Real-time classification of national and international students for ICT and mobile technology: an experimental study on Indian and Hungarian University

Chaman Verma\textsuperscript{1}, Zoltán Illés\textsuperscript{2}, and Veronika Sttofová\textsuperscript{3}

\textsuperscript{1,2}Eötvös Loránd University, Budapest, Hungary
\textsuperscript{3}Trnava University, Trnava, Slovakia

E-mail: chaman@inf.elte.hu

Abstract: This paper focused on the classification of student national-level status such as national and international. A primary survey was conducted in the academic year 2017-2018 to analyze the circumstances of trending ICT and Mobile Technology (MT) in Indian and Hungarian higher education. The main objective was to identify the student's answers provided in the survey based on their national-level status. For the classification tasks, we used Logistic regression (LR), Support Vector Machine, Multilayer perceptron (MLP) and Random Forest (RF) on both balanced and unbalanced datasets with K-fold Cross-Validation (KCV), Leave One Out (LOO), and Hold Out (HO) methods. Also, Xtreme Gradient Boosting (XGB) classifier was also implemented to enhance the classification accuracy of existing classifiers. The findings of the study showed that the XGB classifier outperformed others with the highest accuracy of 95% with 18 significant features. Also, class balancing improved significantly the accuracy of classification. Further, the authors recommended this predictive model to be implemented as a real-time function utility on the website of the university.

Keywords: Real-Time, National Status, Classification, Machine Learning

1. Introduction

The utilization of machine learning in educational data mining is trending and many of data scientists and pedagogical experts are involving eagerly to applying Artificial Intelligent (AI) techniques to explore new learning and teaching methods. Also, the AI techniques are using most frequently with machine learning algorithms as well. A supervised machine learning is an approach of AI in which algorithms always learn from the training dataset and classification is a type of supervised machine learning. In this paper, the authors applied supervised machine learning to classify the national status of a student towards the answers provided in the survey conducted. Traditionally, statistically, the analysis was not more than enough to analyze the data patterns and special for prediction tasks [1] [2] [3] [4] [5]. Recently, a few kinds of literature depict the predictive modeling of student's demography features as a preliminary study. To analyze the latest technological impact on students, the authors need to developed predictive models accordingly. Hence, the student's gender [6], [7], [8] and teacher's gender [9], principal's gender [10] was predicted for the real-time module development in university level. In addition of gender, the age-group [11], national identity [12], residence country [13], nationality status [14], study level [15] was also classified for the real-time system [16]. Also, the attitude of students [17], technological awareness level [18], availability and development of ICT and MT [19] was classified with the latest machine learning algorithms. Lately, the institution of the student was also classified with some feature filtering method to identify the student's answers in the survey [20]. To predict the locality feature, the rural-urban [21], national-international [22], [23], the
machine learning with ensemble methods played a significant role. This paper focused to improve the prediction accuracy of the national status of work done [23] using XGB classifier which is based on gradient boosted decision trees specially considered for speedup and performance of the classification data patterns. Usually, it fits a model on the gradient loss produced from the preceding step. Hence, modification in the traditional gradient boosting algorithm to fit with any differentiable loss function.

2. Research Techniques

This section holds four major subsections to explore the state of art of the research.

Dataset Preprocess

Table 1 shows the short description of each dataset Initial, S1 and S2 used in the present research. Initially, the authors have 331 instances and 37 features belong to four ICT and MT parameters with spotting 6 missing values in it. These 6 missing values are handled Replace Missing Value filter using substituting mean and mode values of the training dataset. Also, it can be seen from Table 2 that the initial unbalanced dataset is balanced using the help of the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm twice. Hence, after the first run of SMOTE, the authors named dataset S1 which increased instances of the minor class (International).

| Table 1. Initial Dataset. |
|---------------------------|
| Type | Instances | Attributes | Missing values |
| Initial | 331 | 37 | 06 |
| S1 | 410 | 37 | 00 |
| S2 | 568 | 37 | 00 |

| Table 2. Instances Distribution. |
|----------------------------------|
| Type | National | International |
| Initial | 252 | 79 |
| S1 | 252 | 158 |
| S2 | 252 | 316 |

To increase the samples of minor class, random data samples are selected within the dataset and in the feature space, k nearest neighbor points are assumed. Afterward, to produce the novel virtual data point selected the vector between one of the k nearest neighbors, and the existing data point. After that, the multiplication of the selected vector is done by a random number x in the range of 0 to 1. Now, the addition of this value in the existing data point may lead to the new, virtual data point which is called SMOTE point. Table 3 also reflects 4 major ICT and MT features of the survey which are taken into consideration.

| Table 3. Features. |
|--------------------|
| Development & Availability | Usability | Educational Benefit | Attitude |
| DA1 | UICTMT1 | EBICTMT1 | AICTMT1 |
| DA2 | UICTMT2 | EBICTMT2 | AICTMT2 |
| DA3 | UICTMT3 | EBICTMT3 | AICTMT3 |
| DA16 | UICTMT6 | EBICTMT9 | AICTMT6 |

Training and Testing

To the train and test the dataset separately, we used there testing methods KCV, LOO, and HO. Firstly, in KCV, the value of k is set as 10 which means each dataset is divided into k number of subsets in which k-1 is test and the rest of all are considered as train tests. Secondly, in the LOO method, the authors used k as equal to the number of balanced instances (k=n) to improve the classification accuracy as well. Lastly, the authors tested each dataset using with HO technique with 60:40 ratios.
This ratio signifies that datasets are divided into two parts: training data and testing data. From each dataset, 60% are considered as training data and 40% are considered as testing data.

**Algorithms and tool**

To classify the responses of students towards ICT and MT, the authors used four popular supervised machine learning algorithms such as binary LR, SVM, an artificial neural network with a multilayer perceptron, RF and Xtreme Gradient Boosting (XGB) tree. All the experiments are conducted in a standard machine learning tool. Further, SVM, RF, LR, and MLP algorithms are implemented in Weka 3.9.3 tool and the XGB algorithm is executed in IBM SPSS Modeler 18.2.

**Performance Measurement**

*To justify the results of binary classification we used the below important matrices.*

(a) Confusion matrix: It shows actual values versus predicted values.
(b) ROC: It is the receiver operating characteristic curve that holds TPRate Vs FPRate.
(c) AUC: It is an area under the curve value which signifies the ROC curve as well.

3. **Experimental Analysis**

This section has three subsections that discuss the results of each experiment performed.

**Experiment-1**

In the first experiment, the authors considered the KCV method in which the value of k is set as 10 for splitting datasets and later on each classifier are executed on each dataset (S1 and S2) with different instances. Figure 1 displays the accuracy provided by each classifier using with KCV testing method. It can be seen that balanced dataset S2 with 568 instances leads to the highest accuracy as compared to S1. Also, the RF classifier found as winner one from others using the KCV technique with k=10. The highest accuracy is provided by RF which is measured as 89.7% with S2-568 and 81.4% with S2-410. The LR is found as a weak classifier in the classification task of the national status of the student.

![Figure 1. Testing using KCV.](image)

**Experiment-2**

In the second experiment, the LOO method is applied which is a complete K-fold (n-fold) cross-validation method. One time the value of k is set as 410 and second-time k is set as 568 to test and train dataset deeply. Later, classifiers are modeled on S1 and S2 and the accuracy is checked out. Figure 2 shows the accuracy provided by each classifier using with LOO method. It is found that class balancing helps to improve classification accuracy. The uppermost accuracy is providing by RF which is measured as 89.1% with S2-568 and 81.1% with S2-410.
Experiment-3
In this experiment, the authors used the HO method with 60:40 training ratio. They trained and tested both datasets named S1 with 410 instances and S2 with 568 instances enhanced by SMOTE. Figure 3 shows the RF classifier outperformed LR, SVM, and MLP only. Also, the authors applied the XGB classifier to enhance the accuracy and extract the important features from the dataset as well. In this, they used the exact greedy method approach to build a boosting tree. Also, the total number of boost rounds is set to 200 which associates with 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. The max delta step parameter is set to 0.0 due to the balanced dataset S2. The XGB objective function is set as binary: logistic for the binary classification. Also, to avoid overfitting subsample value is set to 0.8 to elect 80% instances randomly. Fortuitously, we found that XGB performed superbly in the classification of the national status of the student towards ICT and MT. Hence, the XGB classifier outperformed others in classification accuracy.
Figure 4 displays 18 important features filtered out the whole features. The XGB classifier observed 07 features from the DA, 05 features from EBICTMT, 03 from UICTMT and 03 from AICTMT parameters. In these features, the maximum features of DA and EBICTMT played an energetic role. On one hand, the maximum prediction value of UICTMT1 is calculated as 65 and another hand worst prediction value of DA9 is found 2. Consequently, 18 features are proposed by XGB to classify the national status of the student.

Evaluation
To validate the result of the XGB experiment, the ROC curve is most appropriate to reflect the binary classification impact. Y-axis contains values of TP rate and X-axis having a value of FP rate at varying cut-offs. As Figure 5 shows that the curve is above the benchmark (red line) and the model is sensing at 0.6 and goes up to 0.98 highest. At a very low cut-off point 0.2, the TP rate is 0.98 which is found maximum and seems most significant in the classification task of the national-level status of the student.

Table 4 shows that the maximum prediction count ratio of 506/568 is obtained by RF with KCV on the S2-568 dataset. Also, the second winner classifier is found MLP with an accurate ratio of 498/568. The maximum misclassification 140 is achieved by LR which is insignificant for classification.

Table 5 displays joint confusion matrices of HO method classification with a balanced dataset S2-568. We found that the XGB classifier outperformed others incorrectly classification of instances which is 545 and it is uppermost up to now. Therefore, it is concluded that the XGB classifier predicted in an excellent way with significant feature filtering.
4. Conclusion

In this paper, the authors performed three classification experiments with the five most popular supervised machine learning algorithms. Three distinct testing methods were also implied to test the classification accuracy on a balanced and unbalanced dataset. The results of the first experiments conclude that RF classification accuracy (89.7% with S2-568 and 81.4% with S1-410) is higher than of others using with KCV method. The findings of the second experiment inferred that the uppermost accuracy is provided by RF with the LOO method which is calculated 89.1% with S2-568 and 81.1% with S2-410. The output of the third experiment also revealed that RF also outperformed others with 86.6% accuracy. Instead of it authors also found that XGB classifier defeated RF in classification accuracy by scored 95% accuracy on S2-568. Finally, it is concluded that class balancing improves the classification accuracy of classifiers and XGB classifier outperformed others in the classification of the national status of both country students. Further, 18 most significant features are suggested by the XGB classifier to predict accurately national and international class. Also, the author proposed this XGB model to be implemented as a real-time website to identify the answers of students during filling up a survey towards ICT and MT. Future work may lead to applying more feature filter methods to boost up the accuracy as discovered.

Acknowledgments

The first author thanks Tempus Public foundation of Hungary to sponsoring his Ph.D. study and 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, Sanjay Dahiya and 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] Chaman Verma and Sanjay Dahiya. Gender difference towards information and communication technology awareness in indian universities. SpringerPlus, 5(370):1–7, 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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
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.

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).

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.

Chaman Verma, Ahmad S. Tarawneh, Zoltán Illés, Veronika Stoffová, and Mandep 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.

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.

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).

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).

Chaman Verma, Veronika Stoffová and 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.

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.

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).

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).

Chaman Verma, Zoltán Illés, Veronika Stoffová, and Mandep Singh. ICT and Mobile Technology features predicting the university of Indian and Hungarian student for the real-time. In: IEEE System
[21] 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.

[22] 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, 2019, In Press.

[23] 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.