Gender Prediction of Indian and Hungarian Students Towards ICT and Mobile Technology for the Real-Time

Chaman Verma, Zoltán Illés, Veronika Stoffová

ABSTRACT: ***The present paper focused on the prediction of the university student's gender towards Information Communication Technology (ICT) and Mobile Technology (MT) in Indian and Hungary.In this paper, four experiments were performed on dataset using three popular classifiersnamed Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest (RF) with three numerous testing technique such as K-fold Cross Validation (KCV), Hold Out (HO) and Leave One Out (LOO).Three different applications named Explorer, Experimenter and KnowledgeFlow (KF) of Weka 3.9.1 are used for predictive modeling. The class balancing has been also applied using Synthetic Minority Over-Sampling (SMOTE) to enhance the prediction accuracy of each algorithm. Further, a significant difference among classifier's accuracies has also been tested using T-test at the 0.05 confidence level. Also, CPU user time has been calculated to train each model to justify to present real-time prediction of gender towards ICT and MT.The results of the study inferred that the CPU time is significantly differed in between RF (0.18 Seconds), SVM (0.06 seconds) and ANN (4.40 seconds).Also, the RF classifier (89.4%) outperformed others with LOO method in terms of accuracy.The authors recommended these predictive models to be deployed as an online prediction for the gender of the student towards ICT and MT at both universities to track technological activities.***

Keywords: Gender Prediction, Machine Learning, LOO, Prediction Accuracy, HO, KCDV, SMOTE.

I. INTRODUCTION AND RELATED WORK

To predict data pattern in educational datasets, traditionally statistical analysis was the most trending. Many investigators applied statistical techniques to predict data patterns and significant differences among variables towards ICT [1], [2], [3], [4], [5]. Now a day, Educational Data Mining (EDM) is now trending in higher educational institutions to predict data patterns accordingly. EDM is the process of analyzing hidden educational patterns of data according to different perspectives for categorization into useful information,

Revised Manuscript Received on July 25, 2019.

Chaman Verma, Zoltán Illés, Veronika Stoffová,
Faculty of Informatics,EötvösLoránd University,Budapest,Hungary.
Faculty of Education,TrnavaUniiversity,Trnava, Slovakia.
(Corresponding author: Chaman Verma)
(Received 09July 2019, Revised 30 July 2019, Accepted 15 August 2019) (Published by Research Trend, Website: www.researchtrend.net)

which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.Predictive modeling in the educational domain is now being popular. Using machine learning, identify the data patterns and classification is important. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. The supervised machine learning algorithms are appropriate for predictive modeling. Many researchers have applied machine learning classifiers on educational datasets. The attitude and awareness level of students towards ICT and MT were predicted [6] [7]. Further, the demographic features of students were also predicted using machine learning classifiers such as locality, residence country [8] and state of residence [9]. The gender of European school's students and teachers were predicted with best accuracy [10], [11],[12]. In the real-time prediction of the age group of university students, machine learning plays a significant role [13].Also, the national identity of students [14] and student's locality based on gender and country [15] were also predicted with the help of supervised machine learning algorithms. The concept of real-time prediction of development and availability of ICT and MT at university were suggested [16]. Also, a smart approach to automated gender prediction in real time is also provided [17]

In this paper, the authors modeled three machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) with the primary dataset.RF is an ensemble learning methodology which is a collection of Classification and Regression Trees (CART) like trees for growing, combination, testing and Post-processing. RF is growing while training on a sample obtained from the training set via bagging without replacement [18]. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [19].Support Vector Machine (SVM) is a supervised learning model introduced for binary classification in both linear and nonlinear versions [20]. An SVM performs classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories [21], [22], [23].Artificial neural network (ANN) is simple mathematical models defining a function f: X→Y or distribution over X or both X and Y, but sometimes models are also intimately associated with a learning algorithm or learning rūl̄ē [24], [25].

Fig.1 shows the count of student participated from an

Indian university and Hungarian university. From Hungary, a total of 169 students from EötvösLoránd University has participated in this research study and the total 162 students belong to the Chandigarh University of India. The imbalance count of gender has been found during a survey held in the academic year 2017-2018. The total male students were 265 and the total 66 female students have participated in it.

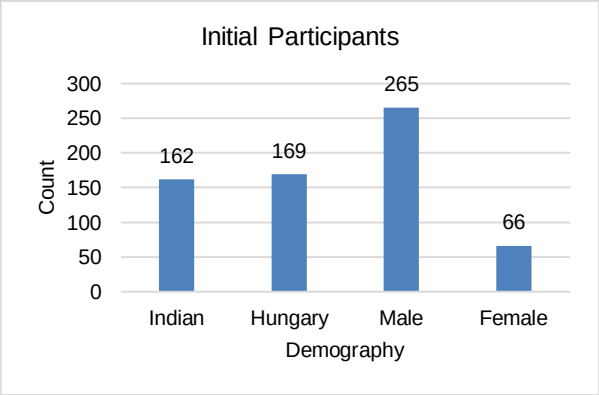

Fig. 1: Student participation

Before the classification of the dataset, it needs to be preprocessed dataset. Only 6 missing values are handled with ***RepalceMissingValue*** filter which replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data. The ***Normalize***filter is used to normalizesall numeric values in the given dataset apart from the class attribute which is gender. The default scale of normalization [0,1] is selected for the data uniformity. Gender has been marked as a class variable which has two values Male-01 and Female-02. The authors performed a binary class balancing operation. Hence, SMOTE algorithm which resamples a dataset by applying the Synthetic Minority Over-sampling Technique. The original dataset must fit entirely in memory. The amount of SMOTE and the number of nearest neighbors may be specified. Fig. 2 shows that initial unbalanced dataset has 66 female and 265 male students who need to be balanced for significant classification. At first run of SMOTE instances belong to a female class are enhanced by 132 and the second run of SMOTE the gender class gets significantly balanced. Now we have a total of 529 instances to be trained and tested for the prediction.

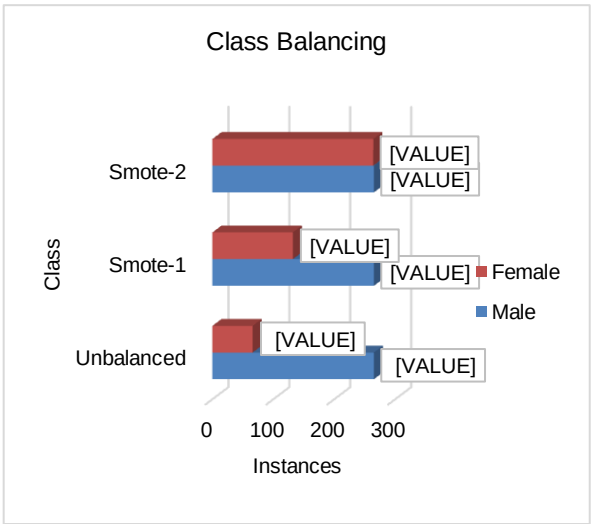

Fig. 2: Gender Class Balancing

II.EXPERIMENTS AND RESULTS

A. *Experiment-1*

In this experiment, the authors performed HO testing methods in which is the simplest form of k-fold cross-validation. This method randomly assigns data points to two sets named training set and the test set, respectively. If we let *k*be some integer less than (or equal to) *n*where *n*is the sample size and we partition the sample into *k*unique subsamples, and in HO validation is really just 2-fold ($k = 2$) cross-validation.The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data.We used various training ratio such as 50:50, 60:40 and 70:30. (see in Fig. 3)

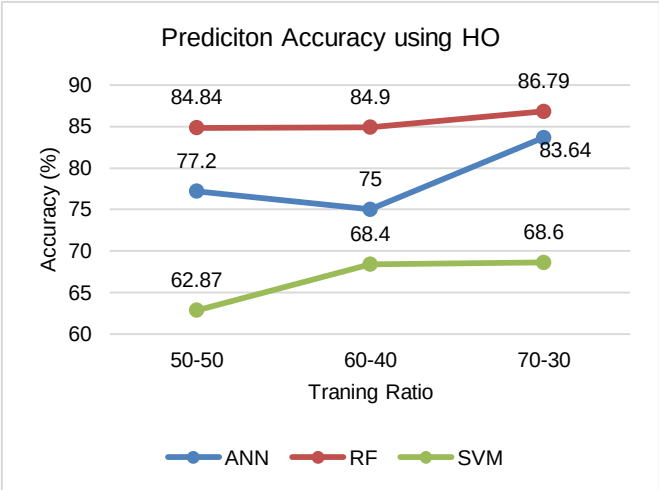

Fig. 3: Prediction Accuracy using HO Method

It can be seen from Fig. 3 that the maximum prediction accuracy (86.79%) is achieved by RF at training ratio 70:30 and worst prediction accuracy (62.87%) is provided by SVM at ratio 50:50. The prediction accuracies of RF and SVM are directly proportional to the training ratio. It is revealed that using HO method the RF classifier outperformed others.

B. *Experiment-2*

This experimentshows that KCV testing methods the original dataset is randomly partitioned into k equal-sized subsets and a single subset is retained as the validation data for testing the model, and the remaining *k−1* subsets are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data.Fig. 4 shows that the dynamic folds of KCV testing. It can be seen that at k=40, KCV enhanced the prediction accuracy of ANN classifier by 0.10 and at k= 50 and k=30 the prediction accuracy of RF is also improved by 2.05%.

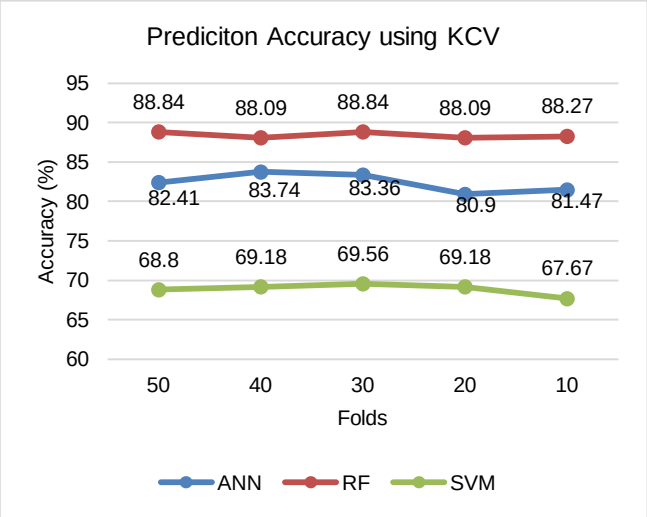

Fig. 4: Prediction Accuracy using KCV Method

C. *Experiment-3*

In thisexperiment LOO testing method is applied with k=529 which is an enhanced version of KCV. In this, the value of k is equal to n where n is the sample size. Likewise, KCV method, LOO also splits randomly dataset but the splitting size is *k=n* equal-sized subsets and a single subset is retained as the validation data for testing the model, and the remaining *k−1* subsets are used as training data. Fig. 5 shows that the at the static value of k=529, One hand the prediction accuracy of ANN classifier is reduced by 1.7% and another hand RF's accuracy is improved by 0.57% and SVM's accuracy is also enhanced by 0.94% as compared to other testing methods. Hence, LOO testing proved most significant for RF and SVM classifier as compared to the ANN classifier.

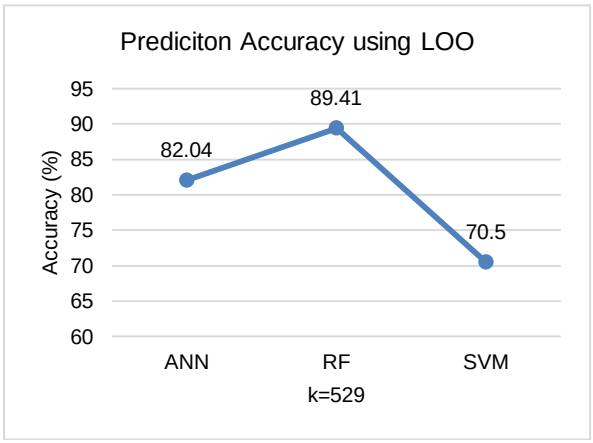

Fig. 5: Prediction Accuracy using LOO Method

D. *Experiment-4*

In this experiment, the statistical t-test is applied at 0.05 significant level to compare the CPU training time modeled by each classifier. For this Weka experimental environment application is used with the KCV testing method and the value of k is set to 30 and the number of iterations are set to 5. The type of the experiment is considered as classification. Fig. 5 shows the joint results obtained provided by T-test at the 0.05 confidence level.

```
Dataset                  (1) functio | (2) func (3) tree
-----------------------------------------------------------
BOTH_GENDER_SMOTE        (2645   4.40 |   0.06 *   0.18 *
-----------------------------------------------------------
                               (v/ /*) |  (0/0/1)  (0/0/1)

Dataset                  (1) function | (2) funct (3) trees
-----------------------------------------------------------
```

```
Dataset                  (1) functio | (2) func (3) tree
-----------------------------------------------------------
BOTH_GENDER_SMOTE        (2645   4.40 |   0.06 *   0.18 *
-----------------------------------------------------------
                               (v/ /*) |  (0/0/1)  (0/0/1)

Dataset                  (1) function | (2) funct (3) trees
-----------------------------------------------------------
BOTH_GENDER_SMOTE        (2645  82.57 |  62.12 *  88.17 v
-----------------------------------------------------------
                               (v/ /*) |  (0/0/1)  (1/0/0)
```

Fig. 6: Accuracy Vs. CPU Time Training comparing using T-test.

The symbol (*) shows a statistically significant difference among classifiers. It can be seen that T-test signifies the statistical difference between the CPU time for training ANN (function) model and SVM (func). Also, the same significant difference is found between the CPU time taken by ANN (function) and RF (tree). It is also revealed that ANN induced maximum time (4.40 seconds) as compare to others to train the gender predictive model. One hand, the T-test also found a significant difference between the accuracy of ANN's and SVM and another no significant difference is found between the accuracy of ANN and RF classifier.Accordingly, RF classifier has achieved the highest accuracy as compared to others to predict the gender of both countries.

III. PERFORMANCE MEASURES

This section describes the experimental results with important performances measures considering binary classification. Table1 presents the seven major performance measures for each algorithm using LOO testing method with k=529. It can be seen that the RF algorithm outperformed others with prediction accuracy and it attained the lowest error rate. The excellent association of instances to predict the gender of the student is scored by RF which is kappa static having value 0.78.The average sensitivity or True Positive Rate (TPR) of male and female is calculated 0.90 by RF which is significant in gender prediction. The False Positive Rate (FPR) [9] is found very less which is 0.10 of RF. Hence, Receiver Operating characteristic area [10] of RF classifier is bigger than other classifiers which signifies the strength of gender predictive model.

Table 1: Prediction Measures.

Measures / Classifier	RF	ANN	SVM
Accuracy	89.4	82.04	70.5
Error	10.6	17.96	29.5
Kappa	0.78	0.64	0.41
TPR	0.90	0.82	0.70
FPR	0.10	0.30	0.30
ROC	0.94	0.87	0.71
F-score	0.89	0.82	0.70

As we know, the F-score is a harmonic mean of precision and recall which also states the significance of predictive models and it is found 0.89 which is also meaningful for

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

the prediction.

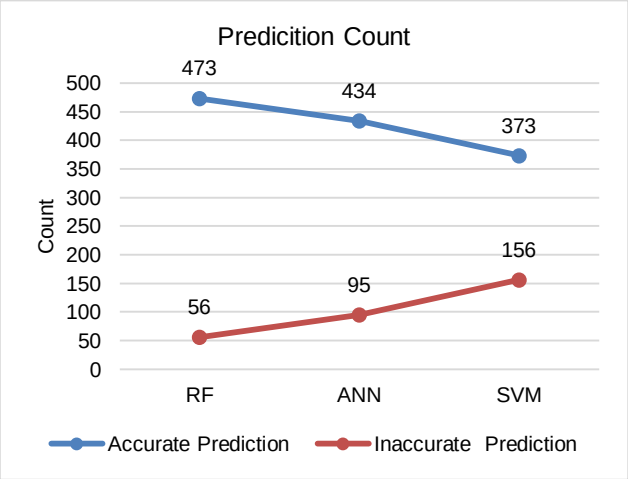

Fig. 6: Prediction Count using LOO Method with k=529.

Data from Fig. 6 infers that the out of total 529, the cumulative (male and female) maximum accurate prediction count 473 is achieved by RF classifier and lowest accurate count is given by SVM classifier. Hence, it is concluded that LOO testing methods significantly boosted the prediction count as compared to others.

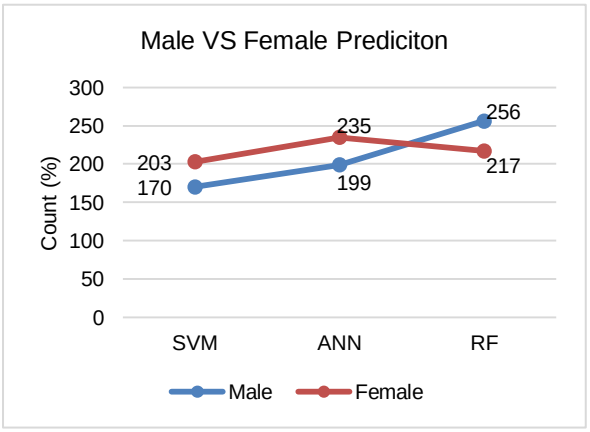

Fig. 7: Accurate gender wise Prediction Count using LOO Method with k=529.

Data from Fig. 7 shows that the maximum right male count of prediction is 256 and the accurate female student count is 217 calculated by RF. The SVM classifier performed insignificant in the prediction of gender ascompared to others.

IV. VALIDATION

To validate each model, Total 8 componentsof KF are used. The*CSVLoader* is used to load the balanced dataset in a Comma Separated File (CSV). The *ClassAssigner* is used to set a target variable named the gender as a class variable.First time male class is picked using the *ClassValuePicker*and second time female class.The *CrossvalidationFoldMaker* is used to select LOO method with set k=529.The*ModelPerfromanceChart*areusedtoprintthe cumulative ROCcurve.

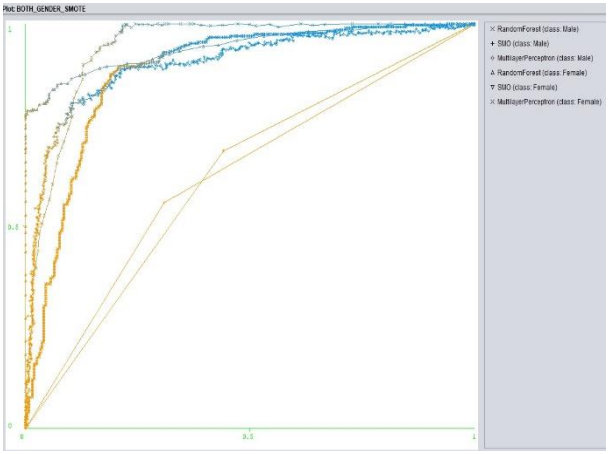

Fig. 8: ROC of gender predictive model.

Data from Fig. 8shows the cumulative ROC curve for each class separately produced using KFE for validating the predictive model which compared the sensitivity with 1-specificity of the real-time model at various thresholds for the student's gender prediction. The ROC shows the sensitivity with 1-specificity for both classes male and female at varying cutoffs and compared each classifier on the dataset.It can have been seen from ROC that RF model is sensing from 0.80 and ends to 0.99 with updating thresholds. Also, can be seen at thresholds 0.5 the sensitivity is high 0.99 and the FP rate is 0.01 which reveals the significance of the real-time genderprediction. Hence, the RF model outperformed others in gender prediction towards ICT and MT.

V. CONCLUSIONS AND FUTURE SCOPE

In this present study, predictive modeling of the gender of Indian and Hungarian students has been performed with three machine learning classifiers (RF, ANN, and SVM) on the primary dataset. In addition to class balancing, this dataset is trained and tested using three different testing methods to enhance the prediction accuracy of each classifier. It is concluded thatusing LOO method accuracy of each classifier are improved. Also, in the prediction of gender, RF classifier outperformed others.The results of the study also proved that there is a significant difference between RF, SVM, and ANN classifier. The authors recommended these predictive models to be deployed as an online prediction for the gender of a student towards ICT and MT at both universities to track technological activities. Further, T-test also found a major difference between the accuracy of ANN's and SVM.

The future work is suggested to apply feature filtering with RGBoost and Deep learning models to boost the prediction accuracy of the RF model. Further, the authors also recommended applying ensemble learning such as bagging, adaptive boosting, and stacking with SVM, RF and ANN classifier to make the predictive model more reliable and accurate. Future, researchers can deploy this predictive model on any other type of survey in any field to identify gender. Also, ICT coordinator of target university may implement this model to track the academic activities of students towards technology. For this, they may develop the real-time web site with a specified deadline of the time to predict the student's gender.

ACKNOWLEDGMENT

The first author thanks Tempus Public foundation of Hungary to sponsoring his Ph.D. study related to this research work. Also, this paper is sponsored by the Hungarian Government and Co-financed by the European Social Fund under the project "Talent Management in Autonomous Vehicle Control Technologies (EFOP-3.6.3-VEKOP-16-2017-00001)".

REFERENCES

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

[2]. Chaman Verma, Sanjay Dahiya and Deepak Mehta (2016). An analytical approach to investigate state diversity towards ICT: A study of six universities of Punjab and Haryana. Indian Journal of Science and Technology,**9** (31): 1–5.

[3]. Chaman Verma, VeronikaStoffová and ZoltánIllés (2018). Perception difference of Indian students towards information and communication technology in context of university affiliation. Asian Journal of Contemporary Education, **2** (1): 36-42.

[4]. Chaman Verma (2017). Educational data mining to examine mindset of educators towards ICT knowledge. Int. J. Data Min. Emerg. Technol.,**7** (1): 53-60.

[5]. Chaman Verma, VeronikaStoffová and ZoltánIllés (2018). 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.

[6]. Chaman Verma and ZoltánIllés (2019). Attitude Prediction Towards ICT and Mobile Technology for The Real-Time: An Experimental Study Using Machine Learning. In: Proceedings of The 15th International Scientific Conference eLearning and Software for Education, **3**, pp. 247–254.

[7]. Chaman Verma, VeronikaStoffová and ZoltánIllés (2019). 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): 1-9.

[8]. Chaman Verma, VeronikaStoffová and ZoltánIllés (2019). Prediction of Residence Country of Student towards Information, Communication and Mobile Technology for Real-Time: Preliminary Results. Procedia Computer Science, Elsevier :1-11.

[9]. Chaman Verma, Ahmad S. Trananweh, VeronikaStoffová and Zoltán. Illés (2018). Forecasting residence state of Indian student based on responses towards information and communication technology awareness: A primarily outcomes using machine learning. In: Proceedings of IEEE International Conference on Innovations in Engineering, Technology and Sciences, In Press.

[10]. Chaman Verma, Ahmad S. Trananweh, VeronikaStoffová, Zoltán. Illés and Sanjay Dahiya (2018). Gender prediction of the European school's teachers using machine learning: Preliminary results. In: Proceedings of IEEE International Advance Computing Conference, pp. 213–220.

[11]. Chaman Verma, VeronikaStoffová, ZoltánIllés (2018). 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.

[12]. Chaman Verma, VeronikaStoffová, ZoltánIllés and Sanjay Dahiya (2018). Binary logistic regression classifying the gender of student towards Computer Learning in European schools. In: Proceedings of THE 11th Conference of Ph.D students in computer science, pp. 45. Szeged University, 2018.

[13]. Chaman Verma, VeronikaStoffová, ZoltánIllés (2019). Age group predictive models for the real time prediction of the university students using machine learning: Preliminary results. In: Proceedings of IEEE International Conference on Electrical, Computer and Communication, In Press.

[14]. Chaman Verma, Ahmad S. Trananweh, VeronikaStoffová, Zoltán. Illés and Mandeep Singh (2019). National identity predictive models for the real time prediction of European schools students: preliminary results. In: Proceedings of IEEE International Conference on Automation, Computational and Technology Management, UK. In Press.

[15]. Chaman Verma, VeronikaStoffová, ZoltánIllés (2019). Prediction of Locality Status of the student based on gender and country towards ICT and Mobile Technology for the real time. In: Proceedings of XXXII-DIDMATTECH, Slovakia, In Press.

[16]. Chaman Verma, ZoltánIllés, VeronikaStoffová (2019). Real-time prediction of development and availability of ICT and Mobile Technology in Indian and Hungarian University. In: Proceedings of InternationalConference on Recent Innovations in Computing, Springer, In Press.

[17]. YatishBathla, Chaman Verma and Neerendra Kumar, "Smart Approach for Real Time Gender Prediction of European School's Principal Using Machine Learning," In: Proceedings of International Conference on Recent Innovations in Computing, Springer, In Press.

[18]. G.Priya and N.Venkatesan (2015). A study of random forest algorithm with implementation using weka. International Journal of Innovative Research in Computer Science and Engineering,**1**(6): 156-162.

[19]. Sushilkumar Ramesh pant Kalmegh (2015). Comparative analysis of weka data mining algorithm random forest, random tree and lad tree for classification of indigenous news data. International Journal of Emerging Technology and Advanced Engineering, **5**(1): 507-517.

[20]. Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin (2003). A practical guide to support vector classification.

[21]. Ali.S (2005). Automated Support Vector Learning Algorithms. PhD dissertation, Monash University, Australia.

[22]. Christopher JC Burges (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, **2** (2): 121-167.

[23]. Ruey Long Cheu, Dipti Srinivasan, and Eng Tian The (2003). Support vector machine models for freeway incident detection. In: Proceedings of IEEE Intelligent Transportation Systems, vol. 1, pp. 238-243.

[24]. Marvin Minsky and Seymour APapert (2017). Perceptrons: An introduction to computational geometry. MIT press, 2017.

[25]. EthemAlpaydin (2009). Introduction to machine learning. MIT press.
