ICIC Express Letters
Volume **13**, Number **3**, March **2019**
 ISSN 1881-803X
pp. **181–189**

AN ANALYSIS OF SUPERVISED LEARNING METHODS FOR PREDICTING STUDENTS' PERFORMANCE IN ACADEMIC ENVIRONMENTS

Liana Maria Crivei, Vlad-Sebastian Ionescu and Gabriela Czibula

Faculty of Mathematics and Computer Science
Babeş-Bolyai University
Str. Mihail Kogalniceanu, nr. 1, Cluj-Napoca 400084, Romania
{ liana.crivei; ivlad; gabis }@cs.ubbcluj.ro

Received October 2018; accepted December 2018

Abstract. *Educational data mining is a challenging interdisciplinary research domain that brings the data mining perspective into the educational field. The major goal of EDM is to provide additional comprehension of the students' learning process. This paper introduces a regression model for predicting the academic performance of students and analyzes the performance of two supervised regressors (random forests and artificial neural networks) in both classification and regression scenarios. Three real experiments are conducted on data collected from Babeş-Bolyai University, Romania. We investigate the effectiveness of different supervised machine learning models and aim to predict the students' final grade at a certain academic discipline based on their performance during the semester.*
Keywords: Educational data mining, Machine learning, Classification, Regression

1. **Introduction.** Applying *machine learning* techniques in education [2] is continuously attracting researchers from the *educational data mining* (EDM) domain, with the major goal of uncovering meaningful patterns from data that come from various educational environments. One purpose of EDM is to offer additional insights into the students' learning process and thus to offer a better comprehension of the educational related activities.

Various applications using data mining techniques have been developed, so far, in the EDM domain. *Machine learning* (ML) methods are extensively investigated, both from a *supervised* and *unsupervised* perspective. ML models as data mining techniques are used for: predicting the students' performance for courses, detecting what type of learners are the students, grouping students according to their similarities or assisting instructors in the educational process [9]. *Supervised learning* (SL) techniques are greatly applied nowadays in various domains for building the so called predictive models which are able to make predictions about what will happen in the future based on some historical data used for training.

Students' performance prediction. Extracting relevant patterns from the educational processes could be effective for understanding students and their learning methods, as well as improving the educational outcomes (e.g., learning outcomes). EDM has received lately considerable attention from the research community since extracting hidden knowledge from educational data is of particular interest for the academic institutions and also useful for improving their teaching methodologies and learning processes [13].

We briefly review, in the following, several recent machine learning based approaches which have been developed for assessing the performance of students in educational environments.

DOI: 10.24507/icicel.13.03.181

Oyedotun et al. [14] attempt to predict, using neural networks, students' performance by the number of times they are likely to retake a certain course. The research is based on 30 attributes that are believed to be possible factors influencing students' performance related to the course itself, the instructor or the particular student. The classification techniques used are *artificial neural networks* (ANNs) [18] and *radial basis function networks* (RBFN). The data set used in the experiments consists of 30 attributes from 5820 students from Gazi University, Ankara, Turkey. The best obtained classification accuracy (0.863) was provided by an ANN classifier, while RBFN provided 0.848 accuracy of the classification process.

Ahmed et al. direct their research in [1] on investigating the factors that influence students' academic performance and achievements, with the main goal of improving the quality of the educational system. In order to determine the best performing algorithm, multiple classifiers such as J48 decision tree, Naïve, Bayes, multilayer perceptron, and sequential minimal optimization are evaluated. The highest classification accuracy of 0.848 is given by J48.

Pal and Pal perform in [15] several experiments in order to determine the best classifier for predicting students' achievement in a computer application examination. The data mining techniques used are: ID3, C4.5 and Bagging. The results are useful for identifying students that need counseling to better their examination results. The data set used in the experiments was obtained from colleges affiliated with VBS Purvanchal University, Jaunpur, India and contained information related to students taking the BCA course (Bachelor of Computer Application) and had 200 instances. An F-score of 0.8 was obtained using a C4.5 classifier.

Tran et al. [19] conduct a study on PSP (predicting student performance) using regression and rating prediction in recommendation systems. To improve performance in the regression case additional course-related skills are taken into account. The authors also propose a hybrid method which is a linear combination of the two presented methods and aims to improve the overall performance. The methods used in [19] are regression methods: *linear regression*, ANNs, *decision trees*, *support vector machines* (SVMs) [6]. The data set used contains scores obtained by the IT students from Vietnam National University. A root mean squared error (RMSE) of 1.705 was obtained using an SVM classifier.

We conduct in this paper a study upon applying two regression models, *random forests* (RFs) and ANNs, for predicting the academic performance of students. For each regressor, two computational models (a regression one and a classification one) are introduced for predicting the final examination grade for a student based on his/her grades received during an academic semester. The results obtained on three real data sets collected from Babeş-Bolyai University, Romania reveal that supervised regressors are helpful for identifying means to increase the quality of the educational processes. To the best of our knowledge, a study similar to ours has not been conducted in the EDM literature so far. In addition, RF classifiers have been applied in the literature, but in other tasks than the one considered in this paper.

To summarize, the purpose of the study conducted in this paper is to answer the following research questions:

RQ1 What is the potential of supervised regression models to predict the final examination grade of students based on the grades they received during the semester?

RQ2 How do the supervised learning models used in this paper compare to other related work from the literature in terms of performance prediction?

The remainder of the paper is structured as follows. Section 2 briefly describes the supervised learning models used for our study. Section 3 introduces our methodology, while the experimental results and a comparison to related work are conducted in Section

4. Section 5 presents the conclusions of our study and outlines directions for future improvements.

2. **Supervised Learning Models Used.** This section presents ANNs and RFs as the main supervised learning models used in our study.

ANNs are widely used as supervised learning models for various applications such as pattern recognition, speech recognition, prediction, system identification and control. Similarly to the biological neural systems, the ANNs [12] consist of a densely interconnected set of computational units, called *neurons*. An ANN [17] is an adaptive system that learns a mapping (an input/output function) from data, by autonomously adjusting the system's parameters during the *training phase*. The ANNs parameters obtained after the training was completed are further used to solve the problem at hand (the *testing phase*).

RFs [3] are an ensemble learning method consisting of combinations of several tree predictors using bootstrap aggregation. During the building process of each of the individual tree, only a random subset of features and a random subset of training examples are considered for analysis. In this way, overfitting is avoided and better stability is achieved for generalization. The generalization error for forests converges to a limit as the number of trees in the forest becomes large, thanks to the law of large numbers. This error depends on the strength of the individual trees in the forest and the correlation between them. Being built on decision trees, random forests can be used in classification and regression problems.

3. **Methodology.** Let us consider the following theoretical model. We denote by $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ a data set in which each *instance* (student) s_i describes the performance of a student, during the academic semester, at a given course $\mathcal{C}$. Each instance s_i is described by a set of *attributes* $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ denoting features which were identified as relevant for measuring the performance of the student for the given course (such as the grades obtained by the student during the semester evaluations). Thus, each s_i is visualized as an m-dimensional vector $s_i = (s_{i1}, s_{i2}, \ldots, s_{im})$, s_{ij} representing the value of attribute a_j for student s_i. For a student s_i, g_i denotes the *final examination grade* obtained by the student at course $\mathcal{C}$.

We are focusing on a supervised learning task, with the goal to predict, based on some historical data, the final examination grade g of a student at a certain academic discipline. A student is visualized as a multidimensional vector $(s_{i1}, s_{i2}, \ldots, s_{im})$, whose elements represent the student's grades obtained during the academic semester. Two experiments will be conducted and analyzed.

(1) **E1**. In this experiment the aim is to predict, using a regression model, a real value for g.
(2) **E2**. In this experiment the focus is to classify students in **3** classes of final grades (< 5, $[5, 7]$, > 7). The final classification is made based on the value predicted for g through regression.

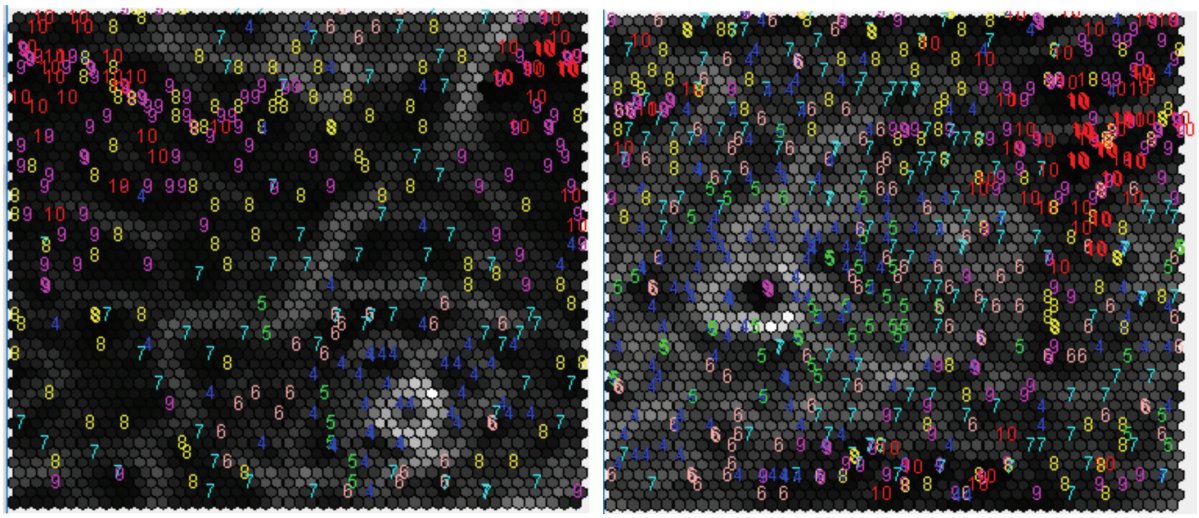
Two regressors will be used in our experiments: RFs and ANNs.

3.1. **Data sets.** Three real data sets collected from Babeş-Bolyai University, Romania will be used in our experiments. The complete data sets are available at [5].

The *first data set* (D1) contains the grades obtained by students at a computer science undergraduate course offered in the *first* semester at Babeş-Bolyai University in a time frame of one academic year (2016-2017). There are 384 instances (students) characterized by 3 attributes (features): laboratory score (a_1), practical test score (a_2) and the *final examination grade* (a_3). The *second data set* (D2) contains the grades obtained by students at a Computer Science undergraduate course offered in the *third* semester at Babeş-Bolyai University in a time frame of four academic years (2014-2018). D2 contains 867 instances

characterized by 5 attributes a_1, a_2, a_3, a_4, a_5: seminar score (a_1), laboratory score (a_2), first practical test score (a_3), second practical test score (a_4) and the *final examination grade* (a_5). The *third data set* (D3) used in our experiment contains the grades obtained by students at a Computer Science undergraduate course offered in the *second* semester at Babeş-Bolyai University in a time frame of six academic years (2012-2018). There are a total of 1169 instances characterized by 5 attributes, denoted by a_1, a_2, a_3, a_4, a_5. The first four attributes represent scores obtained by students during the academic semester: seminar score (a_1), project score (a_2), project status score (a_3), written test score (a_4) and the *final examination grade* (a_5).

Before applying the supervised learning models, the data is analyzed for assessing the complexity of the learning task. First, a *self-organizing map* (SOM) [10] is used as an unsupervised learning model for highlighting how data are organized. The data sets used for building the unsupervised SOM model do not include the target attribute (i.e., the written examination grade obtained by students in the exams session), since we want to test if there is a certain correlation [7] between the grades obtained by a student during the semester and the final examination grade. Thus, the target attribute will be used only for visualization purposes, without being used for building the unsupervised SOM models. Figures 1(a), 1(b) and 2 depict the U-matrix [11] of the SOMs trained on data sets D1, D2 and D3. For the SOM we used our own implementation and the following parameters: 200000 training epochs and a learning rate of 0.1.



(a) (b)

FIGURE 1. SOM visualization for data set D1 (left) and data set D2 (right)

Figures 1(a), 1(b) and 2 highlight, for each of the considered data sets, the difficulty of predicting the students' final grade based on the grades they received during the semester. As expected, one observes that the prediction is more difficult for data set D1, since only 2 attributes are used. For the data set D2, the SOM visualization reveals a better mapping, but still there is no clear separation between the grades. The best mapping is observed on Figure 2, for data set D3, where we observe a cluster of students with the final grades 4, 5, 6 which is well enough delimited and one containing the grades 7, 8, 9, 10. Inside the first cluster, we observe a well distinguishable subclass containing students with the final grade 4.

As a second data analysis method, a statistical analysis was performed. For analyzing the relevance of attributes, for each data set, the Pearson correlation coefficients between the features and the target attribute (the final examination grade) were computed. We observed that, for all data sets, the features are well enough correlated with the final

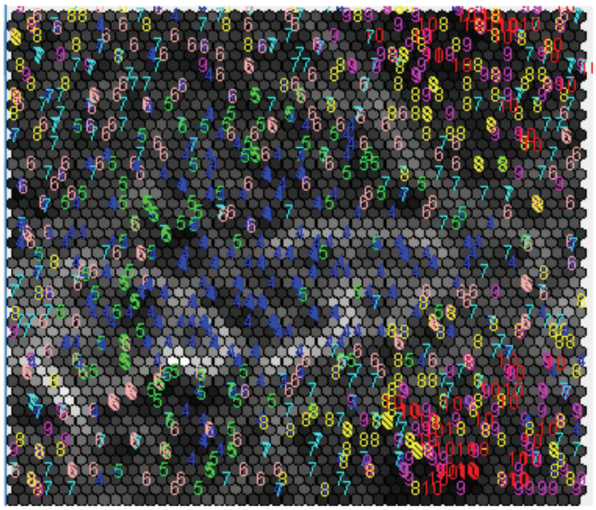


FIGURE 2. SOM visualization for data set D3

examination grade. Correlations between 0.492 and 0.679 were obtained (the smallest correlation of 0.492 is for attribute a_3 and D3, while attribute a_1 in D3 has the highest correlation of 0.679).

3.2. **Performance measures.** For measuring the performance of the regressors (experiment **E1**), the *root mean squared error* (RMSE) is used, while the performance of the classification task (experiment **E2**) will be estimated using the F-score measure.

3.2.1. *RMSE.* The RMSE measures the differences between the values predicted by a regressor and the true values. It is defined as the square root of the average of squared errors, $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$, where $\hat{y}_i$ is the value predicted by the regressor and y_i is the true value.

The RMSE values are non-negative, a value of 0 indicating a perfect fit to the data. For obtaining better regressors, the RMSE should be minimized.

3.2.2. *F-score.* The F-score of a multiclass classifier is computed as the average of the F-score values for all possible classes. In our case, for experiment E2, there are $n = 7$ classes, $C_1, C_2, \ldots, C_7$ corresponding to the grades $4, 5, \ldots, 10$. The generalized confusion matrix for our multiclass classification problem is denoted by $S = (s_{ij})_{\substack{i=\overline{1,n} \\ j=\overline{1,n}}}$ [16], where a_{ij} is computed as the number of instances which are actually belonging to class C_j and were predicted in class C_i.

The *precision* for class C_i ($\forall 1 \leq i \leq n$), denoted by $Prec_i$, is defined as $Prec_i = \frac{s_{ii}}{\sum_{j=1}^{n} s_{ij}}$. The *recall* for the class C_i ($\forall 1 \leq i \leq n$), denoted by $Recall_j$, is computed as $Recall_i = \frac{s_{ii}}{\sum_{j=1}^{n} s_{ji}}$. The F-score for a class C_i, denoted by F-score_i is computed as the harmonic mean between its *precision* and *recall*, i.e., $\text{F-score}_i = \frac{2 \cdot Prec_i \cdot Recall_i}{Prec_i + Recall_i}$. The overall F-score for the multiclass classifier is computed by averaging all F-score_i values, $\text{F-score} = \frac{\sum_{i=1}^{n} \text{F-score}_i}{n}$. The F-score measure ranges from 0 to 1 and should be maximized in order to obtain better classifiers.

3.3. **Experimental methodology.** For each data set described in Section 3.1 and both experiments E1 and E2, a randomly selected subset of 70% instances is used for training the models (RF and ANN) and the rest of 30% will be used for testing (i.e., evaluating the performance of the models considering the RMSE and F-score measures).

For a more precise evaluation, a cross-validation will be used. The 30-70 split is repeated 20 times and the average performance measure (RMSE or F-score) over the 20 runs will be reported, together with the 95% confidence interval (CI) [4] of the mean.

4. **Results and Discussion.** We present in this section the experimental results obtained following the methodology introduced in Section 3 with the goal of answering research question RQ1 stated at the beginning of the paper.

Table 1 depicts the experimental results obtained by applying the experimental methodology described in Section 3.3. We note that, excepting data set D1 and experiment E1, RF performed better in all experiments. Besides, for both E1 and E2 and all data sets, the results on data set D1 are better than those obtained on D2 and D3. A graphical illustration of the obtained experimental results is depicted in Figure 3.

TABLE 1. Experimental results obtained on data sets D1, D2 and D3. A 95% CI is used for the results.

Data set	Regressor	Experiment	Performance measure	Result
D1	**ANN**	E1	RMSE	**1.165 ± 0.037**
		E2	F-score	0.867 ± 0.012
	RF	E1	RMSE	1.221 ± 0.039
		E2	F-score	**0.893 ± 0.012**
D2	**ANN**	E1	RMSE	1.646 ± 0.058
		E2	F-score	0.788 ± 0.006
	RF	E1	RMSE	**1.497 ± 0.018**
		E2	F-score	**0.817 ± 0.007**
D3	**ANN**	E1	RMSE	1.609 ± 0.062
		E2	F-score	0.754 ± 0.005
	RF	E1	RMSE	**1.303 ± 0.013**
		E2	F-score	**0.800 ± 0.007**

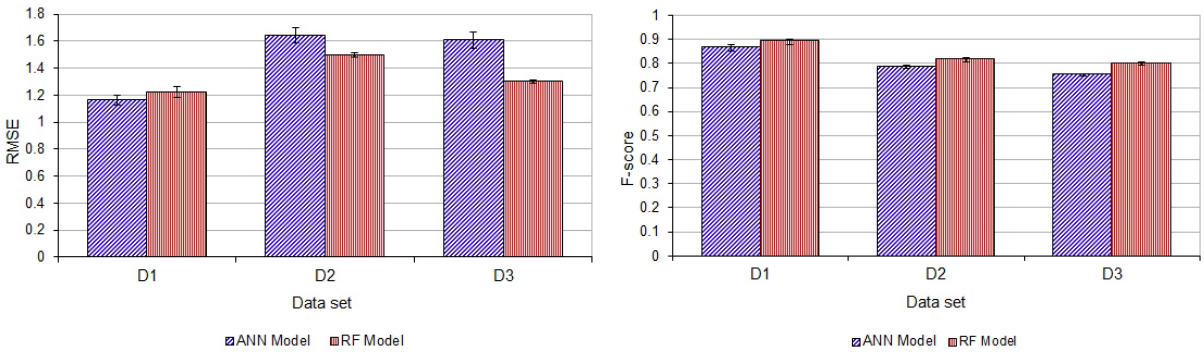


FIGURE 3. Experimental results for experiments E1 (left) and E2 (right) and all data sets. 95% CIs are provided.

4.1. **Comparison to related work.** In order to answer research question RQ2, Table 2 compares our results obtained using RF (mostly better than those using ANN) with the performance of several related approaches. For an accurate comparison, we applied the techniques from the related work on our data sets D1, D2 and D3, employing the same experiments and experimental methodology as described in Section 3.3. The best obtained results are highlighted.

From Table 2 we observe that for experiment E1 and all data sets, both the linear SVM and the SVM with RBF kernel [19] have a slightly better performance than the proposed

TABLE 2. Comparison to related work

Data set	Experiment	Performance measure	Regressor	Result
D1	**E1**	RMSE	**Our RF**	1.221 ± 0.039
			DT [8]	1.482 ± 0.076
			Linear SVM [19]	**1.098 ± 0.032**
			SVM with RBF kernel [19]	1.153 ± 0.030
	E2	F-score	**Our RF**	**0.893 ± 0.012**
			DT [8]	0.879 ± 0.010
			Linear SVM [19]	0.877 ± 0.007
			SVM with RBF kernel [19]	0.873 ± 0.007
D2	**E1**	RMSE	**Our RF**	1.497 ± 0.018
			DT [8]	1.819 ± 0.035
			Linear SVM [19]	**1.437 ± 0.022**
			SVM wih RBF kernel [19]	1.471 ± 0.017
	E2	F-score	**Our RF**	**0.817 ± 0.007**
			DT [8]	0.783 ± 0.008
			Linear SVM [19]	0.778 ± 0.007
			SVM with RBF kernel [19]	0.785 ± 0.009
D3	**E1**	RMSE	**Our RF**	1.303± 0.013
			DT [8]	1.615± 0.022
			Linear SVM [19]	**1.221 ± 0.015**
			SVM with RBF kernel [19]	1.263 ± 0.012
	E2	F-score	**Our RF**	**0.800 ± 0.007**
			DT [8]	0.760 ± 0.007
			Linear SVM [19]	0.762 ± 0.006
			SVM with RBF kernel [19]	0.761 ± 0.007

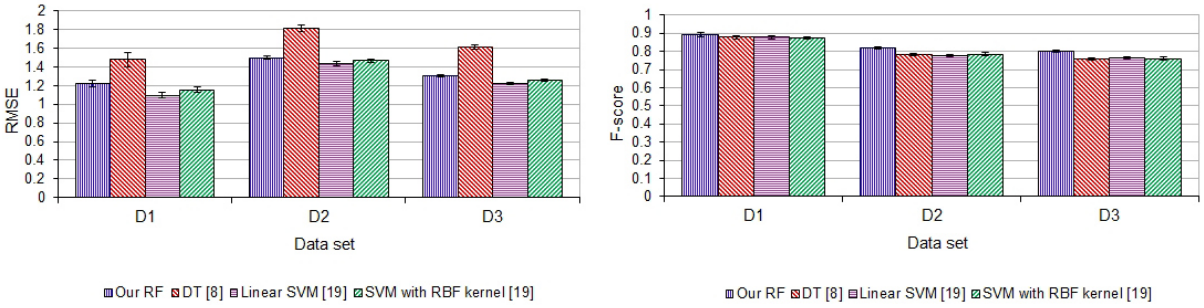


FIGURE 4. Comparison to related work for experiment E1 (left) and E2 (right) and all data sets. 95% CIs are depicted for the results.

RF model. For experiment E2 and all data sets, our RF model outperforms all the related approaches. Overall, out of **18** comparisons, our RF proposal wins the comparison in **12** cases and loses the comparison in **6** cases. Figure 4 summarizes the comparison to related work.

In addition to the results depicted in Table 2, we also note that for all data sets, our RF model outperformed the best results from the literature, i.e.,

- For regression, an average RMSE of **1.34 ± 0.023** compared to an RMSE of 1.705 [19].

- For classification, an average F-score of **0.837 ± 0.009** compared to an F-score of 0.8 [15].

Analyzing the experimental results from Table 1 we may conclude that the grades received by the students during the semester may be relevant in predicting their final examination grade. However, the results regarding the performance of the prediction reveal the following. (1) There are not enough attributes in order to be able to accurately predict the student's final grade. (2) The students' learning process is not continuous during the academic semester, that is why the prediction is difficult. The number of evaluations during the semester should be increased, for constraining the students to study during the semester and not only for the final examination. (3) It is very likely that the instructors from the laboratory and seminars activities do not have the same evaluation standards.

All the previously mentioned aspects have to be further analyzed and considered in the educational process in order to improve its quality and to increase the effectiveness of learning to predict the students' performance.

5. **Conclusions and Future Work.** The study from this paper was conducted with the aim to highlight the effectiveness of supervised regression models in predicting the academic performance of students.

The experiments conducted on three data sets containing real academic data collected from a Romanian University highlighted that generally RF is the best regression model for predicting the final examination grade of students, based on the grades received during an academic semester. For experiment E1 we observed that the SVM regressor provides slightly better results than RF. The study also revealed the difficulty of the students' performance prediction task and the importance of increasing the number of features used in the learning process.

Future work will be made in order to add more features to our learning tasks as well as to apply preprocessing techniques for detecting outliers in data. We also aim to investigate the use of other learning models for predicting students' final performances, such as a classifier based on *relational association rule mining*.

REFERENCES

[1] A. M. Ahmed, A. Rizaner and A. H. Ulusoy, Using data mining to predict instructor performance, *Procedia Computer Science*, vol.102, pp.137-142, 2016.
[2] A. Bogarín, R. Cerezo and C. Romero, A survey on educational process mining, *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, vol.8, no.1, 2018.
[3] L. Breiman, Random forests, *Mach. Learn.*, vol.45, no.1, pp.5-32, 2001.
[4] L. D. Brown, T. T. Cat and A. DasGupta, Interval estimation for a proportion, *Stat. Science*, vol.16, pp.101-133, 2001.
[5] L. M. Crivei, *Academic Data Sets*, http://www.cs.ubbcluj.ro/∼liana.crivei/AcademicDataSets, 2018.
[6] F. E. Gunawan, Improving the reliability of F-statistic method by using linear support vector machine for structural health monitoring, *ICIC Express Letters*, vol.12, no.12, pp.1183-1193, 2018.
[7] T. Hiraoka, T. Katayama and K. Urahama, Generation of parallel-fine-curve-line images by iterative calculation using correlation coefficient, *ICIC Express Letters*, vol.12, no.11, pp.1131-1136, 2018.
[8] N. Hajizadeh and M. Ahmadzadeh, Analysis of factors that affect the students academic performance – Data mining approach, *CoRR*, 2014.
[9] S. T. Jishan, R. I. Rashu, N. Haque and R. M. Rahman, Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, *Decision Analytics*, vol.2, no.1, 2015.
[10] K. Kamei and S. Kawakami, Suppression of isolated clusters occurrence in self-organizing map considering the distance on data space, *ICIC Express Letters*, vol.12, no.5, pp.449-456, 2018.
[11] S. Kaski and T. Kohonen, Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world, *Neural Networks in Financial Engineering. Proc. of the 3rd International Conference on Neural Networks in the Capital Markets*, pp.498-507, 1996.
[12] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[13] S. K. Mohamad and Z. Tasir, Educational data mining: A review, *Procedia – Social and Behavioral Sciences*, vol.97, pp.320-324, 2013.

[14] O. K. Oyedotun, S. Tackie, E. Olaniyi and A. Khashman, Data mining of students' performance: Turkish students as a case study, *International Journal of Intelligent Systems Technologies and Applications*, vol.7, no.9, pp.20-27, 2015.

[15] A. K. Pal and S. Pal, Analysis and mining of educational data for predicting the performance of students, *Intern. Journal of Electronics, Communication and Computer Engineering*, vol.4, no.5, 2013.

[16] D. Picca, B. Curdy and F. Bavaud, Non-linear correspondence analysis in text retrieval: A kernel view, *Proc. of JADT 2006*, pp.741-747, 2006.

[17] R. Rojas, *Neural Networks: A Systematic Introduction*, Springer, 1996.

[18] T. Sun, Y. Peng, F. Yang, H. Sun, N. Fan and B. Zhang, A self-adaptive text classification method based on multiple word embedding and neural network, *ICIC Express Letters*, vol.11, no.6, pp.1133-1141, 2017.

[19] T.-O. Tran, H.-T. Dang, V.-T. Dinh, T.-M.-N. Truong, T.-P.-T. Vuong and X.-H. Phan, Performance prediction for students: A multi-strategy approach, *Cybernetics and Information Technologies*, vol.17, no.2, pp.164-182, 2017.