National Identity Predictive Models for the Real Time Prediction of European School's Students: Preliminary Results

Chaman Verma
Dept. of Media and Edu. Informatics
Eötvös Loránd University
Budapest, Hungary
chaman@inf.elte.hu

Ahamd S. Tarawneh
Dept. of Algorithm and Their Application
Eötvös Loránd University
Budapest, Hungary
Ahmadtr@caesar.elte.hu

Zoltán Illés
Dept. of Media and Edu. Informatics
Eötvös Loránd University
Budapest, Hungary
illes@inf.elte.hu

Veronika Stoffová
Dept. Mathematics and Computer Sci.
Trnava University
Trnava, Slovakia
nikaStoffova@seznam.cz

Mandeep Singh
Dept. of Computer Sci. and Engg.
Chandigarh University
Chandigarh, India
mandeeptinna@gmail.com

***Abstract*—An experimental study is conducted to predict the real time national identity (national or immigrants) of the students based on their responses in information and communication technology (ICT) survey held in European schools. All the experiments are conducted in SPSS IBM modeler version 18.1. The target datasets were collected by ESSIE (SMART 2010/0039) during the big survey at levels 3 of schools ISCED (International Standard Classification of Education) in the year 2011. The auto classifier node selected 5 supervised machine learning classifiers filtering out of 8 classifiers. To predict the national identity of students in academic school, the highest accuracy 96.6% is achieved by decision tree C5 with filtering of 46 features out of total 156 and to predict national identity of students in vocational school, the uppermost accuracy 94.3% is achieved by Tree-AS with reduction of total 41 features out of total 172. Hence, to predict national identity, self-reduction and auto classifier stabilized only 46 features for C5 Tree and 41 features for Tree-AS. The findings of paper also signify that C5 classifier outperformed the Logistic Regression (LR) and Tree-AS after feature reduction at academic schools. Further, Tree-AS also outperformed the Bayesian network (BN), linear support vector machine (LSVM) and LR after feature reduction at vocational schools.**

***Keywords*—Classification, Feature reduction, National identity prediction, Supervised machine learning.**

I. Introduction

Data Mining is a powerful set of methods that used for extract hidden information from large datasets. Trending data mining tools are performing a significant role in prediction tasks in various domains. Data Mining tasks can be classified into two categories: Descriptive and Predictive. Descriptive mining tasks characterize properties of the data in a target dataset [1], [2]. Predictive mining tasks perform induction on the current data to make predictions [3]. In data mining, data is accumulated during the learning process and then study can be done with the techniques from statistics [4], machine learning, and other data mining concepts [5]. In supervised learning, we train and test input with preconceived output, having the idea that there is a relationship between the input and the output. It can be categorized into "regression" and "classification" problems. In regression problems, we try to predict results within a continuous output while in classification problems, we try to map input variables into discrete categories. In fact, a lot of work have done by using supervised machine learning classifiers on many different educational datasets according to their own targets. The Bulgarian university student's academic performance has been predicted by using machine learning classifiers [6] and it was found that a higher prediction accuracy of 73.6% is achieved by ANN as compare to the Decision Tree model 72.7% and KNN 70.5%. The study courses selection has been also predicted with the highest accuracy of 97.3% and 95.9% using support vector machine (SVM) and ANN, respectively [7]. The binary logistic regression achieved 62% accuracy to predict the gender of the European school's students towards ICT responses [8] and prediction accuracy also improved by [9] using SVM and RF classifier with 76% accuracy. The gender predictive models of European school teachers are also presented by [10]. The residence state forecasting model of Indian students is also presented [11]. Further, to predict the progress of student based on demographic features, logistic regression (LR) has been also proved better than linear regression, yielding more stable estimates about the presence of ill-fitting patterns [12]. The factors such as the financial status of the students, motivation to learn and gender were discovered to affect the performance of the students. The maximum 66.8% of the students were predicted to have passed while 33.2% were predicted to fail using CHAID tree [13]. Therefore, various variants of decision tree were also applied to solved classification problems [14], [15]. Regression analysis revealed that the predictive validity of school placement decisions was affected by nationality and found that Luxembourgish students were more likely to keep on the track than immigrant students [16]. To improve the accuracy of the classifier in machine learning, feature engineering is the process of selecting or creating features (variables) in a data set to improve machine learning results. Feature selection can include removing unnecessary or redundant features. The process of removing unnecessary variables requires assessing the [17]. Hence, we used the auto classifier node of IBM modeler with option rank models by accuracy or fields. Machine learning plays a vital role in real time prediction as well. A study on the age group prediction University's students for the real time system is also conducted by [18]. According to [19], the 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 prediction of Indian and Hungarian university student's attitude for real-time was also conducted [20]. By implementing the presented predictive models, we can also query across the entire dataset online or query a subset of the dataset for suitable match in real time prediction of national identity of stakeholder. Therefore, based on the national identity of stakeholders, the presented models can also be applied on real time prediction on the following basis: 1. Monitoring Online ICT access and ICT based activities. 2. Controlling Online ICT material and ICT obstacles facing by stakeholders. 3. Learning activities of stakeholders. 4. Providing a sufficient feedback system under ICT Experience which can help to improve ICT support system for national and immigrants. 5. Significant to avoid false identification of the students if they use their wrong national identity anonymously.

II. RESEARCH METHODS AND TECHNIQUES

A. *Dataset Preparation and Feature Reduction*

To solve binary classification problem on a big dataset of European Survey of Schools: ICT in Education (known as ESSIE), authors tested two datasets named dataset1 belongs to academic schools and dataset2 belongs to the vocational schools of 11th grade at levels-3 ISCED (International Standard Classification of Education) schools. Initially, the dataset1 has 50478 instances and 156 features and dataset2 has 37248 instances and 172 features, out of which 10 features belong to indexing and 15 features do not relate with responses. The features of dataset relate to Experience with ICT, Support to ICT use, ICT access, ICT based activities, ICT material and obstacles in ICT use, Learning activities, Teacher skills and Teacher opinions and attitudes etc. The total 190,000 filled questionnaires were examined and more than 2500 schools, from 27 European countries have participated in the survey held in 2011. Using self-reduction, we eliminated 25 features from dataset1 and 26 features from dataset2 because they were belonged to indexing or mean scores. It is very essential to deals with missing values and improve the data quality in datasets before use [21]. Therefore, the total 3,05,769 missing values in dataset1 and 3,86,487 missing values are handled with *MissingValue* filter of Weka 3.8.1 tool. Due to different scaled of numeric data measurement, we normalized dataset from 0 to 1 scale using Normalize filter of Weka tool. We considered the variable named ST20Q01 as target or response variable; and encoded as Nationality status (1- National and 2- Immigrants) of a student in both datasets. We used *NumericToNominal* filter to convert target class (National Identity) to nominal variable.

Feature extraction plays an important role in ensuring effectiveness and stability of the results [23]. Therefore, after self-reduction, the total of 130 features from dataset1 and 145 features from dataset2 are considered as final input to auto classifier algorithm for test and train various classifiers. The auto classifier node plays a vital role to provide more significant features to predict the target variable with maximum accuracy. To extract significant features from both datasets, auto classifier is used with 8 popular supervise machine learning classifiers such as ANN, KNN, LR, Bayesian network (BN), Linear support vector machine (LSVM) and Decision tree variants such as (C5), QUEST and Tree-AS with feature reduction. The Auto classifier trains and tests both dataset separately. It provides the best model by specifying the criteria used to compare and rank models such as overall accuracy, no. of features, area under the ROC curve, profit and lift [15]. We tested both datasets by using overall accuracy and no. of feature options.

The simulation results of auto classifier node filtered out with 145 significant features out of 172 using LR, BN, LSVM and 45 significant features are filtered out for Tree-AS. One side it is clear from Fig.1(a) that the overall accuracy of LR-145 is 93.9%, BN-145 is 92.2%, LSVM-145 is 93.4% and Tree-AS-41 is 94.3% for stabilization of dataset2 for vocational schools to predict the national identity of students. Another side, In Fig.1 (b) for academic school's dataset auto classifier filtered out 130 features out of 131 with less than ¡90% accuracy discard policy and maximum accuracy is given by decision tree C5 with 46 features and Tree-AS also attained second highest accuracy with 45 features. The LR performed also well with 95.4% accuracy with 130 features which covers maximum features of dataset1. The confidence level of filtered significant features in both types of datasets are lies between 95%-100%.

B. *Classifiers and Performance Metrics*

For prediction, 5 supervised machine learning algorithms are used at 10-fold cross-validation. The discard policy of auto classifier is set up as less than ¡90% accuracy with 0.8 AUC (Area Under Curve). Out of total 08 machine learning classifiers, the auto classifier algorithm suggested 5 best models LR, BN, LSVM and Tree-AS at vocational school's dataset2 and 3 best models LR, C5 and Tree-AS at academic school's dataset1. Although, to predict national identity of students from both of datasets based on considered predictors, 5 optimal supervised machine learning algorithms Tree-AS and C5 with feature reduction and LR, BN and LSVM without feature reduction are selected by applying auto classifier algorithm.

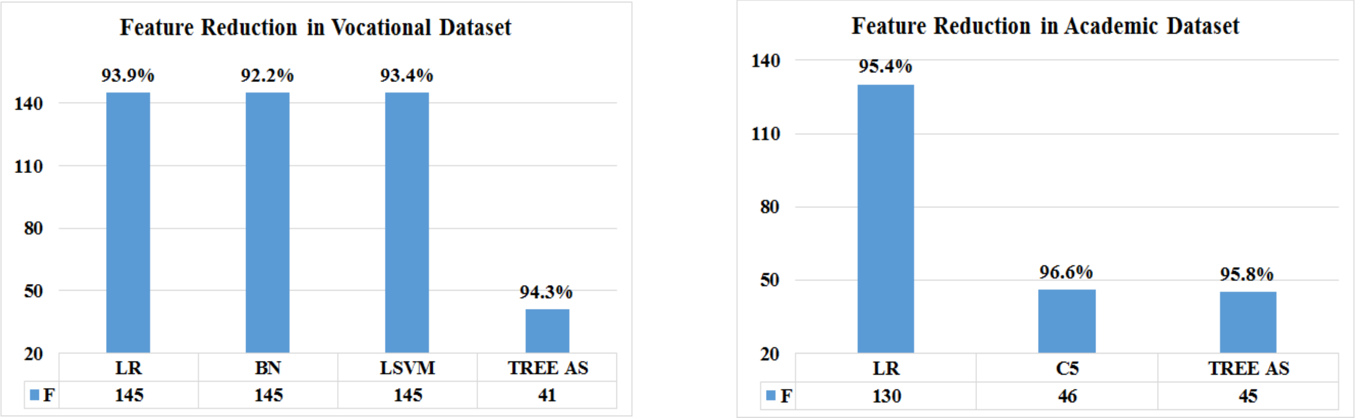

Fig. 1. Feature reduction with highest accuracy in (a) vocational school's dataset2 (b) academic school's dataset1 using Auto Classifier

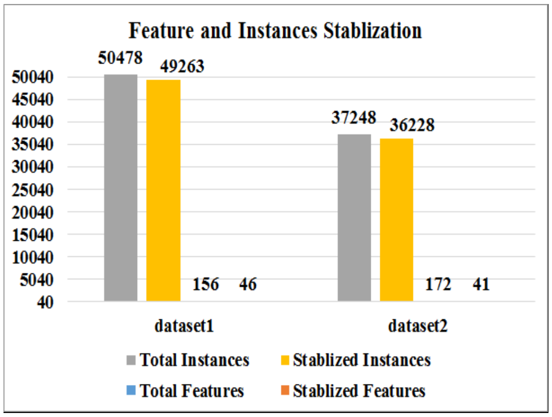

Fig. 2. Stabilized Datasets.

We used performance metrics to measure the performances of the predictive models: (a) Classification Accuracy: The number of correct predictions of student national identity from over all predictions. (b) AUC: To show the accuracy of models' area under the curve of ROC is also appropriate. (c) ROC: Receiver operating characteristics curve presents the graphical evaluation of models which shows the true positive rate (TP or Sensitivity) on the y-axis and false positive rate (FP or 1-Specificity) at x-axis with various thresholds. (d) Gini: It is calculated by subtracting the sum of the squared probabilities of each class from one which computes the inequality among values of a frequency distribution. To evaluate the results, IBM modeler Analysis algorithm (node) is applied to find accuracy, Gini and AUC. Also, the evaluation algorithm (node) is applied to produce ROC curves.

III. Results and Analysis for Academic Schools

This section describes the outcomes of the prediction of the national identity of students of European academic schools. To predict the national identity of students, dataset1 having 49, 263 instances with 130 features are tested and trained using 10-Fold cross-validation using IBM Auto classifier node. The less than 90% accuracy discard policy with best 3 models are fixed to execute 8 classifiers with big dataset1. The outcomes of the auto classifier are LR130, Tree-AS-45, and C5-46.

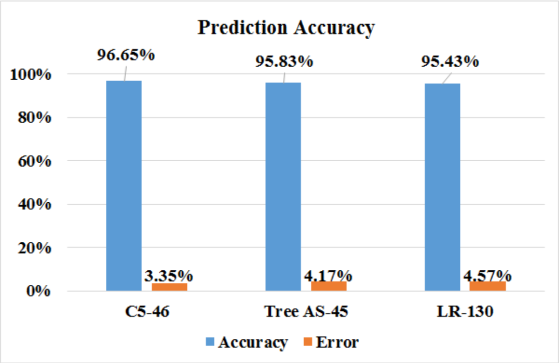

Fig. 3: Accuracy Vs Error for academic schools

From Fig.3 Decision Tree (C5) with 46 features have scored highest accuracy (96.65%) as compared to the LR with 130 features (95.43%) and Tree-AS with 45 features (95.83%). The data-set is also analyzed using 10-Fold cross-validation methods. The maximum error is gained by LR-130 which is calculated as 4.57%.

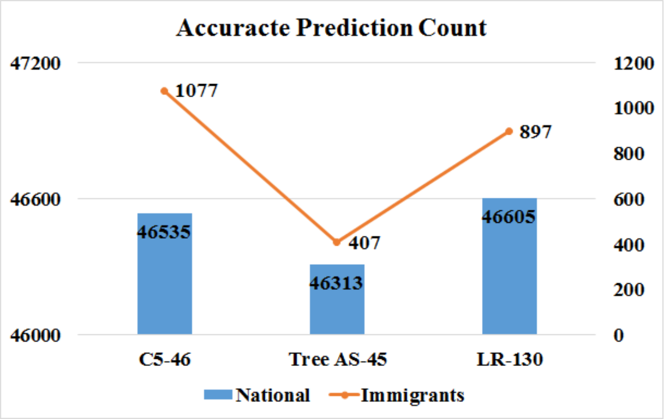

Fig. 4: Accurate Prediction Count of academic schools students.

From Fig.4 the maximum prediction of national students of academic schools is 46605 which is achieved by LR-130 and highest accurate prediction count of immigrants is provided by C5-46. The lowest accurate prediction count of Immigrants is 407 is achieved by Tree-AS-45.

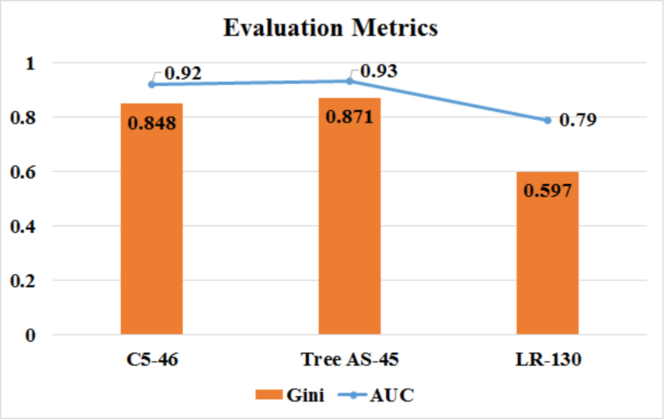

Fig. 5: Performance of classifiers for academic schools.

As we know AUC is the scalar representation of the expected performance of a classifier, Fig.5 reveals that there is no significant difference between AUCs of Tree-AS-45 and C5-46. It can also see that AUC of LR-130 is lowest than other two classifiers. Another alternative performance metric suggested by IBM is Gini Index and the formula to calculate Gini(G)=2*AUC-1. So, the maximum Gini index 0.871 is found of classifier Tree-AS-45 and the lowest is gained by the LR-30.

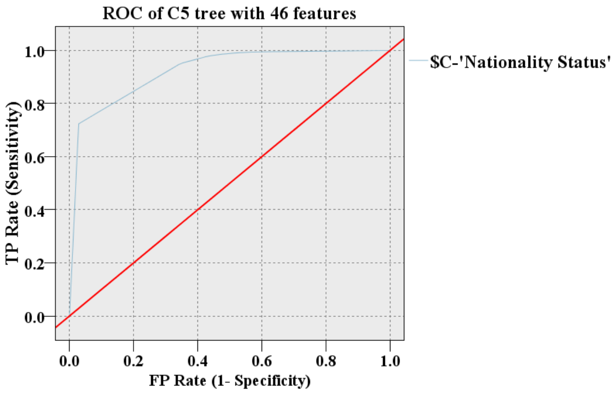

Fig. 6: Best Model Evaluation to predict nationality belongs to academic schools students.

Fig.6 shows decision tree model validation using ROC which reflects significant TP rate starts from 0.70 and ends to 0.99 with updating thresholds. Also, can be seen at thresholds 0.2 the sensitivity is high 0.83 and FP rate is 0.03 which reveals the significance of the predictive model. As thresholds reach 0.5, the model sensing at point 0.98 and FP rate is 0.2. Thus, decision tree C5 with 46 features outperformed the LR-130 and Tree-AS-45 to predict the national identity of students of academic schools towards ICT responses.

IV. Results and Analysis for Vocational Schools

This section explores the comparative results of prediction given by 3 classifiers to predict the national identity of students belongs to European vocational schools. We trained and tested the dataset2 with the total 36,228 instances with 146 features using 10-Fold cross-validation using of the auto classifier with LR145, Tree-AS-41, and LSVM-145.

As Fig.7 shows that there is no major significant difference between accuracy gained by Tree-AS-41 and LR-145 to predict the nationality of vocational schools students. It can be seen also after reduction of 104 features from dataset2, the accuracy of Tree-AS is improved by 0.4% which concludes feature reduction works fit. We also found significantly different between accuracy of LR-145 and LSVM-145 with the same number of features in prediction task. The maximum error of 6.6% is given by LSVM-145 and minimum error 5.6% is shown by Tree-AS-41 with feature reduction. Hence, Tree- AS outperformed the other two classifiers in the prediction of national identity towards ICT responses in vocational schools.

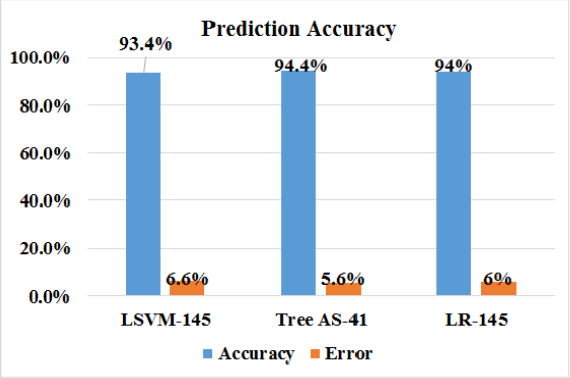

Fig. 7. Accuracy vs Error for vocational schools.

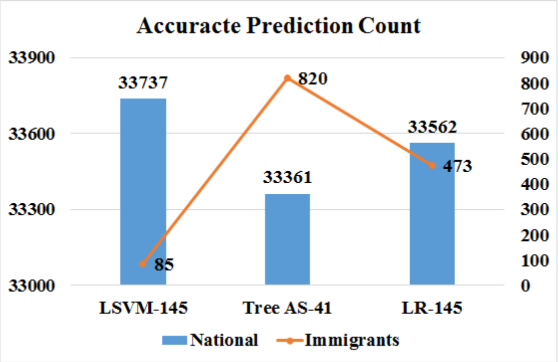

Fig. 8. Accurate Prediction Count of vocational schools students.

It can be seen from Fig. 8 the maximum prediction of an accurate count of the National student is 33737 attained by LSVM with 145 features. Unfortunately, it failed to predict enough number of immigrants. The highest immigrant students and 3rd highest national students are predicted by Tree- AS with 41 features. The second highest prediction count (National-33562, Immigrants-473) is achieved by LR with 145 features. As the balanced prediction is performed by Tree-AS with 41 features. Hence, It is concluded that the Tree-AS-41 has outperformed the LSVM-145 and LR-145 in the prediction of the national identity of vocational schools students. It is also revealed that LR-145 has also performed well as compare to the LSVM-145 in accuracy provided.

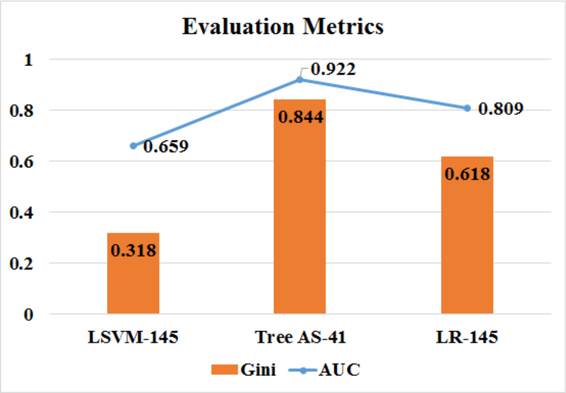

Fig. 9. Performance of classifiers for vocational schools.

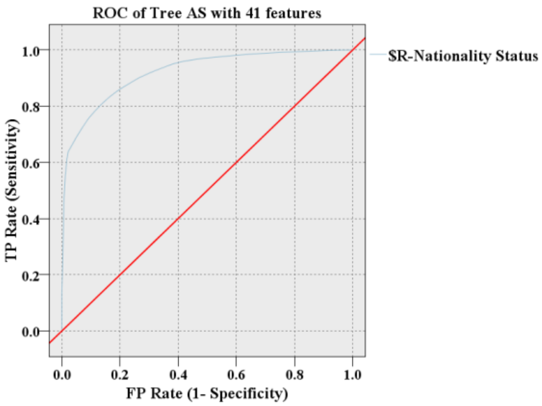

Fig. 10. Best Model Evaluation to predict nationality belongto vocational school's students.

Fig.9 shows the winner classifier Tree-AS-41 has the highest AUC (0.922). The LR with 145 features has also greater AUC (0.809) than LSVM-145. It is concluded that Tree-AS with 41 features outperformed the other two classifiers. The lowest Gini index 0.318 is achieved by LSVM-145 which proves that Linear support vector machine did not perform significantly in the prediction of the national identity of students of vocational schools. Fig.10 shows ROC curve reflecting the results of Tree-AS with extracted 41 significant features which depict significant TP rate starts from 0.62 and ends to 0.98 with dynamic cut-offs. The predictive Model also sensing high at 0.85 with thresholds 0.2 and FP rate is 0.15 which proves the significance of the predictive model. As thresholds reach 0.6, the model is sensing at point 0.98 and FP rate is 0.2. Therefore, Tree-AS with 41 features is playing best role in prediction of the national identity of vocational school's students towards their answers provided in the survey.

V. Conclusion

Supervised machine learning with feature extraction played a vital role to predict the national identity of European academic and vocational schools of 11th grade. The decision tree C5 attained a maximum accuracy of 96.6% to predict the national identity of students in vocational school dataset2. On the hand, feature reduction of auto classifier impacted on Tree- AS. Therefore, it outperformed with 94.3% accuracy the others (BN & LSVM) after feature reduction at vocational schools dataset2 in the prediction of national identity. It also got the second position to provide the highest accuracy of 95.8% in the prediction of the national identity of students in academic school's dataset1. Another hand, the C5 tree is also affected by feature extraction of the auto classifier of IBM Modeler. With only 46 features, the decision tree C5 outperformed the LR and Tree-AS with 96.6% accuracy in nationality prediction of students in academic school dataset1. Therefore, presented predictive models stabilized 46 features for academic schools dataset1 with 49263 instances and 41 features for vocational schools dataset2 with 36228 instances to predict the national identity of students. On the one hand, the maximum immigrants' prediction for academic schools (1077/2419) and for vocational schools (820/2397) are acquired by classifiers C5-46 and Tree-AS-41 respectively. On another hand, the highest nationals' prediction for the academic schools (46535/46844) and for the vocational schools (33737/33831) are acquired by classifiers C5-46 and LSVM-145 respectively. Future work, to develop the real time website by European union to the identify, to the monitor, to the control of the various ICT parameters which may be beneficial for both national and immigrant students as we discussed into the introduction section.

Acknowledgment

The first author thanks to Tempus Public foundation of Hungary to sponsoring his Ph.D study related to this research work. Also, the first and second author's projects are 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 Sanjay Dahiya, Deepak Mehta. An analytical approach to investigate state diversity towards ict: A study of six universities of punjab and haryana. Indian Journal of Science and Technology, 9:1–5, 2016.

[2] Chaman Verma. Educational data mining to examine mindset of educators towards ict knowledge. International Journal of Data Mining and Emerging Technologies, 7(1):53–60, 2017.

[3] Kamber M. Han J. Data Mining: Concepts and Techniques. 2006.

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

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

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

[7] 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.

[8] Chaman Verma, Veronika Stoffová, Zoltán Ill´es, 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.

[9] Chaman Verma Zoltán Illés, 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.

[10] Chaman Verma Ahmed S. Tarawneh Veronika Stoffová Zoltán Illés, 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. IEEE, 2018.

[11] Chaman Verma Ahmed S. Tarawneh Veronika Stoffová, Zoltán 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 In Press, 2018.

[12] Maria Teresa C. Maria Noel R. Prediction of university students' academic achievement by linear and logistic models. The Spanish Journal of Psychology, 2(1):275–288, 2015.

[13] Kolo D. et.al. A decision tree approach for predicting students' academic performance. International Journal of Education and Management Engineering, 5:12–19, 2015.

[14] 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.

[15] 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.

[16] Schaltz P. Klapproth F. The prediction of students track appropriateness in school. International Journal of Psychology (JPsych), 1(1).

[17] Pedro Domingos P. A few useful things to know about machine learning. Communications of the ACM, 55(10).

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

[19] Chaman Verma Veronika Stoffová, Zoltán Illés. Rate-monotonic vs early deadline first scheduling: A review. In Proceeding of Education Technology-Computer science in building better future, pages 188–193. University of Technology and Humanities, Poland, 2018.

[20] Chaman Verma Zoltán Illés, Veronika Stoffová. Attitude Prediction Towards Ict And Mobile Technology for The Real-Time: An Experimental Study Using Machine Learning. The 15th International Scientific Conference eLearning and Software for Education, In Press, Romania, 2019.

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

[22] Ahmad S. Tarawneh Dmitry Chetverikov Chaman Verma, Ahmad B. Hassanat. Stability and reduction of statistical features for image classification and retrieval:preliminary results. In 2018 9th International Conference on Information and Communication Systems (ICICS), pages 117–121. IEEE, Jordan, 2018.