ANALYZING STUDENTS' ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING

Sana
Mehran University of Engineering and Technology. Jamshoro (Pakistan)
E–mail: sanabhutto163@hotmail.com

Isma Farrah Siddiqui
Mehran University of Engineering and Technology. Jamshoro (Pakistan)
E–mail: isma.farah@faculty.muet.edu.pk

Qasim Ali Arain
Mehran University of Engineering and Technology. Jamshoro (Pakistan)
E–mail: qasim.arain@faculty.muet.edu.pk

Recepción: 05/03/2019 **Aceptación:** 01/04/2019 **Publicación:** 13/05/2019

Citación sugerida:
Sana, Siddiqui, I. F. y Arain, Q. A. (2019). Analyzing Students' Academic Performance through Educational Data Mining. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición especial, Mayo 2019*, pp. 12–43. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.15–14

Suggested citation:
Sana, Siddiqui, I. F. & Arain, Q. A. (2019). Analyzing Students' Academic Performance through Educational Data Mining. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019*, pp. 12–43. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.15–14

ABSTRACT

Predicting students' performance is a very important task in any educational system. Therefore, to predict the learner's behavior towards studies many data mining techniques are used like clustering, classification, regression. In this paper, new student's performance prediction model and new features are introduced that have a great influence on student's academic achievement i.e. student absence days in class and parents' involvement in the learning process. In this paper, considerable attention is on the punctuality of students and the effect of participation of parents in the learning process. This category of features is concerned with the learner's interaction with the e–learning management system. Three different classifiers such as Naive Bayes, Decision Tree, and Artificial Neural Network are used to examine the effect of these features on students' educational performance. The accuracy of the proposed model achieved up to 10% to 15% and is much improved as compared to the results when such features are removed.

KEYWORDS

Educational data mining, Students' performance prediction model, Artificial neural network.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

1. INTRODUCTION

In the discipline of data mining and its well–known application Knowledge Discovery in Databases (KDD) , one of the new evolving fields now–a–days is Education Data Mining (EDM) that emphasizes on discovering the useful knowledge and mining the useful patterns from educational information systems such as, course management system (Moodle, Blackboard etc.) , online learning management system , registration systems , admissions systems, and so on which help out students at each stage of their studies like from primary to higher education. Romero and Ventura (2007) proposed that the data can be obtained through manual traditional surveys. The further investigation on education data mining (Romero, Ventura & Garcia, 2008) concluded that data can be gathered from many sources such as databases of academic institutes, online learning management system. In this field, a major focus of concern is to analyze and discover meaningful rules and patterns to either encourage students to manage their education and deliverables in a better way and enhances their performance or to give educational institutes direction to maintain the policies for the betterment of students. Abu Tair and El–Halees (2012) analyzed student's data by creating the decision trees, making an association or sequential mining rules and classifying students for enhancing their performance and taking fruitful decisions in the fascinating research area. Romero and Ventura (2010) concluded that many data mining techniques used to generate specific patterns, rules, classification and prediction to help students in the future. In this paper student performance model is introduced which focus on important features i.e parents' participation in the learning process and student absence days. The dataset is obtained from Kalboard 360 e–learning system. The performance model applies different classifiers such as decision tree, naive bayes, and artificial network to examine the effect of such features on students' academic performance. For building the student's performance model source of data is obtained from http://www.kaggle.com, this is an educational dataset of e–learning website, the dataset contains 500 records and having 17 different features. Then, we applied three of the data mining algorithms. Finally, the results are evaluated by using different measures.

2. LITERATURE REVIEW

Educational data mining is used to find potential knowledge that helps in the utilization of active learning in technological aspects. E–learning is becoming one of the most important areas of research in developing countries. So many well–developed countries switched their educational system into fully or partially automated which not only helps students but teachers as well to provide ease of learning. A survey is made where many data mining application is applied to the course management system. It was a tutorial and case study related to the Moodle system to improve the students' learning experience and their courses. Quadri and Kalyankar (2010) shows C4.5 decision tree algorithm to arrange a set of attributes in hierarchical form, this technique is used by many researchers due to its simplicity through which set of classification rules can be formed. Some of the well–known Decision Tree algorithms are J48, C4.5 and CART. Murugananthan and Shiva (2016) proposed a new approach in deriving association rules for optimal learning sequence of tutors and students using a K–means clustering algorithm. An Artificial neural network is one of the most used practices in mining educational data. This is very intelligent algorithm which works based on a neuron that relate to each other and work together to produce the output. Arsad and Buniyamin (2013) used artificial neural network for predicting academic progress of bachelor's degree student. Hien and Haddawy (2007) used Naïve Byes algorithm to predict final Cumulative Grade Point Average (CGPA) at the time of admission which was based on their academic background. The study about students' educational behavior (Amrieh, Hamtini & Aljarah, 2015) proposed framework having a category of a feature called "Behavioral feature" is introduced where they focus on student's behavioral features and their relationship with student's academic success. The authors (Amrieh, Hamtini & Aljarah, 2016) used the same framework to examine student's progress by using ensemble techniques which enhance the overall accuracy of results. So, numerous researches have been conducted so far to predict the students' performance using data mining. But few of them highlighted the important features that affect students' educational performance. In this research, we are going to use the most

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

important category of the features that affect the grades of a student and their overall performance.

3. DATASET AND DATA PREPROCESSING

The dataset for building the proposed student's performance model to anticipate the students' academic performance is acquired from https://www.kaggle.com/aljarah/xAPI–EDu–Data. It is an instructive dataset collected from e–learning system called Kalboard 360. The dataset consists of 500 student records. It has 17 different features.

3.1. E–LEARNING MANAGEMENT SYSTEM

It is an e–learning system that engages learners, track progress and delivers targeted outcomes. Learning is significant, innovative and interactive. Student engagement was defined by ("Kalboard 360 e–learning system", 2000) as "People ENGAGE and INTERACT with it for better understanding and effective learning. That's why the only focus of this system is on custom–made solutions. Core competency lies in their decade of experience, expertise, and creativity of the solutions". The emphasis is on delivering an inspiring and engrossing experience for students. The aim of this system is to build a world where e–learning and development matters. Their main objective is to tackle recent technologies to develop online learning methods for students and educational institutes. Where they can offer several customized courses options related to students demands. As compared to conventional methods like books, PDFs, PowerPoint's, training manuals they have shifted to fully interactive activities based on e–learning procedures. Course designer prepares a fully interactive course layout where audio voice can be included so that student can get desired content in any format.

Table 1. Features of student dataset and their categories.

Feature	Description of features	Category of features
Gender	The student gender i.e masculine or feminine	Demographical Features
Country	A Country student belongs to.	
Birthplace	Born place of student	
Parent Responsible	Parent of the student (dad or mom)	
Levels of Education	Different educational stages of students like high, medium and low level	Academic Background Features
Student Grade	Grade level of student (GL–1, GL–2, GL–3, GL–4, GL–5, GL–6, GL–7, GL–8, GL–8, GL–9, GL–10, GL–11, GL–12)	
ID of Section	Class section A, B or C student belongs to.	
Student semester	Student semester (1st or 2nd)	
Course	Offered courses such as (IT, Math, English, Arabic, Science, Quran)	
Punctuality of student in the class	No. of student available days in class (Below–07 or Above–07)	
Parent involvement	Survey forms provided by tutors is answered by parents or not	Participation of parents on the whole learning process
Satisfaction of Parent	This feature is concerned with the intensity of satisfaction of the parent (Positive or Negative)	
Group Discussions	These all features are concerned with student behavior while interacting with Kalboard 360 e–learning website.	Behavioral Feature
Resources visited by a student		
Raising hands		
Assignments viewed by a student		

After a dataset is collected the most important task is to pre–process data by applying pre–processing techniques. As real data is not complete (inadequate attributes, missing values of interest, having summarized data). So to eliminate noise and outlier data pre–processing is applied which includes cleaning data, transforming data and selection and analysis of appropriate features.

3.2. PRE–PROCESSING DATA

The techniques are applied to convert unstructured data into some conventional format so that it can be easily accepted and used by data mining algorithm.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

3.2.1. DATA CLEANING

Data cleaning is one of the major tasks in preprocessing. Data cleaning is used to remove noisy and inconsistent data and to deal with incomplete values. In this work, we used a dataset of 500 records out of which 20 records contain some missing values from different categories so after cleaning the final dataset becomes 480 records.

3.2.2. DATA TRANSFORMATION

Data transformation is applied to transform the numerical values into nominal values for classification to represent class labels. In Table 2 we distribute the dataset into below–mentioned class intervals lowest level, medium level and highest level based on student's grade or marks.

Table 2. Classes based on their numerical values.

Classes	
Interval Value	Class Label
0–69	Lowest Level
70–89	Medium Level
90–100	Highest Level

3.2.3. FEATURES SELECTION AND ANALYSIS

A research study (Karegowda, Manjunath & Jayaram, 2010) analyzed feature selection as most important task in data preprocessing. The objective of this step is to choose some important and appropriate subset of features from dataset to transform or reduce the number of attributes that can appear in the algorithm, therefore reducing the proportion of feature area so that the repeatable and inappropriate data is removed. In this way, feature selection helps in enhancing the performance of the learning algorithm by improving the data quality. Feature selection methods are divided into two main categories (1) Wrapper Based methods (2) Filter Based methods. Filter method is applied to identify relevant subset of features while avoids the remaining. These methods rank the features by using variable ranking techniques so that highly ranked features can be selected and applied to the learning algorithm. Acharya and Sinha (2014) investigate

many feature ranking techniques such as information gain and gain ratio that is used for feature evaluation. In our work, we applied selection algorithms based on the gain ratio which is filter based approach to examine different feature scores so that the most important features for building students' performance model can be identified. Figure 1 shows the highly ranked features after filter based evaluation.

As shown in Figure 1 student absence days got the highest rank followed by category related to parent's involvement like their answering survey, satisfaction from school and so on. In Figure 1 we have observed that an important subset of features is selected while others are eliminated. In this way, the features we are considering in this research got the highest rank which means that student punctuality and their parents' participation during whole education practice have a great effect on their academic performance.

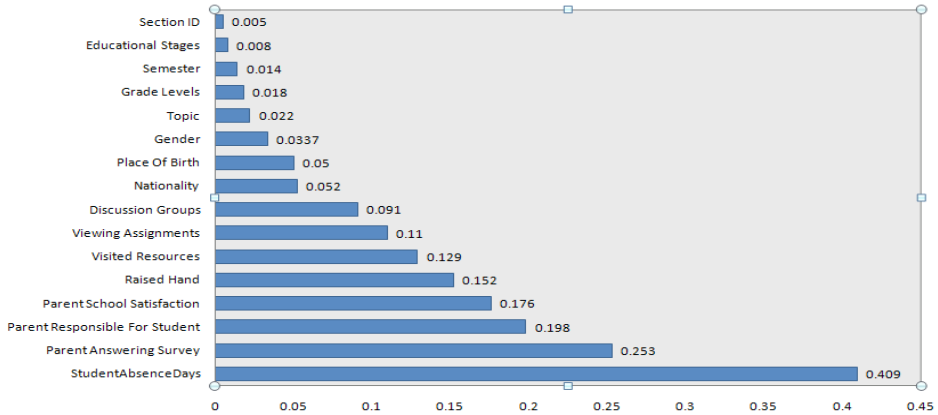

Figure 1. Highly Ranked features after applying filter based evaluation using gain ratio.

4. METHODOLOGY

In this paper, we present students' performance framework using three different classifiers, to assess the subset of features having an effect on students' academic achievement. Figure 2 demonstrates the primary steps in the given framework. This framework begins by gathering information from Kalboard 360 online learning management system referenced in section 3.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

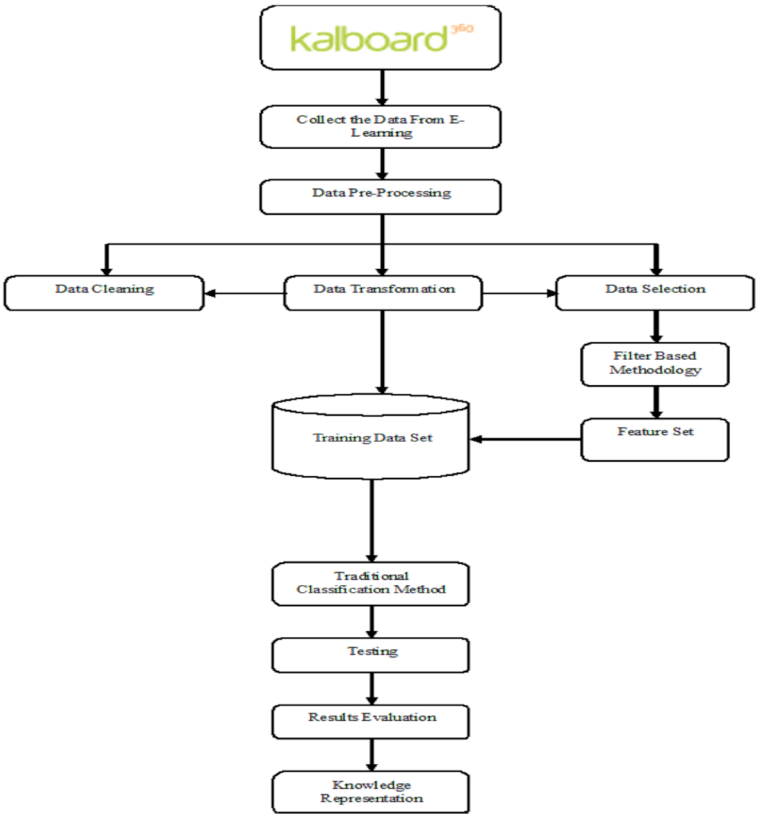

Figure 2. Steps of students' performance prediction model.

This step is trailed by the next step which is pre–processing data related to changing the gathered data into some convenient format. So in this step, first of all, we applied data cleaning technique to remove the irrelevant and redundant data from the dataset. After that, the numerical values are transformed into nominal values for classification to represent class labels. To achieve the task, we distribute the dataset into three class labels (highest level, medium level, and lowest level) based on student's total grade At this step, dataset has a ratio of 199 students at the lowest level, at the middle level there are 248 students and at the highest level there are 33 students. A step onward, feature selection and analysis are used to pick the optimum list of features with highest scores. As appeared in Figure 1, we applied selection algorithms based on the gain ratio which is filter based approach to examine different feature scores. At last, we proposed a

framework for having three classifiers. The classification algorithms are used to get to know about features that may affect students academic achievements. The three different classifiers that are applied to assess the student's performance are Decision Tree (DT), Naïve Bayes (NB) and Artificial Neural Network (ANN).

4.1. NAÏVE BAYES CLASSIFIER

This classifier work on a strategy to evaluate the probabilities of different attributes from training data set for any class after that utilizes these probabilities to characterize new elements. Each level has associated probabilities. With a middle level, it is 0.3, 0.27 with low level and, 0.44 with high level.

4.2. DECISION TREE CLASSIFIER

DT is used to discover rules that characterize the data based on a lot of braches and helps in the decision. For nominal attributes, it gives the best results. Figure 3 demonstrates a J48 pruned tree having 31 number of leaves. Size of the tree is 48.

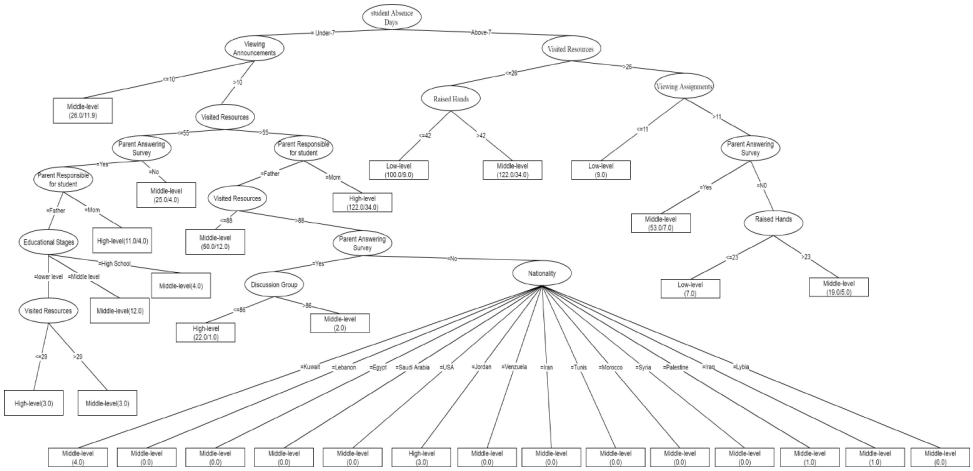

Figure 3. Decision Tree having 31 nodes.

4.3. ARTIFICIAL NEURAL NETWORK CLASSIFIER

In research study (Naser, Zaqout, Ghosh, Atallah & Alajrami, 2015) used ANN which is an approach of neural network prepares data for achieving good accuracy. ANN framework is used to generate patterns and to solve complex

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

prediction problems. It comprises an input layer, the output layer, and a hidden layer. The input is taken by input layer from the user and output to the user is sent by the output layer. Middle layer is between input layer and output layer. The neurons of middle layer are just associated with different neurons and do not straightforwardly interface with the main user application. For knowledge representation patterns and results are assessed.

5. EXPERIMENTS AND EVALUATION OF RESULTS

5.1. SETTING ENVIRONMENT

The experiment is performed on PC having RAM of 8GB, 5 intel core (2.50 GHz). Weka tool in classification algorithms (Arora, 2012) analyzed good accuracy and prediction results. We used Weka tool in our work to evaluate our proposed models, comparisons and results. Training set, cross–validation, supplied test set, and percentage split are few options available for test purpose. The dataset is distributed into a training set and test set using 10 folds cross–validation because this option is widely used one, especially if we have a limited amount of dataset. The dataset is randomly divided into ten subsets. Weka tool uses set 1 for test purpose and remaining 9 sets for training purpose for first training and uses set 2 for testing and rest of 9 sets for training and repeat that in total ten times by interchanging the set each time with next one. In the end, the average success rate is calculated.

5.2. EVALUATION MEASURES

For evaluating the quality of different classification techniques applied on students' academic performance model we use four different measures accuracy, precision, recall, and f–measure. Table 3 demonstrates different calculated measures, it shows confusion matrix comprises of 1,2,3 and 4 equation.

Table 3. Two class confusion matrix

		Predicted	
		Yes	**No**
Actual	**Yes**	TP	FN
	No	FP	TN

Yes is for positive values and No is for negative values whereas TP is for true positive values and FP is for false positive values similarly FN is for false negative and TN is for true negative. Accuracy is calculated as correct classifications divided by a total number of classifications. The Recall is the proportion of rightly classified to total unclassified and rightly classified cases. Precision is the proportion of rightly classified to total misclassified and rightly classified cases. F–measure is also included which is a combination of precision and recall and it is considered the best indicator of the relationship between them.

$$\text{Accuracy} = \frac{\text{TruePositive}+\text{TrueNegative}}{\text{TruePositive}+\text{FalseNegative}+\text{FalsePositive}+\text{TrueNegative}} \quad (1)$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive}+\text{FalseNegative}} \quad (2)$$

$$\text{Precision} = \frac{\text{TruePostive}}{\text{TruePositive}+\text{FalsePositive}} \quad (3)$$

$$\text{F–measure} = 2\,\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

In our case, there are three classes. Table 4 shows the classification confusion matrix based on A, B, and C class.

Table 4. Confusion matrix for more than two classes.

		Predicted		
		A	**B**	**C**
Actual	**A**	TPa	Qab	Qac
	B	Qba	TPb	Qbc
	C	Qca	Qcb	TPc

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

For any class total false negative values is the addition of all values in respective row except true positive and for any class false positive values is the addition of all values in the respective column except true positive values while total true negative values for any class is the addition of all columns and rows except the row and column of that class.

Recall A= $TP_a /(TP_a+Q_{ab}+Q_{ac})$

Recall B= $TP_b/(TP_b+Q_{ba}+Q_{bc})$

Recall C= $TP_c/(TP_c+Q_{ca}+Q_{cb})$

Precision for considered class can be calculated as:

Precision A=$TP_a /(TP_a+Q_{ba}+Q_{ca})$

Precision B=$TP_b /(TP_b+Q_{ab}+Q_{cb})$

Precision C=$TP_c /(TP_c+Q_{ac}+Q_{bc})$

5.3. RESULTS

Different results are examined based on three different classification techniques which are applied to student dataset to predict students' academic performance. Table 5,6,7 shows confusion matrix for three different classifiers i.e DT, NB, and ANN based on which above measures are calculated for A, B, and C class while Accuracy of the overall algorithm is calculated.

Table 5. Confusion matrix for decision tree classifier.

		Predicted		
		A	**B**	**C**
Actual	**A**	143	21	47
	B	27	100	0
	C	29	2	111

$$\text{Accuracy} = \frac{143+100+111}{143+21+47+27+100+0+29+2+111} \text{x}100\%$$

$$= \frac{354}{480} \text{x}100\%$$

$$= 71.1\%$$

Recall for Class A, B and C:

Recall A=143/143+21+47=67.7, Recall B=100/100+27+0=78.7

Recall C=111/111+29+2=78.12

Precision for Class A, B and C:

Precision A=143/143+27+29=71.8, Precision B=100/100+21+2=81.3

Precision C=111/111+47+0=70.3

F–measure for Class A, B and C:

$$\text{F–measure A} = \frac{2 \text{ Precision x Recall}}{\text{Precision+Recall}}$$

$$= \frac{2(71.8 \text{ x } 67.7)}{71.8+67.7}$$

$$= 69.0$$

F–measure B = 80.0

F–measure C = 74.0

Following the above procedure, the results for accuracy, recall, precision, and F–measure is calculated for naïve bayes and artificial neural network by acquainting the data from given respective tables.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x

Table 6. Confusion matrix for naïve bayes classifier.

		Predicted		
		A	B	C
Actual	A	112	40	59
	B	16	111	0
	C	39	2	101

Accuracy = 67.5%

Recall for Class A is 53.1, Recall B is 87.1 and Recall C is 71.1

Precision for Class A is 67.1, Precision B is 72.5and Precision C is 63.1.

F–measure for Class A is 59.3, F–measure B is 79.3 and F–measure C is 66.

Table 7. Confusion matrix for artificial neural network.

		Predicted		
		A	B	C
Actual	A	156	21	34
	B	20	106	1
	C	29	0	113

Accuracy = 78.1%

Recall for Class A is 73.9, Recall B is 83.4 and Recall C is 79.5

Precision for Class A is 76.1, Precision B is 83.5 and Precision C is 76.3

F–measure for Class A is 75.0, F–measure B is 83.5 and F–measure C is 77.9

Table 8 shows results using three data mining algorithms (ANN, NB, DT). Two different classifications results are achieved by each algorithm (1) classification results with highly ranked features (RF) i.e. student absence days and parent's participation (2) classification results without those highly ranked features (WRF). Details of results with a highly ranked feature are given above. The results without those features can be achieved in a similar way. In Table 8, we can see good classification results with highly ranked features as compared with the results without those features this proves there is a great impact of student punctuality in class and their parents' involvement in learning process to students' academic success and achievements.

Table 8. Algorithm results with highly ranked features (RF) and without highly ranked features (WRF).

Evaluation Measures	Decision Tree		Naïve Bayes		Artificial Neural Network	
	RF	WRF	RF	WRF	RF	WRF
Accuracy	71.1	61.4	67.6	58.3	78.1	59.1

Observing the results in Table 8 we notice that ANN outperform other classification algorithms. Artificial Neural Network provides 78.1 accuracies with highly ranked features and 59.1 without ranked features. 78.1 means 375 out of 480 students are correctly classified to correct class label i.e. High Medium and Low and 105 students are incorrectly classified.

6. CONCLUSION

Academic performance of students is a pillar for their successful future and becoming a big area of interest for all academic institutions over the world. Nowadays the use of e–learning management system is increasing rapidly, and many developed countries have shifted their educational system to fully or partially automated systems because this system generates a huge amount of data that contains hidden knowledge and patterns that can be used to generate meaningful knowledge to help students to improve their academic grades and achievements. In this research, we introduce a students' performance model with new categories of features related to student's punctuality in classes and their parents' participation in the learning process. The overall performance of students' academic prediction framework is examined by three different classification algorithms decision tree, naïve bayes, and artificial neural network. The results show that these features have a strong impact on the academic success of a student. The model provides very good accuracy while using these categories of features and is achieved 10 to 15% increased as compared with results when removing such features.

REFERENCES

Abu Tair, M. M. & El–Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*, *2*(2).

Acharya, A. & Sinha, D. (2014). Application of feature selection methods in educational data mining. *International Journal of Computer Applications*, *103*(2). doi: http://dx.doi.org/10.5120/18048–8951

Amrieh, E. A., Hamtini, T. & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X–API for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)* (pp. 1–5). IEEE. doi: http://dx.doi.org/10.1109/AEECT.2015.7360581

Amrieh, E. A., Hamtini, T. & Aljarah, I. (2016). Mining educational data to predict Student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), pp. 119–136. doi: http://dx.doi.org/10.14257/ijdta.2016.9.8.13

Arora, R. (2012). Comparative analysis of classification algorithms on different datasets using WEKA. *International Journal of Computer Applications*, *54*(13).

Arsad, P. M. & Buniyamin, N. (2013). A neural network students' performance prediction model (NNSPPM). In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)* (pp. 1–5). IEEE. doi: http://dx.doi.org/10.1109/ICSIMA.2013.6717966

Hien, N. T. N. & Haddawy, P. (2007). A decision support system for evaluating international student applications. In *2007 37th annual frontiers in education conference–global engineering: knowledge without borders, opportunities without passports* (pp. F2A–1). IEEE. doi: http://dx.doi.org/10.1109/FIE.2007.4417958

Kalboard 360–E–learning system. (2000). Retrieved from http://cloud.kalboard360.com/User/Login#home/index

Karegowda, A. G., Manjunath, A. S. & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), pp. 271–277.

Murugananthan, V. & ShivaKumar, B. L. (2016). An adaptive educational data mining technique for mining educational data models in elearning systems. *Indian Journal of Science and Technology*, *9*(3), pp. 1–5. doi: http://dx.doi.org/10.17485/ijst/2016/v9i3/86392

Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R. & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, *8*(2), pp. 221–228.

Quadri, M. M. & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*. Retrieved from https://computerresearch.org/index.php/computer/article/view/891

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), pp. 135–146. doi: http://dx.doi.org/10.1016/j.eswa.2006.04.005

Romero, C. & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), pp. 601–618. doi: http://dx.doi.org/10.1109/TSMCC.2010.2053532

Romero, C., Ventura, S. & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), pp. 368–384. doi: http://dx.doi.org/10.1016/j.compedu.2007.05.016

Edición Especial Special Issue Mayo 2019
DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.x-x