Proceedings of the SMART–2019, IEEE Conference ID: 46866
8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

ICT and Mobile Technology Features Predicting the University of Indian and Hungarian Student for the Real-time

Chaman Verma[1], Veronika Stoffová[2], Zoltán Illés[3] and Mandeep Singh[4]
[1]*Department of Media and Educational Informatics, Eötvös Loránd University, Budapest, Hungary*
[2]*Department of Mathematics and Computer Science, Trnava University, Trnava, Slovakia*
[3]*Department of Media and Educational Informatics, Eötvös Loránd University, Budapest, Hungary*
[4]*Department of Computer Science and Engineering, Chandigarh University, Chandigarh, India*
E-mail: [1]chaman@inf.elte.hu, [2]nikaStoffova@seznam.cz, [3]illes@inf.elte.hu, [4]mandeeptinna@gmail.com

***Abstract*— Feature extraction has a vibrant part in Machine learning (ML) to identify the data patterns with optimum accuracy. We proposed some significant features to predict the student's institution or university based on their answers in the technological survey. Four experiments were performed in IBM SPSS Modeler version 18.2 using 4 ML to resolve the binary classification problem. In the university prediction problem, the uppermost accuracy of 94.26% is provided by eXtreme Gradient Boosting Tree (XGBT) and suggested 18 significant features out of a total of 37. Further, the Artificial Neural Network (ANN) with boosting scored second maximum accuracy of 93.96% and recommended 10 significant features; Support Vector Machine (SVM) provided third-highest accuracy of 92.45% with the recommendation of 12 features; and Random Tree (RT) attained the least accuracy 92.15% with recommendation of 10 important features. The findings of the paper conclude that the XGBT classifier outperformed others in prediction. Also, a noteworthy dissimilarity was found between XGBT's accuracy and SVM's accuracy, RT's accuracy.**

Keywords: Supervisor Machine learning, Predicator importance, University prediction, Performance Evaluation

I. Introduction

Data Mining (DM) is the way of finding patterns in huge data sets using support of the intersection of statistical analysis and ML, and database systems. It is a potential set of approaches used to pull out anonymous information from big datasets. Many investigators are using various approaches in their respective domains to identify the data patterns. In past, using statistical methods were not enough to analyze the data pattern and differential and inferential analysis on datasets were applied with statistical techniques (t-test, z-test, f-test, ANOVA) [1] [2] [3]. DM is thus a practical field and involves learning in a practical and not in a theoretical sense. ML provides a technical base for data mining but it is not a DM itself. It is value-added applied mathematics and statistics. Now a day, the use of ML in DM are more powerful using significant concepts [4]. Many researchers have used supervised machine learning algorithms on academic datasets. The academic performance of students has been classified using the ML algorithms [5]. The SVM and ANN used to classify the study courses choices of students [6]. European student's and teacher's gender was predicted with supervised ML classifiers [7] [8] [9] [10]. Also, the gender of Indian and Hungarian students towards ICT and MT was predicted using ML techniques [11].

The residence country of students [12], locality (rural and urban) and locality scope (national and international) for the real-time with ensemble methods was presented [13] [14] [15]. The national personality of European school students concerning ESSIE survey responses for real time was also presented [16]. Also, several types of decision trees were realistic to resolve classification problems [17], [18]. The student's age group was classified with the help of machine learning techniques for the real-time system [19]. Further, real-time predictive models to predict the awareness levels and attitudes towards ICT and MT were also presented [20] [21]. The concept of real-time classification of the development and availability of ICT and MT was presented [22]. The presented predictive models might be helpful to query across the online dataset to an appropriate match. Therefore, the deployment of presented models may lead to provide benefit to the centralized administration to the monitoring the technological use in the institutions. Also, it may be useful to inspecting the online ICT material and ICT obstacles facing by stakeholders. Further, university

wise identification of the real educational benefits of ICT to stakeholders may be happened. It would be also easy to manage university wise response system under ICT Knowledge which wires ICT support system for both universities.

II. Research Methodology

A. Dataset

A primary dataset with a stratified random sampling method is used in the experiments. Google form as an instrument and personal visits were conducted by the first author. Initially, the dataset has 337 instances and 46 features out of which 9 features belong to demography. The features of the dataset relate to the attitude (AICTM1-AICTM6), development-availability (DA1-DA16), educational benefits (EBICTM1-EBICTM9) and usability (UICTM1-UICTM6). Using self-reduction, 09 features (demographic) were removed. The preprocessing of data needs to tackle with missing values to enhance accuracy [23]. Weka 3.8.1 tool tackled 6 missing values well with ReplaceMissingValue filter by replacing mode or mean of the training dataset. Because of a hybrid scale of data measurement, normalizations were also performed with the scale from 0 to 1 using the *Normalize* filter. The university was considered as target or response variable which has two values such as ELTE and CU.

B. Classifiers and Performance Metrics

To develop the predictive model, four supervised ML algorithms XGBT, ANN with boosting, SVM and RT are modeled using with the holdout testing method. The holdout ratio 60:40 were applied in which 60% train dataset and 40% considered as test dataset. Below significant measures were used to validate the outcomes of the experiments. (a) Correct: It is the no. of precise predictions of the university from overall predictions. (b) Wrong: It is the no. of erroneous predictions of the university from overall predictions. (c) Coincidence matrix: A tabular form shows the predicted versus actual instances. (d) Performance Evaluation (PE): It shows performance evaluation statistics for models with categorical outputs. (e) Receiver Operating Characteristic Curve (ROC): A comparing graph to see difference between True Positive rate and the False Positive rate at dynamic cutoffs. (f) Area Under Curve (AUC): It is 2-dimensional area underneath the ROC which combine a measure of performance across all dynamic thresholds. (g) Gini: It is a key metric for frequency distribution.

C. Environment

The experiment was performed in the ML tool named IBM SPSS Modeler 18.2. Fig. 1 shows the abstract simulation environment each experiments collectively. Total nine nodes were involved in present experiments. Node-1 is Excel node which is used to load the preprocessed dataset, Node-2 is type node used to set the predictors and target. Node-3 is partition node for holdout testing. Node-4 is for RT, Node-5 is for XGBT, Node-6 is for ANN, Node-7 is for SVM. Node-8 is an analysis node to generate performances metrics for each classifier. Node-9 is evaluation node to produce the ROC curve for each model.

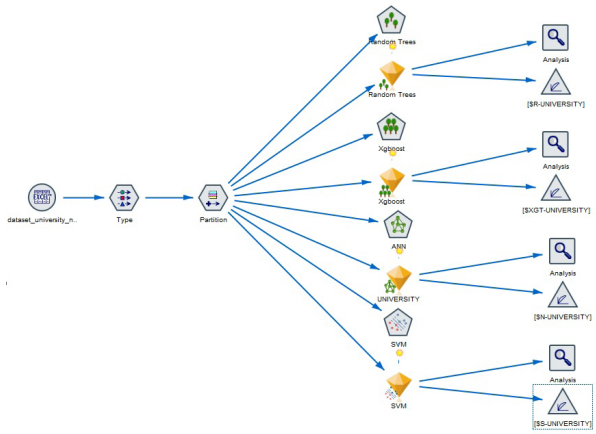

Fig. 1: Experimental Environment

III. Experiments and Results

A. Experiment-1

In this experiment, at 60:40 holdout ratio, 331 instances belong to 37 features are trained and tested. Fig. 2 shows the predictor's performance of RT classifier towards university prediction.

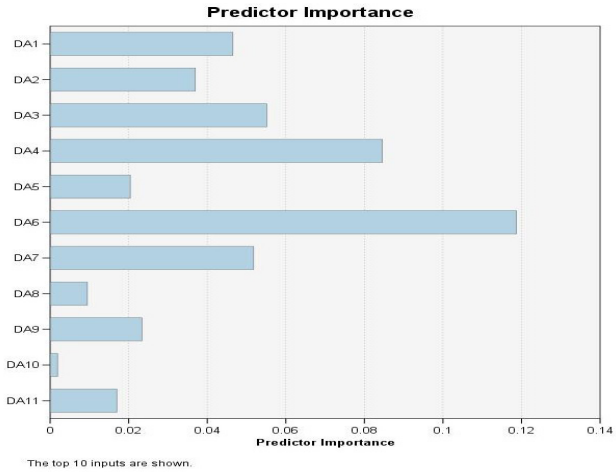

Fig. 2: Predictor Performance of RT

Data from Fig.2 shows only 11 features extracted from the dataset. Also, features relate to DA parameter of the dataset are considered. We found that the DA6 scored highest prediction value 0.12 as compared others. DA10 has worst prediction value 0.01. Therefore, 10 features are recommended in the university prediction of the student.

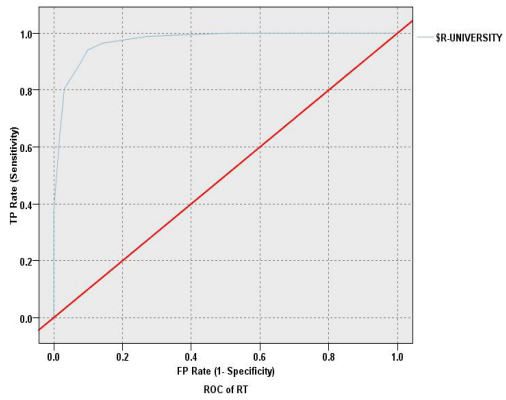

Fig. 3: ROC of RT

In Fig. 3 it can be seen that the sensitivity of RT model is extraordinary 0.96 at 0.2 thresholds and the (1-specificity) is 0.14 which revealed the strong sensing power of the RT predictive model for the university prediction.

B. *Experiment-2*

In this experiment, the SVM classifier is modeled on the dataset with 60:40 holdout ratio.

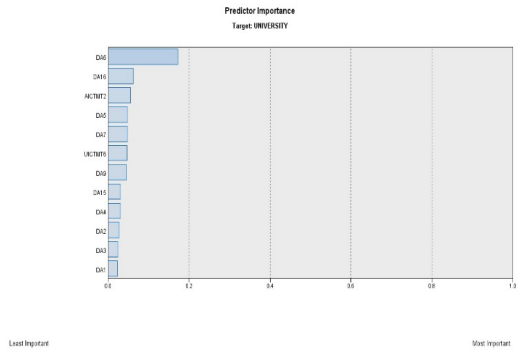

Fig. 4: Predictor Performance of SVM

Fig. 4 shows the predictor's performance of SVM classifier towards university prediction and we found 10 features relate to the DA, and 1 relates to AICTM and 1 from UICTM. Out of total 12, DA6 has the highest prediction value 0.17 and DA1 has the worst prediction value 0.02. Hence, the above 12 features are suggested to predict the University of the student towards ICT and MT.

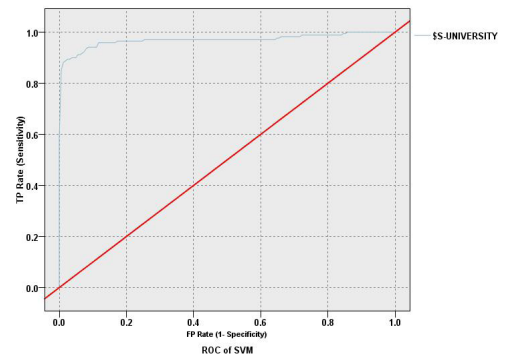

Fig. 5: ROC of SVM.

Fig. 5 shows the ROC curve produced by SVM classifier and the authors found the SVM model sensing high with the value of 0.97 with cutoff value 0.4. Also, the (1-specificity) is found very low which is 0.13. At .08 cutoffs, the sensing rate is 0.98 and the FP rate is calculated 0.02. It means that the SVM model is also significant to predict the University of the student.

C. *Experiment-3*

In this experiment, ANN with a multilayer perceptron function is modeled on the dataset with 60:40 holdout ratio. Also, the boosting technique is ensemble with a multilayer perceptron to enhance the prediction accuracy as well.

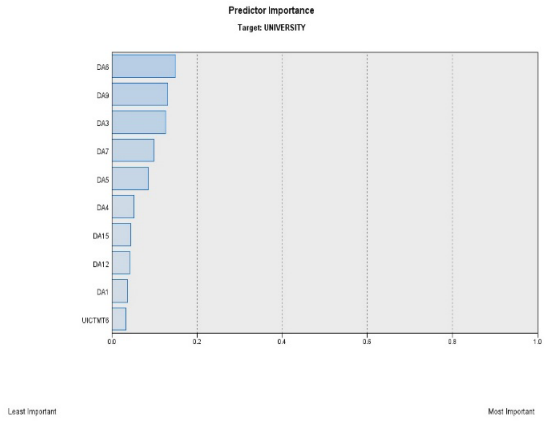

Fig. 6: Predictor Performance of ANN

Data from Fig. 6 shows only 10 most significant features to predict the university of the student. Also, 09 features relate to the DA parameter of the dataset are selected and 01 feature (UICTM6) from the UICTM parameter is suggested. We found that the DA6 scored highest prediction value 0.15 and UICTM6 has the worst prediction value 0.03. Consequently, 10 features are suggested in the university prediction of the student.

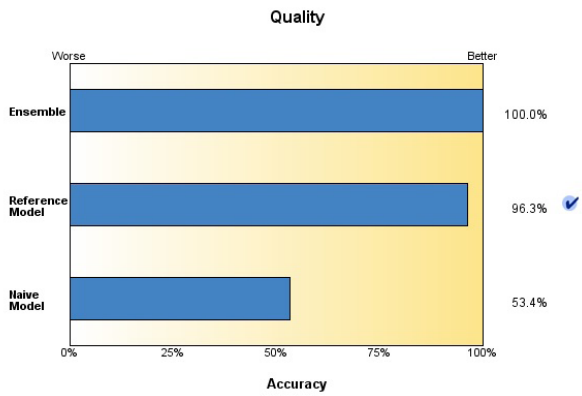

Fig. 7: Ensemble ANN with Boosting.

Data from Fig. 7 signifies the quality of the ANN classifier with using a boosting method. It is observed that the accuracy of the reference model can go up to 96.3% which is measured as better on the scale. Hence, ANN predictive model is significant to use.

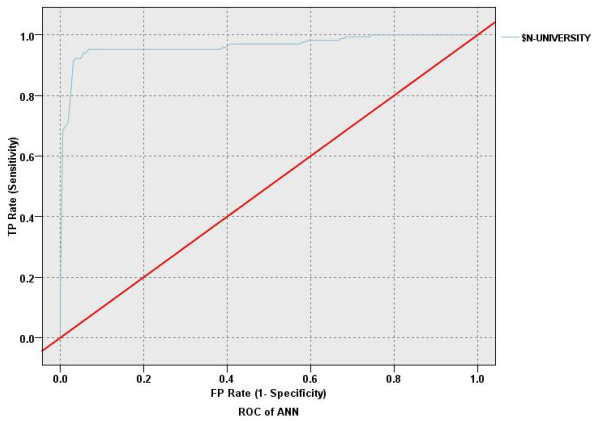

Fig. 8: ROC of ANN with Boosting

From Fig. 8 at thresholds 0.7 the sensitivity is high 0.99 and the FP rate is 0.01. The ANN model starts sensing at 0.1 cutoffs which proves that the ANN model is also weighty in the prediction.

D. Experiment-4

In this, the XGBT classifier [24] [25] is modeled on the dataset to enhance the accuracy of previous SVM and RT predictive models. The dataset is tested with the same 60:40 holdout ratio. The exact greedy method is applied which calculates potential splits in XGBT and the number of boost round is set to 500 which correlates to the number of weak trees to create. The maximum depth of tree growth is set to 6 and minimum child weight is set to 1.0. As the dataset is already balanced although, the max delta step parameter is set to 0.0. The XGBT objective function is set as binary: logistic which meaning is logistic regression of binary classification. Further, the subsample value is set as 0.8 which means that randomly 80% of instances to grow a tree to prevent overfitting.

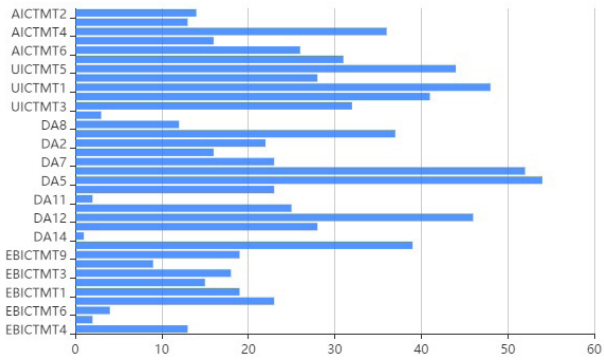

Fig. 9: Predictor Performance of XGBT

Data from Fig. 9 shows that 18 significant features are suggested by XGBT in which 3 features relate to the AICTM parameter; 3 features relate to the UICTM parameter; 07 features concerning the DA, 05 features relate to EBICTM. On one hand, the maximum prediction value of DA5 is calculated as 54 and another hand worst prediction value of DA14 is found 1. Accordingly, these 18 features are suggested to predict the university of the student towards ICT and MT.

Fig. 10 displays the XGBT's ROC curve with values of TP rate Vs FP rate at varying cutoffs. It can be seen that noteworthy The XGBT is sensing from 0.66 and terminates at 0.99 with changing thresholds. It can be seen that the XGBT Model has the highest TP rate of 0.98 at threshold 0.2 and at the same point, the (1-specificity) is 0.02 which proves the implication of the classification model.

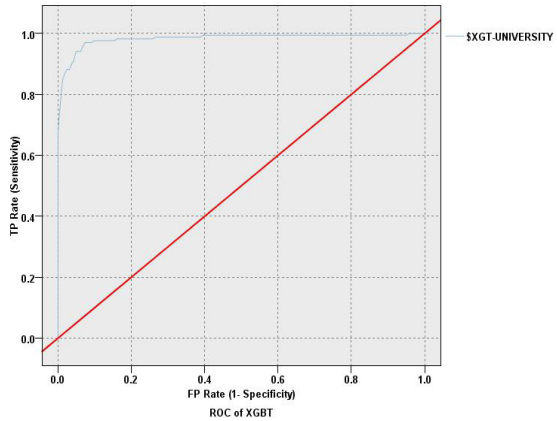

Fig. 10: ROC of XGBT

IV. Performance Measures

This section discusses important performance measures of each predictive model presented in experiments. From Table 1, the highest university prediction accuracy (94.26%) is achieved by XGBT as compare to others. Also, the worst prediction accuracy (92.155) is given by RT and it has the highest prediction error (7.85%).

Table 1: Performance Measures

	Accuracy (%)	Error (%)	AUC	Gini
XGBT	94.26	5.74	0.982	0.964
ANN	93.96	6.04	0.962	0.924
SVM	92.45	7.55	0.968	0.936
RT	92.15	7.85	0.972	0.945

The maximum AUC value (0.982) is calculated by XGBT and the lowest AUC value (0.962) is found by ANN. The highest Gini value is measured 0.964 by XGBT which proves the better performance of it. Hence, it is inferred that XGBT is the winner classifier with 18 features as compared to others.

Table 2: Class Performance Evaluation

Class	ELTE	CU
Classifier	PE	PE
XGBT	0.601	0.669
ANN	0.611	0.651
SVM	0.587	0.643
RT	0.576	0.648

From Table 2, the XGBT favored class CU with maximum PE (0.669) as compare to ELTE class. Also, the ANN classifier performed well for ELTE class with the highest PE value which is 0.611. Hence, these two predictive models are most significant for prediction. A joint coincidence matrix in Table 3 shows the entire number of correctly predicted students is 312 out of 331 by XGBT which is maximum. Also, ANN has correctly predicted student 311; SVM correctly predicted student 306, and RT has predicted 305 students accurately.

V. Conclusion

To predict the student's university towards ICT and MT, 4 experiments were performed in the IBM SPSS Modeler tool. The result of the first experiment evident that the RT classifier achieved 92.15% accuracy and suggested 10 features with the majority of DA parameters for the university prediction. The outcome of the second experiment revealed that the SVM classifier obtained 92.45% prediction accuracy and suggested 12 features with the DA parameter. In the third experiment, ANN with boosting significantly enhanced the prediction accuracy of 93.96% and 10 features were suggested in the university prediction. In the last experiment, the XGBT classifier outperformed others with the best accuracy of 94.26% with 18 most valuable features provided. These features cover almost every parameter (AICTM, DA, EBICTM, and UICTM) of the dataset which is indispensable. On one hand, the findings of the paper concluded no meaningful difference has been found between SVM's and RT's accuracy and another hand, XGBT performance significantly differs from both SVM and RT. The future work is suggested to apply feature extraction methods likewise gain ratio, Info gain, principal component analysis, etc. Also, the use of ensemble methods such as bagging, blending with applied algorithms or with another may enhance the prediction accuracy. The authors recommended this XGBT model to be implemented online to the interpretations of the student's willingness, attitude, and awareness towards technology. Also, the present model may help to identify the usability, availability and educational benefits of ICT and MT at university.

Table 3: Coincidence Matrices

	XGBT		ANN		SVM		RT	
University	CU	ELTE	CU	ELTE	CU	ELTE	CU	ELTE
CU	150	12	152	10	148	14	146	16
ELTE	7	162	10	159	11	158	10	159

Acknowledgment

The first author's Ph.D. is associated with the Stipendium Hungaricum scholarship provided by the Tempus Public Foundation of Hungary and further, the Hungarian Government supported this paper with the grant (EFOP-3.6.3-VEKOP-16-2017-00001)" under the project "Talent Management in Autonomous Vehicle Control Technologies" of European Social Fund.

References

[1] Chaman Verma and Sanjay Dahiya. Gender difference towards information and communication technology awareness in indian universities. SpringerPlus, 5(370):1–7, 2016.

[2] Chaman Verma, Veronika Stoffová and Zoltán Illés. Analysis of Situation of Integrating Information and Communication Technology in Indian Higher Education: International Journal of Information and Communication Technologies in Education, 7(1): 24–29, 2018.

[3] Chaman Verma, Veronika Stoffová and Zoltán Illés. Perception Difference of Indian Students Towards Information and Communication Technology in Context of University Affiliation: Asian Journal of Contemporary Education, 2(1): 36–42, 2018.

[4] Nisbet R. et.al. Handbook of statistical analysis and data mining applications. 2009.

[5] Kabakchieva D. Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research, 1(4):686–690, 2012.

[6] Agarwal S. et.al. Data mining in education: Data classification and decision tree approach. International Journal of e-Education, e-Business, e-Management and e-Learning, 2(2): 140–144, 2012.

[7] Chaman Verma, Veronika Stoffová, Zoltán Illés, and Sanjay 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, page 45, 2018.

[8] Chaman Verma Zoltán Illés and Veronika Stoffová. An ensemble approach to identifying the student gender towards information and communication technology awareness in European schools using machine learning. International Journal of Engineering and Technology, 7(4):3392–3396, 2018.

[9] Chaman Verma, Ahmed S. Tarawneh, Veronika Stoffová, Zoltán Illés, and Sanjay Dahiya. Gender prediction of the European school's teachers using machine learning: Preliminary results. In: Proceeding of 8th IEEE International Advance Computing Conference, pages 213–220, 2018.

[10] Yatish Bathla, Chaman Verma and Neerendra Kumar. Smart Approach for Real Time Gender Prediction of European School's Principal Using Machine Learning. In: The 2nd International Conference on Recent Innovations in Computing, Lecture Notes in Electrical Engineering (LNEE), pages 1–12 Springer. In Press. (2019).

[11] Chaman Verma Zoltán Illés and Veronika Stoffová. Gender Prediction of Indian and Hungarian Students Towards ICT and Mobile Technology for the Real-Time. International Journal of Innovative Technology and Exploring Engineering, 8(9S3): 1260–1264, 2019.

[12] Chaman Verma, Veronika Stoffová and Zoltán Illés. Prediction of Residence Country of Student towards Information, Communication and Mobile Technology for Real-Time: Preliminary Results. In: International Conference on Computational Intelligence and Data Science, pages 1–12, Procedia Computer Science, Elsevier. In Press, 2019.

[13] Chaman Verma, Veronika Stoffová and Zoltán Illés. Real-Time Prediction of Student's Locality towards Information Communication and Mobile Technology: Preliminary Results. International Journal of Recent Technology and Engineering, 8(1):580–585, 2018.

[14] Chaman Verma, Veronika Stoffová and Zoltán Illés. Ensemble Methods to predict the Locality Scope of Indian and Hungarian students for the real time. In: 4th International Conference on Advanced Computing and Intelligent Engineering, Advances in Intelligent Systems and Computing, pages 1–13, Springer, In Press, 2019.

[15] Chaman Verma, Veronika Stoffová and Zoltán Illés. Prediction of locality status of the student based on gender and country towards ICT and Mobile Technology for the real time. In: XXXII International Scientific Conference, DIDMATTECH 2019, pages 1–10, Slovakia, In Press, 2019.

[16] Chaman Verma, Ahmad S. Tarawneh, Zoltán Illés, Veronika Stoffová, and Mandeep Singh. National identity predictive models for the real time prediction of European school's students: preliminary results. In: IEEE International Conference on Automation, Computational and Technology Management, pages 418–423, 2019.

[17] R.S. Bichkar and R.R. Kabra. Performance prediction of engineering students using decision trees. International Journal of Computer Applications, 36(11):8–12, 2011.

[18] Eun Sung Lee and Jae Sung Lee. Exploring the usefulness of a decision tree in predicting peoples' locations. In 2nd World Conference on Psychology and Sociology, PSYSOC 2013, Procedia-Social and Behavioral Sciences, volume 140, pages 447–451. Elsevier, 2014.

[19] Chaman Verma, Veronika Stoffová and Zoltán Illés. Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results. In: 2019 IEEE Third International Conference on Electrical, Computer and Communication. pages 1–7, In Press, 2019.

[20] Chaman Verma and Zoltán Illés. 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, pages 247–254, Romania, 2019.

[21] Chaman Verma, Veronika Stoffová and Zoltán Illés. Prediction of students' awareness level towards ICT and mobile technology in Indian and Hungarian University for the real-time: preliminary results: Heliyon, Elsevier, 5 (6): 1–7 (2019).

[22] Chaman Verma, Zoltán Illés and Veronika Stoffová. Real-Time Prediction of Development and Availability of ICT and Mobile Technology in Indian and Hungarian University. In: The 2nd International Conference on Recent Innovations in Computing, Lecture Notes in Electrical Engineering (LNEE), pages 1–12 Springer. In Press. (2019).

[23] Jiawei Han et.al. Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems). Elsevier, 2006.

[24] Chaman Verma, Zoltán Illés and Veronika Stoffová. Real-Time Classification of National and International students for ICT and Mobile Technology: An experimental study on Indian and Hungarian University. In: The First International Conference on Emerging Electrical Energy, Electronics and Computing Technologies, pages 1–8 Journal of Physics, IOP Science, UK. Accepted. (2019).

[25] Chaman Verma, Zoltán Illés and Veronika Stoffová. Study level prediction of Indian and Hungarian students towards ICT and Mobile Technology for the real-time. In: IEEE International Conference on Computation, Automation and Knowledge Management, pages 1–6, UAE. Accepted. (2019).