute. He has over a decade of teaching experience and has been supervising various masters and PhD. Research scholars.

