International Journal of Engineering & Technology, 7 (4) (2018) 3392-3396

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.14045
Research paper

An Ensemble approach to identifying the student gender towards information and communication technology awareness in European schools using machine learning

Chaman Verma [1] *, Veronika Stoffová [2], Zoltán Illés [3],

[1,3] Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
[2] Department of Mathematics and Computer Science, Faculty of Education, Trnava University in Trnava, Trnava, Slovakia
**Corresponding author E-mail: chaman@inf.elte.hu*

Abstract

Data mining and machine learning play an important role in both research estimation and learning. The present study is conducted to identify the gender of student according to their answers given in survey related to information and communication technology (ICT) in European schools. The student dataset which consists of a total number of 156 attributes and 50478 instances are tested to identify student gender. To develop the ensemble predictive model after comparing prediction accuracy achieved by various supervised machine learning classifiers such as Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Artificial Neural network (ANN) and J48 tree with various k-fold cross-validation. The K-nearest neighbor (IbK or KNN) is also trained with data-set with varying value of k at 8-fold cross-validation. The dichotomous variable is gender and 131 predictors belong to ICT in education are taken into consideration after applying feature reduction methods. Findings of the study reveal that the maximum prediction is gained by SVM (76%) at each fold as compared to others. The total number (23535) of correct females are identified by RF at 6-fold and correct perdition of males is 14678 which is achieved by SVM at 2-fold. The authors also found lowest accuracy for prediction is achieved by NB classifier at each fold. Finally, the ensemble predictive model is presented by joining the best classifier such as SVM at 2-fold, ANN at 2-fold and RF at 6-fold to accurate identification of student gender over data-set. The ensemble confusion matrix also concludes the maximum prediction of the female student as compared to male student towards their response given to survey.

Keywords: *Binary Classification; Confusion Matrix; FPR; TPR.*

1. Introduction

Educational Data mining has emerged as the very important area of research to reveal presentable and applicable knowledge from large educational data repositories. Data mining algorithms are used to obtain the hidden information and desired benefits from these large data repositories [6]. Recently, analysis of educational data, for instance, learning analytics, academic analytics, educational data mining, predictive analytics and learners' analytics has emerged as an innovative area of research [7]. Machine learning (ML) is the process of estimating unknown dependencies or structures in a system using a limited number of observations and it is used in data mining applications to retrieve hidden information and used in decision-making [1]. The ML methods are rote learning, learning by analogy, and inductive learning, which includes methods of learning by examples and learning by experimentation and discovery [12]. According to [11] for classification, and regression problem various classifiers can be used for learning decision trees, rules, Bayes networks, artificial neural networks and support vector machines and different knowledge representation models can be used to support decision-making methods. Multiple, ensemble learning models have been theoretically and empirically shown to provide significantly better performance than single weak learners, especially while dealing with high dimensional, complex regression and classification problems [12]. Below 50% classification accuracy was obtained by OneR, J48 and Naïve Bayes (42.9%) technique to classify student age classification against ICT attitude [9]. Artificial Neural Networks (ANN) has a large generalization capability, and can approximate functions used for both regression and classification [3] and according to [13,5] SVM has discriminatory methods that learn boundaries between classes and performing a binary classification based on the separation of hyperplanes; a separator is chosen to maximize the distances of these hyperplanes and the nearest formation vectors, which are called support vectors. According to [14]. In KNN each sample data is assigned to the majority class of its k closest neighbors where k is a parameter. The training data samples are vectors in a multidimensional feature space, each with a given target class label. Logistic regression was applied to develop the model for the early and reliable prediction of students pass or fail status at the undergraduate level [8]. The key demographic variables and assignment marks in the supervised machine learning algorithms (decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines) to predict student's performance at the Hellenic Open University [10]. The gender is one of the principal determinants of the probability of dropping out. In the binomial probit model they used, males have a higher probability of dropping out relative to the reference group of females [2]. In addition, experi-

ence with computers and the World Wide Web were dichotomous variables; logistic regression was used for those variables [15]. The demographical features such as gender of teacher and residence state was also predicted [19],[20]. The student performance was predicted by two variable absenteeism and misconduct in using Binary Logistic Regression in secondary schools of Tanzania [16]. Fig. 1 presents the pictorial view of conducted research.

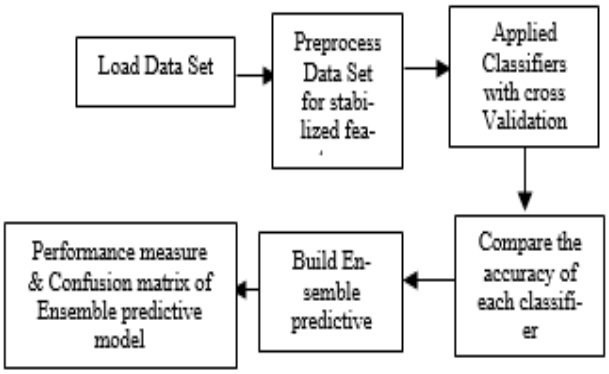

Fig. 1: Conceptual View of Research Design.

2. Method and techniques

This experimental study is performed in Weka workbench using various machine learning classifiers on the student dataset which consists of a total of 156 attributes and 50478 instances. The dataset is available on the website of European Commission [18]. More than 2500 schools, 27 countries, and more than 50000 students have participated in the survey in the academic year 2011 and survey was conducted by European Schoolnet on primary schools those were studying at ISCED level 3 (upper secondary level of education). The survey questionnaire was divided into 3 parts for school head, teachers, and students. We focused on student and data-set belongs to 3rd part is considered. The scale of measurement was mixed such as Likert type, nominal, ordinal and interval. Before tested the data, pre-processing is performed on the data set. To preprocess data and dealing with missing values in data-set unsupervised attribute filter named Replace Missing Value is used. This filter replaced all missing values for nominal and numeric attributes in data-set with modes and mean from training data. The total 25 attributes which are related to identification such as country name, school id, student id and so on and which are not significant to our results. Hence, after data reduction (self) without losing any critical information, the total 131 attributes are taken into consideration. Out of 131, we normalized 130 attributes using an unsupervised Normalized filter and applied Numeric to Nominal filter on Gender which is set as the class attribute (Response). The Gender has three types in data-set; 1 for male, 2 for female and 3 for Missing. After removal instances belong to the missing category, only 49429 were retained. Out of these, 27963 were female and 21466 were the male student. The authors have converted original dataset available in comma separated value (CSV) format to Attribute-Relation File to (ARF).

To predict the gender, various methods for learning Implicit Knowledge [1] five popular supervised machine learning classifiers Support Vector Machine (SVM), Naïve Bayes (NB), Artificial Neural network (ANN), J48 tree and K-nearest neighbor (IbK or KNN) are used with various cross-validation. Further, we used Random Forest (RF) of learning and combining redundant classifiers or ensembles are one approach for increasing prediction accuracy models on unseen examples, which is the most important generalization property [2]. SVM is strong classifier which is used in the mapping of learning examples from input space to a new high dimensional, potentially infinite dimensional feature space in which examples are linearly separable [4]. Except for KNN, five classifiers are tested at the interval of [2] such as 2,4,6,8 and 10. Similarly, KNN is tested with the interval of [2] with different k at 8-fold cross-validation. These algorithms are implemented based on Waikato Environment for Knowledge Analysis (WEKA) is an open source software provided by University of Waikato, New Zealand. This tool has a rich library of various machine learning algorithms allows mining, trending and modeling the data [17]. After successful modeling data-set by all classifiers, the ensemble predictive model is presented to predict most accurately student gender based on survey answers. The results of individual base classifiers are compared with the majority vote classifier and it is determined through experiments that the new approach achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifiers. Later, the performance of ensemble model is evaluated by the following Table 1 important parameters for gender prediction:

Table 1: Research Parameters for Evaluation

S.No.	Performances metrices for Model	Performance metrices for Gender class
1.	Classification Accuracy (CA)	True positive rate (TPR)/ Sensitivity.
2	Classification Error (CE)	False Positive rate (FPR)/ 1-Specificity
3.	Correctly classified Instances (CCI)	Precision (P)
4.	Incorrect classified Instances (ICI)	Recall (R)
5.	Kappa Static (KS)	F-Measure (F)
6.	Mean Absolute error (MBE)	Prevalence (PP)

3. Experimental results and discussion

The authors presented stabilized features for prediction of student gender based on their answers (numeric). The present study is classifying the student gender of European schools towards ICT in school education and presented a significant ensemble predictive gender classification model for accurate prediction of student gender based on their provided response during the survey. The experimental part of the study is divided into the 2 sections. Section 3.1 described prediction of gender using five classifiers with k-fold cross-validation with the interval of two. Subsequently, it presents the misclassification error achieved by them also. It is also elaborating how is KNN classifier predicting the gender with different K-value with 8-Fold cross-validation using. After comparing the accuracy achieved by all classifiers section 3.2 presented maximum prediction description about gender and section 3.3 presented the ensemble predictive model for better prediction of student gender based on their answers in the survey. In addition, it also presents the significant measures with confusion matrices which played the vital role in the prediction of gender.

3.1. Predictive modelling using classifiers

To predict gender, we applied different classifiers on student dataset of European Commission. This experiment is conducted by training five classifiers using various at various level of k-fold cross-validation techniques on the dataset. We used basically two test cases in order predict the gender of the student.

3.1.1. Test case 1

K-fold cross-validation always divides then the number of samples into k mutually exclusive subsets of equal size. During this test case, we performed the evaluation of classification algorithms by test split at various k-fold cross-validation. The validation is achieved by applying [2] to 10-fold cross-validation to validate the models. The data set is tested and trained using by dividing into n (2, 4, 6, 8, 10) equal sized folds. Further, the evaluation process repeats according to n (2,4,6,8,10) times, each time a single different fold will act as a test set (holdout set) while remaining 9 folds are combined and used for training to improve classification accuracy. During this test case, we trained dataset by applying five supervised machine learning algorithms. Fig. 2. It can be seen the

maximum accuracy is 75% and lowest accuracy is 68% achieved from machine learning classifiers.

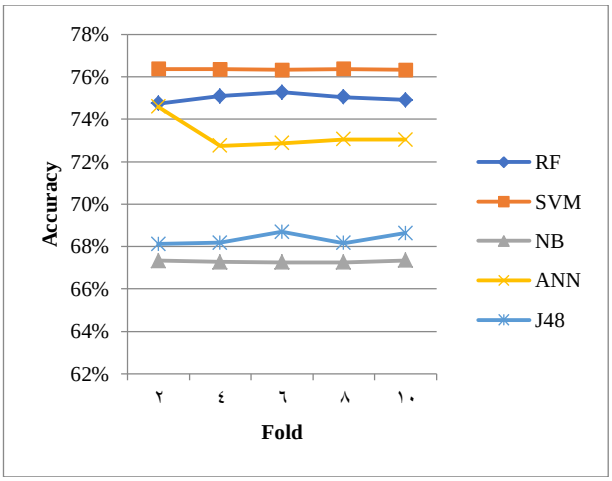

Fig. 2: Classifier Accuracy at Various Folds.

It can be seen from Fig. 2 that the maximum accuracy 76% is gained by SVM at each fold. The second highest accuracy 75% is achieved by RF at each fold and by the ANN 75% at 2-fold. At the rest of folds, the ANN accuracy is 73% measured. The classifier J48 has achieved 69% accuracy at 6 and 10 folds cross-validation and NB performed also well with 67% accuracy at various folds.

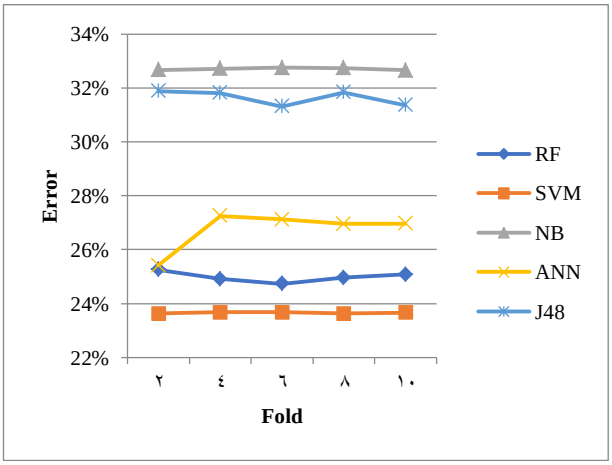

Fig. 3: Misclassification by Classifiers at Various Folds.

Data from Fig. 3 reveals that the maximum error is calculated 33%. As previously discussed, SVM played the better role in the prediction of gender so it has 24% misclassification error at each fold. The misclassification error given by RF and NB is 25% and 33% at each fold. ANN gave 25% error at 2-fold and 27% error at remaining folds. J48 classifier given 32% error at 2,4,8 and 31% error at 6 and 8-fold. It can be concluded that maximum error is given by NB and minimum error is obtained by SVM.

3.1.2. Test case 2

In this, the KNN classifier is also applied on data-set and tried to classify the target variable gender based on predictors. The objective is to learn a function (f):X|→|Y in which predictors f(x):X can confidently be predicting the corresponding target Y which is gender. The Euclidean distance (d) between unseen observation x is calculated by the equation:

$$D(X, Y) = \text{sqrt}((X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_n - Y_n)^2) \quad (1)$$

Next equation of KNN Algorithm for prediction of gender is defined, the K-value is varying from 1 to 9. The testing of KNN with varying value of k at 8-fold cross-validation.

$$\sum_{i=1}^{k=1\ to\ 9}(X^n - Y^n) \quad (2)$$

The maximum accurate prediction is achieved at K=9 and Fold=8. It can be seen from the Figure 4 accuracy directly proportional to the k value.

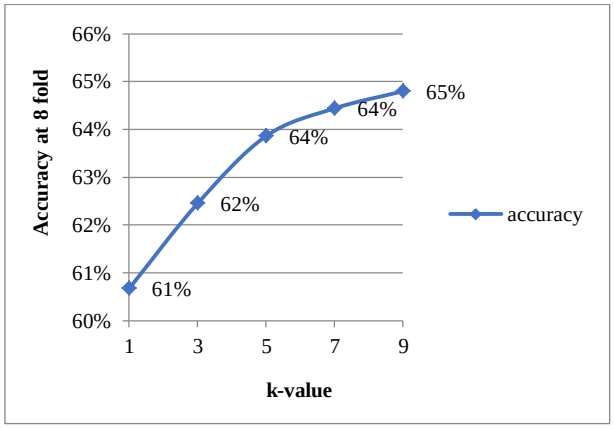

Fig. 4: Accurate Gender Prediction Using KNN at 8-Fold.

Data from Fig. 4 shows KNN accuracy at 8-fold is above 60%. It is also revealed that the accuracy is increasing (accuracy|∞|k) according to growing value of k. Almost same number of instances are correctly classified by KNN at k|=|5 and k|=|7.

3.2. Prediction count

Previously, after comparing the accuracy of all classifiers at various k-fold levels, Figure 5 presents the maximum number of accurate prediction of student gender. It will lead to deciding to select the best classifiers for ensemble predictive model. At accuracy 76%, SVM predicted 23073 females and 14678 males (total=37751) with 2-fold cross-validation. Secondly, ANN classifier at 2-fold provides 75% accuracy with maximum correctly predication of female as compared to male (total=36867).

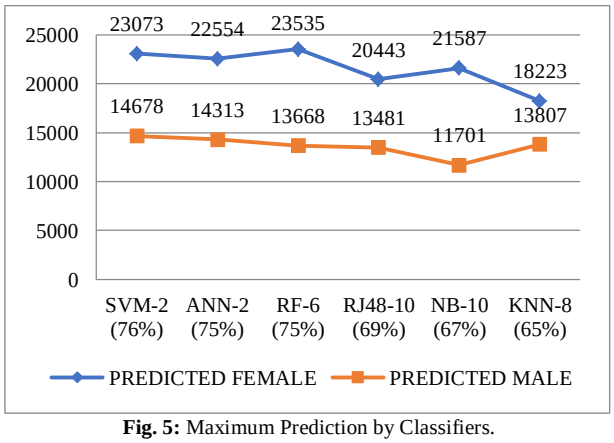

Fig. 5: Maximum Prediction by Classifiers.

RF classifier succeeds better than all to predict 23535 females with the accuracy of 75% at 6-fold. Subsequently, with 10-fold cross validation, NB and RJ-48 classifier achieved 67% and 69% accuracy respectively. The minimum accuracy (65%) level of gender prediction is gained by KNN (k=9, fold=8). It is concluded that winner classifier is SVM at 2-fold and second winners are ANN at 2-fold and RF at 6-fold cross-validation. Therefore, the maximum prediction of female student is achieved by RF at 6-fold and maximum prediction of the male student is provided by SVM at 2-fold. This may lead to present ensemble model for gender prediction which is discussed in the subsequent section.

3.3. Ensemble predictive model

Multiple, ensemble learning models have been theoretically and empirically shown to provide significantly better performance than single weak learners, especially while dealing with high dimensional, complex regression and classification problems (R. Caves, 1982). Previously, as we see every classifier has classified instances accurately over data-set and achieved more than 60% accuracy. To predict student gender with better accuracy, the authors presented ensembled predictive model. This model is the collection of three supervised machine learning predictive models. Hence, authors called ensembled predictive model which proved best in predicting of student gender over data-set.

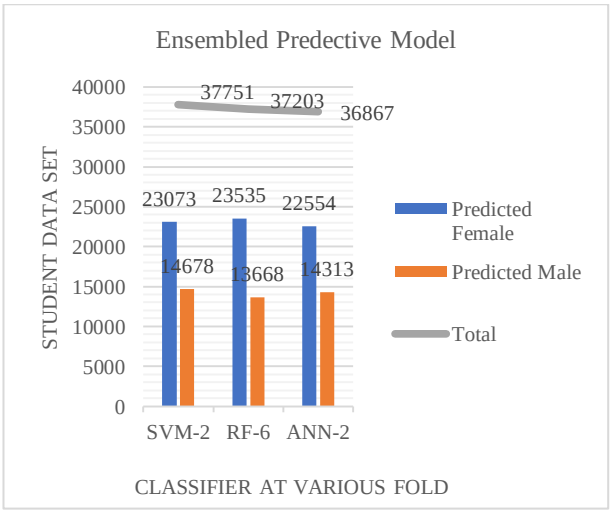

Fig. 6: Ensemble Model Prediction.

Fig. 6 presented the accurate prediction count of student gender based on their responses towards ICT. It can see that ensembled model collectively providing more than 70% accuracy for gender prediction. The maximum count of total predictive gender for SVM-2, RF-6, and ANN-2 is 37751, 37203 and 36867 respectively. The maximum right predicted females if given by RF-6 and right predicted males by SVM-2. The ANN classifier also performed well for prediction as it predicted total 36867 of which 22554 are female and 14313 are males. The performance of ensemble predictive model is shown in Table 2. The accuracy is percentage number of appropriately classified instances from overall data-set. Highest classification accuracy is achieved by SVM at 2-fold and ANN and RF has same at the different fold and minimum CE.

Table 2: Ensemble Model Performance Metrics

	CA (%)	CE (%)	CCI	ICI	KS	MBE
SVM-2	76	24	37751	11678	0.51	0.24
RF-6	75	25	37203	12226	0.49	0.39
ANN-2	75	25	36867	12562	0.48	0.26

Similarly, highest prediction of gender in terms of correctly classified instances (CCI) is provided by SVM-2 and second runner-up is RF-6 and third position is achieved by ANN-2. The lowest incorrect classification of instances (ICI) is given by SVM at 2-fold. The Cohen's Kappa statistic is calculated 0.51, 0.49 and 0.48 for SVM-2, RF-6 and ANN-2 respectively which signifies the presented ensemble predictive model of identification of gender. The mean absolute error for SVM-2 and ANN-2 is lower than RF-6.

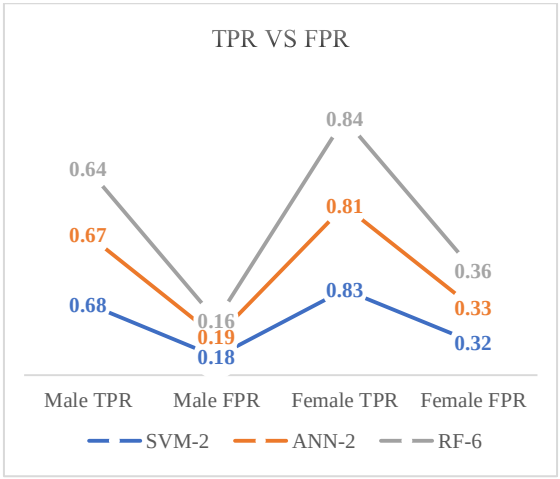

Fig. 7: TPR vs FPR.

Fig. 7 shows gender class wise performance metrics for the model evaluation. One hand, the sensitivity or true positive rate (TPR) for females is calculated 0.84, 0.83 and 0.81 by RF-6, SVM-2 and ANN-2 respectively and another hand, TPR for males is calculated 0.68, 0.67 and 0.64 by SVM-2, ANN-2 and RF-6 respectively. Therefore, the sensitivity of presented model is more significant for prediction of gender. Further, the lowest false positive ratio (FPR) is calculated 0.32 for females by SVM-2 and 0.16 for males by RF-6. Therefore, identification of females is more than the male student.

SVM at 2-fold

Predicted	Gender	Actual	
		Male (1)	Female (2)
	Male (1)	14678	6788
	Female (2)	4890	23073

RF at 6-fold

Predicted	Gender	Actual	
		Male (1)	Female (2)
	Male (1)	13668	7798
	Female (2)	4428	23535

ANN at 2-fold

Predicted	Gender	Actual	
		Male (1)	Female (2)
	Male (1)	14313	7153
	Female (2)	5409	22554

Fig. 8: Ensemble Confusion Matrices.

Data from Fig. 8 shows three confusion matrices for ensemble predictive model. The precision value of each classifier in the model is calculated by dividing the total number correct predicted instances by the total number of the actual instances for that class. The Recall is a measure of classifier completeness, so it is calculated by dividing the total number of correct predicted instances by the overall total number of instances. The F-Measure (F) conveys the balance between the precision and the recall. In SVM-2, the calculated F- value for the female is 0.798 and for the male is 0.715 which significant proved the accurate prediction of gender. The Precision value is also calculated 0.773 for female and 0.750 for male also reflects exactness of classification. In RF-6, the female has highest F-value (0.794), precision (0.751) and recall (0.842) and male also has F-value (0.691), precision (0.755) and recall (0.637). In case ANN-2, the female has highest F-value (0.782), precision (0.759) and recall (0.807) and male also has F-value (0.695), precision (0.726) and recall (0.667). Further, the prevalence (PP) of the model is calculated by dividing the total no.

of the actual female by total no. of students. In case of SVM-2, the PP value is calculated is 0.604; and for RF6, the PP value is calculated 0.63; for ANN-2, the PP value is calculated 0.601. These significant prevalence values also prove the occurrence of the female in data-set. Therefore, ensemble predictive model is proved significant to accurately identify the gender of the student from large data-set. Hence, the positive predictive value of presented gender classification model is significant due to more than 50% prevalence. The Cohen's Kappa statistic 0.51 also proved the significance of model accuracy to correctly identify the gender of the student.

4. Conclusion

This experimental study is conducted to predict the gender of the student based on their answers given in survey to analysis the ICT awareness in European schools. An ensemble predictive model is introduced after testing various supervised machine learning algorithm on data-set. It is concluded that the maximum accuracy is achieved 76% by SVM at each fold and 75% is achieved by RF at each fold and by the ANN 75% at 2-fold. The SVM classifier played the better role in the prediction of gender due to less error 24% as compared to others. Further, KNN classifier achieved 60% accuracy at 8-fold with k=9. Therefore, accuracy is found directly proportional to k value and almost same number of gender prediction is achieved by KNN at 8-fold with k|=|5 and k|=|7. On one hand, the maximum number of corrected prediction of females is 23535 given by RF at 6-fold cross-validation. Another hand, the total number of correct perdition of males is 14678 which is achieved by SVM at 2-fold cross-validation. Further, lowest accuracy for perdition is provided by NB classifier. Hence, ensemble predictive model is constituted by combining the first winner classifier is SVM at 2-fold and second winners ANN at 2-fold and RF at 6-fold. Further, at one side, the model sensitivity for female prediction is 0.84 by RF at 6-fold, 0.83 by SVM at 2-fold and 0.81 by ANN at 2-fold also proved the largest correct prediction of female students. At another side, model sensitivity for males is calculated 0.68, 0.67 and 0.64 by SVM-2, ANN-2 and RF-6 respectively. It is proved that the sensitivity of presented model is more significant for prediction of gender. The positive predictive value of classification model is significant due to more than 50% prevalence which itself revealed ensemble predictive model is meaningful for future real-time applications for ICT survey for identification of gender based on their answers.

Acknowledgement

The present study is funded by Eötvös Loránd University and sponsored by Tempus Public Foundation, Budapest, Hungary.

References

[1] A. Bonnaccorsi, "On the Relationship between Firm Size and Export Intensity", Journal of International Business Studies, vol. 23, no. 4, pp. 605 – 635, 1992. https://doi.org/10.1057/palgrave.jibs.8490280.

[2] G. Boero, "An econometric analysis of student withdrawal and progression in post-reform Italian universities", Centro Ricerche Economiche Nord Sud, CRENoS Working Paper, 2005.

[3] C. M. Bishop, "Neural Networks for Pattern Recognition", New York, NY, USA: Oxford Univ. Press, 1995.

[4] C. E. Brodley and P. E.Utgoff, "Multivariate decision trees", Machine Learning, vol. 19, no. 1, pp. 45 – 77, 1995. https://doi.org/10.1007/BF00994660.

[5] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining Knowledge Discovery, vol. 2, no. 2, pp.121 – 167, 1998. https://doi.org/10.1023/A:1009715923555.

[6] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert .systems with applications, vol. 1, no. 33, pp.135 – 146, 2007. https://doi.org/10.1016/j.eswa.2006.04.005.

[7] G., Siemens and P Long, "Penetrating the fog: Analytics in learning and education", EDUCAUSE Review, vol. 5, no. 46, 2011.

[8] Gerard J.A. and Baarsa et.al. "A Model to Predict Student Failure in The First Year of the Undergraduate Medical Curriculum", Health Professions Education, pp.5 – 14, 2017.

[9] Javier Bravo et.al. "Exploring the influence of ICT in online students through data mining tools", *eighth International conference on educational data mining*, Spain, 2015.

[10] Kotsiantis.S, et.al. "Predicting students' performance in distance learning using machine learning techniques", Applied Artificial Intelligence, vol.18, pp.411 – 426, 2014. https://doi.org/10.1080/08839510490442058.

[11] M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization", *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp.1951– 1957, 1999. https://doi.org/10.1109/CEC.1999.785513.

[12] R. Caves, "Multinational Enterprise and Economic Analysis", Cambridge University Press, Cambridge, 1982.

[13] S. Alghowinem ET. al., "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors", *IEEE Trans. Affective Computing*, vol. 1, no. 9, pp. 1 – 14, 2016. https://doi.org/10.1109/TAFFC.2016.2634527.

[14] R. Singh and M. Kumar, "Gender Classification Techniques-From Machine Learning to Deep Learning", International Journal of Control Theory and Applications, vol. 9, no. 41, pp.77– 88, 2016.

[15] J. Sara and J. Czaja, et.al. "Factors Predicting the Use of Technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE)", Psychol Aging, vol. 21, no. 2, pp.333 –352, 2006. https://doi.org/10.1037/0882-7974.21.2.333.

[16] Salyungu Mabula, "Modeling student performance in mathematics using Binary Logistic Regression at selected secondary schools: A case study of malware municipality and Ilemela district", Journal of Education and Practice, vol. 6, no. 36, pp. 96 – 103, 2015.

[17] W. Chun, P. and Tzung. Et al., "An Integrated MFFP tree Algorithm for Mining Global Fuzzy Rules from Distributed Databases", vol. 19, no. 4, pp. 521 – 538, 2013.

[18] ESSIE Survey on EC https://ec.europa.eu/digital-single-market/news/ict-education-essie-survey-smart-20100039, Accessed on 14 February 2018.

[19] Chaman Verma, Ahmed S. Tarawneh, Veronika Stoffov´a, Zolt´an Ill´es and Sanjay Dahiya, "Gender prediction of the european school's teachers using machine learning: Preliminary results", 8th IEEE International Advance Computing Conference. IEEE In Press, 2018.

[20] Chaman Verma, Ahmed S. Tarawneh, Veronika Stoffov´a and Zolt´an Ill´es. Forecasting residence state of Indian student based on responses towards information and communication technology awareness: A primarily outcomes using machine learning", International Conference on Innovations in Engineering, Technology and Sciences. IEEE In Press, 2018.