International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Real-Time Prediction of Student's Locality towards Information Communication and Mobile Technology: Preliminary Results

Chaman Verma, Zoltán Illés, Veronika Stoffová

Abstract*: The present paper used supervised machine learning to predict the student's locality towards Information and Communication Technology (ICT) and Mobile Technology (MT) of Indian and Hungarian University. For this, a primary dataset is used with 331 instances and 38 features which are related to the 4 major ICT parameters belongs to the attitude, development and availability, educational benefits and usability of modern ICT resources and mobile technology used in education. To predict the locality, three machine learning classifiers multi-layer perceptron (ANN), K-nearest neighbor (KNN or Idk) and Random Forest (RF) are used with hold out method, Leave One Out and K-Fold cross-validation methods. Further, to enhance the prediction accuracy, RF used CorrelationAttributeEval, ANN used InfoGainAttributeEval and KNN used OneRAttributeEval Feature Selection techniques. The outcome of the study reveals that the Feature Selection algorithm significantly improved the prediction accuracy of each classifier. To compare the accuracies of the extracted dataset, T-test at 0.05 significant level was also used. T-test did not find a significant difference between RF and KNN towards CPU training time and another hand, a significant difference is found between ANN and KNN; ANN and RF classifier. It is also proved that KNN classifier has outperformed others in stabilized accuracy and induced optimum time in locality prediction of students.

Index Terms*: Classifier, *Feature Selection, Locality Prediction, Leave One Out, K-Fold, Hold Out, Real-Time.

I. INTRODUCTION AND RELATED WORK

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The use of data mining in the educational domain is called Educational Data Mining is in trends to explore new ideas in education technology and methods of student learning. The study on Information and Communication Technology (ICT) with respect to various parameters were conducted by various scholars with the help of statistical analysis [1], [2]. In addition, Machine Learning (ML) is also so popular in predicting various attributes in educational datasets. The performance of students was predicted using with the help of educational data mining [3] and ML algorithms were also used in the prediction of student placement [4]. Further, student and teacher demographic features such as gender, state of residence were also predicted against ICT responses towards various parameters with the help of supervised machine learning classifiers [5], [6], [7], [8]. The prediction task should not be confined to offline datasets on parameters. The authors have a major focus on real-time prediction of demographic features of the student towards ICT. in the education domain. According to [9], real-time tasks are produced due to the occurrence of either internal or external events. In real-time systems, the absolute deadline for task begins with time zero and the relative deadline is with respect to the task released time. The concept of real-time prediction was suggested to predict student's age towards various ICT parameters with the help of ML [10]. The real-time predictive models of European school students' nationality in monitoring online ICT access and ICT based activities was also suggested [11]. Experiments with ML in the prediction of student's attitude towards ICT and Mobile Technology (MT) in real-time was also conducted [12]. Supervised ML is also applied to present predictive models for the real-time prediction of the development and availability of ICT and MT in University education [13]. The concept of real-time prediction and automatic process of the data sets was presented with the help of ML and web server [14]. In this paper, the authors have conducted four experiments using supervised ML classifiers to predict the student's locality towards the ICT and MT the University in real time. To achieve this objective, the authors compared predictive models using T-test and CPU time is also calculated. Before training dataset, the authors also applied normalization and class balancing. In addition, the dataset is also trained with Leave One Out, K-Fold and Hold out methods. The authors also reduced a few insignificant features using dimension reduction methods. By deploying the presented real-time locality predictive model on the real-time website, ICT coordinator can know about the attitude of Rural (R) and Urban (U) students towards ICT and MT at various parameters. The present predictive model may also help to diagnose the requirements, knowledge, and awareness about the technology of rural and urban students. Also, a locality wise detection may also a piece of significant advice about the availability and development of the latest ICT and MT resources at University. This paper is categorized into eight major sections. Section I is about introductory theory. Section II discusses the research design and methodology.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Chaman Verma*, Department of Media and Educational Informatics, Eötvös Loránd University, Budapest, Hungary.
Zoltán Illes, Department of Media and Educational Informatics, Eötvös Loránd University, Budapest, Hungary.
Veronika Stoffová, Department of Mathematics and Computer Science, Trnava University, Trnava, Slovakia.

The experimental work retains between Section III and Section VI. Section VII evaluates the results of experiments with significant discussions. Section VIII concludes the essence of the present study with a significant recommendation for future work.

II. RESEARCH METHODOLOGY

A. Dataset Preprocess

A Stratified random sampling was used to collect primary samples using google form and offline using personal visits. A structured questionnaire was designed with four major sections, one is demographic and four belongs to ICT parameters. The scale of sampling was hybrid in nature such as nominal, binary, ordinal etc. Due to this, Normalization was applied on a dataset from 0 to 1 scale using ***Normalize filter***. The dataset has 46 feature and 331 instances. The unsupervised filter named ***ReplaceMissingValues*** has been used to handle missing values. It replaced 3 missing values in a dataset with the modes and means from the training data. The locality feature has been considered as the class variable and the rest of all considered as predictors. The locality class has two sub-classes named Rural and Urban. Initially, the class rural has 84 and urban has 247 instances. To make class balancing, supervised filter Synthetic Minority Oversampling Technique (SMOTE) has been applied on a dataset which significantly balanced both classes for prediction. Smote has doubled the instances of minority class rural. Now, the final count of instances belong to the rural class is 168 and urban has 247instances. Total of 415 instances was trained and tested for binary classification

B. Feature Selection

In present study, 9 features related to the demographic characteristics such as age, sex, locality, country etc. were removed using self-reduction. The features perform well under dimensionality reduction and for different validation approaches and several classification algorithms [15]. For this, ***AttributeSelectedClassifier*** algorithm has been applied on the datasets. In this, ***CorrelationAttributeEval*** algorithm suggested 25 significant features for RF classifier (RF-25). Basically, it evaluates the worth of a feature by measuring the Pearson's correlation between it and the class. To enhance the accuracy of ANN, ***Info-GainAttributeEval*** algorithm provided 35 significant features (ANN-35) which evaluate the worth of a feature by measuring the information gain with respect to the class. For the KNN, the ***OneRAttributeEval*** algorithm provided 33 significant features (KNN-33). All these three Feature Selection methods used Ranker search technique which ranks features by their individual evaluations. Finally, in addition to Initial dataset D0, we have extracted dataset D1 to train and tested for locality prediction towards ICT and MT.

C. Train, Test, and Validation with Classifiers

In Hold-Out method, splitting task performed randomly in two distinct subsets whereas the first subset from where classifier tries to extract knowledge and the second set is used to the tested extracted information. Firstly, datasets D0 and D1 were tested using hold out method. The considered training ratio for the train: test splitting is 50:50 and 60:40. Secondly, K-Fold cross-validation has been applied with varying k as 2, 4, 6 and 8. In this approach, we usually divide the dataset in k number of subset among which one is used as (k-1) test set and rest of sets shall be (k) train sets. Thirdly, leave one out method has been also applied with k=415. Further, the prediction accuracies of each classifier were compared using T-test at 0.05 significant level on D1dataset. To predict the locality of students against their responses, D0 and D1 were tested with 3 supervised machine learning classifiers such as ANN, RF, and KNN with 3 different type of testing modes such as Hold Out, K-Fold and Leave one out.

D. Performance Measures

(a) Confusion Matrix: A joint matrix which reflects actual rural-urban with the rows and predicted rural-urban defined by columns. (b) Accuracy: The percentage of classified locality counts of a student from overall prediction counts. (c) Error: The percentage of unclassified locality counts of a student from overall prediction counts. (d) Receiver operating characteristics curve (ROC): Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a decision threshold. (e) Kappa Static: The Cohen's Kappa is statistical which determines the agreement among instances in the dataset. (f) F-score: A harmonic mean of precision and recall which states the significance of the predictive model. It can be calculated using the formula F = 2 * (Recall * Precision) / (Recall +Precision).

E. Real Time

To propose the real-time prediction of student's locality towards ICT and MT, authors have compared the ***USER CPU TRAINING TIME AND USER CPU TESTING TIME*** in seconds on D1 dataset to test unseen instances for prediction. We need to calculate the process (CPU) time to training the model. Weka's experimental application is much significant to measure and compare both important times for the predictive model.

III. EXPERIMENT -I

In this experiment, the authors have used to hold out a method to test and train unextracted dataset D0 and, for validation testing, K-Fold is used with updating k values. Fig. 1 shows that RF attained the highest prediction accuracy of 80.1% at 60:40 and 79.2% at 50:50 training ratio. Hence, it outperformed other classifiers in the locality prediction of students towards ICT and MT.

Data from Fig. 2, it is also clear that the highest prediction accuracy of 79% at 10-fold are achieved by KNN. The second highest accuracy of 78.8% is provided by RF at 6-fold and 8-fold. Also, the number of folds is directly proportional to the accuracies of RF and KNN. Therefore, KNN classifier outperformed others in the locality prediction of students towards ICT and MT.

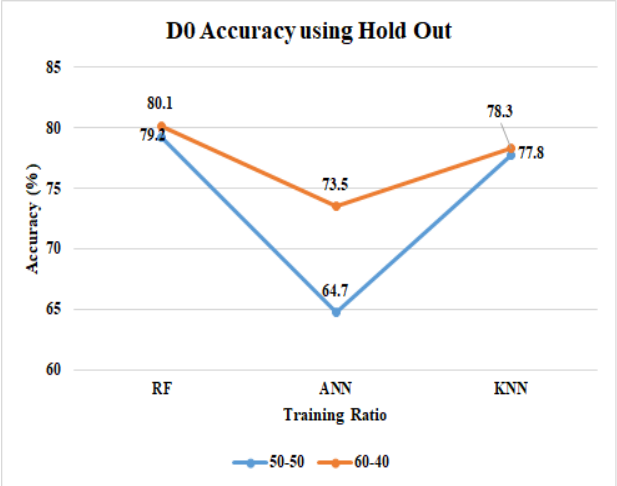

Fig. 1 Dataset D0 testing using Hold out.

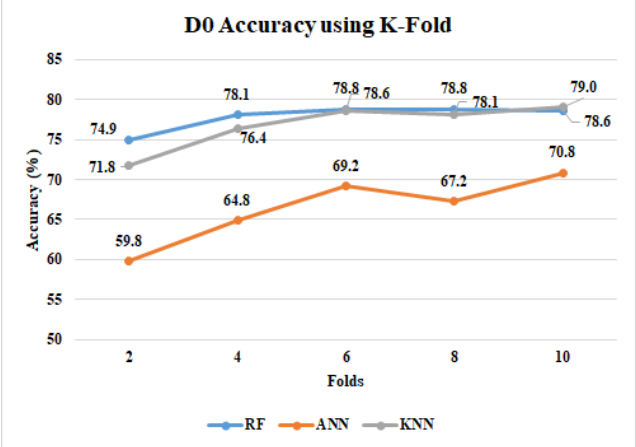

Fig. 2 Dataset D0 testing using K-Fold.

IV. EXPERIMENT –II

In this experiment, the extracted dataset D1 are trained, tested and validated for unseen values. The authors have used K-Fold method with varying the k values and hold out method as well. The Feature Selection mechanism with hold out method significantly improved the prediction accuracies of each classifier as compare to Fig. 1.

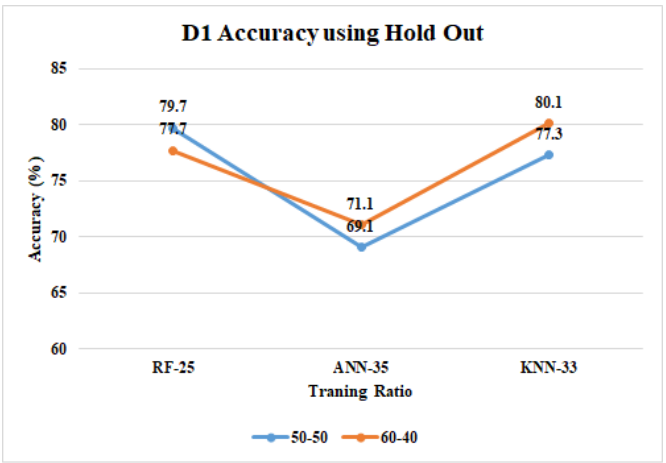

Fig. 3 Dataset D1 testing using Hold out.

Data from Fig. 3, the KNN-33 achieved a maximum prediction accuracy of 80.1% with increase 2% accuracy at 60:40 training ratio and second highest accuracy of 79.7% with 0.5% enhancement at 50:50 training ratio is achieved by RF classifier. ANN-35 has also increased accuracy by 4.4% at 50:50 training ratio.

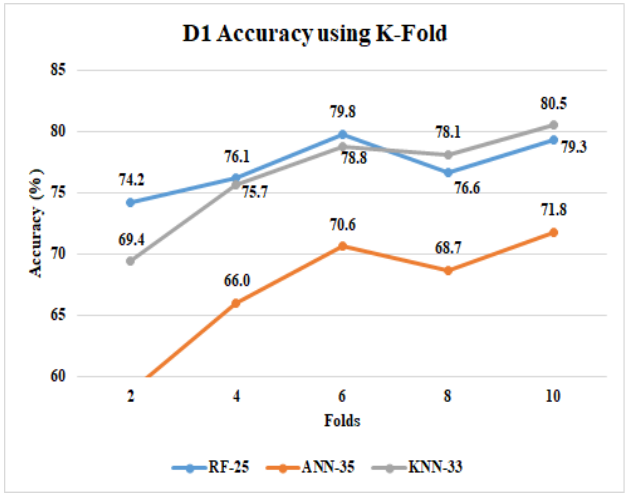

Fig. 4 Dataset D1 testing using K-Fold.

Data from Fig. 4, the highest accuracy of KNN-33 is 80.5% at 10-fold. The RF-25 has achieved 79.3% and ANN-35 has attained the accuracy of 71.8% at 10-fold. In this, the Feature Selection methods with K-fold significantly enhanced the prediction accuracies of each classifier as compare to Fig. 2. The accuracy of RF-25 has enhanced the accuracy by 0.7%, KNN-33 has improved the accuracy by 1.5% and ANN-35 has increased the accuracy by 1%.

V. EXPERIMENT –III

In this experiment, the extracted dataset D0 and D1 are trained, tested and validated using Leave One Out method with the k=415. The value of k is equal to the total number of training instances. Further, accuracies are compared with previous experiments.

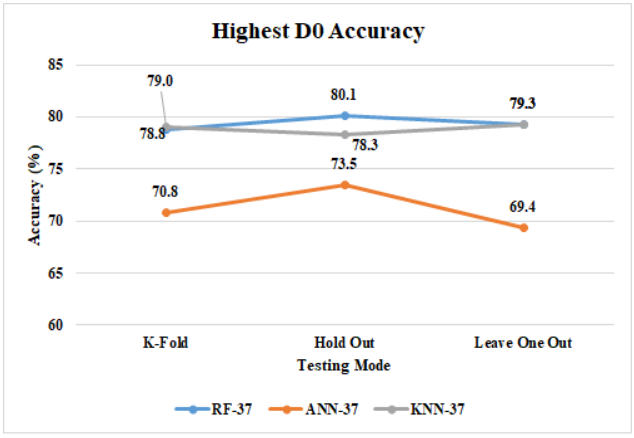

Fig. 5 Dataset D0 testing using Leave One Out.

Data from Fig. 5, in D0 dataset, the Leave one out method provided identical and maximum accuracies 79.3% for both KNN-37 and RF-37.

One hand, for both RF-37 and ANN-37, hold out method is appropriate in prediction accuracy and on another hand, Leave One Out method is suitable for KNN-37.

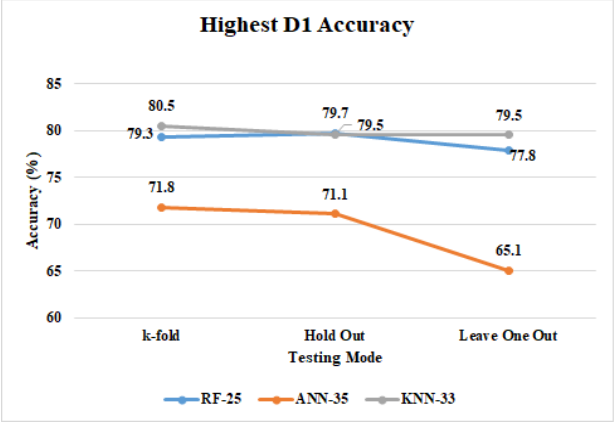

Fig. 6 Dataset D1 testing using Leave One Out.

Data from Fig. 6, in D1 dataset, the Leave one out method provided a maximum accuracy of 79.3% for KNN-33. One hand, for RF-25, Hold Out method is suitable in prediction accuracy and on another hand, for both ANN-35 and KNN-33, K-Fold is appropriate due to the highest accuracies.

VI. EXPERIMENT –IV

On the dataset D0, T-test at 0.05 significant is applied using Feature Selection methods with each classifier at training ratio 66:44 and K-Fold with k=10. It is evident from Fig. 3, that T-test found a significant difference between the ANN and RF; ANN and KNN in locality prediction. It is also found that the blue V shape graph is higher than the orange V shape graph. It reflects that the K-fold method played a significant role to enhance the prediction accuracy of initial dataset D0.

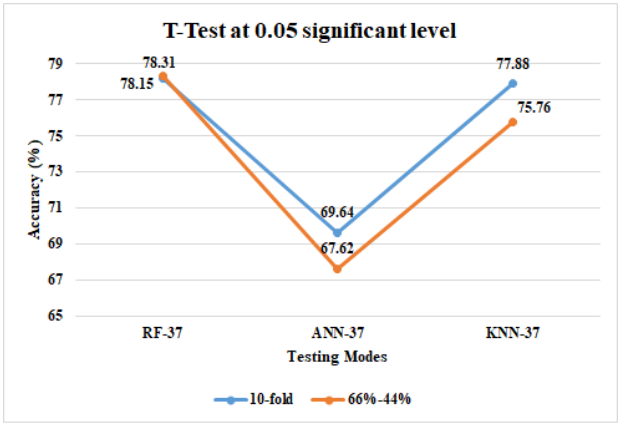

Fig. 7 Classification accuracy comparison using T-test.

Further, we tested and validated extracted dataset D1 using T-test at 0.5 significant level to keep in view CPU training time in seconds. In **Fig. 8**, For both of classes Rural and Urban, the stabilized and highest prediction count prediction is found by KNN in minimum time 0.08 seconds.

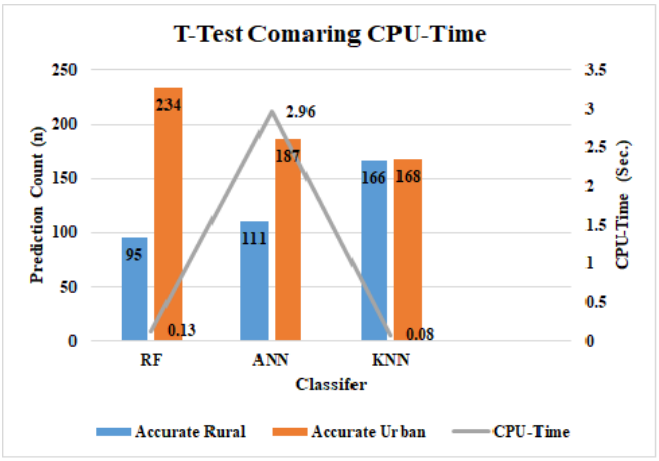

Fig. 8 Prediction Count at Training CPU-Time.

Out of 415, a total of 334 instances are predicted correctly by KNN at 10-Fold. The ANN classifier has taken maximum time 2.96 seconds in the prediction of total of 228 instances. It is evident that KNN classifier outperformed others in stabilized accuracy with optimum training time. One hand, T-test did not find a significant difference between RF and KNN towards CPU training time and another hand, a significant difference is found between ANN and KNN; ANN and RF classifier.

VII. RESULTS EVALUATION AND DISCUSSIONS

This section discusses four performance metrics which are significant for binary classification. In **Table 1**, the KNN classifier shown the strongest bonding among instances due to highest kappa static 0.62 which infers the strength of locality prediction.

The utmost F-score 0.80 for rural and 0.81 for urban is calculated by KNN which also proved strong balanced between the precision and the recall. The ROC value of KNN is also found maximum as compared to other classifiers which proved the sensitivity and specificity of the model in a good manner. The KNN-33 has minimum 19.5% and ANN-35 has maximum 28.2 prediction error.

Table. 1 Performance metrics.

Classifier	Kappa Static	Error (%)	ROC value		F-score	
			R	U	R	U
RF-25	0.54	20.7	0.82	0.82	0.69	0.85
ANN-35	0.42	28.2	0.76	0.76	0.66	0.76
KNN-33	0.62	19.5	0.83	0.83	0.80	0.81

Fig. 9 shows that the Feature Selection methods improved locality prediction accuracies of each classifier.

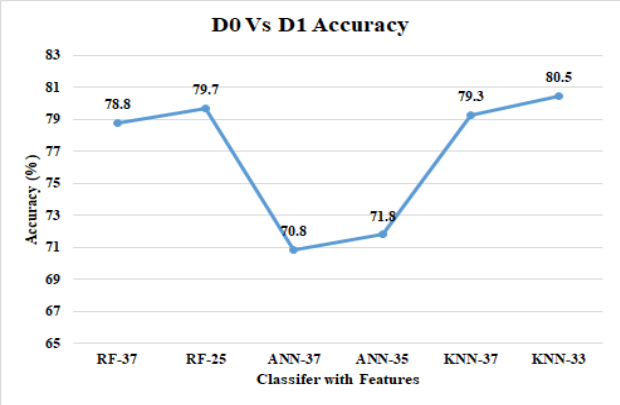

Fig. 9 Feature Selection Accuracy Comparison.

Data from Fig. 9 the locality prediction accuracy of RF-37 is enhanced by 0.9% with a reduction of 13 features. It is also evident that the prediction accuracy of ANN-37 is improved by 1% with a reduction of 02 features. With reducing 4 features, the accuracy of KNN-37 is also significantly improved by 1.2%. Hence, KNN-33 is proved as best and optimum prediction classifier to predict the student's locality towards ICT and MT in both countries.

VIII. CONCLUSION

To present real-time locality predictive model, the authors performed 4 experiments in Weka 3.9.1. For this, three testing methods were applied to both initial and extracted datasets. For Feature Selection, three different types of filter methods were used. Also, accuracy comparison is performed by T-test at 0.0.5 significant level. During experiment I, while testing of initial dataset D0, it is found that the RF classifier outperformed others in locality prediction of students towards ICT and MT at training ratio 60:40 and the KNN classifier outperformed others at 10-fold. During experiment II, while testing extracted dataset D1, it is evident that the KNN-33 achieved a maximum prediction accuracy of 80.1% at 60:40 training ratio and it is also enhanced by 0.4% at 10-fold. In experiment III, the Leave one out method provided identical and extreme accuracies 79.3% for both KNN-37 and RF-37. One hand, for both RF-37 and ANN-37, hold out method is appropriate in prediction accuracy and on another hand, Leave One Out method is suitable for KNN-37. During experiment IV, while testing D0, the T-test has found the significant difference between the accuracies of ANN and RF; ANN and KNN in locality prediction. Further, in the testing of D1, the KNN outperformed others and it stabilized and attained the highest prediction count in optimum time 0.08 seconds. It is revealed that *CorrelationAttributeEval* has enhanced the accuracy of RF by 0.9%, *InfoGainAttributeEval* has improved the accuracy of ANN by 1%, and *OneRAttributeEval* has increased the accuracy of KNN by 1.2%. Future work is recommended to apply more wrapper methods with other classifiers to enhance the prediction accuracy in real time. Also, the academic institutions are advised to deploy this predictive model as real-time website to measure the student's attitude, development and availability of ICT and MT, educational benefits of ICT and use of modern ICT resources and MT in higher education.

ACKNOWLEDGMENT

The present study relates to the first author's Ph.D. study sponsored by Tempus Public Foundation, Budapest, Hungary. Also, this work is supported by the Hungarian Government and co-financed by the European Social Fund under the project entitled "Talent Management in Autonomous Vehicle Control Technologies (EFOP-3.6.3-VEKOP-16-2017-00001)".

REFERENCES

1. C. Verma, S. Dahiya, "Gender difference towards information and communication technology awareness in indian universities," in *SpringerPlus*, vol. 5, pp. 1–7, 2016.
2. C. Verma, S. Dahiya, D. Mehta, "An analytical approach to investigate state diversity towards ict: A study of six universities of Punjab and Haryana," in *Indian Journal of Science and Technology*, vol. 9, pp. 1–5, 2016.
3. R. Zhang, C. Shi, Li. Yang, "Predicting students' performance in educational data mining,". in *International Symposium on Educational Technology*", IEEE, 2015.
4. K. Sreenivasa Rao N. Swapna, P. Praveen Kumar, "Educational data mining for student placement prediction using machine learning algorithms," in *International Journal of Engineering and Technology Sciences*, Vol. 7, Issue 1.2, pp. 43–46, 2018.
5. C. Verma, S. Ahmad, V. Stoffová, Z. Illés, S. Dahiya, "Gender prediction of the European school's teachers using machine learning: Preliminary results," in *International Advance Computing Conference*, pp. 213–220. IEEE, 2018.
6. C. Verma, S. Ahmad, V. Stoffová, Z. Illés, "Forecasting residence state of indian student based on responses towards information and communication technology awareness: A primarily outcomes using machine learning," in *International Conference on Innovations in Engineering, Technology and Sciences*, IEEE, to be published, 2018.
7. C. Verma, S. Ahmad, V. Stoffová, Z. Illés, S. Dahiya, "Binary logistic regression classifying the gender of student towards Computer Learning in European schools," in *THE 11th Conference of Ph.D students in computer science*, pp. 45. Szeged University, 2018.
8. C. Verma, V. Stoffová, Z. Illés, "An ensemble approach to identifying the student gender towards information and communication technology awareness in european schools using machine learning," in *International Journal of Engineering and Technology*, vol. 7, pp. 3392–3396, 2018.
9. C. Verma, V. Stoffová, Z. Illés, "Rate-Monotonic Vs Early Deadline First Scheduling: A Review," in *International Conference on education technology and computer Science in building better future*, pp. 188–193. University of Technology and Humanities, 2018.
10. C. Verma, V. Stoffová, Z. Illés, "Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results,". in *International Conference on Electrical, Computer and Communication*, IEEE, to be published, 2019.
11. C. Verma, S. Ahmad, V. Stoffová, Z. Illés, M. Singh, "National identity predictive models for the real time prediction of European schools students: preliminary results," in *International Conference on Automation, Computational and Technology Management*, IEEE, to be published, 2019.
12. C. Verma, Z. Illés, V. Stoffová, "Attitude Prediction Towards ICT and Mobile Technology for The Real-Time: An Experimental Study Using Machine Learning," in *The 15th International Scientific Conference eLearning and Software for Education*, vol. 3, pp. 247–254, 2019.
13. C. Verma, Z. Illés, V. Stoffová, "Real-Time Prediction of Development and Availability of ICT and Mobile Technology in Indian and Hungarian University," *International Conference on Recent Innovations in Computing*, Springer, to be published, 2019.
14. Y. Bathla, C. Verma, N. Kumar, "Smart Approach for Real Time Gender Prediction of European School's Principal Using Machine Learning," *International Conference on Recent Innovations in Computing*, Springer, to be published, 2019.

15. S. Ahmad, D. Chetverikov, C. Verma, “Stability and Reduction of Statistical Features for Image Classification and Retrieval: Preliminary Results”, 9th International Conference on Information and Communication Systems (ICICS), pp. 117–121. IEEE, 2018.

AUTHORS PROFILE

Chaman Verma is a Research Scholar of the Ph.D. program at Faculty of Informatics at Eötvös Loránd University, Hungary. He has completed his first Ph.D. in Computer Science and Engineering from Shri JJT University, India during the academic year 2014-2016. He received his M-Tech (Computer Science and Engineering) from Ch. Devi Lal University, Sirsa, Haryana during the academic year 2008-2010. He has completed a Master of Science in IT in 2007-2008 from PTU Jalandhar. He was a student of National Institute of Electronics Information Technology, New Delhi in duration 2005-2007. He remained Assistant Professor for 3 Year in Computer Science and Engineering department of JCD Memorial college of Engineering, Sirsa, Haryana (2013-2017). He stayed as Assistant Professor in the Department of Computer Science and Engineering in Eternal University, Baru Sahib, Himachal Pradesh during 2010-2012. He is a member of editorial board of various International Journals and a life member of ISTE, New Delhi. He is also a senior member of IACSIT Singapore and member of IAENG Hong Kong. His area of interest is AI, Data Mining, Machine Learning, Deep learning, Real Time systems, Data analytics, ICT and its implementations.

Zoltán Illés is an associate professor in the department of Media and Educational Informatics, faculty of informatics, Eötvös Loránd University, Hungary. His area of interest is a real-time system, Educational Informatics, Embedded system, and Programming. He is author or co-author of more than 100 publications (articles in reviewed journals, proceedings of conferences, chapters in monographs, textbooks, etc.). He is a doctoral study supervisor in computer science at Eötvös Loránd University in Budapest, Hungary.

Veronika Stoffová is a university professor at the Faculty of Education, Trnava University in Trnava (Slovakia) and has a part-time job at Palacký University in Olomouc (Czech Republic). She received her Ing. degree (1974) and Ph.D. degree (1982) in Technical cybernetics at Slovak Technology University in Bratislava. The degree of dr. hab. which she received in Technical cybernetics at Army Academy in Brno (the Czech Republic, 1984) and in Teaching mathematics at Constantine the Philosopher University in Nitra (2000). She received the degree of university professor after an inauguration process in Teaching mathematics at the last-mentioned university (2003). She worked as a university teacher at Slovak Technology University in Bratislava (1974-1981), at Army Academy in Liptovský Mikuláš (1981-1987), at Constantine the Philosopher University in Nitra (1997-2005) and at J. Selye University in Komárno (2004-2015). Her main research interests are oriented towards the computer modeling, simulation, and visualization of systems in different fields of science. She also deals with the use of artificial intelligence in computer modeling of the teaching and learning process. She is author or co-author of more than250 publications (articles in reviewed journals, proceedings of conferences, chapters in monographs, textbooks, etc.). She is doctoral study supervisor in computer science at Eötvös Loránd University in Budapest, at Faculty of Education of Palacký University in Olomouc and at Faculty of Mathematics Physics and Computer Science of Comenius University in Bratislava.