Research Article

Student Performance Measure by Using Different Classification Methods of Data Mining

Ashutosh Mishra[a], and Neha Chaudhary[b]
a,b
Department of Computer Science and Engineering, Thapar Institute of Engineerng and Technology. Patiala, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online:28April 2021

Abstract: The assessment in outcome based learning is very vital and significant approach toward measuring the student's performance. There are many traditional methods existing in this context. The data mining is one of the intelligent computing methods which are having widely accepted features that enable the idea of its usage in assessment. Much work has been done to measure the student performance by using different methodologies and modern technologies. In this work, we have gone through the current datasets of students of the university and different classification methods of data mining are used to measure the accuracy of student performance. Based on the analysis of the result, it has been concluded that accuracy and the other measures of SVM is more than the other classification methods.

Keywords: Student performance, classification, machine learning, educational data mining

1. Introduction

New technologies are being developed in the field of data management and analysis due to large supply of data being present in several companies, including both private and public. The main aim of the techniques of data mining is to discover hidden and insignificant links within the information having diverse characteristics. Various techniques of data mining are being used in different fields including the educational environment.

In the sector of education, educational data mining is an emerging discipline which is very recent and its practice is preconceived to identify and extract new and valuable knowledge from the data [1].The aim is to resolve problems of research areas of education and improve the whole educational process using various statistical techniques, data mining algorithms and machine learning programming. Educational data Mining (EDM) is a prospering discipline that can be used for analysis and visualization of data, predicting student performance, student modelling, grouping of students etc [2].

There are numerous challenges in the higher education like increase in the number of students, global competitive education market, rising student expectations, a demand and need for new technologies, significant reductions in government funding, etc[3]. Therefore, educational data mining assists to develop methodologies that allow to improve the overall process of education. It has been observed that mostly data mining techniques involve large data sets to work with. But in the ambience of education, we are usually encountered with relatively small data sets containing small groups of students. In addition, it is helpful to the administrators in decision making. From several years, various data mining classification and clustering models have been constructed and executed to analyze and measure the performance of students. For instance, AHP has been employed successfully to predict the student course selections in higher education and the outcomes show that the accuracy of the student's course prediction is high [4]. Shahiri et al. [5] have also provided an overview on several techniques of data mining that were applied to predict and analyze performance of students, concentrating on the identification of most valuable attributes in a student's data by employing the prediction algorithm**.** Osmanbegović and Suljić [1] applied three supervised data mining algorithms on the assessment data of first year students to predict favourable outcome in a course and evaluating the performance based on certain factors like convenience, accuracy and approach of learning**.** A model has also been developed based on some selected input attributes assembled through questionnaire method [6]. Goga et al. [7] designed a tool by using .NET framework to predict student's grade by providing various parameters as input. Models based on the student's enrollment records were developed by using ten classifications trees (OneR, Random forest, ZeroR, random tree, Decision stump, REPTree, JRip, J48, PART, and Decision table) and a multilayer perceptron (Artificial Neural Network) learning algorithms by operating on WEKA(Waikato Environment for Knowledge Analysis). Prediction model is developed based on the participation of the students through Genetic Programming by integrating educational data mining and learning analytics [8].

The recent research by Natek and Zwilling[9] compares two tools of data mining applied to data sets of small size related to institutes of higher education and summarizes that the results will encourage the institutions to incorporate the methods of data mining to be an essential segment of higher education institutes and intelligence management systems. Another research has developed an admission system based on ANN techniques using several machine learning methods to design the valid criteria to the selection admission for higher education [14]. Our research work proposes an effective methodology for measuring the student's overall performance by using the grades of each semester. Different results of the classification algorithms of data mining are analyzed and final outcome is made based on the accuracy of the model.

2. Methodology

Many universities adopt the grading system to estimate and decide the performance of students in academics. The approach adopted by us also uses the grades for the analysis and measurement of performance.

A. Workflow

The first and foremost step is to collect the dataset required for the study. The methodology is applied to a factual data having information about the students who did their graduation in Computer Science and Engineering at Thapar University (India).

Once the data is fetched, we then transform the data into an appropriate and required form for the mining process, which is known as pre - processing phase. This task is accomplished by using a specific mining method, algorithm or technique. This phase usually consumes 60 to 90 % of the time, training, efforts and resources employed in the complete knowledge discovery process. R. Asif et al. [10] highlights the importance of this step.

After the data is pre-processed, we then identify the incomplete, incorrect, irrelevant and inadequate data from our dataset and remove this erroneous and improperly formatted data. This phase is known as data cleaning phase as shown in Figure1. When the data is complete and consistent in all respects, the next step is to filter the data according to our requirements. Since all the information is in one file and is jumbled, so separate the data of each semester with same attributes. The major attributes are described in Table1.

Now the data is ready for the data mining methods to be applied and generate the model so that it can be further used for analysis and prediction. The algorithms proposed are: SVM, Bayesian Network and Decision Tree (C5.0). For the purpose of this study, IBM SPSS Modeler is employed which is an extensive predictive analytics platform and is used for predictive intelligence decision making. It includes a wide range of advanced algorithms and techniques that helps to make the right decisions [11].

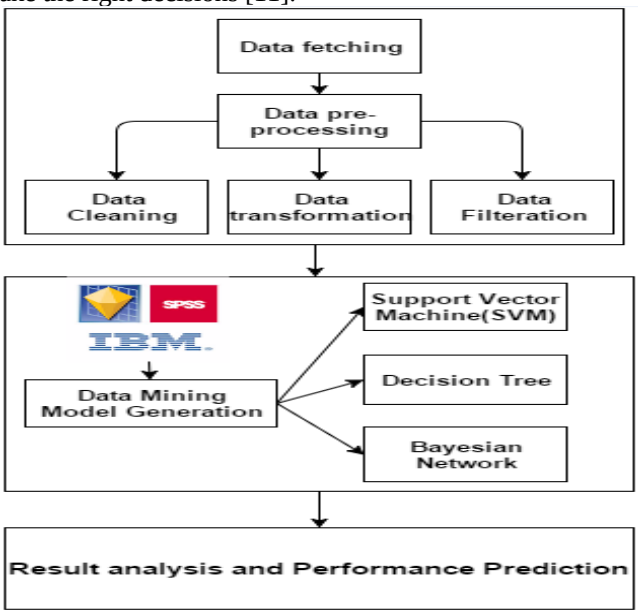

Figure1. Workflow of Study

Table1: Attributes and their description

ATTRIBUTES	DESCRIPTION
Exam code	The code of the exam, whether odd or even semester
Academic year	The year of studying the subject
Subject code	The unique code of the subject
Semester	The semester in which the student is studying
Subject	The subject of study
Enrollment no	Distinct roll number of the student
Student name	Name of the student
Grade	The grade obtained by the student, whether A, B, C, D or E

3. Implementation

The data mining classification algorithms are different in many aspects such as: the learning rate, performance, speed, robustness, accuracy, etc [1].In this research, we investigated the impact of three algorithms for performance prediction: SVM, Bayesian network and C5.0 decision tree algorithm.

Support Vector Machines (SVM) is the newest technique for supervised machine learning. SVMs spin about the notion of the margin-any side of a hyper plane that splits two data classes [12]. Our study includes four kernels of SVM namely, RBF (Radial Basis Function), polynomial, sigmoid and linear. Bayesian Network is the directed

acyclic graph (DAG) portraying dependence and independence between variables. Decision tree algorithm is a tree like framework that classifies the instances beginning from source node to the leaf node by electing the variables at each level so that the set of items are perfectly separated [13].In our study, the decision tree generated by C5.0 is used for the purpose of classification.

Procedure

Step 1: Semester-wise collect all the grades attained by each student in same sequence of subjects, as shown in Figure2.

ENROLLMENTNO	GRADES(1)	GRADES(2)	GRADES(3)	GRADES(4)
101003001	A,D,B,D,D,C	D,B,B,D,C,D	C,C,C,C,E,E,D	E,C,B,B,D,C,D
101003003	C,B,B,B,C,B	C,C,C,C,C,C	B,B,B,C,C,D,B	B,C,A,B,C,B,A
101003004	D,C,C,B,B,C	D,D,E,D,D,C	B,B,C,C,C,C,D	C,C,C,C,B,D,B
101003005	C,D,B,C,C,B	D,C,D,B,C,B	C,D,D,B,B,C,C	A,B,C,B,C,B,C
101003006	B,B,A,A,B,A	A,A,B,B,A,A	A,B,A,A,A,A,B	B,A,B,A,B,A,A
101003007	C,C,E,C,D,D	C,D,D,D,F,D	D,C,E,D,D,E,E	D,D,D,D,C,C,C
101003008	B,B,B,B,B,B	B,B,B,C,B,B	A,B,C,B,B,B,B	A,C,B,B,C,B,B
101003009	C,D,D,C,D,B	D,D,F,D,D,D	C,D,D,D,C,E,D	C,D,C,E,E,C,E
101003010	A,A,B,B,A,A	B,A,B,B,C,A	A,A,B,A,A,B,B	A,B,B,B,B,A,C
101003012	B,C,D,D,C,C	D,D,D,F,D,C	C,D,D,D,C,C,C	D,C,E,C,C,D,D

Figure2. Collection of grades for four semesters

The grades are collected as GRADES (i), where 'i' is semester and $1 \leq i \leq 4$.

$S_i = \{GS_1, GS_2, GS_3 ... GS_j\}$, Where, GS=grade of the subject

S_i=semester, $1 \leq i \leq 4$

j=number of subjects corresponding to i^{th} semester. j=6 when i (i.e. semester) is 1 or 2 and j=7 when i is 3 or 4.

e.g. S_2= {B, A, B, B, C, A} is the sequence of grades of the student in second semester having enrolment no. 101003010 and studying six subjects as highlighted in Figure2.

The grades and their corresponding performance criteria are shown in Table2. This performance criteria is used for prediction in the final outcome i.e. OVERALLGRADE (F), where 'F' stands for final.

Step 2: Prepare the logic table (n^{th} level logic predicate) on the basis of GRADE in Table II.

Level-wise logic order:

L_0: A – A → A
D - D → D
B – A → A
B – B → B
A – B → A

L_1: A – C → B
B – D → C
C – E → D

L_2: A – D → B
B – E → C
D – A → C
E – B → D

L_3: A – E → C

Table2: Performance on the basis of grades

GRADE	PERFORMANCE	MARKS
A	Excellent	9-10
B	Good	8-7
C	Average	6-5
D	Poor	4-3
E	Fail	2-1

e.g. We have the sequence of grades {B, A, B, B, C, A} for the student having enrolment no. 10603010.While applying the step 2, take two consequent grades together and then compute the output in the way as shown below in Figure3 using Table2. .Like, B and A gives output A, then take this output as input for the next step. Now, A and B gives output A. These steps are done using the above level-wise logic order until we reach to our final grade.

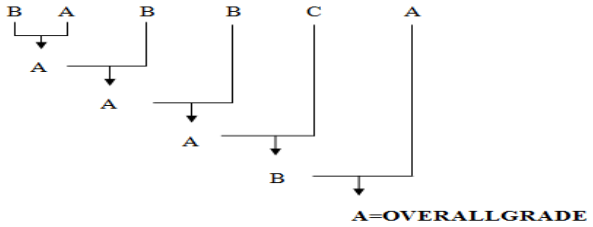

Figure3. Processing of grades

This shows that the performance of the student is **Excellent**, on the basis of Table 2 and is highlighted in Figure4.

Step 3*:* Therefore, OVERALLGRADE (i), $1 \leq i \leq 4$ is computed for each semester (in the same manner as used above), where *i* is the semester. E.g. OVERALLGRADE (2) is the overall grade computed for second semester.

ENROLLMENTNO	GRADE(1)	OVERALL GRADE(1)	GRADE(2)	OVERALL GRADE(2)	GRADE(3)	OVERALL GRADE(3)	GRADE(4)	OVERALL GRADE(4)
101003001	A,D,B,D,D,C	C	D,B,B,D,C,D	C	C,C,C,C,E,E,D	D	E,C,B,B,D,C,D	C
101003003	C,B,B,B,C,B	B	C,C,C,C,C,C	C	B,B,B,C,C,D,B	B	B,C,A,B,C,B,A	A
101003004	D,C,C,B,B,C	B	D,D,E,D,D,C	C	B,B,C,C,C,C,D	C	C,C,C,C,B,D,B	B
101003005	C,D,B,C,C,B	B	D,C,D,B,C,B	B	C,D,D,B,B,C,C	B	A,B,C,B,C,B,C	B
101003006	B,B,A,A,B,A	A	A,A,B,B,A,A	A	A,B,A,A,A,A,B	A	B,A,B,A,B,A,A	A
101003007	C,C,E,C,D,D	C	C,D,D,D,F,D	D	D,C,E,D,D,E,E	D	D,D,D,D,C,C,C	C
101003008	B,B,B,B,B,B	B	B,B,B,C,B,B	B	A,B,C,B,B,B,B	B	A,C,B,B,C,B,B	B
101003009	C,D,D,C,D,B	B	D,D,F,D,D,D	D	C,D,D,D,C,E,D	D	C,D,C,E,E,C,E	C
101003010	A,A,B,B,A,A	A	B,A,B,B,C,A	A	A,A,B,A,A,B,B	A	A,B,B,B,B,A,C	B
101003012	B,C,D,D,C,C	C	D,D,D,F,D,C	C	C,D,D,D,C,C,C	C	D,C,E,C,C,D,D	C

Figure4. Overall grade computation for each semester

Step 4: Similarly, OVERALLGRADE (F) for each of the 126 student is computed as the final performance result. First ten results are shown in Figure5.

ENROLLMENTNO	OVERALL GRADE(1)	OVERALL GRADE(2)	OVERALL GRADE(3)	OVERALL GRADE(4)	OVERALL GRADE(F)
101003001	C	C	D	C	C
101003003	B	C	B	A	A
101003004	B	C	C	B	B
101003005	B	B	B	B	B
101003006	A	A	A	A	A
101003007	C	D	D	C	C
101003008	B	B	B	B	B
101003009	B	D	D	C	C
101003010	A	A	A	B	A
101003012	C	C	C	C	C

Figure5. Final performance computed using overall grades of four semesters

Step 5: Now various data mining methods are applied to OVERALLGRADE (F), which is the measured performance of each student.

4. Experimental Setup

The data for the model was collected for four semesters of Computer Science Engineering students, batch 2016-20 studying in Thapar University, Patiala. After eliminating the incomplete and unwanted data, the sample comprised 126 students having 'j' subjects. There are six subjects in first two semesters and seven subjects in semester 3 and 4.So,for semester1 and semester 2, j=6 and for semester 3 and semester 4, j=7.

Eq. 1 is the total number of records

$$\sum_{j=1}^{i} S_j * N \tag{1}$$

Eq. 2 is total number of records for each student

$$\sum_{j=1}^{i} S_j \tag{2}$$

Here, *'i'* is the number of semesters. *'S_j'* is the number of subjects corresponding to i^{th} semester and *'N'* is the total number of students.

So, for 126 students there are 6*126 + 6*126 + 7*126 + 7*126 = 3276 data records. Each student is associated with 6+6+7+7=26 records. The outcome of each model is the student's predicted final result, which is then compared with our manual predicted performance result.

The three algorithms and their implementation using IBM SPSS Modeler are illustrated in the figures. There are four inputs: OVERALLGRADE(1) i.e. overall grade of first semester, OVERALLGRADE(2),

OVERALLGRADE(3) and OVERALLGRADE(4) and the target being OVERALLGRADE(F).The type of all the input and output fields is nominal.

Figure6 and Figure7 represent the implementation of four kernels of SVM: RBF (Radial Basis Function), polynomial, sigmoid and linear. Each kernel elucidates the predictors in a different way so the output is taken in the form of tables. Figure8 shows the network model generated and Figure9 displays the rule set generated for overall performance using grades obtained in various semesters.

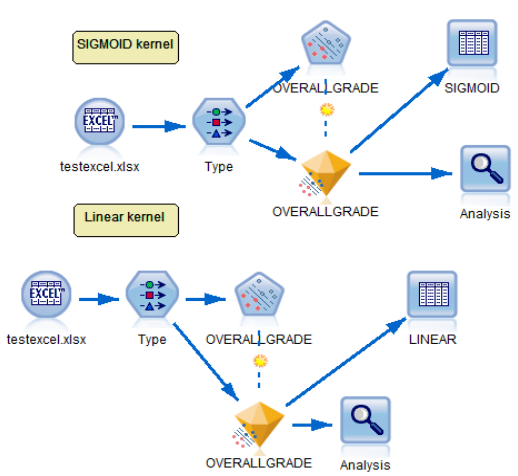

Figure6. RBF and polynomial implementation of SVM kernels of SVM kernels

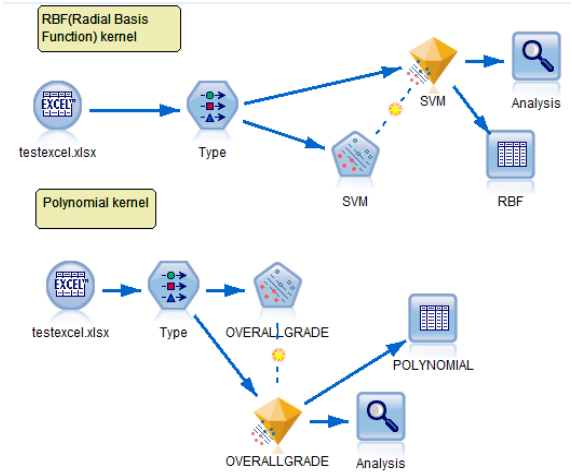

Figure7. Sigmoid and Linear implementation of SVM kernels

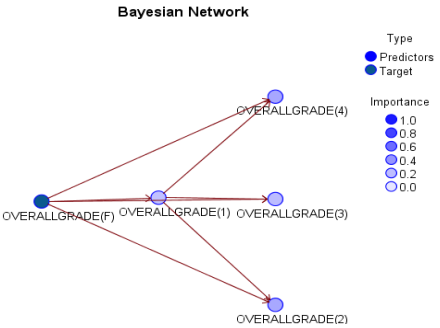

Figure8. Directed Acyclic Graph(DAG)

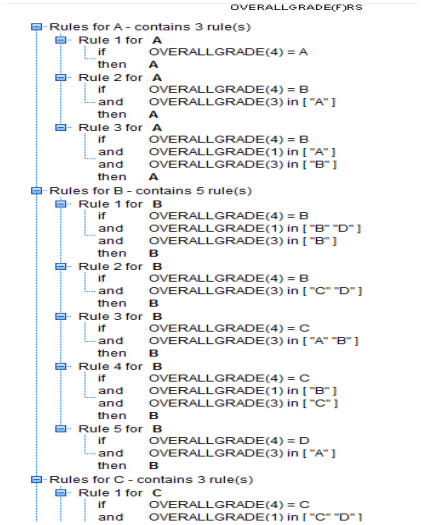

Figure9. Rule set generated by C5.0

5. Results

The four algorithms provide different accuracy levels, i.e. each of them interprets the relevance of attributes in a different way. There are different evaluation criteria based on the classification algorithms used.

Table3 compares the correct and incorrect instances obtained along with the prediction accuracy, when the three algorithms are applied.SVM has the highest accuracy as compared to other classifiers. The graph in Figure10 clearly shows the accuracy levels of the three algorithms, SVM being the most accurate. Similarly, the SVM kernels output is described in Table IV in which two out of four kernels namely, RBF and polynomial show same number of correct instances.

Table3: Comparison of the three classifiers

EVALUATION CRITERIA	**CLASSIFIERS**		
	SVM	**C5.0**	**Bayesian Network**
Correctly classified instances	123	121	121
Wrongly classified instances	3	5	5
Prediction accuracy	97.62%	96.03%	96.03%

A report based on the confidence values is shown in Table5. All the four kernels of SVM and C5.0 and Bayesian Network values are observed based on their predicted values, as generated by the respective models.

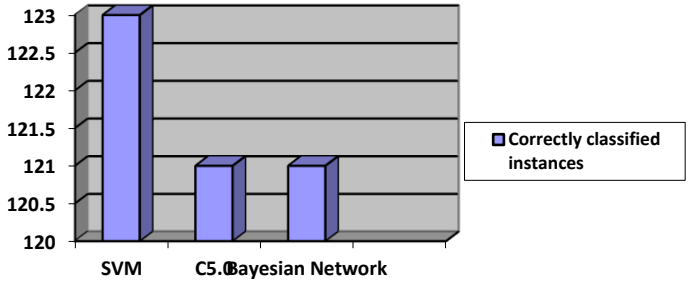

Figure10 Graph of correctly classified instances

Table4. Comparison of different kernels of SVM Model

EVALUATION CRITERIA	**SVM KERNELS**			
	RBF	**Polynomial**	**Sigmoid**	**Linear**
Correctly classified instances	123	123	80	120
Wrongly classified instances	3	3	46	6
Prediction accuracy	97.62%	97.62%	63.49%	95.24%

Table5. Confidence values report

EVALUATION CRITERIA	**CLASSIFIERS**					
	SVM Kernels				**C5.0**	**Bayesian Network**
	RBF	Polynomial	Sigmoid	Linear		
Range	0.583 - 0.989	0.629-0.982	0.288-0.943	0.607 - 0.994	0.8-1.0	0.513-1.0
Mean correct	0.918	0.922	0.662	0.848	0.964	0.94
Mean incorrect	0.888	0.867	0.528	0.82	0.861	0.674

6. Conclusion and Future Work

1. This work has been done on factual and real data.
2. From the above analysis, we have concluded that SVM produces the best prediction results as compared to Bayesian network and C5.0. SVM exhibits higher accuracy i.e. 97.62%.
3. The results indicate that RBF and polynomial SVM kernels perform better than sigmoid and linear. The proposed methodology can be adopted to help the teachers as well as the students to enhance the quality of learning and student's performance by taking significance decision at right time.

In future work, the study can be enhanced by including various demographic factors and more distinguishing attributes like SSC marks, HSC marks, projects undertaken etc. to obtain more accurate student performance and to determine student behaviour. Also, the work could be carried out with other classification algorithms of data mining to acquire a wider approach and more reliable outputs.

7. Acknowledgements

Authors are greatful to Thapar University for providing the data to carried out this research work**.**

References

1. E. Osmanbegović and M. Suljić. 2012. "Data mining approach for predicting student performance", *Economic Review – Journal of Economics and Business,* vol., no. 1, pp. 3-12.
2. D. P. Nithya, B. Umamaheswari, and A. Umadevi. 2016. "A Survey on educational data mining in field of education," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 5, no. 1, pp. 69–78.
3. R.Srivastava, M.Gendy, M.Narayana, Y. Arun, & J.Singh. 2012. University of the future — "A thousand year old industry on the cusp of profound change", Melbourne, Australia: Ernst & Young (Retrieved from

http://www.ey.com/Publication/vwLUAssets/University of the future/$FILE/University of the future 2.12.pdf).

4. Boylan, H. R., Bliss, L. B., & Bonham, B. S. (1997). Program components and their relationship to student performance. Journal of Developmental Education, 20, 2-9.
5. Van der Berg, S. E. R. V. A. A. S., & Louw, M. (2006, September). Lessons learnt from SACMEQII: South African student performance in regional context. In conference on Investment Choices for Education in Africa (pp. 19-21).
6. S. Natek and M. Zwilling.2014. "Student data mining solution–knowledge management system related to higher education institutions," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6400–6407.
   I. Ognjanovic, D. Gasevic, and S. Dawson.2016. "Using institutional data to predict student course selections in higher education," *The Internet and Higher Education*, vol. 29, pp. 49–62.
      A. M. Shahiri, W. Husain, and N. A. Rashid.2015. "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422.
7. V. Ramesh, P. Parkav and K. Rama.2013. "Predicting Student Performance: A Statistical and Data Mining", *International Journal of Computer Applications*, vol. 63, no. 8, pp. 35-39.
8. M. Goga, S. Kuyoro, and N. Goga.2015. "A Recommender for improving the student academic performance," *Procedia - Social and Behavioral Sciences*, vol. 180, pp. 1481–1488.
9. W. Xing, R. Guo, E. Petakovic, and S. Goggins.2015. "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in Human Behaviour*, vol. 47, pp. 168–181.
10. R. Asif, A. Merceron, and M. K. Pathan.2014. "Predicting student academic performance at degree level: A case study," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 1, pp. 49–61.
11. IBM SPSS Modeler, IBM.2016. Avialable : http://www.01.ibm.com/software/analytics/spss/products/modeler.
12. Mendell, M. J., & Heath, G. A. (2005). Do indoor pollutants and thermal conditions in schools influence student performance? A critical review of the literature. Indoor air, 15(1), 27-52.
13. Cotton, K. (1996). School size, school climate, and student performance.
14. Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13(1), 61-72.
15. K. Kaur and K. Kaur.2015. "Analyzing the effect of difficulty level of a course on students performance prediction using data mining", *Next Generation Computing Technologies (NGCT), 2015 1st International Conference*, pp. 756-76.
16. P. Guleria, N. Thakur and M. Sood. 2014. "Predicting student performance using decision tree classifiers and information gain," *Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on*, Solan, 2014, pp. 126-129.
17. Mengash, Hanan. 2020. Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. IEEE Access, vol.8.