Chapter 2
An Ensemble-Based Semi-Supervised Approach for Predicting Students' Performance

Ioannis E. Livieris, Konstantina Drakopoulou, Tassos Anastasios Mikropoulos, Vassilios Tampakas, and Panagiotis Pintelas

Introduction

Educational data mining (EDM) is a growing academic research area, which aims to gain significant insights on student behavior, interactions, and performance and to improve the technology-enhanced learning methods in a data-driven way by applying data mining methods on educational data (Bousbia & Belamri, 2014). During the last decade, research has been focused to enhance the learning experience and institutional effectiveness by merging the computer-assisted learning systems and automatic analysis of educational data. EDM can offer opportunities and great potentials to increase our understanding about learning processes to optimize learning through educational systems. These opportunities have been strengthened by a huge shift in the availability of the data resources, which constitute an inspiring motivation for growing research in this academic research area. In this regard, EDM can be utilized to inform and support learners, teachers, and their institutions and therefore help them understand how these powerful tools can lead to huge benefits in learning and success in educational outcomes, through personalization and adaptation of education based on the learner's needs (Greller & Drachsler, 2012).

I. E. Livieris (✉) · V. Tampakas
Department of Computer and Informatics Engineering (DISK Lab),
Technological Educational Institute of Western Greece, Patras, Greece
e-mail: livieris@teiwest.gr; vtampakas@teimes.gr

K. Drakopoulou · P. Pintelas
Department of Mathematics, University of Patras, Patras, Greece
e-mail: kdrak@math.upatras.gr; pintelas@upatras.gr

T. A. Mikropoulos
The Educational Approaches to Virtual Reality Technologies Laboratory,
University of Ioannina, Ioannina, Greece
e-mail: amikrop@uoi.gr

T. A. Mikropoulos (ed.), *Research on e-Learning and ICT in Education*,
https://doi.org/10.1007/978-3-319-95059-4_2

In Greece, like in most countries, secondary education takes place after 6 years of primary education and may be followed by higher education or vocational training. Its main objectives are to engender a balanced and all-round development of the students' personality at a cognitive and emotional level. It comprises two main stages: Gymnasium and Lyceum. Gymnasium covers the first 3 years with the purpose to enrich students' knowledge in all fields of learning and support the development of composite and critical thinking. The next 3 years are covered by Lyceum which further cultivates the students' personalities while at the same time prepares them for admission in higher education. Essentially, Lyceum acts like a bridge between school education and higher learning specializations that are offered by universities.

In the end of the first grade of Lyceum (A′ Lyceum), the students are obligated to select between three directions: humanity, science, and technology. This selection establishes the courses, which the students will attend in the Panhellenic national examinations in order to proceed to the higher education. In this regard, the students' entry into a specific higher educational institution is mainly based on the orientation and group chosen. Therefore, the ability to predict students' performance in the final examinations of A′ Lyceum is considered essential not only for students but also for the educators and the educational institutes. More comprehensively, the "knowledge discovery" can assist students to have a first evaluation of their progress and possibly enhance their performance and teachers to conduct their classes better, identifying difficulties and improving their teaching methods. Thus, it is of major importance to closely monitor the students' performance in order to identify possible retardation and proactively intervene towards their academic enhancement through the assignment of extra learning material, small group training, etc. Nevertheless, the early identification of students who are likely to exhibit poor performance is a rather difficult and challenging task, and even if such identification is possible, it is usually too late to prevent students' failure (Livieris, Drakopoulou, Kotsilieris, Tampakas, & Pintelas, 2017; Livieris, Drakopoulou, & Pintelas, 2012; Livieris, Mikropoulos, & Pintelas, 2016).

A workable solution to prevent this trend is to analyze and exploit the knowledge acquired from students' academic performance records. In this context, many researchers in the past have conducted studies on educational data in order to cluster students based on academic performance in examinations. However, most of these studies examine the efficiency of supervised classification methods, while the ensemble methods (Gandhi & Aggarwal, 2010; Kotsiantis, Patriarcheas, & Xenos, 2010; Livieris et al., 2016, 2017) and semi-supervised methodologies (Kostopoulos, Kotsiantis, & Pintelas, 2015; Kostopoulos, Livieris, Kotsiantis, & Tampakas, 2017) have been rarely applied to the educational field. Semi-supervised methods and ensemble methods are two important machine learning techniques. The former attempt to achieve strong generalization by exploiting unlabeled data, while the latter attempt to achieve strong generalization by using multiple learners. Although both methodologies have been efficiently applied to a variety of real-world problems during the last decade, they were almost developed separately. Recently, Zhou (2011) presented that semi-supervised learning algorithms and ensemble learning algorithms are indeed beneficial to each other, and more efficient and robust classification algorithms can be developed. More specifically, semi-supervised methodologies could be useful to ensemble methodologies since:

1. Unlabeled data can enhance the diversity of individual classifiers.
2. The lack of labeled examples can be exploited by utilizing unlabeled ones.

Furthermore, the combination of individual classifiers could assist semi-supervised methods since:

1. An ensemble of classifiers could be more accurate than an individual classifier.
2. The performance of the ensemble classifier could be significantly improved using unlabeled data.

In this work, we propose a new ensemble-based semi-supervised learning algorithm for predicting the students' performance in the final examinations of Mathematics at the end of academic year of A′ Lyceum. The specific course has been selected since it has been characterized as the most significant and most difficult course of the Science direction. Our objective and expectation is that this work could be used as a reference for decision-making in the admission process and to provide better educational services by offering customized assistance according to students' predicted performance.

The remainder of this chapter is organized as follows: Section "A Review of Semi-supervised Machine Learning Algorithms" presents a brief discussion of the semi-supervised learning algorithms utilized in our framework. Section "Literature Review on Educational Data Mining" reviews the related work of other researchers in the area of machine learning algorithms for prediction and classification in education. Section "Proposed Methodology" presents the educational dataset utilized in our study and our proposed ensemble-based semi-supervised learning algorithm, which is compared with the most popular classification algorithms by conducting a series of tests. Finally, the last section considers the conclusions and some further research topics for future work.

A Review of Semi-supervised Machine Learning Algorithms

Semi-supervised learning (SSL) consists of a mixture of supervised and unsupervised learning, aiming to obtain better classification results and performance by exploiting the explicit classification information of labeled data and the information hidden in the unlabeled data. SSL algorithms have become a topic of significant research as an alternative to traditional methods of machine learning, which exhibit remarkable performance over labeled data but lack the ability to be applied on large amounts of unlabeled data. The general assumption of SSL algorithms is that data points in a high-density region are likely to belong to the same class and the decision boundary lies in low-density regions (Zhu, 2006). Therefore, these methods have the advantage of reducing the effort of supervision to a minimum while still preserving competitive recognition performance.

More specifically, SSL methods utilize only a small proportion of the whole amount of data to be labeled for accomplishing their task. This attribute known as labeled ratio R is defined by

$$R = \frac{\text{Number of labeled instances}}{\text{Number of all instances}}$$

and it is usually provided in percentage values (%). Next, after the labeled ratio is defined, all the available data are split into two distinct subsets: the labeled and the unlabeled set.

In the literature, several semi-supervised algorithms have been proposed so far with different philosophy and performance and have been successfully applied in many real-world applications (Chapelle, Scholkopf, & Zien, 2009; Kostopoulos et al., 2015, 2017; Levatic, Dzeroski, Supek, & Smuc, 2013; Liu & Yuen, 2011; Sigdel et al., 2014; Triguero, Saez, Luengo, Garcia, & Herrera, 2014; Wang & Chen, 2013; Zhu, 2006, 2011). Based on their experimental results, many researchers have stated that the classification accuracy can be significantly improved if a large number of unlabeled data are used together with a small number of labeled data. We refer the reader to Pise and Kulkarni (2008), Triguero and Garcıa (2015), and Zhu (2006) and the references therein, for an overview on semi-supervised learning methods and their applications.

In this study, we investigate the classification accuracy utilizing the most famous and frequently used semi-supervised learning techniques: self-training, co-training, and tri-training, which constitute the most representative SSL algorithms.

Self-Training

Self-training is a wrapper-based semi-supervised approach which constitutes an iterative procedure of self-labeling unlabeled data. It has been established as a very popular algorithm due to its simplicity, and it is often found to be more efficient and more accurate than other semi-supervised algorithms (Kostopoulos et al., 2015; Roli & Marcialis, 2006; Sigdel et al., 2014). According to Ng and Cardie (2003), "self-training is a single-view weakly supervised algorithm." Initially, an arbitrary classifier is trained with a small amount of labeled data, which have been randomly chosen from the training set. Subsequently, the training set is iteratively augmented gradually using a classifier trained on its own most confident predictions. More specifically, each classified unlabeled instance that has achieved a probability value over a defined threshold c is considered sufficiently reliable to be added to the training set for the following training phases. Finally, these instances are added to the initial training set, increasing in this way its efficiency and robustness. Therefore, the retraining of the classifier is done using the new enlarged training set until stopping criteria are satisfied.

An important reason why performance may fluctuate compared with supervised algorithms' performance is the fact that, during the training phase of the former, some of the unlabeled examples will not get labeled, since the termination of the algorithm will have been preceded (Schwenker & Trentin, 2014). However, since the success of the self-training algorithm is heavily dependent on its own predictions, its weakness is that erroneous initial predictions will probably lead the classifier to generate incorrectly labeled data (Zhu & Goldberg, 2009).

Co-training

Co-training is a semi-supervised algorithm which can be considered as a different variant of self-training technique (Blum & Mitchell, 1998). The underlying assumptions of the co-training approach are that feature space can be split into two different conditionally independent views and that each view is able to predict the classes perfectly (Du, Ling, & Zhou, 2011; Sun & Jin, 2011). Under these assumptions, co-training algorithm assumes that it is more effective to predict the unlabeled instances by dividing the features of data into two separable categories. In this framework, two classifiers are used. One classifier is trained on each subset, and then the classifiers teach each other with a respective subset of unlabeled examples with the highest confidence predictions. Subsequently, each classifier is retrained with the additional training examples given by the other classifier, and the process is repeated.

Blum and Mitchell (1998) analyzed the classification performance and effectiveness of co-training and disclosed that if the two views are conditionally independent, the predictive accuracy of an initially weak learner can be boosted to arbitrarily high using unlabeled data by co-training. Nevertheless, the assumption about the existence of sufficient and redundant views is a luxury hardly met in most scenarios; several extensions of this algorithm have been developed such as tri-training.

Tri-training

Tri-training algorithm has been originally proposed for solving the problem of co-training since it requires neither two views nor special learning algorithms. This algorithm attempts to exploit unlabeled data utilizing three classifiers. However, such a setting tackles the problem of determining how to efficiently select most confidently predicted unlabeled examples to label. Therefore, in order to make the three classifiers diverse, the original labeled set is bootstrap sampled (Efron & Tibshirani, 1993) to produce three perturbed training sets, on each of which a classifier is then generated and avoids estimating the predictive confidence explicitly. Subsequently, in each tri-training round, if two classifiers agree on the labeling of an unlabeled instance while the third one disagrees, then these two classifiers will label this instance for the third classifier. It is worth noticing that the "majority teach minority strategy" serves as an implicit confidence measurement, which avoids the use of complicated time-consuming approaches to explicitly measure the predictive confidence, and hence the training process is efficient.

However, sometimes the performance of tri-training degrades; hence three other issues must be taken into account (Guo & Li, 2012):

1. Estimation of the classification error is unsuitable.
2. Excessively confined restrictions introduce further classification noise.
3. Differentiation between initial labeled example and labeled of previously unlabeled example is deficient.

Literature Review on Educational Data Mining

During the last decade, the application of data mining for the development of accurate and efficient decision support systems for monitoring students' performance is becoming very popular in the modern educational era. A large proportion of these studies examines the efficiency of supervised classification methods, while ensemble and SSL methodologies have been rarely applied to the educational field. Some excellent reviews (Baker & Yacef, 2009; Pena-Ayala, 2014; Romero & Ventura, 2007, 2010) provide a comprehensive resource of papers on EDM, which present a detailed description of the mining learning data process, covering the application of EDM from traditional educational institutions to web-based learning management systems and intelligently adaptive educational hypermedia systems. Moreover, they present how EDM seeks to discover new insights into learning with new tools and techniques, so that those insights impact the activity of practitioners in all levels of education, as well as corporate learning. A number of rewarding studies have been carried out in recent years and some of them are presented in this section.

Kotsiantis, Pierrakeas, and Pintelas (2003, 2004) studied the accuracy of six common machine learning algorithms in predicting students that tend to drop out from a distance learning course in the Hellenic Open University. Based on previous works, Kotsiantis et al. (2010) proposed an online ensemble of supervised algorithms to predict the performance on the final examination test (pass/fail) of students attending distance courses in higher education. The proposed ensemble of classifiers outperformed classical well-known algorithms and could be utilized as a predictive tool from tutors during the academic year to underpin and boost low performers.

Thai-Nghe, Janecek, and Haddawy (2007) attempted to predict the performance of undergraduate and postgraduate students at two academic institutes using machine learning techniques. Along this line, Thai-Nghe, Busche, and Schmidt-Thieme (2009) presented an extensive study to deal with the class imbalance problem in order to improve the prediction results of academic performances. Firstly, they balanced the datasets and then they used both cost-insensitive and costsensitive learning with a support vector machine for the small datasets and decision tree for the larger datasets which provided satisfactory classification results.

Cortez and Silva (2008) predicted the student grades for two core classes (Mathematics and Portuguese) from two secondary schools. The data were extracted from school records, as well as provided by the students through questionnaires. They applied four classification algorithms on three data setups, with different combinations of attributes, trying to find out those with more effect on the prediction. Based on their numerical experiments, the authors concluded that the students' achievements are more related with their performance in the past years and less correlated with their social and cultural characteristics.

Gandhi and Aggarwal (2010) presented a methodology based on the assessment of their past performance as well as on their respective learning curves constructed over time to predict the future performance of students. More specifically, they applied the Rasch model technique to capture the effects of student level proficiency and steps' level difficulty. They demonstrated robust validation results from hybrid ensemble of logistic regression models and also discussed the scope of improved models with segmentation analysis.

Ramaswami and Bhaskaran (2010) presented the CHi-squared Automatic Interaction Detector (CHAID) prediction model, which was utilized to analyze the interrelation between variables that were used to predict the performance at higher secondary school education. The CHAID prediction model of student performance was constructed with seven class predictor variables. Their study showed that features, which constitute the strongest indicators, are marks in written assignments and tests, school location, living area, and the type of secondary education.

Independently, Ramesh, Parkav, and Rama (2013) tried to identify the factors influencing the students' performance in final examinations based on a dataset including questionnaire data and students' performance details. Their primary task was identifying the essential predictive variables, which affect the performance of higher secondary students, predict the grade at higher examinations, and determine the best classification algorithm. Their comparative study revealed that parent's occupation and possibly financial status plays a major role in the students' performance. Furthermore, their numerical experiments showed that the multilayer perceptron exhibited the best classification accuracy.

Livieris et al. (2012) introduced a software tool for predicting the students' performance in the course of "Mathematics" of the first year of Lyceum. The proposed software is based on a neural network classifier, which exhibits more consistent behavior and illustrates better accuracy than the other classifiers. Along this line, Livieris et al. (2016) presented a user-friendly decision support software for predicting students' performance, together with a case study concerning the final examinations in Mathematics. Their proposed tool is based on a hybrid predicting system, which combines four learning algorithms utilizing a simple voting scheme. In more recent works, Livieris et al. (2017) presented an updated version, which is based on a novel two-level classification algorithm, which achieves much better classification performance than any single classifier. The motivation and the primary task of their works was to support the academic task of successfully predicting the students' performance in the final examinations of the school year. Based on their preliminary results and on the comments made by the high school educators, the authors concluded that the application of data mining can provide significant insights into student progress and performance.

Recently, semi-supervised methods have been applied to predict the student's future progression and identity their characteristics, which induce their behavior and performance. More specifically, Kostopoulos et al. (2015) examined the effectiveness of semi-supervised methods for predicting students' performance in distance higher education. Several experiments were conducted using a variety of semi-supervised learning algorithms compared with well-known supervised methods, which revealed some very promising results, especially the self-training and the tri-training algorithm. Based on the previous works, Kostopoulos et al. (2017) examined and evaluated the effectiveness of SSL algorithms for the prognosis of high school students' grade in the final examinations at the end of the school year. Their numerical experiments demonstrated the efficiency of semi-supervised methods compared to familiar supervised methods.

Proposed Methodology

The motivation for this study is to develop a methodology for predicting the students' performance in the final examinations of A′ Lyceum, exploiting the effectiveness of semi-supervised methods. Apparently, this methodology is not restricted to A′ Lyceum but extends to any final examinations. For this purpose, we propose the following methodology which consists of three stages.

The first stage of the proposed methodology concerns the data collection and data preparation for this research. In the next stage, we present our proposed ensemble-based SSL algorithm. Finally, in the third stage, we compare our proposed ensemble-based semi-supervised algorithm with the most popular SSL algorithms by conducting a series of tests.

Data Collection and Preparation

In this study, we have utilized a dataset concerning the performance of 799 students in courses of "Mathematics" which have been collected by the *Microsoft showcase school* "Avgoulea-Linardatou" during the years 2012–2016. At this point, we recall that we have selected the course of "Mathematics" since it has been characterized as the most significant and most difficult course of the Science direction. Table 2.1 presents eleven (11) attributes, which characterize the performance of each student in each class of the first 4 years of high school. They are based on several written assignments and frequent oral questions, which assess students' understanding of important mathematical concepts and topics daily.

Table 2.1 Attributes description for each class

Attribute	Type	Values
Oral grade of the first semester	Integer	[0, 20]
Grade of the first test of the first semester	Real	[0, 20]
Grade of the second test of the first semester	Real	[0, 20]
Grade of the final examination of the first semester	Real	[0, 20]
Grade of the first semester	Integer	[0, 20]
Oral grade of the second semester	Integer	[0, 20]
Grade of the first test of the second semester	Real	[0, 20]
Grade of the second test of the second semester	Real	[0, 20]
Grade of the final examination of the second semester	Real	[0, 20]
Grade of the second semester	Integer	[0, 20]
Grade in the final examinations	Ordinal	"Fail," "good," "Very good," "excellent"

The first 10 values are time-variant attributes and refer to the students' performance on both academic semesters, utilizing a 20-point grading scale, where 0 is the lowest grade and 20 is the perfect score. Many related studies have shown that such attributes have a significant impact in students' success in the examinations (Cortez & Silva, 2008; Livieris et al., 2012, 2016; Ramaswami & Bhaskaran, 2010). The assessment of students during the academic year consists of oral examination, two 15-min pre-warned tests, a 1-h exam, and the overall semester performance of each student in the first and second semester. The 15-min tests include multiple-choice questions and short-answer problems, while the 1-h exams include several theory and multiple-choice questions, as well as a variety of difficult mathematical problems requiring arithmetic skills, solving techniques, and critical analysis. The overall semester performance of each student addresses the personal engagement of the student in the course and his progress. Finally, the last attribute concerns the students' performance in the final examinations (2-h exam) utilizing a four-level classification, according to the classification scheme used in students' performance evaluation in the Greek schools, namely:

- "Fail" stands for student's performance between 0 and 9.
- "Good" stands for student's performance between 10 and 14.
- "Very good" stands for student's performance between 15 and 17.
- "Excellent" stands for student's performance between 18 and 20.

Figure 2.1 presents the class distribution which depicts the number of students who are classified as "Fail" (178 instances), "Good" (202 instances), "Very good" (178 instances), and "Excellent" (241 instances).

Furthermore, similar to Livieris et al. (2012, 2016, 2017), since it is of great importance to predict students' performance at the final examination of A′ Lyceum as soon as possible, two datasets have been created based on the attributes presented in Table 2.1:

- $DATA_1$: It contains the attributes which concern the students' performance in A′, B′, and C′ Gymnasium (3 × 11 attributes + class).

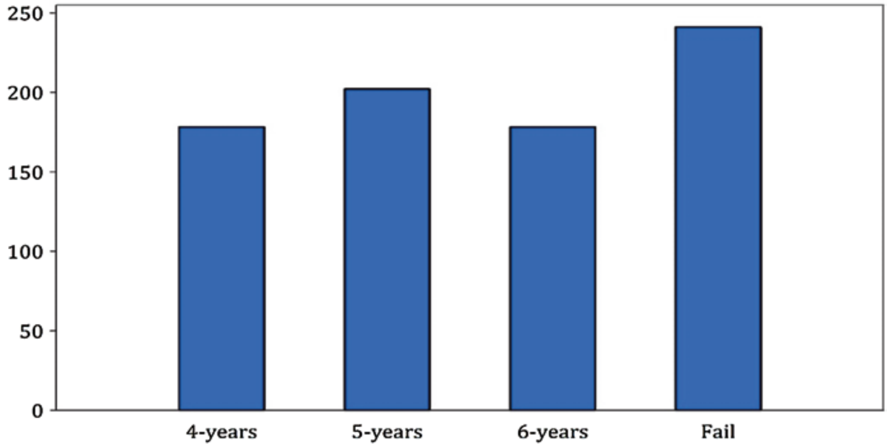

Fig. 2.1 Class distribution

- $DATA_2$: It contains the attributes which concern the students' performance in A′, B′, and C′ Gymnasium and A′ Lyceum (3×11 attributes + 10 attributes+ class).

The Proposed Ensemble-Based Semi-supervised Classifier

Our goal is to develop a classifier with strong classification ability by hybridization of ensemble learning and semi-supervised learning. We recall that SSL algorithms could be useful to ensemble learning algorithms since unlabeled data can enhance the diversity of individual classifiers and the lack of labeled examples can be exploited by utilizing unlabeled ones. Furthermore, ensemble learning methodologies could assist SSL since the combination of classifiers could be more accurate than an individual classifier and the performance of the ensemble classifier could be significantly improved using unlabeled data.

On the basis of this idea, we consider utilizing an ensemble of classifiers as a single base learner, instead of a single classifier, in each SSL algorithm. Generally, the development of an ensemble of classifiers consists of two steps: selection and combination. The selection of the appropriate component classifiers is considered to be an essential step towards obtaining highly accurate classifier systems (Zhou, 2011). A commonly used approach is to generate an ensemble of classifiers by applying diverse learning algorithms (with heterogeneous model representations) to a single dataset (see Merz, 1997, 1999; Todorovski & Džeroski, 2002). Furthermore, the combination of the individual predictions of learning algorithms takes place through several methodologies (see Dietterich, 2001; Re & Valentini, 2012; Rokach, 2010).

In this regard, our proposed ensemble-based classifier combines the individual predictions of three learning algorithms via a simple majority voting; hence the ensemble output is the one made by more than half of them. This selection constitutes the simplest and easiest implementation methodology for combining the individual predictions of component classifiers. The advantages of this technique are that it exploits the diversity of the errors of the learned models by utilizing different learning algorithms (Merz, 1997, 1999) and it does not require training on large quantities of representative recognition results from the individual classifiers. Moreover, several studies have reported that majority voting usually exhibits very good classification performance, developing highly accurate classifiers (Lam & Suen, 1997; Livieris et al., 2016; Matan, 1996).

Table 2.2 presents a high-level description of our proposed scheme, which utilizes an ensemble-based learner in any SSL algorithm.

Experimental Results

In this section, we conduct a series of tests in order to evaluate the performance of the SSL algorithms self-training, co-training, and tri-training deploying the most popular supervised classifiers as base learners. The selected supervised classifiers are the Naive Bayes (NB) (Domingos & Pazzani, 1997), the multilayer perceptron (MLP) (Rumelhart, Hinton, & Williams, 1986), the sequential minimal optimization (SMO) (Platt, 1999), the logistic model tree (LMT) (Landwehr, Hall, & Frank, 2005), and the PART (Frank & Witten, 1998) as the representative of the classification rules. Finally, 3-NN algorithm was selected as instance-based learner (Aha,

Table 2.2 Ensemble-based semi-supervised learning algorithm

Input:	*D*—Initial training dataset
	R—Ratio of labeled instances along *D*
	C_i—User selected classifiers, $i = 1, 2, 3$
/* Initialization phase */	
1:	Set of labeled training instances *L*
2:	Set of unlabeled training instances *U*
3:	Set the ensemble-base classifier *E*, using majority vote of individual classifiers C_1, C_2, C_3
/* Training phase */	
4:	Repeat
5:	Train *E* as base learner on *L* using any SSL algorithm
6:	Apply *E* on the unlabeled data *U*
7:	Add selected newly labeled data from *U* to the training set *L*
8:	Until some stopping criterion is met
Output:	Use trained ensemble *E* to predict class labels of the test cases
Remarks:	In step 5, the selected SSL algorithm is one of self-training, co-training, and tri-training

1997). Several studies have shown that the above classifiers constitute some of the most effective and frequently utilized data mining algorithms (Wu et al., 2008).

The classification accuracy of all learning algorithms was evaluated utilizing the standard procedure called stratified tenfold cross-validation, i.e., the data were separated into folds so that each fold had the same distribution of grades as the entire dataset. Furthermore, the implementation code was written in JAVA, using WEKA Machine Learning Toolkit (Hall et al., 2009), and all the base learners were utilized with default parameter settings.

Tables 2.3, 2.4, and 2.5 present the classification performance of each test algorithm utilizing 10%, 20%, and 30%, respectively, as labeled data ratio, and the best accuracy among the different algorithms in each experiment is highlighted in bold style. The aggregated results presented in Tables 2.3, 2.4, and 2.5 show that LMT exhibits the best classification performance utilized as base classifier followed by SMO and PART, relative to all SSL algorithms.

Table 2.3 Comparison of accuracy of self-training algorithms

		Self-training algorithm					
Dataset	Ratio	(NB)	(MLP)	(SMO)	(LMT)	(PART)	(3NN)
$DATA_1$	10%	69.90%	72.88%	70.98%	**81.47**%	74.30%	69.05%
	20%	69.16%	73.65%	74.39%	**81.85**%	76.23%	71.01%
	30%	70.67%	72.54%	72.09%	**82.62**%	76.98%	67.99%
$DATA_2$	10%	76.31%	78.23%	**80.77**%	79.22%	79.30%	75.46%
	20%	77.08%	76.99%	78.19%	**81.51**%	79.69%	75.87%
	30%	77.46%	78.56%	77.41%	**78.83**%	73.99%	71.65%

Table 2.4 Comparison of accuracy of co-training algorithms

		Co-training algorithm					
Dataset	Ratio	(NB)	(MLP)	(SMO)	(LMT)	(PART)	(3NN)
$DATA_1$	10%	70.66%	72.11%	67.19%	**81.50**%	75.44%	70.24%
	20%	69.10%	73.33%	71.42%	**77.35**%	75.88%	69.47%
	30%	71.03%	71.74%	72.45%	**80.30**%	76.24%	67.65%
$DATA_2$	10%	75.61%	78.58%	76.30%	**78.50**%	76.99%	72.11%
	20%	75.94%	77.76%	73.21%	**79.19**%	76.65%	72.81%
	30%	75.19%	76.99%	75.53%	**80.36**%	77.41%	74.36%

Table 2.5 Comparison of accuracy of tri-training algorithms

		Tri-training algorithm					
Dataset	Ratio	(NB)	(MLP)	(SMO)	(LMT)	(PART)	(3NN)
$DATA_1$	10%	69.90%	70.68%	73.99%	78.19%	**78.39**%	68.38%
	20%	69.53%	70.64%	70.28%	**81.47**%	74.33%	70.60%
	30%	69.90%	73.29%	72.52%	**81.10**%	75.85%	72.05%
$DATA_2$	10%	76.32%	78.48%	**78.87**%	78.57%	76.24%	73.58%
	20%	76.71%	77.05%	78.87%	**79.60**%	79.26%	75.87%
	30%	75.20%	78.50%	77.74%	**81.17**%	79.27%	77.35%

Subsequently, we evaluate the performance of our proposed SSL algorithm, which utilizes an ensemble as base classifier (denoted as Vote). The ensemble-based learner combines the individual predictions of three classifiers (LMT, PART, and SMO) using majority vote. Notice that these classifiers have been selected since they exhibit the best classification performance, regarding both datasets. Moreover, the performance of the proposed algorithm is compared against the best reported performance of all base learners (denoted as Best) for each SSL algorithm. As before, the accuracy measure of the best performing algorithm is highlighted in bold for each base learner and on each dataset. Additionally, a more representative visualization of the classification performance of the compared base learners for each SSL algorithm is presented in Figs. 2.2, 2.3, and 2.4.

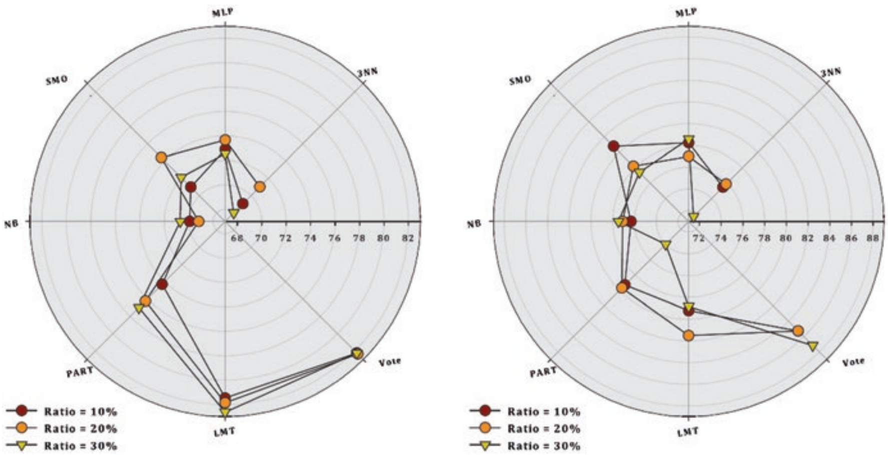

Fig. 2.2 Comparison of average accuracy of self-trained classifiers on $DATA_1$ and $DATA_2$

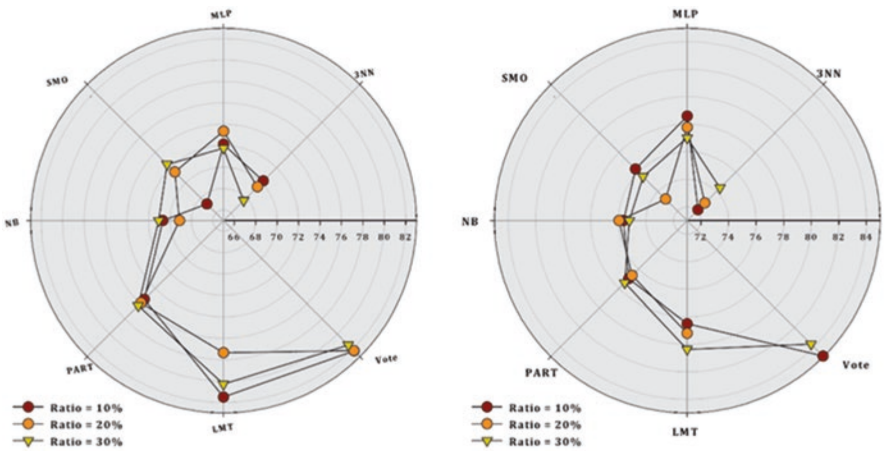

Fig. 2.3 Comparison of average accuracy of co-trained classifiers on $DATA_1$ and $DATA_2$

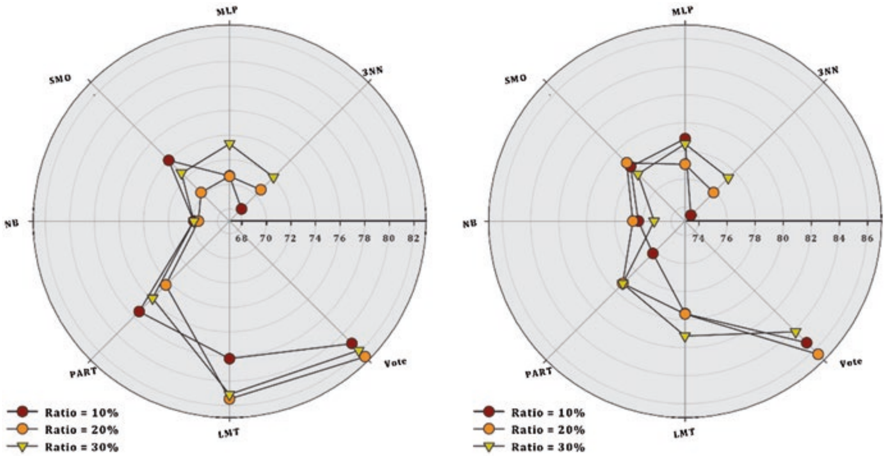

Fig. 2.4 Comparison of average accuracy of tri-trained classifiers on $DATA_1$ and $DATA_2$

Table 2.6 Comparison of accuracy of SSL algorithms

Dataset	Ratio	Self-training (Best)	Self-training (Vote)	Co-training (Best)	Co-training (Vote)	Tri-training (Best)	Tri-training (Vote)
$DATA_1$	10%	81.47%	**82.24%**	81.50%	**82.21%**	78.39%	**81.07%**
	20%	81.85%	**82.34%**	77.35%	**82.19%**	81.47%	**82.59%**
	30%	**82.62%**	82.24%	80.30%	**81.45%**	81.10%	**81.87%**
$DATA_2$	10%	80.77%	**85.26%**	78.50%	**84.93%**	78.87%	**85.24%**
	20%	81.51%	**85.24%**	79.19%	**85.66%**	79.60%	**86.40%**
	30%	78.83%	**87.15%**	80.36%	**83.70%**	81.17%	**84.13%**

The interpretation of Table 2.6 reveals that Vote presents by far the best classification results utilized as base classifier in all cases except the one when self-training algorithm utilized LMT as base learner with a labeled ratio of 30%. Furthermore, tri-training (Vote) and self-training (Vote) exhibit the best performance relative to $DATA_1$ and $DATA_2$, respectively. An interesting point, which is highlighted in Figs. 2.2, 2.3, and 2.4 is that all the SSL algorithms, which utilize Vote as base classifier, report similar classification results independent of the utilized ratio of labeled data and dataset, assuring their robust behavior.

The statistical comparison of multiple algorithms over multiple datasets is fundamental in machine learning, and usually it is typically carried out by means of a statistical test (Kostopoulos et al., 2015, 2017) Therefore, we utilized the non-parametric Friedman Aligned Ranking (Hodges & Lehmann, 1962) test in order to evaluate the rejection of the hypothesis that all the classifiers perform equally well for a given level. Since the test is non-parametric, it does not require commensurability of the measures across different datasets, it does not assume normality of the sample means, and it is robust to outliers.

Table 2.7 Friedman aligned ranks test (significance level of 0.05)

Self-training		Co-training		Tri-training	
Base learner	Friedman ranking	Base learner	Friedman ranking	Base learner	Friedman ranking
Vote	5.00	Vote	3.83	Vote	4.33
LMT	9.33	LMT	9.83	LMT	9.67
PART	18.67	PART	16.17	PART	17.00
SMO	24.00	MLP	21.83	SMO	24.17
MLP	24.17	SMO	30.83	MLP	26.67
NB	32.00	NB	30.83	NB	33.83
3NN	37.33	3NN	37.17	3NN	34.83

Table 2.7 presents the SSL algorithms ranked from the best performer to the worst. The proposed voting scheme illustrates statistically better classification results among all tested algorithms. More specifically, the base learner Vote reports the best performance due to better probability-based ranking and higher classification accuracy in all SSL algorithms.

Conclusions

In this work, we propose a new ensemble-based SSL method for predicting the students' performance in the final examinations at the end of academic year of A′ Lyceum. Our experimental results reveal that our proposed method is proved to be effective and practical for early student progress prediction as compared to some existing semi-supervised learning methods. Our objective and expectation is that this work could provide prognosis for better educational support by offering customized assistance according to students' predicted performance and be used as a reference for decision-making in the admission process.

Acknowledgments The authors are grateful to the private high school "Avgoulea-Linardatou" for the collection of the data used in our study and valuable comments which essentially improved our work.

References

Aha, D. (1997). *Lazy learning*. Dordrecht: Kluwer Academic Publishers.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *11th Annual Conference on Computational Learning Theory* (pp. 92–100).

Bousbia, N., & Belamri, I. (2014). Which contribution does EDM provide to computer-based learning environments? In *Educational data mining* (pp. 3–28). Berlin: Springer.
Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning. *IEEE Transactions on Neural Networks, 20*(3), 542–542.
Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference* (pp. 5–12).
Dietterich, T. (2001). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems* (Vol. 1857, pp. 1–15). Berlin: Springer.
Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. In *Machine learning* (Vol. 29, pp. 103–130).
Du, J., Ling, C., & Zhou, Z. (2011). When does co-training work in real data? *IEEE Transactions on Knowledge and Data Engineering, 23*(5), 788–799.
Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
Frank, E., & Witten, I. (1998). Generating accurate rule sets without global optimization. In *15th International Conference on Machine Learning* (pp. 144–151).
Gandhi, P., & Aggarwal, V. (2010). Ensemble hybrid logit model. In *Proceedings of the KDD 2010 Cup: Workshop Knowledge Discovery in Educational Data* (pp. 33–50).
Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society, 15*(3), 42.
Guo, T., & Li, G. (2012). Improved tri-training with unlabeled data. In *Software engineering and knowledge engineering: Theory and practice* (pp. 139–147). Berlin: Springer.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletters, 11*, 10–18.
Hodges, J., & Lehmann, E. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics, 33*(2), 482–497.
Kostopoulos, G., Kotsiantis, S., & Pintelas, P. (2015). Predicting student performance in distance higher education using semi-supervised techniques. In *Model and data engineering* (pp. 259–270). Berlin: Springer.
Kostopoulos, G., Livieris, I., Kotsiantis, S., & Tampakas, V. (2017). Enhancing high school students' performance prediction using semi-supervised methods. In *8th International Conference on Information, Intelligence, Systems and Applications (IISA 2017)*. Piscataway: IEEE.
Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems, 23*(6), 529–535.
Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-based intelligent information and engineering systems* (pp. 267–274). Berlin: Springer.
Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18*(5), 411–426.
Lam, L., & Suen, S. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27*(5), 553–568.
Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59*(1–2), 161–205.
Levatic, J., Dzeroski, S., Supek, F., & Smuc, T. (2013). Semi-supervised learning for quantitative structure-activity modeling. *Informatica, 37*(2), 173.
Liu, C., & Yuen, P. (2011). A boosted co-training algorithm for human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(9), 1203–1213.
Livieris, I., Drakopoulou, K., Kotsilieris, T., Tampakas, V., & Pintelas, P. (2017). DSS-PSP – A decision support software for evaluating students' performance. In *Engineering applications of neural networks* (pp. 63–74). Berlin: Springer.
Livieris, I., Drakopoulou, K., & Pintelas, P. (2012). Predicting students' performance using artificial neural networks. In *Information and communication technologies in education* (pp. 321–328).

Livieris, I., Mikropoulos, T., & Pintelas, P. (2016). A decision support system for predicting students' performance. *Themes in Science and Technology Education, 9*, 43–57.

Matan, O. (1996). On voting ensembles of classifiers. In *Proceedings of AAAI-96 Workshop on Integrating Multiple Learned Models* (pp. 84–88).

Merz, C. (1997). Combining classifiers using correspondence analysis. In *Advances in neural information processing systems* (pp. 592–597).

Merz, C. (1999). Using correspondence analysis to combine classifiers. *Machine Learning, 36*, 33–58.

Ng, V., & Cardie, C. (2003). Weakly supervised natural language learning without redundant views. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (Vol. 1, pp. 94–101). Stroudsburg: Association for Computational Linguistics.

Pena-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432–1462.

Pise, N., & Kulkarni, P. (2008). A survey of semi-supervised learning methods. In *Proceedings of the 2008 International Conference on Computational Intelligence and Security* (Vol. 2, pp. 30–34). Washington, DC: IEEE Computer Society.

Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems* (pp. 557–563). Cambridge, MA: MIT Press.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues, 7*(1), 135–146.

Ramesh, V., Parkav, P., & Rama, K. (2013). Predicting student performance: A statistical and data mining. *International Journal of Computer Applications, 63*(8), 35–39.

Re, M., & Valentini, G. (2012). Ensemble methods: A review. In *Advances in machine learning and data mining for astro-nomy* (pp. 563–594). Boca Raton: CRC Press.

Rokach, L. (2010). *Pattern classification using ensemble methods*. Singapore: World Scientific Publishing Company.

Roli, F., & Marcialis, G. (2006). Semi-supervised PCA-based face recognition using self-training. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 560–568).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135–146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 40*(6), 601–618.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.

Schwenker, F., & Trentin, E. (2014). Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters, 37*, 4–14.

Sigdel, M., Dinç, I., Dinç, S., Sigdel, M., Pusey, M., & Aygun, R. (2014). Evaluation of semi-supervised learning for classification of protein crystallization imagery. In *Southeastcon 2014* (pp. 1–6). IEEE.

Sun, S., & Jin, F. (2011). Robust co-training. *International Journal of Pattern Recognition and Artificial Intelligence, 25*(07), 1113–1126.

Thai-Nghe, N., Busche, A., & Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. In *9th International Conference on Intelligent Systems Design and Applications (ISDA'09)* (pp. 878–883).

Thai-Nghe, N., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Proceeding of 37th IEEE Frontiers in Education Conference* (pp. 7–12).

Todorovski, L., & Džeroski, S. (2002). Combining classifiers with meta decision trees. *Machine Learning, 50*(3), 223–249.

Triguero, I., & Garcıa, S. (2015). Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowledge and Information Systems, 42*(2), 245–284.

Triguero, I., Saez, J., Luengo, J., Garcia, S., & Herrera, F. (2014). On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing, 132*, 30–41.

Wang, Y., & Chen, S. (2013). Safety-aware semi-supervised classification. *IEEE Transactions on Neural Networks and Learning Systems, 24*(11), 1763–1772.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37.

Zhou, Z. (2011). When semi-supervised learning meets ensemble learning. In *Frontiers of electrical and electronic engineering in China* (Vol. 6, pp. 6–16). Berlin: Springer.

Zhu, X. (2006). *Semi-supervised learning literature survey* (Technical Report 1530). Madison: University of Wisconsin.

Zhu, X. (2011). Semi-supervised learning. In *Encyclopedia of machine learning* (pp. 892–897). Berlin: Springer.

Zhu, X., & Goldberg, A. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*(1), 1–130.