Automatic forecasting of student’s province towards Information and communication technology awareness

Chaman Verma¹, Deepak Kumar², Zoltán Illés³ and Veronika Stoffová⁴
¹² Eötvös Loránd University, Budapest, Hungary
³ Bharat Group of Colleges, Punjab, India
⁴ Trnava University, Trnava, Slovakia

Abstract: An experimental study is conducted to forecast the residence state of the students based on their response provided in the ICT survey held during the academic year 2015-2016 at two universities of India. The dataset consists of 560 instances and 59 features. We considered the state as the response variable and 35 features as predictors after self-reduction. The dataset is trained and tested with k-fold cross validation using three popular supervised machine learning classifiers named Artificial Neural Network (ANN), Sequential Minimal Optimization (SMO) and random forest (RF) in the Weka 3.8.1 workbench. The outcomes of the study reveal that RF classifier outperformed with highest accuracy (83.39%) the ANN and SMO at 6-Fold of Cross-Validation (CV). Finally, the authors presented state forecast models which accurately forecasting the state of the student based on their answers during the survey with stabilizing the value of k=6 of CV. The maximum accurate forecasting count for Punjab student is found 239 out of 282 and for Haryana student is found 228 out of 278 with k=6. The findings of study prove a significant difference between the RF and the ANN, the SMO in accuracy and there is no significant difference between accuracy gained by SMO and ANN for state forecasting against student’s response during the survey.

Keywords: Classification, Cross-validation, Supervised Machine learning, Forecast state.

1. INTRODUCTION

Nowadays, data mining using machine learning is trending in the educational field and the fundamental principle of data mining is to analyze the data from the different angles, categorize it and finally to summarize it [1]. Data mining is also known as KDD (knowledge discovery in databases), data pre-processing, pattern recognition, clustering [2], classification [3] are the popular technologies in data mining. There are many numbers of data mining tools are available in markets such as rapid miner, Orange, KNIME, Weka and many more. The Waikato Environment for Knowledge Analysis (Weka) tool is used by many of researchers for decision making in various fields [4], [5], [6]. Support Vector Machine (SVM) is a supervised learning model introduced for binary classification in both linear and nonlinear versions [7]. An SVM performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories [8], [9], [10]. The Sequential Minimal Optimization (SMO) is the extension of SVM to solve the problem of handling large datasets in SVM [11]. Random Forest (RF) is proposed by Breiman father of CART at the University of California, Berkley [12]. 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 [13]. 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 [14]. Artificial neural network (ANN) is simple mathematical models defining a function \( f: X \rightarrow Y \) or a distribution over \( X \) or both \( X \) and \( Y \), but sometimes models are also intimately associated with a learning algorithm or learning rule [15], [16]. A common use of the phrase ANN model really means the definition of a class of such functions. A multilayer perceptron in Weka is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output [17]. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions and is more powerful than the perceptron in that it can distinguish a not linearly separable data, or separable by ahyperplane. The ANN has achieved 80% accuracy as compare to SVM 70% in Credit Card Approval dataset [18]. Supervised machine learning classifiers were used on different educational datasets to predict demographical features of students [19-29]. These supervised machine learning classifiers are applied to the dataset using two ways. One way is splitting the dataset according to various training ratio and another way is portioning the dataset using CV. The K-Fold CV assess the performance of predictive models by dividing randomly the original dataset into two type of subsets such test (k) and train (k-1) [4]. The train set is used to train the model, and test set validates the models. In this paper, we used popular performance metrics which are evaluating the strength of forecasting models. The Cohen’s Kappa is a statistical measure of interrater reliability which determines the agreement among instances in the dataset. The calculation formula of kappa value is \( \text{Kappa} = (\text{Calculated accuracy}-\text{expected accuracy})/(1-\text{expected accuracy}) \). The sensitivity or true positive rate (Tpr) of the specific class in a model is calculated by dividing the total number correct predicted instances by a total number of actual instances related to specific class. The false positive rate (Fpr) or (1-specificity) is calculated by dividing the total number incorrectly predicted instances by the total number of actual instances of the specific class. The precision value is measured by dividing the total number of correct predicted instances/ total number of predictions. The specificity of the model is calculated by dividing the total number of incorrect predicted instances/ total number of actual predictions. As we know, the F-score is a harmonic mean of precision and recall which also states the significance of forecasting models is calculated by the formula \( F=2*(\text{Recall}*\text{Precision})/((\text{Recall}+\text{Precision}) \).

2. RESEARCH METHODS AND TECHNIQUES

The present research used the primary real dataset which is consisted of a total 59 features out of which 24 are dependent and 35 are independent. Out of 35, the 10 features belong to the availability of ICT resources, 8 features belong to the usability of ICT resources, 3 features belong to ICT problems, 5 features belong to solutions of problems and 09 features belongs to the ICT opportunities. The first authors have conducted survey to analyse the ICT scenario in the higher education system of India and collected 560 samples from six northern Indian universities. In this paper, we considered 35 features and one categorical variable (Residence State). The missing values are handled by Replace Missing Value filter and normalized using Normalize filter. The target of the response variable is converted to nominal using NumericToNominal filter [30]. Figure. 1 shows that almost equal ratio of student participants in this research study. Out of total 560, the 282 students belong to Punjab and 278 students belong to the Haryana state of India. All the students were studying in either a bachelor or master level of the study program.
In this paper, the Weka 3.8.1 tool is used, which is an open-source machine learning tool, implemented in JAVA and developed by The University of Waikato, New Zealand [31]. We used knowledge flow and explorer environments of Weka to test the dataset. Three supervised machine learning classifiers such as random forest (RF), sequential minimal optimization (SMO) and artificial neural network (ANN) are trained and tested on the dataset to forecast the residence state of student based on responses. RF is an ensemble learning approach based on multiple numbers of decision trees which mostly used in classification and regression problems. The basic procedure to build RF model with training dataset D is as follows. Suppose, the training dataset:

\[ D = \{(f_i, C_i) | i = 1, N \} \quad (1) \]

is given, where \( f_i \) is a feature, \( C_i \) is the set of classes and \( N \) denotes the number of training samples. Sample the training set \( D \) with replacement to create bagged samples \( D_1, D_2, ..., D_p \) and each decision tree are grown from these bagged sample set. In each decision tree, for every node, we consider a random and separate subset of forecasting features as candidate feature for splitting the node. The class forecasting of RF model with \( p \) number of trees can define as follows. Let us assume \( C_p \) be the forecasting of tree \( T_p \) given input \( f \).

\[ C = \{ \text{Majorityvoting} \{C_p\} \mid 1 \leq p \leq p \} \quad (2) \]

SMO is employed as a learning algorithm to train Support Vector Machine with a linear kernel. SMO is preferred because of its better scaling abilities for large and complicated SVM problems with less computational time than standard SVM [32]. We have also used most popular neural network architecture multilayer perceptron (MLP) which has many inputs and has one or more hidden layers with any number of units. It uses generally sigmoid activation functions in the hidden layers. MLP is especially suitable for approximating a classification function which sets the example determined by the vector attribute values into one or more classes. MLP trained with backpropagation algorithm is used for data mining [33]. An artificial neural network (ANN) can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining [34]. The experimental setup is divided into the following two major sections. The first section presents the holdout method of the test-train dataset using classifiers and second section explores the way to the test-train dataset using K-Fold CV method using classifiers.
3. EXPERIMENTS

Experiment-1

In percentage split (holdout method), the database is randomly split into two separate datasets whereas the first set from where classifier tries to extract knowledge and second set is used to test the extracted information. Usually, a random split occurs on data set under the mining task of test data [35]. Training ratio is initially significant to sampling the dataset for evaluation purpose. The portioning dataset is essential to test the accuracy of classifiers. Some portion of the dataset is needing to be trained with the help of testing portion of the dataset. During this test case, we performed classification using percentage split method on the dataset at various training ratio. For this, we divided the entire data set into two separate tests: one for training and other for testing. In this experiment, the considered training ratio for the train: test splitting is 20:80, 40:60, 60:40 and 80:20. Afterward, we applied three classifiers at various training ratio of the whole dataset to forecast the better accuracy. It can be seen from Figure 2 that the highest accuracy is achieved by the RF at maximum split 80:20 (88%) and minimum split 20:80 (73%) for the train test. The SMO is the second winner to gain accuracy 71% at the minimum split and 79% at the maximum split. The ANN classifier gave 79% accuracy at the maximum split.

![Classification with Splitting](image)

**Figure 2.** Classification using training ratio.

Experiment-2

CV is another validation approach to sampling the dataset in k number of subset among which one is used as (k-1) test set and rest of sets shall be (k) train sets. It is much meaningful to evaluate the classifier performance to perform during the classification process. During this experiment, we applied K-Fold CV method with varying k value. For this, we selected k value from 2, 4, 6 and 8 for the dataset and applied classifiers to examine classification accuracy. The following algorithm we used to the test and train dataset with various machine learning classifiers in Weka tool.

**Algorithm 1:** Test and train models using K-fold CV.

```
Step 1: Set k=6 (Splitting real dataset into the train set into 6 equal subsets or Fold named as f1; f2;...............; fk; 
Step 2: For i = 1 to i = k; 
Step 3: Let fi as a test set and retain all the remaining k-1 folds as a train set in CV; 
Step 4: Apply RF, SMO, and ANN to train 5-Fold (train-set) against 1-Fold (validation-set) and build a learning model 
```
and compute the accuracy of forecasting the models;

**Step 5:** Select the final forecast model with the highest Accuracy.

Data from Figure 3 presents the classification accuracy at various fold using cross-validation approach. Classification accuracy is increasing up to 6-fold and afterward it gets down. So, maximum accuracy (83%) is achieved by RF at 6-Fold and SMO and ANN both are at the same level to provide accuracy (78%) with 6-Fold.

![Classification using K-Fold](image)

**Figure 3.** Classification using K-Fold.

### 4. RESULTS AND DISCUSSIONS

Previously, based on the classification using K-Fold cross-validation, we presented state forecast models which accurately forecasting the state of students based on 35 features. The value of k=6 represents that dataset divided into subsamples in which one set of samples considered as test and remaining five samples set as train dataset. Figure 4 reflects the maximum count of accurate forecasting of students belongs to the Punjab state. The RF classifier at 6-Fold forecasts 239 Punjab student out of 282 and 228 Haryana students out of 278. The ANN classifier at 6-Fold forecasts 225 Punjab student and 211 Haryana students. The SMO classifier at 6-Fold forecasts less count of the student as compared to the ANN and the RF.

![Accuracy at 6 Fold](image)

**Figure 4.** Accurate forecasting count by models.
Data from Figure 5 shows that minimum forecasting error (17%) is achieved by RF classifier at 6-Fold as compare to other Folds. The second winner classifiers are both SMO (22%) and ANN (22%) due to achieving almost the same percentage of less misclassification error. The maximum forecasting error (24%) is given by both ANN at 8-Fold and SMO at 2-Fold.

![Figure 5. Forecasting error by models using K-Fold.](image)

Figure 6 shows the forecasting count of state wise student as provided by forecasting models. RF classifier has the minimum forecasting count 43/282 for Punjab state student and 50/278 for Haryana state student. One hand, ANN classifier has maximum forecasting count 67/278 for Haryana state student, on another hand, SMO has maximum forecasting count 59/282 for Punjab state student. Hence, to evaluate the performance with respect to less error, RF outperformed the other classifiers at 6-Fold CV.

![Figure 6. Inaccurate forecasting count by Models.](image)
Figure 7. ROC curves for Punjab Student with 6-Fold CV.

In Figure 7 The receiver operating characteristic (ROC) curve is showing the comparison of RF, SMO and ANN at 6-Fold CV for Punjab student and Figure 8 presents the ROC for Haryana student. We know that ROC curve plots sensitivity or Tpr across y-axis and Fpr or 1-specificity across x-axis with various thresholds (cut-offs). For Punjab student, at 0.5 thresholds, the RF gained maximum accuracy nearby 82% and SMO and ANN have achieved accuracy less than 80%.

Figure 8. ROC curves for Haryana Student with 6-Fold CV.

In the case of Haryana student, RF gives accuracy around 81%, SMO and ANN provide less than 79% accuracy. Therefore, thresholds are directly proportional to the accuracy and at more than 50% accuracy is gained by each almost every classifier at 0.2 thresholds.
Figure 9. Overall Forecasting Count.

Table 1. Performance measures of classifiers using 6-Fold.

| Classifier | RF | SMO | ANN |
|------------|----|-----|-----|
| Class      |    |     |     |
| Punjab     |    |     |     |
| student    |    |     |     |
| Haryana    |    |     |     |
| student    |    |     |     |
| Kappa static | 0.67 | 0.55 | 0.56 |
| RMSE       | 0.37 | 0.47 | 0.43 |
| Sensitivity (TPR) | 0.85 | 0.82 | 0.79 | 0.76 | 0.80 | 0.76 |
| 1-Specificity (FPR) | 0.18 | 0.15 | 0.24 | 0.21 | 0.24 | 0.20 |
| Recall     | 0.85 | 0.84 | 0.80 | 0.76 | 0.80 | 0.76 |
| Precision  | 0.83 | 0.84 | 0.77 | 0.78 | 0.77 | 0.79 |
| F-Measure  | 0.67 | 0.67 | 0.78 | 0.77 | 0.78 | 0.77 |
| ROC area   | 0.91 | 0.71 | 0.84 |

Table 1 presents the various types of performance measures of forecasting models at 6-Fold. The RF, SMO, and ANN classifiers have good Cohens Kappa static values such as 0.67, 0.55 and 0.56 respectively which signifies the forecasting power of models. The minimum RMSE is 0.37 given by RF as compare to SMO (0.47) and ANN (0.43) for forecasting the residence state of the student. The highest sensitivities of RF classifier for Punjab student and Haryana student are calculated 0.85 and 0.82 respectively and the minimum Fprs are calculated 0.18 and 0.15 respectively which signifies the strength of the forecasting model. The strongest recall value for RF is 0.85 for Punjab student and 0.84 for Haryana student signifies the model performance too. Both classifiers SMO and ANN have equal recall value which is 0.80 for Punjab student and 0.76 for Haryana student also good for these classifiers. The highest precisions values are obtained by RF (0.83) for Punjab student and (0.84) for Haryana student, and the ANN has equal precisions values 0.77 for Punjab students which proves the significance of forecasting models. Likewise, precision values calculated by RF, SMO, and ANN for Haryana student are 0.84, 0.78, 0.79 respectively also signifies the power of forecasting model. More than 65% F-measure also proves that each classifier has performed well in forecasting of state of students based on their responses provided during the survey.
5. CONCLUSION

The present experimental study is conducted to forecast the residence state of university students based on their answers provided during the ICT survey held during the academic year 2015 in two states of India. To present the forecast model, author have trained and tested the dataset using holdout and CV methods to model three popular classifiers SMO, RF, and ANN. The authors have concluded that RF classifier outperformed the others at 6-Fold CV. The maximum forecast count is 239/282 for Punjab student and 228/278 for Haryana student attain by RF with accuracy (83.39%). If k=8, the accuracy gets down by each classifier. Hence, the authors have stabilized the value of k=6. Further, maximum forecasting error (24%) is given by both ANN at 8-Fold and SMO at 2-Fold. It is also revealed that ANN classifier at 6-Fold forecasts 225 Punjab Student and 211 Haryana Student out of total 560. The minimum forecasting is performed by SMO classifier as compare to ANN and RF. The maximum area under the curve is 0.89 for Punjab student and 0.91 for Haryana student in ROC curve also proves the significant forecasting for both of class in the dataset. The author also found a significant difference between the accuracy gained by random forest (RF) and artificial neural network (ANN), Sequential Minimal Optimization (SMO) towards forecasting of residence state. It is also concluded that there is no significant difference between accuracy gained by SMO and ANN for state forecasting against student’s responses.

ACKNOWLEDGMENTS

The first author’s work 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] Swati Singhal, Monika Jena. A study on weka tool for data preprocessing, classification and clustering. In: International Journal of Innovative technology and exploring engineering, 2(6): 250–253, 2013.
[2] Ritu Chauhan, Harleen Kaur, M. Afshar Alam. Data clustering method for discovering clusters in spatial cancer databases. International Journal of Computer Applications, 10 (6.9), 14, 2010.
[3] Salzberg S.L. . C4.5: Programs for machine learning by j. ross quinlan. Machine Learning, morgan kaufmann publishers, 16 (3), 235–240, 1994.
[4] Oduandike Georgina N, Audu Isah, John Alhasan. Analytical study of some selected classification algorithms in weka using real crime data. International Journal of Advanced Research in Artificial Intelligence, 4(12), 44–48, 2015.
[5] Kumar Mandal S. Performance analysis of data mining algorithms for breast cancer cell detection using naïve bayes, logistic regression and decision tree. International Journal of Engineering and Computer Science, 6(2), 20388–20391, 2017.
[6] Bharat Deshmukh, Ajay S. Patil, B. V. Pawar. Comparison of classification algorithms using weka on various datasets. International Journal of Computer Science and Information Technology, 4(2), 85–90, 2011.
[7] C.W. Wei Hsu, C.C. Chang and C.J. Lin. A practical guide to support vector classification, 2003.
[8] Ali.S. Automated Support Vector Learning Algorithms. PhD dissertation, Monash University, Australia, 2005.
[9] Christopher J.C Burges. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121–167, 1998.
[10] Ruey Long Cheu, D. Srinivasan, Eng Tian Teh. Support vector machine models for freeway incident detection. International Conference on Intelligent Transportation Systems, 238–243.
IEEE, 2003.

[11] Jingmin Wang and Kanzhang Wu. Study of the smo algorithm applied in power system load forecasting. In Pacific Rim International Conference on Artificial Intelligence, 1022–1026, Springer, 2006.

[12] Leo Breiman. Random forests. Machine learning, 45(1), 5–32, 2001.

[13] G.Priya and N.Venkatesan. A study of random forest algorithm with implementation using weka. International Journal of Innovative Research in Computer Science and Engineering, 1(6), 2394–6364, 2015.

[14] S.A. Ghogare, Dr.S.R.Kalmegh. Comparative analysis of weka data mining algorithm Randomforest, Randomree and LADtree for classification of indigenous news data. International Journal of Emerging Technology and Advanced Engineering, 5(1), 507–517, 2015.

[15] Minsky M, and Seymour A Papert. Perceptrons: An introduction to computational geometry. MIT press, 2017.

[16] Ethem Alpaydin. Introduction to machine learning, MIT press, 2009.

[17] Chandra R, Chaudhary K and Kumar A. The combination and comparison of neural networks with decision trees for wine classification. School of sciences and technology, University of Fiji, 2007.

[18] Kumar T.D. A comparative study of classification algorithms for credit card approval using weka. GALAXY International Interdisciplinary Research Journal, GIIRJ, 2(3), 165–174, 2014.

[19] Chaman Verma, Zoltán Illés, Veronika Stoffová. Real-time classification of national and international students for ICT and mobile technology: an experimental study on Indian and Hungarian University, Journal of Physics: Conference Series, 1432 012091, 1-8, 2019.

[20] Chaman Verma, Ahmad 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, 2018.

[21] 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.

[22] Chaman Verma, Zoltán Illés, 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(983):1260–1264, 2019.

[23] 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.

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

[25] Chaman Verma, Zoltán Illés, Veronika Stoffová. 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, 2019.

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

[27] Chaman Verma, Zoltán Illés. Attitude Prediction Towards ICT and Mobile Technology for The Real-Time: An Experimental Study Using Machine Learning. The 15th International
Chaman Verma, Veronika Stoffová, 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, Veronika Stoffová. Real-Time Prediction of Development and Availability of ICT and Mobile Technology in Indian and Hungarian University. Proceedings of ICRIC-2019, Lecture Notes in Electrical Engineering (LNEE), 605–615, Springer, 2020.

The University of Waikato, https://www.cs.waikato.ac.nz/ml/weka, note = Accessed: 2020-03-09.

Wikipedia, https://en.wikipedia.org/wiki/weka/machinelearning, note = Accessed: 2020-03-09.

BiswaRanjan Samal, Anil Kumar Behera, Mrutyunjaya Panda. Performance analysis of supervised machine learning techniques for sentiment analysis. In IEEE Third International Conference Sensing, Signal Processing and Security (ICSSS), 128–133, 2017.

Suman Sebastian. Performance evaluation by artificial neural network using weka. International Research Journal of Engineering and Technology (IRJET), 3(3), 2395–0056, 2016.

S. M. Kamruzzaman, A. M. Jehad Sarkar. A new data mining scheme using artificial neural networks. Sensors, 11(5), 4622–4647, 2011.

Trilok Chand Sharma, Manoj Jain. Weka approach for comparative study of classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(4), 1925–1931, 2013.