Forecasting students' performance using an ensemble SSL algorithm

Ioannis E. Livieris
Department of Computer Engineering & Informatics,
Technological Educational Institute of Western Greece,
Antirrio, Greece, GR 263-34.
Email: livieris@teiwest.gr

Vassilis Tampakas
Department of Computer Engineering & Informatics,
Technological Educational Institute of Western Greece,
Antirrio, Greece, GR 263-34.
Email: vtampakas@teimes.gr

Niki Kyriakidou
Department of Mathematics,
University of Patras,
Patras, Greece, GR 265-00.
Email: kyriakidou@upatras.gr

Tasos Mikropoulos
The Educational Approaches to Virtual Reality Technologies Lab,
University of Ioannina, Greece.
Email: amikrop@uoi.gr

Panagiotis Pintelas
Department of Mathematics,
University of Patras,
Patras, Greece, GR 265-00.
Email: ppintelas@gmail.com

***Abstract*—Educational data mining is a growing academic research area which aims to gain significant insights on student behavior, interactions and performance by applying data mining methods on educational data. During the last decades, a variety of accurate models has been developed to monitor students' future progress, while most of these studies are based on supervised classification methods. In this work, we propose an ensemble semi-supervised algorithm for the prediction of students' performance in the final examinations at the end of academic year. The experimental results demonstrate the efficiency and robustness of the proposed algorithm compared to some classical classification algorithms, in terms of accuracy.**

I. Introduction

Educational Data Mining (EDM) is an essential process where intelligent methods are applied to extract knowledge hidden in data of students records in order to provide significant insights on student behavior and assist educational decision making support. The importance of EDM is founded on the fact that it allows to extract useful conclusions from sophisticated and complicated questions such as "*find the students who are at-risk in failing the examinations*" or "*find the students who will exhibit excellent performance*" in which traditional database queries cannot be applied [22]. Therefore, EDM is mainly concentrated on the development of accurate models that predict students' characteristics and performance, offering opportunities and great potentials to increase our understanding about the learning processes and students' behavior.

Secondary education in Greece takes place after six years of primary education and may be followed by higher education. It comprises of two main stages: *Gymnasium* and *Lyceum*. Gymnasium covers the first three years with the purpose to support the development of composite and critical thinking and enrich students' knowledge in all fields of learning. Lyceum covers the following three years aiming in further cultivating the students' personalities while at the same time prepares them for the National examinations in order to proceed to higher education. Essentially, Lyceum acts like a bridge between school education and higher learning specializations that are offered by universities [20]–[22]. In the end of the 1st grade of Lyceum (A' Lyceum) the students are obligated to select between three possible directions: *Humanity*, *Science* and *Technology* in order to establish the courses which the students will attend in the National examinations. Clearly, this selection constitutes a significant and decisive factor in the life of any student since it defines their future entry in a specific higher academic institution.

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. More specifically, the "knowledge discovery" can assist students to have a first evaluation of their progress and possibly enhance their performance and teachers to identify slow learners and learning difficulties. Hence, it is of major importance to closely monitor the students' performance in order to identify possible retardation and proactively intervene towards their academic enhancement. Nevertheless, the early identification of students at risk of exhibiting 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 [18], [20]–[23].

Over the last decades, educational institutes have managed to accumulate a large amount of data about their students. Machine learning and data mining techniques constitute a significant prediction tool, offering a first step and a helping hand in analyzing and exploiting the knowledge acquired from students' records. In this context, many researchers in the past have conducted studies on educational data in order to cluster students based on their academic performance. Nevertheless, most of these studies examine the efficiency of supervised classification methods, while ensemble methods [11], [20], [22] and semi-supervised methods [15]–[17] have been rarely applied to the educational field. Generally, ensemble methods and semi-supervised methods are two important machine learning techniques. The former attempt to built powerful and accurate predictive models by using multiple learners while the latter attempt to achieve strong generalization by exploiting unlabeled data. 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 [41] presented that ensemble learning algorithms and semi-supervised learning algorithms are indeed beneficial to each other and more efficient classification models can be developed by a combination of diverse classifiers and by leveraging unlabeled data. More analytically, ensemble algorithms could assist semi-supervised algorithms since the combination of classifiers could be more accurate than an individual classifier [7] while semi-supervised algorithms could be useful to ensemble algorithms since unlabeled data can significantly enhance the diversity of a classifier [41], [44].

In this work, we propose a new ensemble 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. The proposed ensemble algorithm combines the predictions of three semi-supervised algorithms, based on a voting methodology. The efficiency of the proposed algorithm is evaluated in terms of classification accuracy using several base learners, while our experiments illustrate its efficacy.

The remainder of this paper is organized as follows: Section II reviews some recent studies of data mining applications in education while Section III presents a brief discussion of the semi-supervised learning algorithms. Section IV presents our proposed ensemble semi-supervised learning algorithm and Section V presents the educational dataset utilized in our study. In Section VI, we present a series of tests in order to evaluate the accuracy of proposed algorithm with the most popular classification algorithms. Finally, the last section considers the conclusions and some proposal for future research.

II. Related work

During the last decade, the application of machine learning and data mining for the development of efficient and accurate prediction models for monitoring students' behavior and future performance has become very popular in modern educational era. Some excellent reviews [3], [28], [35] have described in detail the most accurate models and developments utilized for gaining significant insights on students' behavior, interactions and progress and summarized the diverse factors which influence students' future performance. A number of rewarding studies have being carried out in recent years and some of them are presented below:

Cortez and Silva [6] conducted a performance study on secondary school students for two core classes (Mathematics and Portuguese). The data were extracted from school records, as well as provided by the students through questionnaires. Four classification algorithms were applied on three data setups, with different combinations of attributes, trying to find out those effecting on the prediction. Based on their numerical experiments, the authors concluded that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Moreover, they stated that students achievements are highly influenced by past evaluations and in some cases there are other relevant features such as social and cultural characteristics of the students which affect students performance.

Ramesh et al. [32] tried to predict the grade of higher secondary students in the examinations and identify the essential predictive variables which affect the performance. Their motivation consists of determine the best classification model and identify the factors influencing the students' performance in final examinations based on a dataset including questionnaire data and students' performance details. Their numerical experiments showed that neural networks exhibited the best classification accuracy. Furthermore, their comparative study revealed that parent's occupation and possibly financial status plays a significant role in students' performance.

Marquez-Vera et al. [24] studied the serious problem of students failure utilizing data from 670 first year high school Mexican students. Firstly, they applied feature selection techniques to detect the factors that most influence student failure and then they rebalanced the data and applied cost sensitive classification in order to resolve the problem of classifying imbalanced data. Additionally, they proposed a genetic programming model to obtain accurate and comprehensible classification rules for predicting the academic status or final student performance at the end of the course. Their experimental results presented that feature selection, cost-sensitive classification and data balancing can also be very useful for improving the classification accuracy.

Recently, Kostopoulos et al. [15], [16] 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 interesting results. Based on the previous works, Kostopoulos et al. [17] 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.

Livieris et al. [22] presented a decision support software for predicting high school students' performance, together with a case study concerning the final examinations in course of Mathematics. Their proposed software is based on a hybrid predicting system utilizing a simple voting scheme combining the individual predictions of four individual learning algorithms. Along this line, in [20] the authors introduced an updated version of their software which is based on a novel 2-level classification algorithm. Their numerical experiments reveal that their proposed algorithm identifies the students who are at-risk of failing in the examinations and classifies the students who have successfully passed the course with high accuracy. 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. In more recent works, Livieris et al. [21] applied semi-supervised learning methods to predict the student's future progression and identity their characteristics which induce their performance. Based on their preliminary results, the authors concluded that the application of semi-supervised learning methods on educational data can provide significant insights into students' progress and performance.

III. Semi-supervised learning

Semi-supervised learning (SSL) constitutes an amalgamation of supervised and unsupervised learning. Compared to traditional classification approaches, SSL utilizes large amount of unlabeled samples together with labeled samples to build an efficient and accurate classifier. Since the acquisition of sufficient labeled samples is cumbersome and expensive and frequently requires the efforts of domain experts, SSL has been established as a powerful and effective machine learning technique. The general assumption of SSL algorithms is to leverage the large amount of unlabeled data in order to reduce data sparsity in the labeled training data and boost the classifier performance, particularly focusing on the setting where the amount of available labeled data is limited. Hence, these methods have the ability of reducing the supervision to a minimum, while still preserving competitive and sometimes better classification performance (see [4], [12], [17], [19], [21], [42], [43] and the references therein)

In the literature, many SSL algorithms have been proposed with different philosophy on exploiting the information hidden in the unlabeled data. Self-training, Co-training and Tri-training constitute the most representative and commonly utilized from this class of algorithms.

Self-training is generally considered as the simplest and one of the most efficient SSL algorithms. This algorithm is a wrapper based SSL approach which constitutes an iterative procedure of self-labeling unlabeled data. According to Ng and Cardie [27] "*self-training is a single-view weakly supervised algorithm*" which is based on its own predictions on unlabeled data to teach itself. Firstly, an arbitrary classifier is initially trained with a small amount of labeled data, constituting its training set which is iteratively augmented using its own most confident predictions of the unlabeled data. More analytically, each unlabeled instance which has achieved a probability over a specific threshold c is considered sufficiently reliable to be added to the labeled training set and subsequently the classifier is retrained.

Clearly, the success of Self-training is heavily depended on the newly-labeled data based on its own predictions, hence its weakness is that erroneous initial predictions will probably lead the classifier to generate incorrectly labeled data [44].

Co-training is a SSL algorithm which utilizes two classifiers, each trained on a different view of the labeled training set. The underlying assumptions of the Co-training approach is that feature space can be split into two different conditionally independent views and that each view is able to predict the classes perfectly [9], [39]. Under these assumptions, two classifiers are trained separately for each view using the initial labeled set and then iteratively the classifiers augment the training set of the other with the most confident predictions on unlabeled examples.

In essence, Co-training is a "*two-view weakly supervised algorithm*" since it uses the self-training approach on each view [27]. The efficacy of Co-training has been extensively studied by Blum and Mitchell [4] and they concluded that if the two views are conditionally independent, then the use of unlabeled data can significantly improve the predictive accuracy of a weak classifier. Nevertheless, the assumption about the existence of sufficient and redundant views is a luxury hardly met in most real world scenarios.

Tri-training [42] consists of an improved version of Co-training which overcomes the requirements for multiple sufficient an redundant feature sets. This algorithm is a bagging ensemble of three classifiers, trained on the data subsets generated through bootstrap sampling from the original labeled training set. In case two of the classifiers agree on a prediction, then they label the unlabeled example with their prediction and augment the third classifier with the newly labeled example. The efficiency of the training process is based on the "*majority teach minority strategy*" which serves as an implicit confidence measurement, avoiding thereby the use of a complicated time consuming approach to explicit measure the predictive confidence.

In contrast to several SSL algorithms, Tri-training does not require different supervised algorithms as base learners which leads to greater applicability in many real world classification problems [12], [17], [44].

IV. An ensemble SSL algorithm

We recall that our main goal is to develop a classifier with high classification accuracy by the hybridization of ensemble learning and semi-supervised learning. Generally, the development of an ensemble of classifiers consists of two steps: *selection* and *combination*.

The selection of the appropriate component classifiers is considered essential for the efficacy of the ensemble and the key points for its effectiveness is based on the accuracy and the diversity of the component classifiers [41]. A commonly and widely utilized approach is to generate an ensemble of classifiers by applying diverse learning algorithms (with heterogeneous model representations) to a single dataset (see [25], [26], [40]). Furthermore, the combination of the individual predictions of learning algorithms takes place through several methodologies (see [7], [33], [34]) with different philosophy and classification performance.

On this basis, the learning algorithms which constitute the proposed ensemble are: Co-training, Self-training and Tri-training SSL algorithms. These methods are self-labeled methods trying to exploit the hidden information in unlabeled data with complete different way since Co-training is a multi-view method, while Self-training and Tri-training are single-view methods.

Moreover, our proposed ensemble-based classifier combines the individual predictions of the three SSL algorithm via a maximum-probability voting. This combination strategy is considered as the simplest and easiest implementation methodology for combining the individual predictions of component classifiers. Notice that in case the confidence of the prediction of the selected classifier does not meet a predefined threshold, then the prediction is not considered reliable enough. In this case, the output is defined as the combined predictions of three SSL learning algorithms via a simple majority voting, namely the ensemble output is the one made by more than half of them.

An obvious advantages of the utilized combination technique is that it exploits the diversity of the errors of the learned models by utilizing different learning algorithms [25], [26] and it does not require training on large quantities of representative recognition results from the individual classifiers.

Subsequently, we present a high-level description of our proposed Ensemble Semi-Supervised Learning (En-SSL) algorithm.

`Algorithm 1`: **En-SSL**

Input: L - Set of labeled training instances.
U - Set of unlabeled training instances.
Parameters: *ThresLev* - Threshold level.

Output: The labels of instances in the testing set.

/* Training phase */
[1]: Train Self-training(L, U) classifier.
[2]: Train Co-training(L, U) classifier.
[3]: Train Tri-training(L, U) classifier.

/* Testing phase */
[1]: for each x from test set
[2]: Apply Self-training, Co-training, Tri-training classifiers on x.
[3]: Find the classifier SSL* with the highest confidence prediction on x.

[4]: if (Confidence of SSL* $\geq$ *ThresLev*)
[5]: SSL* predicts the label y^* of x.
[6]: else
[7]: Use majority vote to predict the label y^* of x.
[8]: end if
[9]: end for

V. Dataset

The dataset utilized in our study has been provided by the Microsoft showcase high school "*Avgoulea-Linardatou*". For a time period of four years (2012-2015), data of 630 students have been collected concerning the course of "Mathematics".

During the academic year, the educators are required to use a variety of assessment methods including oral examinations, tests, written assignments and exams while the students are obliged to attend the final examinations at the end of the academic year. It is worth noticing that the final exam is marked out of 20 and is of prime importance to the overall final grade.

Table I presents eleven (11) time-variant attributes which characterize the performance of each student in each class of the first four years of high school. The selected attributes 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. The assessment of students during the academic year consists of oral examination, two 15-minutes pre-warned tests, an 1-hour exam, the overall semester performance of each student in the 1st and 2nd semester and his/her performance at the final examinations.

The oral grade is defined by several written assignments and frequent oral questions, evaluating students' understanding of basic mathematical terms and concepts daily. The 15-minutes tests include multiple choice questions and short answer problems. The 1-hour exams cover a wide range of the curricula and include several theoretical and multiple choice questions, as well as a variety of problems requiring arithmetic skills 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-hour exam). Many related studies have shown that such attributes assess students' understanding of important mathematical concepts and topics daily and have a significant impact in students' success in the examinations [6], [22], [23], [31].

Attribute	Type	Values
Oral grade of the 1st semester	integer	$[0, 20]$
Grade of the 1st test of the 1st semester	real	$[0, 20]$
Grade of the 2nd test of the 1st semester	real	$[0, 20]$
Grade of the final examination of the 1st semester	real	$[0, 20]$
Grade of the 1st semester	integer	$[0, 20]$
Oral grade of the 2nd semester	integer	$[0, 20]$
Grade of the 1st test of the 2nd semester	real	$[0, 20]$
Grade of the 2nd test of the 2nd semester	real	$[0, 20]$
Grade of the final examination of the 2nd semester	real	$[0, 20]$
Grade of the 2nd semester	integer	$[0, 20]$
Grade in the final examinations	real	$[0, 20]$

TABLE I
ATTRIBUTES DESCRIPTION FOR EACH CLASS

Furthermore, since the early prediction of the students' performance at the final examination of A' Lyceum is of great importance, similar to [20]–[23] we have created two datasets based on the attributes presented in Table I.

- DATA$_{G}$: It contains the attributes which concern the students' performance in A', B' and C' Gymnasium (33 attributes + class).
- DATA$_{GL}$: It contains the attributes which concern the students' performance in A', B' and C' Gymnasium and A' Lyceum (43 attributes + class).

Finally, the students' were classified based on the performance in the final examinations of A' Lyceum (2-hour exam) utilizing the following four-level classification: 0-9 (poor), 10-14 (good), 15-17 (very good), 18-20 (excellent) as in [20]–[23]. This classification scheme also used in students' performance evaluation in the Greek schools.

VI. Experimental results

In this section, the classification performance of the proposed algorithm was compared to its component SSL algorithms and in particular Self-training, Co-training and Tri-training in terms of classification accuracy. Accuracy is one of the most frequently used measures for assessing the overall effectiveness of a classification algorithm [38] and it is defined as the percentage of correctly classified instances. Furthermore, the most popular and commonly used supervised algorithms were deployed as base learners: Naive Bayes (NB) [8], Multilayer Perceptron (MLP) [37], Sequential Minimum Optimization (SMO) [29], C4.5 decision tree algorithm [30], JRip [5] as a typical rule-learning algorithm and 3-NN [1].

The classification accuracy of all learning algorithms was evaluated utilizing the standard procedure called stratified 10-fold cross-validation i.e. the data was separated into folds so that each fold had the same distribution of grades as the entire data set. Moreover, the implementation code was written in JAVA, using the WEKA Machine Learning Toolkit [13]. The configuration parameters for all SSL algorithms used in our experiments are presented in Table II. Regarding the base learners, the default parameter settings included in the WEKA software were utilized in order to minimize the effect of any expert bias by not attempting to tune any of the algorithms to the specific datasets. In order to study the influence of the amount of labeled data, three different ratios of the training data were used: 10%, 20% and 30%.

Algorithm	Parameters
Self-training	MaxIter $= 40$. $c = 95\%$.
Co-training	MaxIter $= 40$. Initial unlabeled pool $= 75$.
Tri-training	No parameters specified.
En-SSL	*ThresLev* $= 95\%$.

TABLE II
Parameter specification for all the SSL methods employed in our experiments

Tables III-VIII present the classification performance of each SSL algorithm regarding all base learners and the best accuracy among the different algorithms in each experiment is highlighted in bold style. Additionally, a more representative visualization of the average classification performance of the compared SSL algorithms is presented in Figures 1 and 2. Despite the ratio of labeled instances, En-SSL algorithm presents by far the best classification results, outperforming all SSL algorithms, relative to both datasets. It is worth noticing that our proposed algorithm exhibits the best classification accuracy utilizing JRip and C4.5 as base learners. Furthermore, the interpretation of Figures 1 and 2 reveal that En-SSL illustrates the best average classification accuracy, significantly outperforming all SSL algorithms, regarding both datasets.

Dataset	Ratio	Tri-training (NB)	Co-training (NB)	Self-training (NB)	En-SSL (NB)
DATA_G	10%	69.80%	69.40%	**70.57%**	**70.57%**
	20%	**70.56%**	69.42%	70.20%	70.19%
	30%	**70.93%**	70.19%	70.19%	70.56%
DATA_{GL}	10%	77.01%	**77.60%**	77.01%	77.01%
	20%	**77.76%**	77.10%	77.01%	**77.76%**
	30%	77.39%	77.10%	**77.41%**	77.02%

TABLE III
Comparison of SSL algorithms using NB as base learner

Dataset	Ratio	Tri-training (MLP)	Co-training (MLP)	Self-training (MLP)	En-SSL (MLP)
DATA_G	10%	76.45%	78.15%	77.79%	**78.56%**
	20%	74.40%	74.44%	76.30%	**77.45%**
	30%	72.89%	73.28%	75.53%	**77.05%**
DATA_{GL}	10%	77.76%	68.72%	**78.92%**	78.56%
	20%	78.96%	77.41%	77.81%	**80.84%**
	30%	75.87%	75.98%	**78.93%**	**78.93%**

TABLE IV
Comparison of SSL algorithms using MLP as base learner

Dataset	Ratio	Tri-training (SMO)	Co-training (SMO)	Self-training (SMO)	En-SSL (SMO)
DATA_G	10%	71.27%	**73.16%**	64.57%	69.42%
	20%	71.30%	68.35%	69.83%	**71.35%**
	30%	**72.81%**	68.35%	70.20%	71.32%
DATA_{GL}	10%	79.59%	80.03%	75.81%	**80.73%**
	20%	77.36%	75.50%	72.75%	**78.11%**
	30%	79.56%	78.50%	73.90%	**80.70%**

TABLE V
Comparison of SSL algorithms using SMO as base learner

Dataset	Ratio	Tri-training (C4.5)	Co-training (C4.5)	Self-training (C4.5)	En-SSL (C4.5)
DATA_G	10%	78.13%	78.16%	77.41%	**78.89%**
	20%	78.53%	77.76%	**79.29%**	77.78%
	30%	**77.36%**	73.28%	**77.36%**	75.85%
DATA_{GL}	10%	81.20%	81.14%	77.02%	**81.44%**
	20%	79.29%	80.01%	79.30%	**80.04%**
	30%	81.20%	77.46%	81.89%	**83.05%**

TABLE VI
Comparison of SSL algorithms using C4.5 as base learner

Dataset	Ratio	Tri-training (JRip)	Co-training (JRip)	Self-training (JRip)	En-SSL (JRip)
$DATA_G$	10%	78.46%	78.82%	80.34%	**82.64%**
	20%	77.39%	78.11%	75.87%	**80.77%**
	30%	77.36%	79.67%	77.35%	**79.99%**
$DATA_{GL}$	10%	77.72%	79.96%	78.92%	**81.51%**
	20%	78.95%	78.49%	79.66%	**81.92%**
	30%	81.10%	80.33%	80.37%	**81.88%**

TABLE VII
COMPARISON OF SSL ALGORITHMS USING JRIP AS BASE LEARNER

Dataset	Ratio	Tri-training (3NN)	Co-training (3NN)	Self-training (3NN)	En-SSL (3NN)
$DATA_G$	10%	68.38%	70.24%	69.05%	**71.35%**
	20%	70.60%	69.47%	71.77%	**72.11%**
	30%	**72.05%**	67.65%	67.62%	71.74%
$DATA_{GL}$	10%	73.58%	72.11%	**75.46%**	74.73%
	20%	75.87%	72.81%	75.87%	**76.24%**
	30%	**77.35%**	74.36%	69.4 %	75.09%

TABLE VIII
COMPARISON OF SSL ALGORITHMS USING 3NN AS BASE LEARNER

Table IX presents the number of wins of each one of the tested algorithms according to the utilized ratio in the training set, while the best scores are highlighted in bold. Notice that draw cases between algorithms have not been encountered. The above aggregated results show that En-SSL is by far the most effective algorithm, reporting the highest accuracy in 7, 9 and 5 cases, using a labeled ratio of 10%, 20% and 30%, respectively.

Ratio	Tri-training	Co-training	Self-training	En-SSL
10%	0	2	2	**7**
20%	1	0	1	**9**
30%	4	0	1	**5**

TABLE IX
TOTAL WINS OF EACH SSL ALGORITHM

In machine learning, the statistical comparison of multiple algorithms over multiple data sets is fundamental and it is usually carried out by means of a statistical test [17]. Therefore, in order to evaluate the rejection of the hypothesis that all the algorithms perform equally well for a given level and highlight the existence of significant differences between our proposed algorithm and the classical SSL algorithms, we utilized the non-parametric Friedman Aligned Ranking (FAR) [14] test. Notice that, since the test is non-parametric, it does not require commensurability of the measures across different data sets, it does not assume normality of the sample means and it is robust to outliers. Moreover, the Finner post hoc test [10] with a significance level $\alpha = 0.05$ was applied a post hoc procedure to detect the specific differences among the algorithms.

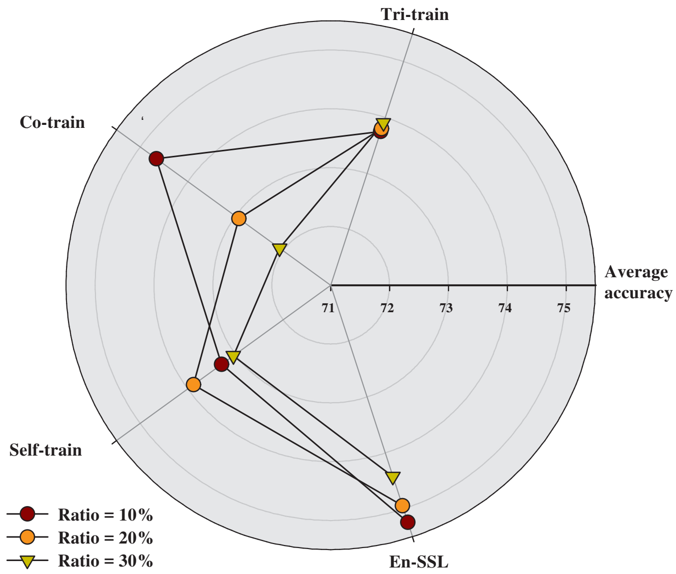

Fig. 1. Average classification accuracy of all SSL algorithms for $DATA_G$

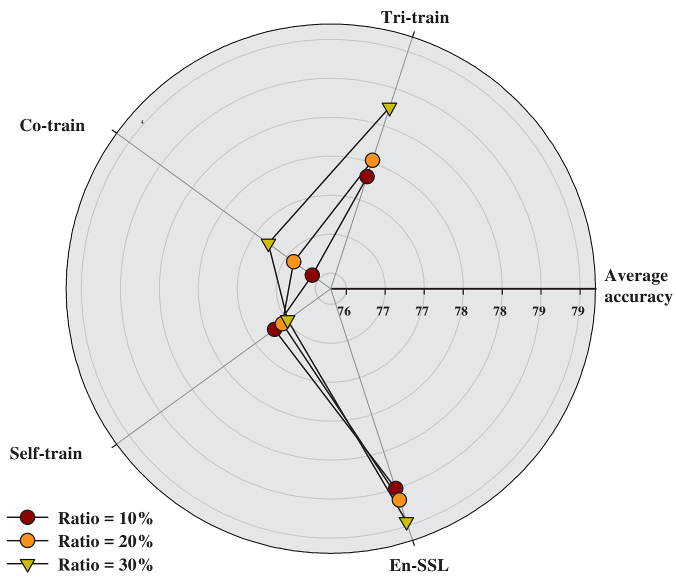

Fig. 2. Average classification accuracy of all SSL algorithms for $DATA_{GL}$

Tables X, XI and XII present the information of the statistical analysis performed by nonparametric multiple comparison procedures over 10%, 20% and 30% of labeled data, respectively. The best(lowest) ranking obtained in each FAR test determines the control algorithm for the post hoc test. Clearly, the proposed algorithm exhibits the best overall performance, illustrating the highest probability-based ranking, presenting statistically better results, relative to all labeled ratio.

Algorithm	Friedman	Finner post-hoc test	
	Ranking	p-value	Null Hypothesis
En-SSL	14.0833	–	–
Co-training	25.9167	0.038415	rejected
Tri-training	27.9167	0.023169	rejected
Self-training	30.0833	0.015280	rejected

TABLE X
FAR TEST AND FINNER POST HOC TEST (LABELED RATIO 10%)

Algorithm	Friedman	Finner post-hoc test	
	Ranking	p-value	Null Hypothesis
En-SSL	11.1250	–	–
Tri-training	23.4167	0.031508	rejected
Self-training	26.9583	0.008390	rejected
Co-training	36.5000	0.000027	rejected

TABLE XI
FAR TEST AND FINNER POST HOC TEST (LABELED RATIO 20%)

Algorithm	Friedman	Finner post-hoc test	
	Ranking	p-value	Null Hypothesis
En-SSL	14.3750	–	–
Tri-training	20.4583	0.287165	accepted
Self-training	29.2917	0.013556	rejected
Co-training	33.8750	0.001935	rejected

TABLE XII
FAR TEST AND FINNER POST HOC TEST (LABELED RATIO 30%)

VII. CONCLUSION & FUTURE RESEARCH

In this work, we propose a new SSL algorithm for predicting the students' performance in the final examinations at the end of the 1st class of Lyceum. Our experimental results illustrated that our proposed classification algorithm is proved to be effective and practical for the early and accurate prediction of students' progress, as compared to some traditional SSL algorithms.

In conclusion, we point out that the students' attributes utilized in our work do not constitute a conclusive list. An extension can introduce new attributes and other criteria which were not in the current database, but are collectable by tutors and may potentially influence the performance and the quality of the prediction of student's performance i.e. students' characteristics (social and cultural), more tests, more written assignments. Nevertheless, the identification of which attributes should be utilized or which have higher impact on the students' performance is still under consideration by many researchers [2], [35], [36].

Furthermore, since our experimental results are quite encouraging, another direction for future research would be to enlarge our experiments with more schools (private and state) and school years and evaluate the proposed algorithm for predicting the students' performance at national level examinations for admission to higher education institutes.

ACKNOWLEDGMENTS

The authors are grateful to the private high school "*Avgoulea-Linardatou*" for the collection of the data used in our study.

REFERENCES

[1] D. Aha. *Lazy Learning*. Dordrecht: Kluwer Academic Publishers, 1997.

[2] R.S. Baker and P.S. Inventado. *Learning Analytics: From Research to Practice*, chapter Educational Data Mining and Learning Analytics, pages 61–75. Springer, 2014.

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

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *11th annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[5] W. Cohen. Fast effective rule induction. In *International Conference on Machine Learning*, pages 115–123, 1995.

[6] P. Cortez and A.M.G. Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, pages 5–12, 2008.

[7] T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857, pages 1–15. Springer Berlin Heidelberg, 2001.

[8] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[9] J. Du, C.X. Ling, and Z.H. Zhou. When does cotraining work in real data? *IEEE Transactions on Knowledge and Data Engineering*, 23(5):788–799, 2011.

[10] H. Finner. On a monotonicity problem in step-down multiple test procedures. *Journal of the American Statistical Association*, 88(423):920–923, 1993.

[11] P. Gandhi and V. Aggarwal. Ensemble hybrid logit model. In *Proceedings of the KDD 2010 cup: Workshop knowledge discovery in educational data*, pages 33–50, 2010.

[12] T. Guo and G. Li. Improved tri-training with unlabeled data. *Software Engineering and Knowledge Engineering: Theory and Practice*, pages 139–147, 2012.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: An update. *SIGKDD Explorations Newsletters*, 11:10–18, 2009.

[14] J.L. Hodges and E.L. Lehmann. Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, 33(2):482–497, 1962.

[15] G. Kostopoulos, S. Kotsiantis, and P. Pintelas. Estimating student dropout in distance higher education using semi-supervised techniques. In *Proceedings of the 19th Panhellenic Conference on Informatics*, pages 38–43. ACM, 2015.

[16] G. Kostopoulos, S. Kotsiantis, and P. Pintelas. Predicting student performance in distance higher education using semi-supervised techniques. In *Model and Data Engineering*, pages 259–270. Springer, 2015.

[17] G. Kostopoulos, I.E. Livieris, S. Kotsiantis, and V. Tampakas. Enhancing high school students performance prediction using semi-supervised methods. In *8th International Conference on Information, Intelligence, Systems and Applications (IISA 2017)*. IEEE, 2017.

[18] S. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37:331–344, 2012.

[19] C. Liu and P.C. Yuen. A boosted co-training algorithm for human action recognition. *IEEE transactions on circuits and systems for video technology*, 21(9):1203–1213, 2011.

[20] I. Livieris, K. Drakopoulou, Th. Kotsilieris, V. Tampakas, and P. Pintelas. DSS-PSP - A decision support software for evaluating students' performance. In *Engineering Applications of Neural Networks*, pages 63–74. Springer, 2017.

[21] I. Livieris, K. Drakopoulou, V. Tampakas, T. Mikropoulos, and P. Pintelas. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research*, 2018.

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

[23] I.E. Livieris, K. Drakopoulou, and P. Pintelas. Predicting students' performance using artificial neural networks. In *Information and Communication Technologies in Education*, pages 321–328, 2012.

[24] C. Marquez-Vera, A. Cano, C. Romero, and S. Ventura. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, 38:315–330, 2013.

[25] C.J. Merz. Combining classifiers using correspondence analysis. In *Advances in Neural Information Processing Systems*, pages 592–597, 1997.

[26] C.J. Merz. Using correspondence analysis to combine classifiers. *Machine Learning*, 36:33–58, 1999.

[27] V. Ng and C. Cardie. 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-Volume 1*, pages 94–101. Association for Computational Linguistics, 2003.

[28] A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[29] J. Platt. Using sparseness and analytic QP to speed training of support vector machines. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in neural information processing systems*, pages 557–563, MA: MIT Press, 1999.

[30] J.R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco, 1993.

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

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

[33] M. Re and G. Valentini. *Advances in Machine Learning and Data Mining for Astronomy*, chapter Ensemble methods: A review, pages 563–594. Chapman & Hall, 2012.

[34] L. Rokach. *Pattern Classification Using Ensemble Methods*. World Scientific Publishing Company, 2010.

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

[36] C. Romero, S. Ventura, S. Pechenizkiy, and M. Baker. *Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[37] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D. Rumelhart and J. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 318–362, Cambridge, Massachusetts, 1986.

[38] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australian conference on artificial intelligence*, volume 4304, pages 1015–1021, 2006.

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

[40] L. Todorovski and S. Džeroski. Combining classifiers with meta decision trees. *Machine Learning*, 50(3):223–249, 2002.

[41] Z.H. Zhou. When semi-supervised learning meets ensemble learning. In *Frontiers of Electrical and Electronic Engineering in China*, volume 6, pages 6–16. Springer, 2011.

[42] Z.H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[43] X. Zhu. Semi-supervised learning. In *Encyclopedia of Machine Learning*, pages 892–897. Springer, 2011.

[44] X. Zhu and A.B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.