PREDICTING STUDENT ACADEMIC PERFORMANCE USING DATA GENERATED IN HIGHER EDUCATIONAL INSTITUTES

Areej Fatemah Meghji
Mehran University of Engineering and Technology, Jamshoro (Pakistan)
E–mail: areej.fatemah@faculty.muet.edu.pk

Naeem Ahmed Mahoto
Mehran University of Engineering and Technology, Jamshoro (Pakistan)
E–mail: naeem.mahoto@faculty.muet.edu.pk

Mukhtiar Ali Unar
Mehran University of Engineering and Technology, Jamshoro (Pakistan)
E–mail: mukhtiar.unar@faculty.muet.edu.pk

Muhammad Akram Shaikh
Scientific and Technological Information Center, Islamabad (Pakistan)
E–mail: akramshaikh@hotmail.com

Recepción: 05/03/2019 **Aceptación:** 21/03/2019 **Publicación:** 17/05/2019

Citación sugerida:
Meghji, A. F., Mahoto, N. A., Unar, M. A. y Shaikh, M. A. (2019). Predicting Student Academic Performance using Data Generated in Higher Educational Institutes. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición Especial, Mayo 2019*, pp. 366–383. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366–383

Suggested citation:
Meghji, A. F., Mahoto, N. A., Unar, M. A. & Shaikh, M. A. (2019). Predicting Student Academic Performance using Data Generated in Higher Educational Institutes. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019*, pp. 366–383. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366–383

ABSTRACT

The analysis of data generated by higher educational institutes has the potential of revealing interesting facets of student learning behavior. Classification is a popularly explored area in Educational Data Mining for predicting student performance. Using student behavioral data, this study compares the performance of a broad range of classification techniques to find a qualitative model for the prediction of student performance. Rebalancing of data has also been explored to verify if it leads to the creation of better classification models. The experimental results, validated using well–established evaluation matrices, presented potentially significant outcomes which may be used for reshaping the learning paradigm.

KEYWORDS

Educational data mining, Student performance prediction, Education, Machine learning.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

1. INTRODUCTION

Educational institutes generate massive amounts of student data which can broadly be categorized as descriptive, behavioral, attitudinal and interactional (Meghji, Mahoto, Unar & Shaikh, 2018). In recent years, there has been growing interest in analyzing this data to better understand student learning behavior (Casey & Azcona, 2017). The prediction and understanding of student performance are essential for the establishment of a student centric learning environment; if educators can predict student performance, they can have mechanisms in place to ensure this performance constantly improves or, at any rate, does not fall beneath an acceptable threshold. Educational Data Mining (EDM) is a field dedicated towards the application of data mining and machine learning strategies on data emerging from educational institutes (Baker & de Carvalho, 2008). The goal of EDM is to explore educational data to gain insights into how individuals learn (Meghji & Mahoto, 2016). Classification is a popularly explored area in EDM for predicting student performance. This method is used to assign items to a class, from a set of pre–specified classes, based on certain known properties of the items (Hämäläinen & Vinni, 2011). The process of classification can be implemented using various algorithms, usually referred to as classifiers (Tan, Steinbach, & Kumar, 2006).

Using behavioral data belonging to students of the Department of Software Engineering (SWE), Mehran University of Engineering and Technology (MUET), Pakistan, this research aims at determining the impact of behavioral attributes on overall student academic performance. Using classifiers belonging to four classification families i.e., rule–based, decision tree, probabilistic and instance–based, this study attempts to find a qualitative model that best predicts student performance based on in–class face–to–face behavioral data. Although the prediction of student performance has been previously addressed, to the best of our knowledge, the specific behavioral attributes considered in this research have not been used in existing scientific literature for student performance prediction. The findings of this study provide opportunities for improved pedagogical decision–making and can be used to enrich the existing teaching practices. This paper has

been organized as follows: Section 2 describes the classification process and the families of classification methods used in this paper, Section 3 presents a literature review of previous studies, Section 4 outlines the experimental setup of this paper, Section 5 presents the results and discussion followed by a conclusion in Section 6.

2. CLASSIFICATION

The process of classification can fundamentally be broken down into two steps. Using certain training data, a classifier first produces a classification model; the classification model is then used to predict the target class for new items – items that were not used during the training process to prepare the classification model (Aggarwal, 2014). The goal of the classification process is identification of a classification model that fits the relation between the known properties and class label of the input data in the best manner. Apart from fitting the data well, the model generated by the classifier should accurately predict the class of new/unseen data items.

Classification methods differ from each other based on their internal mechanism of processing and extracting relevant features from training data for the creation of a classification model. Some popular categories of classification methods include:

Decision Tree: A decision tree represents a tree–like hierarchical structure comprising of a set of conditions. This predictive model consists of nodes and leaves. Each node in the tree represents a logical test; based on the outcome of the test, the node branches to one child or another. New instances of data are classified into classes based on the path of satisfied conditions until a leaf node is reached; the leaf node represents a class label (Witten, Frank, Hall & Pal, 2016). J48 and REPTree are decision tree based algorithms.

Rule–Based: Rule induction comprises of the generation of a set of If–Then relational rules based on a set of training observations (Hand, Mannila & Smyth, 2001). Some algorithms in this category include OneR and PART.

Probabilistic: Rather than predicting the output class of an instance of data, the classifiers in this category predict the probability distribution over the label classes

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

based on the observation of an instance of data. The Bayes theorem is utilized for calculating the probability of an item belonging to a class (Han, Pei & Kamber, 2011). Naïve Bayes and Bayes Net belong to this class of classifiers.

Instance–Based: Unlike the classifiers that create a model/generalized explicit description based on which future data items are to be classified, the classifiers in this category postpone this step until data items need to be classified. The new data item is examined at run–time to find its relationship with the previously stored training data. It is due to this reason that these classifiers are also called memory–based or lazy (Aha, Kibler & Albert, 1991). The IBK belongs to this category of classifiers.

3. LITERATURE REVIEW

Educators have utilized classifiers for predicting different facets of student learning. Working on student demographic data and grades obtained in an introductory level test, decision tree and probabilistic classification algorithms have been used by Sivasakthi (2017) to predict the initial programming performance of students in the first year of bachelor's in computer applications.

Using data of 72 freshman students on parameters based on student background and characteristics exhibited by students during their class, the probabilistic classifier Naïve Bayes has been used by Purwaningsih and Arief (2018) to predict student performance in the subject of English.

Experimenting on data of 231 students, Shah (2012) used several algorithms, including J48, RandomForest, REPTree, Bayes and NaiveBayes to predict student academic performance. This study demonstrated that re–sampling of data has a significant effect on the improvement of prediction accuracy.

Experimenting on attributes relating to student demographics and performance, Alharbi, Cornford, Dolder and De La Iglesia (2016) used a decision tree classifier to predict students in danger of not achieving their honors degree.

Chau and Phung (2013) used sampling with C4.5, Naïve Bayes and Random Forests to devise early predictions of student final–status based performance. Their study suggests that sampling imbalanced data directly influences the improvement of algorithm accuracy.

4. EXPERIMENTAL SETUP

4.1. DATA COLLECTION

This research uses behavioral data of 2nd year students studying B.E in the department of SWE, MUET, Pakistan. The data for this study has been collected through qualitative class observations which were carried out over a period of one semester (i.e., six months). The dataset comprises of 176 student records.

4.2. DATA PREPARATION

Data preparation is an imperative step of the EDM process. The collected data has been processed to remove any erroneous or missing data and transformed into a format that can facilitate maximum extraction of knowledge. Table 1 presents attributes considered in this study. Possible values for attributes 1–8 are excellent, good, average, below average and poor. Possible values for attributes 9–11 are always, mostly, average, rarely and never. Attribute 12 can have the values of front, mid or back. Finally, attribute 13 is the label class with possible values of pass or fail.

Table 1. Data Attributes and their Possible Values

S#	Attribute	Description
1.	class_performance	Overall performance of the student
2.	attention	Student attention towards lectures
3.	interaction_class	Student tendency to clear confusions in–class
4.	interaction_afterclass	Student tendency to clear confusions after–class
5.	note_taking	Student tendency to take and maintain notes
6.	assignment_submission	Assignment submission record of the student
7.	attendance	Attendance record of the student
8.	test_marks	Marks obtained by student in class tests
9.	excuses_leave	Does student make excuses to skip lectures?
10.	assignment_self	Are assignments actually made by the student?
11.	project	How often does the student participate in class projects?

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

S#	Attribute	Description
12.	seating_position	Where does the student sit during lectures?
13.	verdict	Student exam outcome

Next, the considered behavioral attributes were visualized to better understand the distribution of various attribute values (see Figure 1). By breaking down the data in terms of number of students exhibiting various behavioral attributes and the pass and fail ratio within each attribute value, it was observed that the label class (verdict) is not equally represented – most of the data belongs to class 'Pass'. Algorithms are data driven; most state of the art classification approaches are developed with the assumption that the underlying data is evenly distributed (Wang, Xu, Wang & Zhang, 2006). The performance of an algorithm can, thus, greatly vary if it is trained for classification using disproportioned data.

To ensure that algorithms function optimally, the Synthetic Minority Over–sampling TEchnique (SMOTE) has been applied on the dataset to oversample the minority (Fail) class. This technique works by creating new synthetic samples of the minority class – in this case, the 'Fail' class. The newly generated samples are introduced along the line joining all or any of the specified (K) minority class nearest neighbors. The algorithm randomly selects the K nearest neighbors (Chawla, Bowyer, Hall & Kegelmeyer, 2002). As the new synthetic examples are added to the bottom of the dataset, the dataset was shuffled using the randomize filter of WEKA tool. The original student dataset comprised of 176 records (dataset–1) whereas the SMOTE dataset comprises of 221 records (dataset–2).

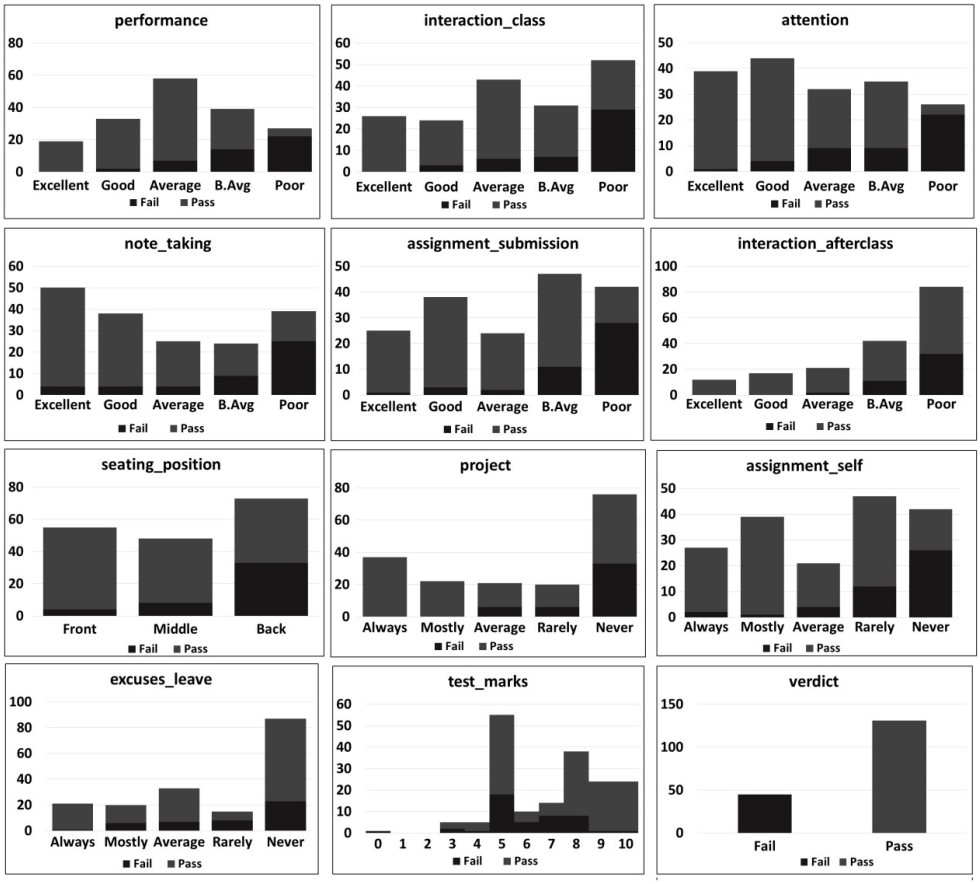

Figure 1. Visualization of the Collected Student Data.

Dataset–2 has also been visually represented to better understand the breakdown in terms of number of students exhibiting various behavioral attributes (see Figure 2).

4.3. DATA ANALYSIS APPROACH

This paper uses prominent algorithms from four classification families for predicting students into two classes – pass and fail. Specifically, the J48, REPTree and Random Forests from decision tree; OneR and PART from rule–based; Naïve Bayes and Bayes Net from probabilistic; and the IBK classifier from the instance–based family of classifiers have been used. The reason for using a diverse array of classifiers is twofold. First, since classification algorithms are data driven with their

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

performance being influenced by the dataset being used, an algorithm that works well with one form of data might not present equally striking results when the underlying data is changed. Second, using a wide–range of algorithms increases the likelihood of finding the better and most efficient classification model in terms of accuracy and allows a better comparison of overall performance.

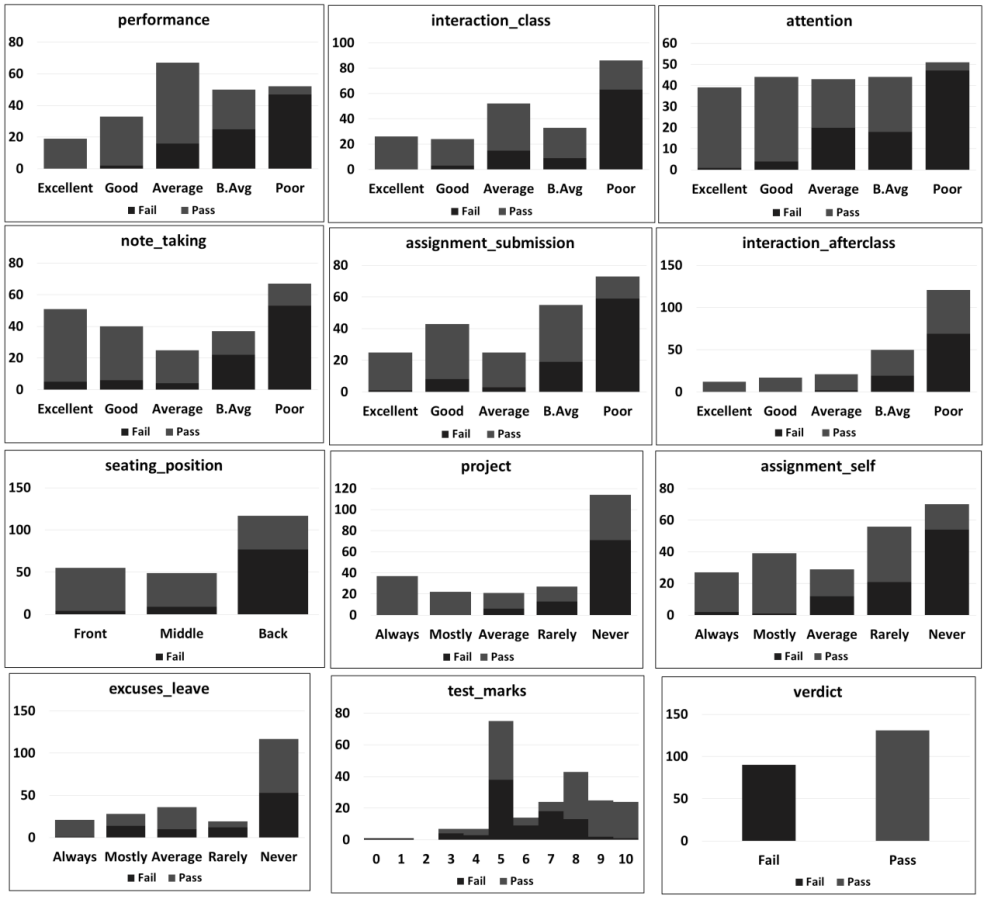

Figure 2. Visualization of Student Data after SMOTE (dataset–2).

4.4. PERFORMANCE EVALUATION

It is essential to evaluate the performance and usefulness of the classifiers before their results can be used to practically predict and/or improve students' performance. The classifiers used in this paper have been evaluated using two evaluation metrics – Accuracy which the measure is, in percentage, of the number of correct predictions made by the

classifier and Kappa Statistic which takes randomness or chance of predictions into considerations and essentially is a measure of how better a classifier is performing when compared to a classifier that simply guesses the target class label (Nasa & Suman, 2012). Kappa is especially useful when working with imbalanced datasets. The value of kappa ranges between 0 and 1, a higher value signifying better performance.

5. RESULTS AND DISCUSSION

The collected data (dataset–1 and dataset–2) has been mined using WEKA open source software (Witten, *et al.*, 2016). WEKA provides a large collection of machine learning algorithms and thus is widely used in EDM research.

Considering that a classifier has to classify items into one of two classes: i) Positive (P) and ii) Negative (N), there are four possible outcomes that the classifier can predict: True Positives (TP) – items that have been correctly classified in to class P, True Negatives (TN) – items that have correctly been classified into class N, False Positives (FP) – items that should have been classified into class N but have been incorrectly classified into class P and False Negatives (FN) – items that should have been classified into class P but have been incorrectly classified into class N. The 10–fold cross validation approach has been used to validate the outcome of the considered classifiers. Table 2 presents a performance measure of classifiers in terms of how successfully they could classify data items.

Table 2. Performance of Classifiers.

Classifier	Classifiers applied on Dataset–1			Classifiers applied on Dataset–2		
	Correctly Classified Instances (TP)	Incorrectly Classified Instances (FP)	Time to Build Model (secs)	Correctly Classified Instances (TP)	Incorrectly Classified Instances (FP)	Time to Build Model (secs)
OneR	146	30	0	169	52	0
PART	143	33	0.09	182	39	0
J48	148	28	0	186	35	0
RF	144	32	0.05	195	26	0.06
REPTree	147	29	0	188	33	0.01
BNet	143	33	0	186	35	0
NB	139	37	0	182	39	0

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

Classifier	Classifiers applied on Dataset–1			Classifiers applied on Dataset–2		
	Correctly Classified Instances (TP)	Incorrectly Classified Instances (FP)	Time to Build Model (secs)	Correctly Classified Instances (TP)	Incorrectly Classified Instances (FP)	Time to Build Model (secs)
IBK	139	37	0	188	33	0

Experimenting on the original dataset, it has been observed that the J48 classifier attained the highest number of correctly classified items, closely followed by the OneR, REPTree and Random Forest classifiers respectively. Experimenting on the balanced dataset, the Random Forest outperformed the remaining classifiers by classifying 195 items correctly and misclassifying only 26 items. The REPTree and IBK also exhibited better results.

The classifiers used in this paper have been evaluated using the evaluation metrics of Accuracy and Kappa Statistic. For results of performance evaluation of the considered classifiers, see Table 3.

Table 3. Comparison based on Evaluation Measures.

Category	Classifier	TP Rate	FP Rate	Precision	Recall	Kappa	Accuracy (%)
Results of Classifiers applied on Dataset–1							
Rule	OneR	0.83	0.423	0.826	0.83	0.4757	82.95%
	PART	0.813	0.342	0.807	0.813	0.4887	81.25%
Decision Tree	J48	0.841	0.39	0.838	0.841	0.5188	84.09%
	RF	0.818	0.34	0.812	0.818	0.5004	81.82%
	REPTree	0.835	0.348	0.827	0.835	0.5293	83.52%
Probabilistic	BNet	0.813	0.196	0.84	0.813	0.5560	81.25%
	NB	0.79	0.204	0.828	0.79	0.5149	78.98%
Instance	IBK	0.79	0.306	0.798	0.79	0.4713	78.98%
Results of Classifiers applied on Dataset–2							
Category	Classifier	TP Rate	FP Rate	Precision	Recall	Kappa	Accuracy (%)
Rule	OneR	0.765	0.252	0.765	0.765	0.51	76.47%
	PART	0.824	0.191	0.823	0.824	0.63	82.35%
Decision Tree	J48	0.842	0.154	0.846	0.842	0.67	84.16%
	RF	0.882	0.130	0.882	0.882	0.75	88.23%
	REPTree	0.851	0.172	0.850	0.851	0.68	85.06%

Category	Classifier	TP Rate	FP Rate	Precision	Recall	Kappa	Accuracy (%)
Probabilistic	BNet	0.842	0.140	0.854	0.842	0.68	84.16%
	NB	0.824	0.153	0.841	0.824	0.64	82.35%
Instance	IBK	0.851	0.137	0.859	0.851	0.69	85.06%

Considering the original dataset, the J48 classifier has the best accuracy value closely followed by the REPTree classifier. The Random Forest classifier exhibited an accuracy of 88.23% after the application of SMOTE followed by the REPTree and IBK classifiers, both exhibiting an accuracy of 85.06%.

It can thus be evidently stated that the accuracy of classifiers greatly improves when they are trained on balanced data. Examining the results of dataset–1, although Bayes based classifiers did not perform that well in terms of accuracy, the BayesNet classifier has better kappa static score followed by OneR and RepTree classifiers respectively. Similar experiments applied on dataset–2 resulted in a significant improvement in kappa score, with the Random Forest achieving a kappa score of 0.75 making it highly significant.

5. CONCLUSION

Academicians are always interested in discovering means through which students may learn in better ways. The abundance of student data and innovative technological breakthroughs has allowed discovering useful patterns. Behavioral features such as note taking, attention, assignment submission, and seating position, extracted from real student data, have been used in this paper for predicting students' performance. Several classifiers exhibited good performance in terms of accuracy and kappa scores. Balancing data with SMOTE greatly improved the performance of the classifiers evident through improved accuracy and kappa scores. The model generated by the Random Forest classifier exhibited significantly better results with an accuracy of 88.23% and a kappa score of 0.75. This research demonstrated that behavioral tendencies depicted by the students in class could be used to predict their semester outcomes allowing the creation of early warning systems. Interventions can be planned to ensure proper

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

guidance is provided to students at the risk of failure. Future line of classifications can explore behavioral factors such as talkative tendency, social interaction, punctuality, participation in extracurricular activities, etc., and combine these with descriptive, behavioral, attitudinal and interactional features.

ACKNOWLEDGEMENTS

This research has been performed under the Institute of ICT Mehran University of Engineering and Technology, Pakistan and funded by the ICT Endowment for Sustainable Development.

REFERENCES

Aggarwal, C. C. (Ed.). (2014). *Data classification: algorithms and applications*. CRC press.

Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance–based learning algorithms. *Machine learning*, *6*(1), pp. 37–66. doi: http://dx.doi.org/10.1007/BF00153759

Alharbi, Z., Cornford, J., Dolder, L. & De La Iglesia, B. (2016). Using data mining techniques to predict students at risk of poor performance. In *2016 SAI Computing Conference (SAI)* (pp. 523–531). IEEE.

Baker, R. & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. In *Educational Data Mining 2008*.

Casey, K. & Azcona, D. (2017). Utilizing student activity patterns to predict performance. *International Journal of Educational Technology in Higher Education*, *14*(1), p. 4. doi: http://dx.doi.org/10.1186/s41239–017–0044–3

Chau, V. T. N. & Phung, N. H. (2013). Imbalanced educational data classification: An effective approach with resampling and random forest. In *The 2013 RIVF International Conference on Computing & Communication Technologies–Research, Innovation, and Vision for Future (RIVF)* (pp. 135–140). IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Mver–sampling Technique. *Journal of artificial intelligence research*, *16*, pp. 321–357. doi: http://dx.doi.org/10.1613/jair.953

Hämäläinen, W. & Vinni, M. (2011). Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, pp. 57–71.

Han, J., Pei, J. & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hand, D., Mannila, H. & Smyth, P. (2001). Principles of Data Mining. The MIT Press. In *A comprehensive, highlytechnical look at the math and science behind extracting useful information from large databases* (Vol. 546).

Meghji, A. F. & Mahoto, N. A. (2016). Using big data to improve the educational infrastructure and learning paradigm. In *Effective Big Data Management and Opportunities for Implementation* (pp. 158–181). IGI Global.

Meghji, A. F., Mahoto, N. A., Unar, M. A. & Shaikh, M. A. (2018). Analysis of Student Performance using EDM Methods. In *2018 5th International Multi–Topic ICT Conference (IMTIC)* (pp. 1–7). IEEE.

Nasa, C. & Suman, S. (2012). Evaluation of different classification techniques for web data. *International journal of computer applications*, *52*(9), pp. 34–40. doi: http://dx.doi.org/10.5120/8233–1389

Purwaningsih, N. & Arief, D. R. (2018). Predicting students' performance in English class. In *AIP Conference Proceedings* (Vol. 1977, No. 1, p. 020020). AIP Publishing. doi: http://dx.doi.org/10.1063/1.5042876

Shah, N. S. (2012). Predicting factors that affect students' academic performance by using data mining techniques. *Pakistan business review*, *13*(4), pp. 631–638.

Sivasakthi, M. (2017). Classification and prediction based data mining algorithms to predict students' introductory programming performance. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 346–350). IEEE.

Tan, P. N., Steinbach, M. & Kumar, V. (2006). Classification: basic concepts, decision trees, and model evaluation. *Introduction to data mining*, *1*, pp. 145–205.

Wang, J., Xu, M., Wang, H. & Zhang, J. (2006). Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *2006 8th international Conference on Signal Processing* (Vol. 3). IEEE.

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

AUTHORS

Areej Fatemah Meghji

Ms. Areej Fatemah is an Assistant Professor in the Department of Software Engineering at MUET, Pakistan. She received her M.E working on Social Networking Analysis from MUET Pakistan in 2011 and is presently pursuing her PhD working on Predictive Analysis of Data emerging from the sector of Higher Education. Her research interests include Knowledge Management, Educational Data Mining, Artificial Intelligence and Data Analytics.

Naeem Ahmed Mahoto

Dr. Naeem Ahmed Mahoto is an Associate Professor and Chairman of the Department of Software Engineering, MUET Pakistan. He received his Master degree in Computer Engineering from MUET, Pakistan and Ph.D in Control and Computer Engineering from Politecnico di Torino, Italy, in 2013. His research interests are focused in the field of data mining and bioinformatics. His research activities are also devoted to summarization of web documents, sentiment analysis, data visualization and data mining.

Mukhtiar Ali Unar

Prof. Dr. Mukhtiar Ali Unar is the Dean Faculty of Electrical, Electronics and Computer Systems Engineering and a meritorious Professor at the Department of Computer Systems Engineering, MUET, Pakistan. He did his B.E in Electronic Engineering from MUET in 1986, M.Sc in Electrical and Electronic Engineering in 1995 and Ph.D in Artificial Intelligence from University of Glasgow, UK in 1999. He also remained the pro vice chancellor of MUET, S.Z.A.Bhutto campus, Khairpur Mir's and Director Institute of Information & Communication Technologies MUET, Pakistan. He has 30 years of teaching, research & management/admin experience. He is the author of more than 60 journal/conference papers of national/international repute.

His research interests include Artifical Intelligence, Control System Design, Digital Signal Processing and Knowledge Discovery. Dr. Unar is a member of IEEE (USA), an affiliate of International Federation of Automatic Control, a member of Pakistan Institute of Engineers and a member of Pakistan Engineering Council.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.366-383

Muhammad Akram Shaikh

Prof. Dr. Muhammad Akram Shaikh is working as Director General in Pakistan Scientific & Technology Information Centre (PASTIC), a subsidiary of Pakistan Science Foundation under Ministry of Science & Technology. He remained Professor in the Department of Software Engineering, & Co–Director Institute of Information & Communication Technologies Mehran University of Engineering & Technology Pakistan. He has received his B.E. from Mehran University Pakistan in 1993, MBA from University of Sindh in 1996, MSc. from University of Huddersfield (UK) in 2001, and Ph.D. from Tsinghua University (China) in 2008. He has 25 years of teaching, research & management/admin experience. He is author of more than 30 journal/conference papers of national/international repute. In addition, he is also attached as editor/ co–editor/ reviewer of national/international journals, Session chair/ PC member of national/international conferences, member accreditation committee PEC, member national curriculum review committees of HEC, member technical review committee of PSF, member executive committee of PSF, member Board of Trustee (BoT) PSF, member HEC digital library advisory board, and member board of studies of various universities of Pakistan.

His research areas of interest include Knowledge Engineering, Scientific & Technological Databases, Information Processing, Data Mining & Data Warehousing, Software Engineering, Automation & Control, Networks, Virtual Reality and Graphics.