Paper ID: 81

COMPARING THE PERFORMANCE OF DATA MINING METHODS IN CLASSIFYING SUCCESSFUL STUDENTS WITH SCIENTIFIC LITERACY IN PISA 2015*

Serap Büyükkıdık[1], Batuhan Bakırarar[2], Okan Bulut[3]
[1]Sinop University, [2]Ankara University, [3]University of Alberta
sbuyukkidik@gmail.com, batuhan_bakirarar@hotmail.com, bulut@ualberta.ca

Abstract: This study aims to classify successful and unsuccessful students in PISA (2015) scientific literacy using the indices and student questionnaire items in the PISA 2015 database. The sample of the study consists of 5895 Turkish students who participated in PISA 2015. In data analysis, Multilayer Perceptron, Logistic Regression, and Support Vector Machine methods were used as data mining methods. The data set was evaluated in three different ways using 80% training-20% test, 70% training-30% test and 10-fold Cross Validation test. Accuracy, F-measure, Precision, Recall, and ROC Area were used as the evaluation criteria. The results showed that the most important variables were found to be environmental awareness scale items in order to classify successful and unsuccessful students in the research. The highest Accuracy value across all conditions was 0.81 for the Support Vector Machine method in the data set tested with 10-fold Cross Validation. The lowest Accuracy value was 0.74 for the Multilayer Perceptron method when the data was split as 80% training-20% test. In the study, the performance measures obtained from the data set tested with 10-fold Cross Validation were found to be the highest in all conditions. Based on the Accuracy criterion, values obtained from Support Vector Machine are the highest in 70% training-30% test and 10-fold Cross Validation data set. Although the performance measures obtained from the other methods used and evaluation criteria are relatively close to each other, it can be seen that they can vary according to the conditions.

***Keywords*: ** PISA, Scientific Literacy, Multilayer Perceptron, Logistic Regression, Support Vector Machine

INTRODUCTION

The Programme for International Student Assessment (PISA) is an international, large-scale assessment, organized by the Organization of Economic Cooperation and Development (OECD). Since 2000, PISA has been given to 15-year-old students every three years to assess their competencies in reading, mathematics, and science. Rather than focusing on the extent to which the students have mastered a specific school curriculum, the PISA assessments focus on the students' ability to use their knowledge and skills to meet real-life challenges. In addition to achievement tests, PISA also uses student questionnaires to collect information from students on various aspects of their home, family and school background, and school questionnaires to collect information from schools about various aspects of organization and educational provision in schools.

Each PISA assessment takes a literacy perspective, focusing on the extent to which students can apply the knowledge and skills from a particular subject area into problems and challenges that they might come across in real life. The latest results, from PISA 2015, focus on student's performance in and attitudes towards science, with reading and mathematics as minor areas of assessment. PISA 2015 also included the assessment of an innovative domain, collaborative problem solving and the assessment of financial literacy which was optional for the participating countries and economies. In PISA 2015, a sample of 540,000 students (representing 28 million 15-year-old students) from 72 OECD countries and economies participated in the assessment.

Reliable and representative results from the tests and questionnaires incorporated in PISA allow researchers from all around the world to investigate various research problems related to the quality of education systems, effectiveness of educational policies, and most importantly students' competency in science, reading, mathematics, and other innovative subject areas. The PISA 2015 framework for science emphasizes the importance of educating all young people to become informed, critical users of scientific knowledge. The assessment tasks focused on three aspects of science: the knowledge of the fundamental scientific ideas, the knowledge and understanding of scientific enquiry, and the ability to interpret data and evidence scientifically (OECD, 2017).

The large and rich datasets from PISA assessments also allow researchers to implement more innovative statistical approaches for investigating research problems. As one of these innovative methods, data mining has been used in several recent studies, focusing on educational problems (e.g., Aksu & Güzeller, 2016; Alan, 2012; Ayesha, Mustafa, Sattar & Khan, 2010; Ayık, Özdemir & Yavuz, 2007; Baradwaj & Pal, 2011; Bilen, Hotaman, Aşkın & Büyüklü, 2014; Gaafar & Khamis, 2009; Liu & Ruiz, 2008; Jormanainen & Sutinen, 2012; Kahveci & Özdemir, 2016; Şen, Uçar & Delen, 2012; Şen & Uçar, 2012; Şengül, 2011; Tsai, Tsai, Hung & Hwang, 2011; Yadav, Bharadwaj & Pal, 2012). With the recent release of the PISA 2015 results, it is possible to apply various data mining methods in the PISA datasets to identify and investigate different classification and clustering problems in the context of science, math, and reading assessments.

As highlighted in PISA 2015, the subject of science plays a gatekeeping role in students' understanding of a variety of issues, ranging from infectious diseases and human cloning to artificial intelligence and climate change. A scientifically literate student would be expected to engage in reasoned discourse about science and technology (OECD, 2017, p. 44). Therefore, measuring scientific literacy and identifying students with high and low scientific literacy is a new challenge for educators and researchers. Using large amounts of science assessment data available, researchers can employ advanced techniques – such as data mining – to better understand students' performance in science and determine what factors contribute to acquisition of scientific literacy.

This study aims to examine the PISA 2015 results to identify variables in the PISA database explaining students' scientific literacy, using various data mining methods. To predict whether the students reached the scientific literacy proficiency (i.e., proficient vs. not proficient classification); three data mining methods were employed: Multilayer perceptron, logistic regression, and support vector machine. The performance of the three methods in classifying student correct was evaluated based on the following criteria: Accuracy, F-measure, Recall, and ROC area. The primary research questions addressed in this study are as follows: (1) Based on the four evaluation criteria (accuracy, F-measure, precision, recall, and ROC area), which data mining method (multilayer perceptron, logistic regression, and support vector machine) performs the best in identifying students who are proficient in scientific literacy from those who are not? (2) What is the impact of splitting the data with 80% training-20% test, 70% training-30% test and 10-fold cross validation on the classification results?

METHODOLOGY

The sample of the study included the Turkish students who participated in PISA 2015. The student population size in Turkey was estimated as 1,324,089 students, while 925,366 of these students were eligible to participate in PISA 2015. Out of this large population, 5895 students were sampled from 187 schools across 61 provinces in Turkey (Taş, Arıcı, Ozarkan, & Özgürlük, 2016). The PISA 2015 database consists of 922 variables in total. For the analyses of these variables, WEKA 3.6 software was used. To reduce the large number of variables in the database and eliminate the variables that would not contribute to the classification of scientific literacy, the InfoGainAttributeEval, GainRatioAttributeEval, and ChiSquaredAttributeEval methods in WEKA were used. The variables that were flagged as "insignificant" from all of the three methods were excluded from the analysis. The final database consisted for 66 variables. To replace the missing values for these variables, mod values were used as a replacement for qualitative variables and the mean values were used as a replacement for quantitative variables.

Two types of variables were used in the data analysis. The first set of variables were the demographic variables obtained from the students (grade compared to modal grade in country - ST001D01T, gender - ST004D01T, out-of-school study time per week - ST071Q01NA, studying for school or homework before going to school - ST076Q02NA, and highest education of parents – HISCED). The second set of variables consisted of latent variables derived from the questions in the PISA 2015 Student Questionnaire. Each latent variable was estimated based on a set of questionnaire questions and the resulting scores were scaled across the participating countries in PISA 2015. Table 1 shows the description, the number of questions, the names of the questions from the PISA 2015 database, response categories for the questions, and the reliability index for each latent variable.

Table 1. Derived variables in the PISA 2015 Student Questionnaire

Description	Questions' number	Questions	Categories	Realibility
Disciplinary climate in science classes	5	ST097Q01TA ST097Q02TA ST097Q03TA ST097Q04TA ST097Q05TA	"every lesson", "most lessons", "some lessons" and "never or hardly ever"	0.893
Inquiry-based science teaching and learning practices	9	ST098Q01TA ST098Q02TA ST098Q03NA ST098Q05TA ST098Q06TA ST098Q07TA ST098Q08NA ST098Q09TA ST098Q10NA	"in all lessons", "in most lessons", "in some lessons", "never or hardly ever"	0.894
Teacher support in a science classes	5	ST100Q01TA ST100Q02TA ST100Q03TA ST100Q04TA ST100Q05TA	"every lesson", "most lessons", "some lessons" and "never or hardly ever"	0.915
Teacher-directed science instruction	4	ST103Q01NA ST103Q03NA ST103Q08NA ST103Q11NA	"never or almost never", "some lessons", "many lessons", and "every lesson or almost every lesson"	0.803
Adaption of instruction	3	ST107Q01NA ST107Q02NA ST107Q03NA	"never or almost never", "some lessons", "many lessons", and "every lesson or almost every lesson"	0.813
Environmental awareness	6	ST092Q01TA ST092Q02TA ST092Q04TA ST092Q05TA	"I have never heard of this", "I have heard about this but I would not be able to explain what it is really about",	0.885

		ST092Q06NA ST092Q08NA	"I know something about this and could explain the general issue", "I am familiar with this and I would be able to explain this well"	
Enjoyment of science	5	ST094Q01NA ST094Q02NA ST094Q03NA ST094Q04NA ST094Q05NA	"strongly agree", "agree", "disagree", and "strongly disagree"	0.852
Interest in broad science topics	5	ST095Q04NA ST095Q07NA ST095Q08NA ST095Q13NA ST095Q15NA	"not interested", "hardly interested", "interested", "highly interested", and "I don't know what this is"	0.856
Instrumental motivation	4	ST113Q01TA ST113Q02TA ST113Q03TA ST113Q04TA	"strongly agree", "agree", "disagree", and "strongly disagree"	0.900
Science self-efficacy	8	ST129Q01TA ST129Q02TA ST129Q03TA ST129Q04TA ST129Q05TA ST129Q06TA ST129Q07TA ST129Q08TA	"I could do this easily", "I could do this with a bit of effort", "I would struggle to do this on my own", and "I couldn't do this"	0.892
Epistemological beliefs	6	ST131Q01NA ST131Q03NA ST131Q04NA ST131Q06NA ST131Q08NA ST131Q11NA	"strongly agree", "agree", "disagree", and "strongly disagree"	0.919

Table 1 shows the description of 11 subscales derived from the PISA 2015 Student Questionnaire. In each subscale, the number of items varies from three to nine. Also, the number of response categories is four for all of the items across the 11 subscales, although the labels of these response categories differ based on the subscales. All of the subscales listed in Table 1 indicated acceptable levels of reliability (> .80 or higher). In addition to the subscales in Table 1, a few indiviual items from the PISA 2015 Student Questionnaire were included in the data analysis. These items were about the grade of students, gender, out-of-school study time per week, studying for school or homework before going to school, and parents' education levels.

In the study, the average scientific literacy score ($\bar{x}$ = 425.00) was obtained by taking average of 10 scientific literacy scores (i.e., estimated plausible values for scientific literacy) with the PVSCIE code as the outcome variable from the dataset in PISA 2015. The resulting average scientific literacy score was used as a cut-off point to determine students who were successful in scientific literacy and those who were not (i.e., students with the average and higher scores are successful, those whose scores are below the average are unsuccessful).

The three data mining methods used in the data analysis are Multilayer Perceptron, Logistic Regression, and Support Vector Machine. The final dataset was evaluated in three different ways using 80% training-20% test, 70% training-30% test and 10-fold Cross Validation test. Accuracy, F-measure, Precision, Recall, and ROC Area were used as the evaluation criteria.

FINDINGS

Table 2 presents the results of the correct classification rates obtained from 66 variables.

Table 2. Multilayer Perceptron, Logistic Regression and Support Vector Machine Results for the dataset that was divided into 70% training-30% test and 80% training-20% test and tested with 10-fold Cross Validation

Methods	**Performance Criteria**				
	Accuracy	**F-measure**	**Precision**	**Recall**	**ROC Area**
70% training-30% test					
Multilayer Perceptron	0.769	0.770	0.772	0.769	0.845
Logistic Regression	0.800	0.800	0.801	0.799	0.875
Support Vector Machine	0.802	0.801	0.800	0.803	0.799
80% training-20% test					
Multilayer Perceptron	0.739	0.747	0.755	0.739	0.838
Logistic Regression	0.804	0.803	0.801	0.806	0.873
Support Vector Machine	0.798	0.797	0.796	0.799	0.794
10-fold Cross Validation					
Multilayer Perceptron	0.775	0.774	0.778	0.771	0.845
Logistic Regression	0.810	0.810	0.813	0.808	0.877
Support Vector Machine	0.811	0.811	0.815	0.805	0.809

Table 2 shows that the highest values are in Support Vector Machine, Logistic Regression, and Multilayer Perceptron, respectively, when Accuracy, F-measure and Recall performance criterion of the dataset separated by 70% training-30% test is examined. When the results obtained with the Precision criterion are examined, Support Vector Machine and Logistic Regression seem to produce similar values as in the previous methods, whereas the values obtained from Multilayer Perceptron are lower. When the values obtained with the ROC Area criterion are examined, the highest classification accuracy is observed in Logistic Regression and the lowest accuracy is observed in Support Vector Machine.

The highest values of all performance criteria for the dataset separated by 80% training-20% test were obtained by Logistic Regression. When the accuracy, F-measure, Precision and Recall performance criteria are examined, the second highest accuracy value is obtained with Support Vector Machine Precision. When the results obtained with the ROC Area criterion are examined, the highest accuracy value is in Logistic Regression and the lowest accuracy value is in Support Vector Machine.

When the results for the Accuracy, F-measure, and Precision performance criteria of the dataset tested with 10-fold Cross Validation are examined, Support Vector Machine and Logistic regression have very similar results and Multilayer Perceptron has relatively low values. When the results obtained with the recall criterion are examined, the values obtained from the Multilayer Perceptron are lower while the Support Vector Machine and the Logistic regression produce similar values as in the preceding methods. When the results obtained with the ROC Area criterion are examined, it is seen that the highest value is in Logistic Regression and the lowest value is in Support Vector Machine.

In order to evaluate the performance of the three test options, the Accuracy criterion is used as the baseline in previous research (Güldoğan, Yağmur, Yoloğlu, Asyalı & Çolak, 2015). Based on the Accuracy criterion, the results of this study are demonstrated in Figure 1.

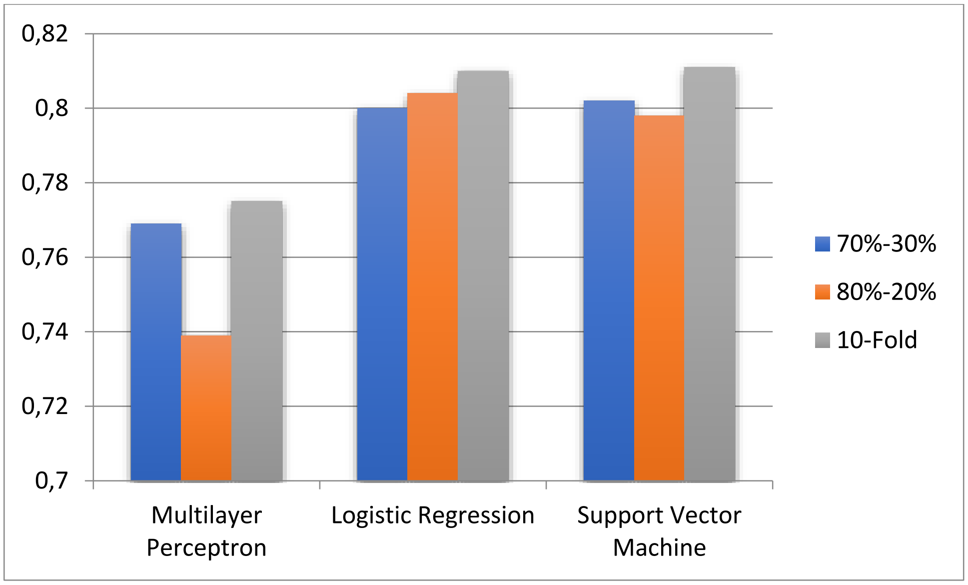

Figure 1. Assessing the correct classification rates of data mining methods under different test options

Figure 1 shows that he best results in the Multilayer Perceptron method were obtained from the dataset tested with 10-fold Cross Validation, while the lowest values were obtained from the dataset, separated by 80% training-20% test. In the logistic regression method, the best results were obtained from the data set tested with the 10-fold Cross Validation, followed by the data set divided into 80% training-20% test and 70% training-20% test, respectively. When the results obtained from the Support Vector Machine method according to the Accuracy scale are examined, it is found that the results are consistent with the findings obtained with the Multilayer Perceptron method. Figure 1 also shows that the results obtained with 10-fold Cross Validation are better than the other data sets in all methods.

CONCLUSIONS

In this study, 922 variables of PISA 2015 data were used. The variables in the data set were reduced to 66 variables by using the InfoGainAttributeEval, GainRatioAttributeEval, and Chi-SquareAttributedEval methods in the WEKA software program and by looking at their values in the data set and removing the variables considered to be meaningless by the three methods from the data set. For all these variables, missing data analysis was performed and the mod values for the missing data in the qualitative variables and the mean values for the missing data in the quantitative variables were assigned. Analyzes can be performed using different methods of missing data assignment in future studies.

When the dataset is divided into 70% training-30% test and analyzed with 10-fold Cross Validation option, the Support Vector Machine method gives the best results in terms of Accuracy and F-measure criteria. Other best practices are Logistic Regression and Multilayer Perceptron, respectively. Only when the dataset is divided into 80% training-20% test, the Logistic Regression gives the best results in terms of Accuracy and F-measure criteria, while the others are Support Vector Machine and Multilayer Perceptron.

The data set used was divided into 70% training-30% test and 80% training-20% test and tested with 10-fold Cross Validation option. Instead of 10-fold Cross Validation, which is frequently used in research, analysis can be performed by choosing k values differently in k-fold cross validation by separating data set differently. In addition, Multilayer Perceptron, Logistic Regression and Support Vector Machine methods were used in this study. Future research can be conducted using different methods of data mining.

*The first author was supported by the Scientific and Technological Research Council of Turkey as part of 2214-A scholarship program.

REFERENCES

Aksu, G., & Güzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *Education and Science, 41* (185), 101-122.

Alan, M. A. (2012). Veri madenciliği ve lisansüstü öğrenci verilerine üzerine bir uygulama. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, (33),* 165-174.

Ayesha, S., Mustafa, T., Sattar, A. R., & Khan, M. I. (2010). Data mining model for higher education system. *Europen Journal of Scientific Research, 43* (1), 24-29.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği işe analizi. *Atatürk Üniversitesi Sosyal Bilimler Dergisi, 10* (2), 441-454.

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. International *Journal of Advanced Computer Science and Applications, 2* (6), 63-69.

Bilen, Ö., Hotaman, D., Aşkın, Ö. E., & Büyüklü, A. H. (2014). LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 İstanbul örneği. *Eğitim ve Bilim, 39* (172), 78-94.

Gaafar, L., & Khamis, M. (2009). Applications of data mining for educational decision support. Proceedings of the 2009 Industrial Engineering Research Conference, 228-233.

Güldoğan, E., Yağmur, J., Yoloğlu, S., Asyalı, M. H., & Çolak, C. (2015). Myocardial infarction classification with support vector machine models. *Turgut Özal Tıp Merkezi Dergisi*, *22*(4), 221-224.

Jormanainen, I., & Sutinen, E. (2012). Using data mining to support teacher's intervention in a robotics class. Paper presented at the Fourth IEEE International Conference On Digital Game And Intelligent Toy Enhanced Learning, Takamatsu, Japan (March 27-30).

Kahveci, F., & Özdemir, A. (2016). Öğrenci bilgi sisteminde değerlendirmenin veri madenciliği ile yapılması. *Yönetim Bilişim Sistemleri Dergisi, 1 (*3), 1-10.

Liu, X., & Ruiz, M. E. (2008). Using data mining to predict K-12 students' performance on large-scale assessment items related to energy. *Journal of Research in Science Teaching*, *45*(5), 554–573.

OECD (2017). *PISA 2015 technical report*. OECD Publishing, Paris. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications, 39* (10), 9468-9476.

Şen, B., & Uçar, E. (2012). Evaluating the achievements of computer engineering department of distance education students with data mining methods. *Procedia Technology, 1,* 262-267.

Şengül, A. (2011). *Türk öğrencilerinın PISA 2009 okuma becerilerini açıklayan değişkenlerin CHAID analizi ile belirlenmesi.*Yayımlanmamış yüksek lisans tezi: Ankara Üniversitesi, Ankara.

Taş, U. E., Arıcı, Ö., Ozarkan, H. B., & Özgürlük, B. (2016). *PISA 2015 ulusal raporu*. Ankara: MEB.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student opinion data with regression and decision tree analysis. *Research in Higher Education, 45* (3), 251-269.

Tsai, C. F., Tsai, C. T., Hung, C. S. , & Hwang, P. S. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology, 27* (3), 481-498.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention. *International Journal of Computer Science and Information Security*, *10* (2), 113-117.