applied sciences

MDPI

Article

Predicting and Interpreting Students' Grades in Distance Higher Education through a Semi-Regression Method

Stamatis Karlos , Georgios Kostopoulos and Sotiris Kotsiantis *

Department of Mathematics, University of Patras, 26504 Rio Patras, Greece; stkarlos@upatras.gr (S.K.); kostg@sch.gr (G.K.)

* Correspondence: sotos@math.upatras.gr

Received: 30 October 2020; Accepted: 24 November 2020; Published: 26 November 2020

Abstract: Multi-view learning is a machine learning app0roach aiming to exploit the knowledge retrieved from data, represented by multiple feature subsets known as views. Co-training is considered the most representative form of multi-view learning, a very effective semi-supervised classification algorithm for building highly accurate and robust predictive models. Even though it has been implemented in various scientific fields, it has not adequately used in educational data mining and learning analytics, since the hypothesis about the existence of two feature views cannot be easily implemented. Some notable studies have emerged recently dealing with semi-supervised classification tasks, such as student performance or student dropout prediction, while semi-supervised regression is uncharted territory. Therefore, the present study attempts to implement a semi-regression algorithm for predicting the grades of undergraduate students in the final exams of a one-year online course, which exploits three independent and naturally formed feature views, since they are derived from different sources. Moreover, we examine a well-established framework for interpreting the acquired results regarding their contribution to the final outcome per student/instance. To this purpose, a plethora of experiments is conducted based on data offered by the Hellenic Open University and representative machine learning algorithms. The experimental results demonstrate that the early prognosis of students at risk of failure can be accurately achieved compared to supervised models, even for a small amount of initially collected data from the first two semesters. The robustness of the applying semi-supervised regression scheme along with supervised learners and the investigation of features' reasoning could highly benefit the educational domain.

Keywords: educational data mining; student grade prediction; semi-regression; early prognosis; interpretation; COREG algorithm

1. Introduction

Educational data mining (EDM) has emerged in the past two decades as a highly-growing research field concerning the development and implementation of machine learning (ML) methods for analyzing datasets coming from various educational environments [1]. The key concept is to utilize these methods, extract meaningful knowledge about students' performance, and improve the learning process enriching the insights that the tutor may obtain on time. These methods are grouped into five main categories [2]: Prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. The main research interest has been centered on predictive problems primarily concerned with three major questions [3]: (1) What outcome of students will be predicted? (2) Which ML methodology is the most effective for the specific problem? (3) How early can such a prediction be made?

Most of the EDM research is mainly focused on implementing supervised methods utilizing only labeled datasets. To this end, a plethora of classification and regression techniques have successfully been applied for predicting various learning outcomes of students, such as dropout, attrition, failure, academic performance, and grades, to name a few. In addition, the main interest concentrates on building efficient predictive models at the end of a course using all available information about students [4]. However, it is of practical importance to provide both accurate and early-step predictions at minimum cost [5]. A review of recent studies and developments in the field of EDM reveals that there is an urgent demand for accurate identification of students at risk of failure as soon as possible during the academic year, since early intervention activities and strategies can be implemented. Preventing academic failure, enhancing student performance, and improving learning outcomes is of utmost importance for higher education institutions that intend to provide high-quality education [6]. Some new directions that have recently been formatted concern the recognition of errors during the composition or the writing of code assessment, usually based on self-attenuation mechanisms for providing high quality automated debugging solutions to undergraduate and post-graduate students, as well as the exportation of remarkable insights about the obstacles that are met by them during such tasks [7].

Apart from supervised methods, semi-supervised learning (SSL) has gained a lot of attention among scientists in the past few years for solving a wide range of problems in various domains [8]. SSL methods exploit a small pool of labeled examples together with a large pool of unlabeled ones for building robust and highly-efficient learning models. However, SSL has not adequately used in the educational domain as easily identified after a thorough literature review. Nevertheless, some notable studies have emerged recently dealing with semi-supervised classification (SSC) tasks, such as student performance prediction or student dropout, while semi-supervised regression (SSR) is uncharted territory. The primal difference between SSC and SSR is that the target attribute is categorical in the former case, while a pure numeric quantity has to be predicted in the latter case. A recent literature review of SSR depicts the most important works in this field [9], separating them into approaches with a common strategy to solve their task, while more related works have been demonstrated on behalf of SSC [10].

Multi-view learning has also attracted the interest of this research community, distilling information from separate views, original or transformed ones, while a search of more appropriate subspaces into the initial feature set always remains a crucial learning task for boosting the performance of SSL methods [11,12]. Adopting ensemble learners has also been an active research territory concerning SSL [13], while some similar works have been demonstrated by our side [14,15]. Although some recent advances have taken place—exploiting graph-based solutions [16–18], or deep learning neural networks (DNNs) [19,20]—attempting to acquire more and more accurate predictions, or even robust ones in case that noisy inputs/labels have violated the ideal case of compact training data [21], such mechanisms introduce some important defeats:

- Increased computational issues regarding the size of the provided datasets;
- Operation under transductive mode with inefficient complexity for most real-life cases rejecting at the same time the extraction of an inductive mechanism as a generic solution;
- Inability to facilitate interpretability of the exported decisions/predictions [22,23].

The main scope of the present study is three-fold. At first, we implement a well-known semi-supervised regression algorithm that is based on multi-view learning, adopting several ML learners into its main kernel, tackling with the early prediction of undergraduate students' final exam grades in a one-year distance learning course. Each student is represented in terms of a plethora of features, which were collected from three different sources, thus producing three distinct sets of attributes: Demographics, academic achievements, and interaction within the course Learning Management System (LMS). Secondly, we investigate the effectiveness of the separate SSR variants that are produced compared with their corresponding supervised performance on the examined EDM task.

In this sense, the proposed model may serve as an early alert tool with a view to providing appropriate interventions and support actions to low performers.

Finally, we apply a well-established framework for acquiring trustworthy reasoning scores per included attribute/indicator into the original dataset. Hence, interpretable models are created, providing carefully computed explanations about the predicted grades ranking the importance of each indicator without any dimensionality reduction trick and avoiding overconsumption of computational resources under specific cases. To the best of our knowledge, this is the first completed study towards this direction [24], which hopefully will provide the basis for further research in the field of EDM, as it is stated in the relevant and conclusory Sections.

The remainder of this paper is organized as follows. In the next section, we discuss the need for explainable artificial intelligence (XAI) solutions to the field of EDM, highlighting some of the most important approaches in interpreting decisions/predictions of various learning models and the assets of the selected interpretability framework. Section 3 presents a brief overview of relevant studies in the EDM field and some recently published works related to the SSR task. The research goal is set in Section 4, together with an analysis of the dataset used in the experimental procedure. The total pipeline for applying a well-known COREG algorithm (CO-training REGressors) [25] as an SSR wrapper along with several ML learners and some DNNs variants is provided in Section 5, also describing the two distinct explaining mechanisms that are based on the computation of Shapley values [26]. The experimental process and results are presented in Section 6. Finally, our conclusions are drawn in Section 7, which also mentions some promising improvements to this seminal work.

2. Interpretability in Machine Learning

Consider the problem of predicting the final exam grade of students enrolled in a distance learning course using ML. In this case, a supervised algorithm is trained over a set of labeled data (the target attribute values are known), and an ML model is produced (supervised learning), which is subsequently deployed for predicting the grade of a previously unknown student for given values of the input attributes (features of students). The predictive model does not know why the student received the specific grade, while, at the same time, it fails to grasp the difference between anticipated grades and actual performance. Decision-makers are often hesitant to trust the results of these models, since their internal functions are primarily hidden in black-boxes [27]. This is quite reasonable, since people outside of the ML field neither can understand the manner that outputs are exported, nor are confident on just consuming some pure decisions without accompanying them with some consistent proofs. There is also a well-known trade-off regarding the predictive ability and the interpretability of ML algorithms, which unfortunately deters the co-existence of both these properties to be highly qualified under the same ML algorithm, in general. Since predictive models play a decisive role in the decision-making process in higher education institutions, the ability to comprehend these models seems to be indispensable. Thus, the interpretability of provided solutions usually needs to be filtered through XAI tools [28,29].

Model interpretability is the process of understanding the predictions of an ML model. In fact, it is the key point to build both accurate and reliable learning models. In traditional ML problems, the objective is to minimize the predictive error, while interpretability is focused on extracting more valuable information from the model [30]. Commonly, it aims to address questions, such as (Figure 1):

- What each attribute represents?
- Why was a specific prediction was made by the model?
- Which are the key factors of a specified prediction?
- Why a specific student was assigned a failing grade?
- Can we describe what the model has learned?
- How confident are we for the decisions of the model?

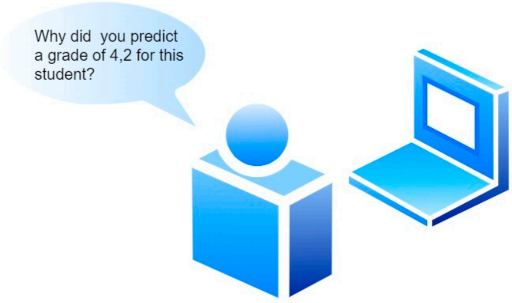

Figure 1. Why was a specific prediction was made by the model?

Although several published works have appeared in the literature of XAI recently, the majority of them make assumptions that are not actually consistent with the specifications of an educational task. For example, dimensionality reduction or feature transformations (e.g., semantic embeddings) may lead to incorrect conclusions or reasoning factors that ignore some of the underlying relationships that may be crucial for the real-life problem [31]. Furthermore, DNNs and their variants that operate by manipulating raw-data directly have highly attracted the interest of the XAI community, leading to solutions that are not applicable to our numerical source data. However, this fact does not exclude DNNs from being used as accurate black-boxes to such kind of problems, adopting mainly some model-agnostic approaches [32]. A representative work was done by Akusok et al. [22] exploiting extreme learning machines (ELM) trained on sampled subsets of the initial training set for increasing the output variance of the learning model, and later, explaining the information gained thought this strategy via proper confidence intervals for specific confidence levels. Both artificial and real-life datasets were evaluated, performing robust behavior without inducing much computational effort.

Besides DNNs, conventional ML algorithms need to overcome the long-standing obstacle of explainable predictions. One of the most popular libraries is LIME (local interpretable model-agnostic explanations) [33], which offers explanations based on local assumptions regarding the contribution of the examined learning model. A proper function that measures the interpretability and the local fidelity is defined, which is optimized using sparse linear models that are fed with perturbed samples from the region of interest. Global patterns are taken into consideration in the [32]. A framework of teacher-student models was proposed in Reference [34], where the corresponding explanations are obtained through adopting some additional models that mimic the behavior of the target black-box model and compare their performance on ground-truth trained models to clarify possible bias factors or reveal cases where the missing information has corrupted the final predictions. Because of the behavior of the adopted models, the confidence intervals are also produced for determining the importance of the detected differences.

Linear models and ensemble of trees were used in the previous work, while a solution that exploits some unsupervised mechanisms internally and focuses on exporting small, comprehensible, and more reliable rules exploiting ensemble of tress was proposed by Mollas et al. [35]. Both quantitative and qualitative investigation of the proposed LionForests approach has been taken place regarding Random Forest (RF) over binary classifications tasks, which is categorized as a local-based one. Another work that investigates classification tasks, but specializes in interpreting convolutional neural networks (CNNs) was recently demonstrated in Reference [36], where the Layer wise Relevance Propagation strategy was applied for extracting meaningful information when usual image transformations of audio signals are given as input. This process has been widely preferred for such networks, trying to propagate the computed weights of the total network to the input nodes, transforming them to important indications.

As it regards the adopted XAI framework by our side in the context of this work, Shapley values that stem from coalitional game theory constitute the basic concept that a more recent approach, named as Shapley additive explanations (SHAP), seems to satisfy better our research scope [37]. First, it is based on well-established theory and operates without violating a series of axioms: Efficiency, symmetry, dummy, and additivity. Without providing any extended analysis, we mention that Shapley values provide helpful insights by measuring the contribution of each feature into the original d-dimensional feature space $F \in \mathbb{R}^d$. Although this process demands quadratic computations regarding the size of F, it is an accurate and safe manner for revealing the actual contribution of each feature taking into consideration all the underlying dependencies of the measured values, thus assigning a combined profile of both local and global explanations. The exact formula for computing the total contribution of a random feature $i \in F$ through all the necessary weighted marginal contributions is given here:

$$contribution_i = \sum\nolimits_{S \subseteq F \backslash i} \frac{|S|!(d-|S|-1)!}{d!} (payout_{S \cup i} - payout_S) \tag{1}$$

$$payout_F = \int model(F) dF_{feature \notin F} - E_F(model(F)) \tag{2}$$

where each pay-out integrates the predictions of the selected model for any feature that belongs to the feature space F, while the rest ones are replaced by their mean value. In total, the Shapley values express the contribution that corresponds to each feature regarding the difference of the predicted value minus the average predicted value. Modifications that are more carefully implemented for obtaining the SHAP values reducing the overhead of the original procedure based on statistical assumptions or exploiting the nature of the base learner. Two such variants were adopted for facilitating the total efficacy of our methodology [26].

3. Related Work

Semi-regression has not been sufficiently implemented in the domain of EDM, as evidenced by a thorough study of the pertinent literature. Apparently, SSL classification algorithms cannot be directly applied for regression tasks, due to the nature of the target attribute, which is a real-valued one. Nevertheless, some recent and notable studies are discussed below.

Nunez et al. [38] proposed an SSR algorithm for predicting the exam marks of fourth-grade primary school students. The dataset comprised a wide range of students' information, such as demographics, social characteristics, and educational achievements. At first, the Tree-based Topology Oriented Self-Organizing Maps (TTOSOM) classifier was employed for building clusters exploiting all available data. These clusters were subsequently used for training the semi-regression model, which proved quite effective for handling the missing values directly without requiring a pre-processing stage. The experimental results demonstrated that the proposed algorithm achieved better results in terms of mean errors, compared to representative regression methods. Kostopoulos et al. [39] designed an SSR algorithm for predicting student grades in the final examination of a distance learning course. A plethora of demographic, academic, and activity attributes in the course Learning Management System (LMS) were employed, while several experiments were carried out. The results indicated the efficiency of the SSR algorithm compared to familiar regression methods, such as linear regression (LR), model trees (MTs), and random forests (RFs).

Bearing in mind the aforementioned studies and their findings, an attempt is made in the present study to implement an SSR algorithm for predicting the grades of undergraduate students in the final exams of a one-year online course offered by the Hellenic Open University. The main contribution of our research concentrates mainly on the following points:

- Semi-regression,
- Early prognosis, and
- Interpretation of features.

We also include some related works that concern the SSR field, which tackle problems from different domains. Besides the COREG algorithm [25], which inspired most of the upcoming SSR works on how to exploit unlabeled data for performing SSL methods for predicting numeric target attributes, the use of a co-training scheme did not found great acceptance for SSR works. We highlight just the direct expansion of COREG designed by Hady et al., via inserting the co-training by Committee for Regression (CoBCReg) scheme [40], which tries to encompass the use of more than one regressors for reducing noisy predictions, as well as the co-regularized least squares regression approach (CoRLSR) [41]. The latter one sets a risk minimization problem on the combined space of labeled and unlabeled data through proper kernel methods, focusing mainly on proposing some variants—a semi-parametric and a non-parametric—that scale linearly on the size of the unlabeled subset. The predictive benefits of adopting the co-training scheme without using any sophisticated feature split, just a random one, were remarkable.

More recently, a local linear regressor was employed by Liang R.-Z. et al. [42], which was iteratively applied for minimizing a joint problem on the neighborhood of each unlabeled examplFDe through sub-gradient descent algorithms. The authors of this work transformed two datasets that stem from unstructured data into structured problems and managed to outperform the compared algorithms regarding each posed performance metric, managing a competitive behavior regarding the time consumption. A multi-target fashion SSR model was presented in Reference [43], where the self-training scheme was combined with an efficient ensemble decision tree-based algorithm. Several modifications of the proposed scheme were examined, differentiated on the manipulation of the decisions that are drawn from the corresponding ensemble learner. Although their approach depends heavily on a reliability threshold which is domain-specific, a qualitative analysis was made over a dynamic selection, managing to outperform the supervised baseline as well as a random strategy for selecting unlabeled data for augmenting the initially collected data. Finally, an SSR method was used before applying an SSL method in the field of optical sensors, where limited data were readily available. In that scenario, a randomized method was used for generating unlabeled artificial data aiming at augmenting the labeled subset, but their annotation with pseudo-values was still crucial [44]. Therefore, a typical SSR strategy was applied before providing the finally created dataset to tackle the classification process.

4. Dataset Description

The dataset used in the research was provided by the Hellenic Open University and comprised records of 1073 students who attended the 'Introduction to Informatics' module of the 'Computer Science' course during the academic year 2013–2014.

These records were collected from three different sources, the course database, the teachers, and the course LMS, thus producing three distinct sets of attributes (Figure 2):

- The demographic set S_1 = {Gender, NewStudent} (Table 1).
  The distribution of male and female students was 76.5% and 23.5%, respectively. In addition, 87.5% of the students had enrolled in the course for the first time, while the rest failed to pass the previous year's final exams.
- The academic performance set $S_2 = \{\mathrm{Ocs_i}, \mathrm{Wri_i}\}_{i=1}^{2}$ (Table 2).
  The attribute named $\mathrm{Ocs_i}$ refers to students' absence or presence in the i-th optional contact session, while the attribute named $\mathrm{Wri_i}$ represents students' grades (ten-point grading scale) in the i-th written assignment, $i \in \{1, 2\}$. Four written assignments should be submitted during the academic year, while a total sum $\sum_{i=1}^{4} \mathrm{Wri_i} \geq 20$ was required for a student to undertake the final exam.
- The LMS activity set $S_3 = \{\mathrm{L_i}, \mathrm{V_{1i}}, \mathrm{V_{2i}}, \mathrm{V_{3i}}, \mathrm{V_{4i}}, \mathrm{V_{5i}}, \mathrm{P_{1i}}, \mathrm{P_{2i}}, \mathrm{P_{3i}}, \mathrm{P_{4i}}\}_{i=1}^{2}$ (Table 3).
  These attributes monitor student activity on the online LMS course (i.e., logins, views, and posts).

Table 1. Demographic attributes.

Attribute Name	Description	Values
Gender	Student gender	male, female
New Student	A student enrolled in the course for the first time	yes, no

Table 2. Academic performance attributes, i ∈ {1, 2}.

Attribute Name	Description	Values
Ocs_i	Student presence in the i-th optional contact session	absence, presence
Wri_i	Student grade in the i-th written assignment	[0, 10]

Table 3. LMS activity attributes in the i-th time-period, i ∈ {1, 2}.

Attribute Name	Description	Values
L_i	Total number of student logins	integer
V_{1i}	Number of student views in the pseudo-code forum	integer
V_{2i}	Number of student views in the compiler forum	integer
V_{3i}	Number of student views in the module forum	integer
V_{4i}	Number of student views in the course forum	integer
V_{5i}	Number of student views in the course news	integer
P_{1i}	Number of student posts in the pseudo-code forum	integer
P_{2i}	Number of student posts in the compiler forum	integer
P_{3i}	Number of student posts in the module forum	integer
P_{4i}	Number of student posts in the course forum	integer

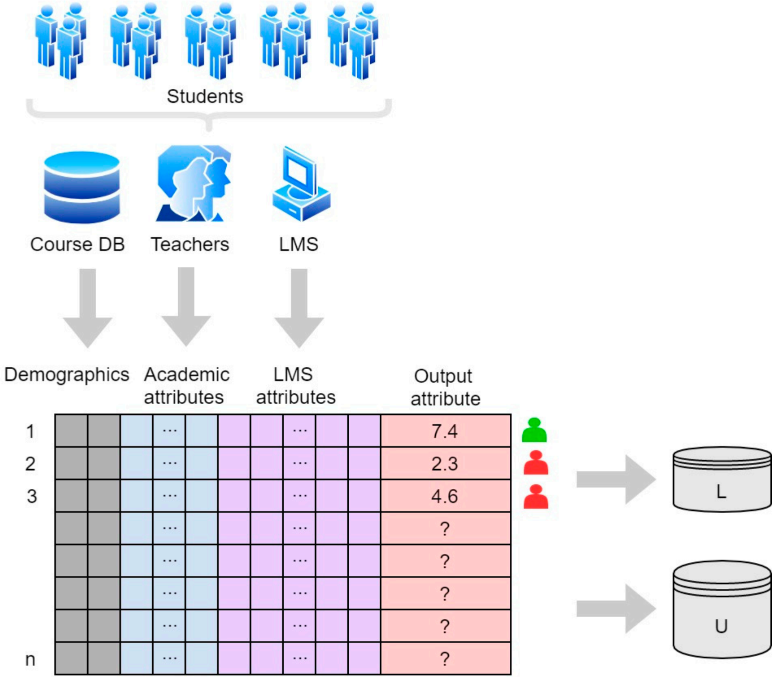

Figure 2. Gathering the data during the academic year.

Each instance of the dataset represents a single student (Figure 2) and is described by a vector of attributes, such as $x = (s_1, s_2, s_3)$, where s_1, s_2, s_3 correspond to the vector attributes of S_1, S_2, S_3 sets, respectively. Since the early prognosis of students at risk of failure is of utmost importance for higher education institutions, the academic year was divided into four time-periods according to each written assignment submission deadline (Figure 3). To this end, the notation V_{1i} denotes the total number of student views in the pseudo-code forum in the i-th period, i ∈ {1, 2}, and so forth. For example, attribute P_{21} refers to the total number of student posts in the compiler forum in the first time-period (i.e., from the beginning of the academic year until the first written assignment submission deadline). Finally, the output attribute $y \in [0, 10]$ represents the grade of students in the final examinations of the course. Note that we examine two distinct scenarios, corresponding to the first one and the first two time-periods, respectively.

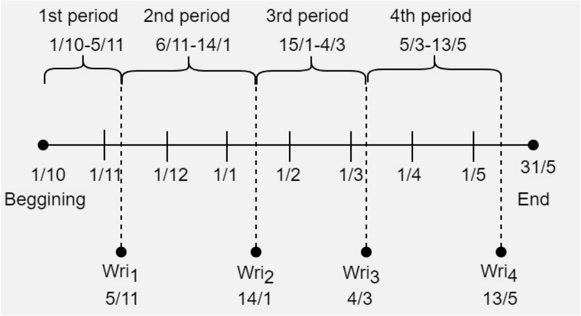

Figure 3. Time-periods of the academic year.

5. Proposed Semi-Supervised Regression Wrapper Scheme

Semi-Supervised Learning (SSL) is a rapidly evolving subfield of ML, embracing a wide range of high-performance algorithms. Typically, an ML model h is built from a training dataset $D = L \cup U$ consisting of a small pool of labeled examples $L = \{x_i, \ y_i\}_{i=1}^{l}$ and a large pool of unlabeled ones $U = \{x_i\}_{i=1}^{u}$, $l << u$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $L \cap U = \varnothing$, without human intervention [45]. Depending upon the nature of the output attribute SSL is divided into two settings [9]:

- Semi-Supervised Classification (SSC).
  The labels y_i of the output attribute are discrete, i.e., $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$.
- Semi-Supervised Regression (SSR).
  The labels y_i of the output attribute are real, i.e., $\mathcal{Y} \subseteq \mathbb{R}$.

In our case, we employed an SSR scheme for exploiting the existence of both labeled and unlabeled data trying to acquire accurate estimations of the target attribute—students' final grade—based on a set of readily available data. Thus, one or more regressors are trained iteratively via selecting the most appropriate unlabeled data and annotating their missing target value in an automated fashion. Of course, the initial hypothesis is formatted on the manually gather the subset of L. Furthermore, the fact that the training set is split into two disjoint subsets, L and U, and that we aim at applying our trained model on another subset—the test set—which does not overlap with the training set leads us to an inductive SSR algorithm.

The most representative algorithm found in the literature that seems to satisfy our ambitions is the COREG that was firstly proposed by Zhou [25]. Actually, this learning scheme constitutes the analog of the co-training scheme also based on disagreement rule in the case of SSC [46], inserting a local-based criterion for measuring the effectiveness of the candidate unlabeled instances into the currently trained model for completing a regression task. Although various criteria have been designed in the context of SSC [47,48], the corresponding essential stage during an inductive SSR algorithm has not been highly studied by the related research community, following variants of the same criterion proposed in the case of COREG or proposing some new metrics that are mainly used under single-view works [44,49,50].

More specifically, the main concern of inductive SSR algorithms during the annotation of unlabeled examples is their *consistency* with the already existing labeled instances. This property is examined by measuring the next formula:

$$Consistency_{x_j} = \sum_{x_i \in L} \Big(f(y_i, h(x_i)) - f\big(y_i, h(\hat{x_i})\big) \Big), \ \forall x_j \in U \tag{3}$$

where f is a suitable performance metric, y_i is the actual value of the x_i labeled example, while $h(x_i)$ and $h(\hat{x_i})$ denote the output of regressor h when is trained solely on the current labeled set and on the augmented labeled set with the currently examined x_j example, respectively. According to the COREG algorithm, a local criterion is inserted for investigating if the consistency of each unlabeled example is beneficial for the current model per iteration. Thus, instead of examining the whole current L subset, only the neighbors of each $x_j \in U$ are considered for measuring the corresponding consistency metric, which is described in Equation (1). As it is discussed in the original work of the COREG, by maximizing this variant—mentioned hereinafter as $\delta_{x_j}\ \forall x_j \in U$—we reach safely either to the maximization of the general consistency metric or we acquire a zero value. In the first case, we pick the j-th unlabeled instance with a greater impact. Otherwise, we do not select any of them.

This strategy is similar to fitting an instance-based algorithm, like the k-Nearest Neighbors (kNN) [51], for selecting the unlabeled instances to augment the current labeled set per iteration, as it was preferred during the COREG approach. However, this fact does not hinder us from applying different kinds of regressors on the augmented labeled set, thus exploiting possible advantages of other learning models for capturing better the underlying relationships of the examined data. Based on our search in the literature, such a study has not yet been done.

Moreover, the already mentioned augmented per iteration labeled subset does not contain exclusively accurate values of the target attribute per its included instance, since during the training stage pseudo-labeled instances are joining the initially labeled examples, and their estimated values may differ from the actual one. This kind of noise into any SSL scheme may heavily deteriorate their total performance, settling them as myopic approaches that cannot guarantee safe predictions and violate the interpretation of the exported results.

Therefore, to alleviate the inherent confidence of COREG, we examine its efficacy on an EDM task that supports the multi-view description, increasing, thus the diversity of the trained regressors. Since the COREG algorithms is based on the co-training scheme, the feature space F of the original problem D is split into two disjoint views: $F = F_1 \cup F_2$. Although the random split has been proven quite competitive in several cases [52,53], co-training scheme should work if these two views are independent and sufficient.

The examined real-world problem brings a multidimensional and multi-view description that encourages us to train each regressor on separate views and get trustworthy predictions that would not harm our learning model regarding neither its predictiveness nor its interpretability despite the limited labeled data. Algorithm 1 presents the pseudocode of the end-to-end SSR pipeline.

Algorithm 1. The extended framework of the COREG algorithm.

```
Framework: Pool-based COREG(D, selector1, selector2, regressor1, regressor2)
Input:
  •  Initially collected labeled $\boldsymbol{L} = \{x_i,\ y_i\}_{i=1}^{l}$ and unlabeled $\boldsymbol{U} = \{x_i\}_{i=1}^{u}$ instances, where $\boldsymbol{D} = \boldsymbol{L} \cup \boldsymbol{U}$ and $\boldsymbol{L} \cap \boldsymbol{U} = \varnothing$
  •  $F_1, F_2$: provide the split of the original feature space F, where $F = F_1 \cup F_2$ and $F_1 \cap F_2 = \varnothing$
  •  Define Max_iter: maximum number of semi-supervised iterations and $f$: performance metric
Main process:
  •  Set iter = 1, consistentSet = ∅
  •  Train selectori, regressori on $\boldsymbol{L}(F_i)\ \forall i \in \{1,2\}$
  •  While iter ≤ Max_iter do
  •     For each $i \in \{1,2\}$ do
  •         For each $x_j \in \boldsymbol{U}$ do
  •            Compute $\delta_{x_j}(f)$ based on selectori $\forall i\ \in \{1,2\}$
  •            If $\delta_{x_j}(f) > 0$ : add j to consistentSet
  •       If consistentSet is empty do
  •            iter:= iter + 1 and continue to the next iteration
  •       else do
  •            Find the index $j^*$ of consistentSet s.t. $j^* = arg\ \max_{j} \delta_{x_j}$
  •            Update $\mathbf{U}:\ \mathbf{U} \leftarrow \mathbf{U} - \{x_{j^*}\}$
  •            For $i$ {1,2} do
  •               Update $\mathbf{Li}:\ \mathbf{Li} \leftarrow \mathbf{Li} \cup \{x_{j^*},\ regressor \sim i\left(x_{j^*}\right)\}$, where ~i means the opposite index of the current
  •            Retrain selectori, regressori on $\boldsymbol{L}(F_i) \forall i \in \{1,2\}$
  •            iter:= iter + 1
Output:
  •  Apply the next rule to each met $x_{test}$ instance:
                 $h_{COREG}(x_{test}) = 1/2 \cdot (regressor1(x_{test}) + regressor2(x_{test}))$
```

6. Experimental Process and Results

To conduct our experiments, we exploited the sci-kit Python library along with its integrated regressors and an implementation of computing the necessary Shapley values [37,54]. In order to systematically examine the efficiency of the extended COREG variant over the problem of early prognosis on student's performance, various choices of instance-based selectors and different learning model for the case of the regressors were chosen. Furthermore, we investigated two separate cases of the total dataset based on the measured indicators: Regarding only the first semester (D_1-first scenario) and only the first two semesters (D_2-second scenario). Thus, our predictions excuse the characterization of the early prognosis task, providing in time predictions using indicators that stem from the initial stages of an academic year. To be more specific, the size and the attributes of each view per dataset-scenario are reported here:

- First scenario:

$$D_1 = F_1 \cup F_2$$

$$|F_1| = 4,\ F_1 = (gender,\ NewStudent, Ocs_1, Wri_1)$$

$$|F_2| = 10, F_2 = (L_1, V_{11}, V_{21}, V_{31}, V_{41}, V_{51}, P_{11}, P_{21}, P_{31}, P_{41})$$

- Second scenario:

$$D_2 = F_1 \cup F_2$$

$$|F_1| = 6, F_1 = (gender,\ NewStudent, Ocs_1, Wri_1, Ocs_2, Wri_2)$$

$$|F_2| = 20, F_2 = (L_1, L_2, V_{11}, V_{12}, V_{21}, V_{22}, V_{31}, V_{32}, V_{41}, V_{42}, V_{51}, V_{52}, P_{11}, P_{12}, P_{21}, P_{22}, P_{31}, P_{32}, P_{41}, P_{42})$$

Besides the multi-view role of our extended COREG framework, the diversity of the SSR algorithm is enriched by the fact that each selector_i cannot select during one iteration the same x_{j^*} instance, while during the initial design of the COREG, randomly selected subsamples of the original U set were selected per iteration. Although we also attempted to implement this strategy, our results were constantly worse than the case of exploiting the full length of the original U set. This is probably due to the relatively small size of our total problem D, which we hope to undertake during the next semesters to enrich our collected data.

As it regards the choice of the investigated selectors and regressors for the extended COREG framework, we mention here all the different variants/models that were included in our experiments:

- (selector1, selector2): We have selected kNN algorithm for detecting the appropriate neighbors and fitting appropriate models. Following the original COREG scheme for injecting further diversity between the two separate views, we kept different power parameter for the internal distance that is exploited for formatting the neighborhood. Thus, we used Euclidean distance and Minkowski of 5th power for first and second selector, respectively. Moreover, we examined four separate cases based on the number of the nearest neighbors that are considered per case: (k_1, k_2) ≡ (1, 1), (1, 3), (3, 1), and (3, 3).
- (regressor1, regressor2): A different pair of same models have been used for this choice. To be more specific, we have used kNN with $k = 3$, a typical Linear regressor (LR), the Gradient Boosting regressor which is an additive model that operates under a forward-stage manner with 2 different loss functions: Least squares regression (ls) and 'huber'—a hybrid between ls and least absolute deviation—which are depicted as GB(ls) and GB(huber) and multi-layer perceptron that optimizes the squared-loss function by using the 'lbfgs' quasi-Newton solver. The last regressor is denoted as MLP, while its default hyperparameters were used: The 'Relu' as activation function and a hidden layer with 100 neurons. Although some further modifications of the internal parameters of each learner were investigated, as well as the combination of same learning models, but distinct regressors per view (e.g., train GB(ls) on $\boldsymbol{L}(F_1)$ and train GB(huber) on $\boldsymbol{L}(F_2)$), but neither this fact

serves our ambitions nor any great improvement was achieved. More information could be found in Reference [41].

As it concerns the rest required information about our evaluations, we set *Max_iter* equal to 100 and the performance metric f ≡ MSE (Mean Squared Error). Moreover, we applied a 5-fold-Cross-Validation (5-fold-CV) evaluation process, while we held 100 instances out of the 1073 as the test set. Consequently, the rest $n = 973$ instances constitute the D set, where the size of the L (l) and the U (u) subsets sum up to n. Thus, we examined four different values of the initially labeled instances: 50, 100, 150, and 200, while all of the rest instances were exploited from the first iteration as the U subset, since, as already mentioned, a possible random sampling of the total U subset per iteration did not favor us. Finally, the scenario under which our selectors exploit kNN algorithm with (k_1, k_2) = (1,1) did not manage to detect instances that satisfy the restriction of consistency as described in Equation (3) in the majority of the conducted experiments, and for this reason, was excluded by our results. The performance of the examined COREG variants based on the mean absolute error (MAE) metric is presented in Tables 4 and 5.

Table 4. Relative improvement of mean absolute error (MAE) metric (±std) of the dataset based only on the first semester during the best iteration per different combination of selector and regressor.

Size (L)	**Selector$_i$**			**Regressor$_i$**		
	(k_1, k_2)	**3NN**	**LR**	**GB (ls)**	**GB (huber)**	**MLP**
50	(1,3)	3.63 (±2.46)	10.06 (±6.42)	6.83 (±0.89)	3.78 (±1.62)	8.93 (±6.00)
	(3,1)	3.38 (±1.47)	10.51 (±8.12)	6.66 (±3.15)	5.07 (±1.09)	14.33 (±8.91)
	(3,3)	3.37 (±3.49)	15.27 (±8.15)	7.22 (±2.72)	4.60 (±2.19)	18.18 (±11.08)
100	(1,3)	2.28 (±3.00)	2.02 (±1.34)	1.81 (±1.66)	5.08 (±1.64)	5.82 (±4.49)
	(3,1)	2.33 (±1.70)	2.89 (±1.33)	1.95 (±1.63)	5.89 (±2.19)	6.14 (±5.11)
	(3,3)	2.26 (±2.76)	3.21 (±2.46)	2.42 (±2.84)	7.05 (±1.87)	9.42 (±4.77)
150	(1,3)	2.52 (±1.47)	0.63 (±0.72)	5.43 (±2.85)	3.32 (±2.14)	3.73 (±2.47)
	(3,1)	5.80 (±3/94)	1.07 (±1.62)	4.86 (±3.02)	3.44 (±1.28)	9.15 (±12.48)
	(3,3)	1.88 (±0.52)	2.07 (±1.53)	6.97 (±5.35)	2.67 (±1.31)	8.97 (±15.04)
200	(1,3)	2.83 (±1.82)	1.16 (±1.06)	3.31 (±2.53)	3.40 (±2.76)	4.95 (±2.46)
	(3,1)	4.04 (±3.41)	0.67 (±0.59)	2.93 (±2.56)	3.52 (±2.56)	3.05 (±1.94)
	(3,3)	0.88 (±1.05)	0.45 (±0.57)	4.58 (±3.34)	5.41 (±2.39)	5.85 (±2.66)

Table 5. Relative improvement of MAE metric (±std) of the dataset based only on the first and second semester during the best iteration per different combination of selector and regressor.

Size (L)	**Selector$_i$**			**Regressor$_i$**		
	(k_1, k_2)	**3NN**	**LR**	**GB (ls)**	**GB (huber)**	**MLP**
50	(1,3)	5.79 (±1.89)	21.85 (±8.10)	4.47 (±2.70)	7.55 (±2.12)	12.51 (±7.92)
	(3,1)	7.26 (±4.22)	22.65 (±7.89)	5.36 (±3.35)	7.70 (±3.99)	11.25 (±5.28)
	(3,3)	2.81 (±2.06)	30.69 (±13.83)	7.28 (±5.17)	8.40 (±3.72)	18.17 (±7.16)
100	(1,3)	5.56 (±1.85)	8.30 (±6.64)	6.00 (±1.18)	4.64 (±4.52)	9.26 (±7.55)
	(3,1)	3.65 (±2.72)	8.04 (±8.18)	6.62 (±2.61)	2.92 (±3.19)	6.91 (±2.8)
	(3,3)	1.91 (±2.16)	11.91 (±13.15)	7.57 (±2.50)	2.76 (±2.57)	10.77 (±8.35)
150	(1,3)	8.00 (±2.54)	6.50 (±2.71)	4.16 (±2.44)	6.64 (±3.00)	16.36 (±9.00)
	(3,1)	6.35 (±5.32)	6.03 (±2.90)	4.72 (±3.25)	6.47 (±3.87)	3.41 (±4.76)
	(3,3)	1.85 (±2.54)	9.46 (±6.59)	5.46 (±3.83)	8.57 (±5.82)	15.41 (±10.79)
200	(1,3)	2.35 (±2.53)	1.48 (±1.19)	3.94 (±2.37)	3.86 (±2.52)	13.25 (±6.64)
	(3,1)	1.80 (±1.02)	1.55 (±1.31)	3.88 (±2.76)	3.94 (±3.33)	7.21 (±6.20)
	(3,3)	1.64 (±1.31)	2.61 (±2.56)	5.91 (±5.13)	5.27 (±2.44)	15.36 (±10.65)

To be more specific, in these tables, we have recorded the relative improvement between the performance of each regressor during the initially provided labeled set, and the iteration that recorded the best performance until the criterion of either exceeding the *Max_iter* or not satisfying the consistency is violated. The results indicate that there is a decrease in the MAE metric, whilst the number of labeled instances is increased, as could be expected. Based only on the information regarding the first semester, it is noticed that the best performers are LR and MLP for size(L) = 50, while the tree-based learners achieved a more stable improvement over all the examined initially labeled subsets. Based on the information regarding both the first and the second semester, it is observed that the best performers are again LR and MLP for size(L) = 50, while they also performed greater improvement in the rest of the examined scenarios against their behavior on the previous case.

Additionally, we observe that as the cardinality of the L subset increases, the relative improvement of the investigated multi-view SSR approaches is decreasing in both cases during the majority of the recorded results. Through this kind of information, we can understand better the benefits of SSR approaches like COREG when multi-view problems are considered even under both limited labeled data are provided, and the volume of the unlabeled data is also highly restricted, reducing, thus the informativeness of this source of knowledge which is crucial for SSL scenario. Hence, the most important asset of transforming the COREG approach into a multi-view SSR variant is the remarkable reduction of the mean absolute error under strict conditions regarding the initially provided labeled instances. Despite the fact that the supervised learning performance in that cases is usually poor, since it heavily depends on the initially labeled data, both the insights that are obtained through the distinct, independent views and the disagreement mechanism that interchanges information between regressors that are fitted to these views lead to superior performance against it. Therefore, we believe that this indication is our most important contribution: Proof that in a real-life scenario, the complementary behavior of two separate views can be a trustworthy solution—even under highly limited labeled instances and not a large pool of unlabeled ones.

Another key is the fact that by mining additional unlabeled instances, we would expect even larger improvements in some cases, something that occurs by observing the fact that some approaches achieved their best performance at the late iterations, while almost none approach recorded its best performance during the early iterations. Thus, we are confident that by providing additional unlabeled instances, even better improvements should be achieved. Another interesting point that should be examined in the future is to insert a dynamic stage for terminating such a learning algorithm, avoiding saturation phenomena. A validation set could be useful, but small cardinality in a real-life dataset does not favor such a strategy.

Furthermore, in the majority of the presented results, we conclude that when the selectors coincide with the two 3NN algorithms, larger improvements of the relative error are recorded, especially for the more accurate models: GB-based variants and MLP. This happens due to the fact that in the majority of the cases that one selector coincides with the 1NN algorithm, this view through its fitted regressor does not detect any unlabeled instance that satisfies the consistency criterion. Hence, the other view is not actually enriched via the existence of annotated unlabeled instances. However, in the case of weaker regressors—3NN and LR—this behavior may be proven beneficial when noisy annotations take place, reducing, thus the chances of degeneration. To be complete with our experimental procedure, all our results are included in the following link: http://ml.math.upatras.gr/wp-content/uploads/2020/11/mdpi-Applied-Sciences-math-upatras-2020.7z, where the index of the best position per examined fold along with the improvement during the arbitrarily selected value of *Max_iter* are recorded per regressor based on the separate views, as well as the finally exported one. Furthermore, the supervised performance of the whole dataset D for both cases and each investigated regressor, as well as their performance on all the four separate initial versions of the L size, are included—facilitating each interested researcher about the efficacy of our approaches.

Regarding the interpretability of our results, we computed the Shapley values of each one of the five distinct regressors. To safely conclude that the COREG scheme can produce trustworthy explanations under the existence of limited labeled data per different learner, we made the next assumptions: We compared the purely supervised decisions of the total dataset evaluated with the aforementioned 5-fold-CV process per learner with the corresponding decisions that are exported by training the same regressor on the finally augmented L subset according to the adopted COREG scheme having fixed the choice of selector to (3NN, 3NN) with the pre-defined distance metrics as mentioned previously into this Section. Hopefully, in all the cases, we obtained similar enough decisions regarding the importance weights assigned to each indicator, while we had a perfect match between the ranking of the indicators. This fact verifies our main scope: To apply a multi-view SSR scheme that can improve the initial predictiveness of the model despite the limited number of the provided instances, acquiring at the same time trustworthy explanations about the importance of each included attribute.

Next, we present through suitable visualizations the SHAP values per case, exploiting the implementation provided by the authors of Reference [55]. Before we step to this stage, a short description is given regarding the two used approaches for computing these explainable weights that approximate the actual, but still computationally hungry Shapley values. First of all, a kernel-based approach was applied over all the five examined regressors (KernelSHAP), which is agnostic regarding the applied learning model and introduces a linear model that is fitted over the sampled pairs of (data, targets) and their generated weights. To generate these weights, several coalitions over the F space is produced, while the marginal distribution instead of the accurate conditional distribution is sampled for replacing the features that are absent during a random coalition. Although the assumptions here may lead to poor results because of the randomly selected coalitions that ignore some feature dependencies, the fact that a linear regression is applied during the last stage of the computation, additional strategies may easily be implemented trying to smoothen possible defects of this approximation (regularization, different learning model). On the other hand, a tree-based approach (TreeSHAP) has been applied in the case of GB-based approaches trying to figure out possible discrepancies between the explanation of this kind of learner. TreeSHAP constitutes an expansion of the KernelSHAP approximations, leading to faster results and facilitating the learners that are based on Decision Trees, integrating aggregating behavior through proper additive properties. Further information is provided in the original work [55].

We present here only the corresponding diagram of GB (ls) with both SHAP explainers, ignoring the similar enough performance of GB (huber), since it is the only tree-based regressor. The SHAP visualization plots (Figures 4–8) illustrate the attribute impact on the output of the produced regression model (the attributes are ranked in descending order from top to bottom) and how the attribute values impact the prediction (red color correlates to positive impact) in the first scenario using the D_1 dataset. Attributes Wri_1 (grade in the first written assignment), Ocs_1 (presence in the first optional contact session), and V_{31} (number of views in the module forum) are the most important ones in all cases regardless of the regressor employed. In addition, these attributes seem to positively influence the target attribute (i.e., student grade in the final examinations). Therefore, high-achieving students in the first written assignment, students with high participation rates in the first optional contact session, and students with high view rates in the module forum achieve a higher grade in the final course exam. Very similar results were produced regarding the second scenario using the D_2 dataset. In this case, attribute Wri_2 (grade in the second written assignment) proved to be the most significant, along with attribute V_{32} (number of views in the module forum).

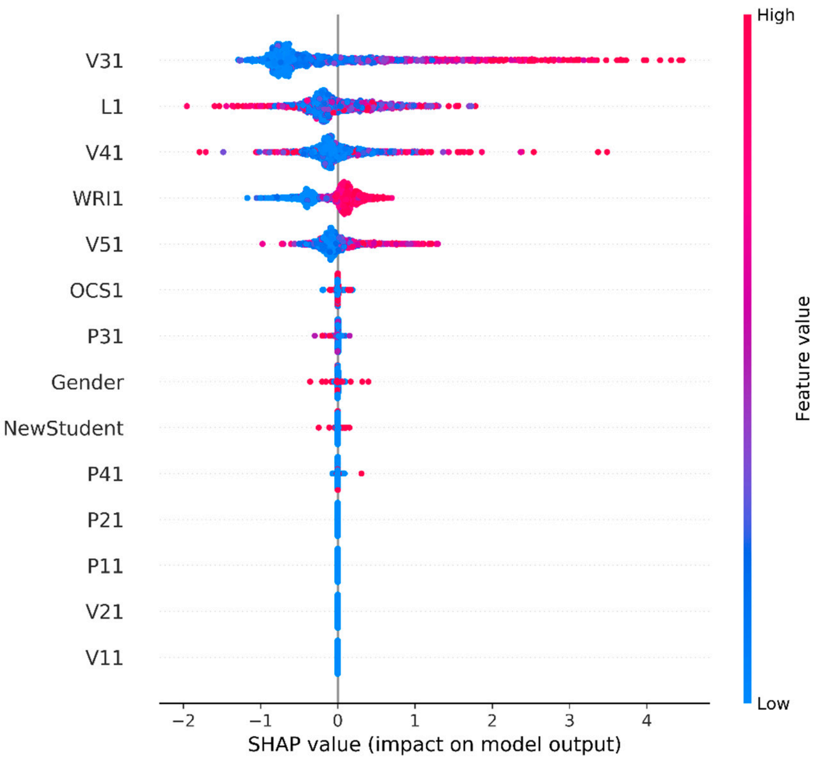

Figure 4. KernelSHAP values of the 5NN regressor (D_1 dataset).

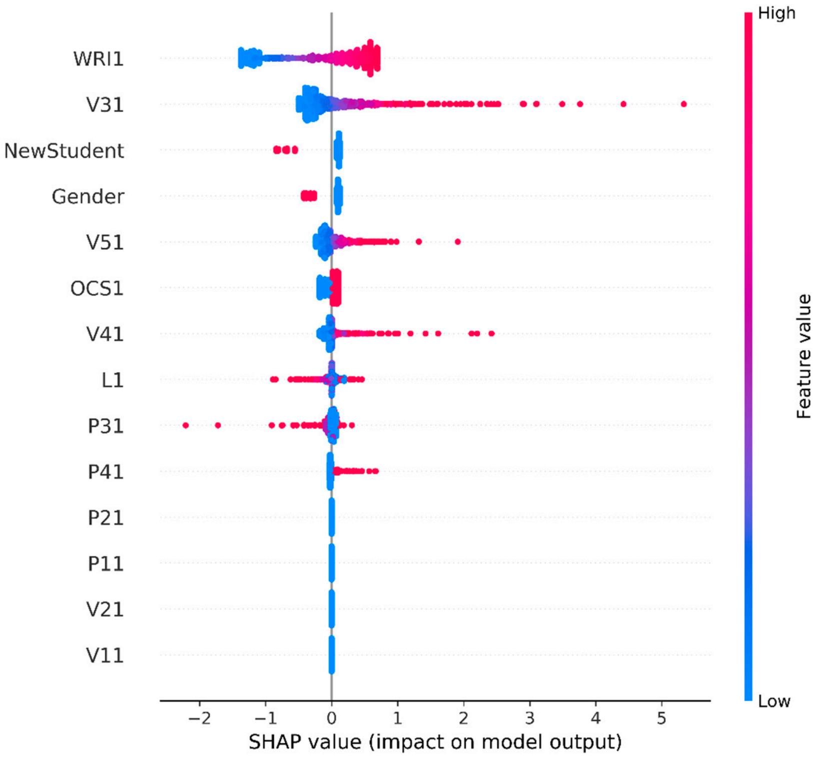

Figure 5. KernelSHAP values of the LR regressor (D_1 dataset).

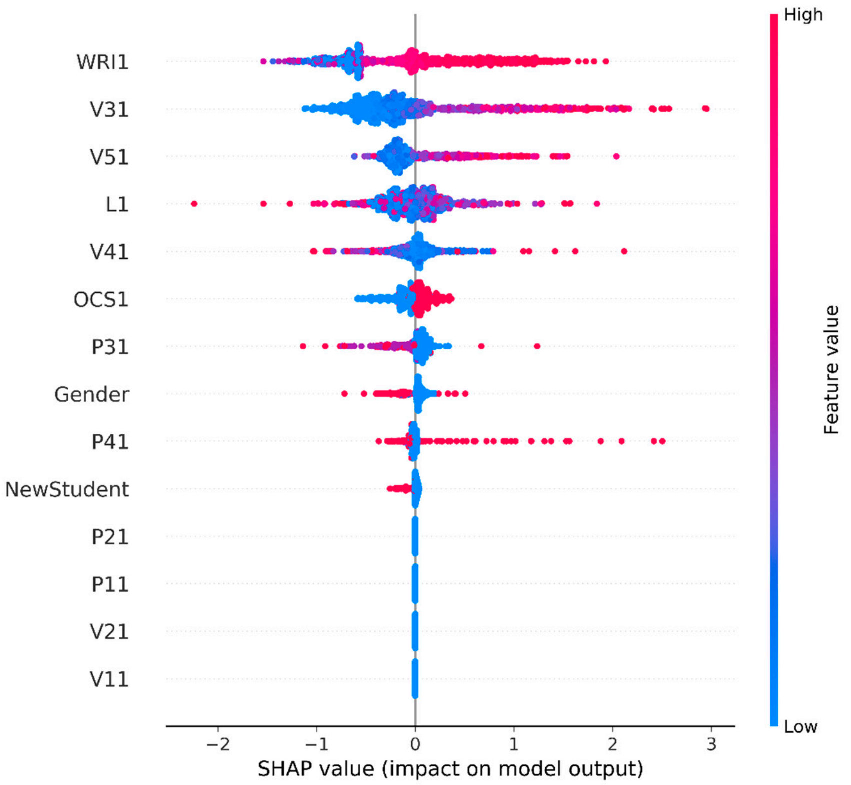

Figure 6. KernelSHAP values of the GB (ls) regressor (D_1 dataset).

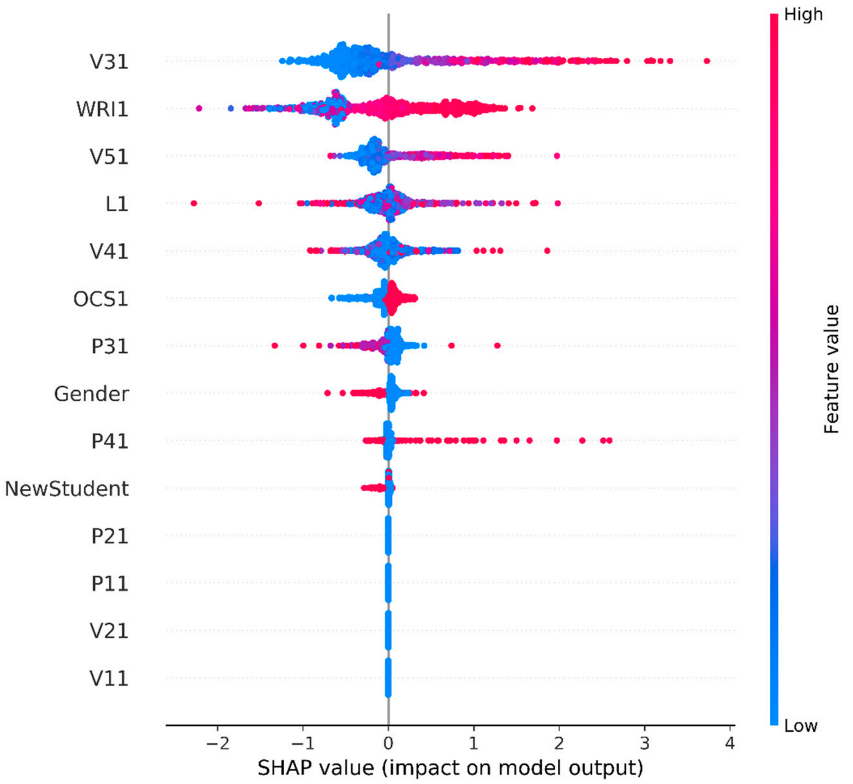

Figure 7. TreeSHAP values of the GB (ls) regressor (D_1 dataset).

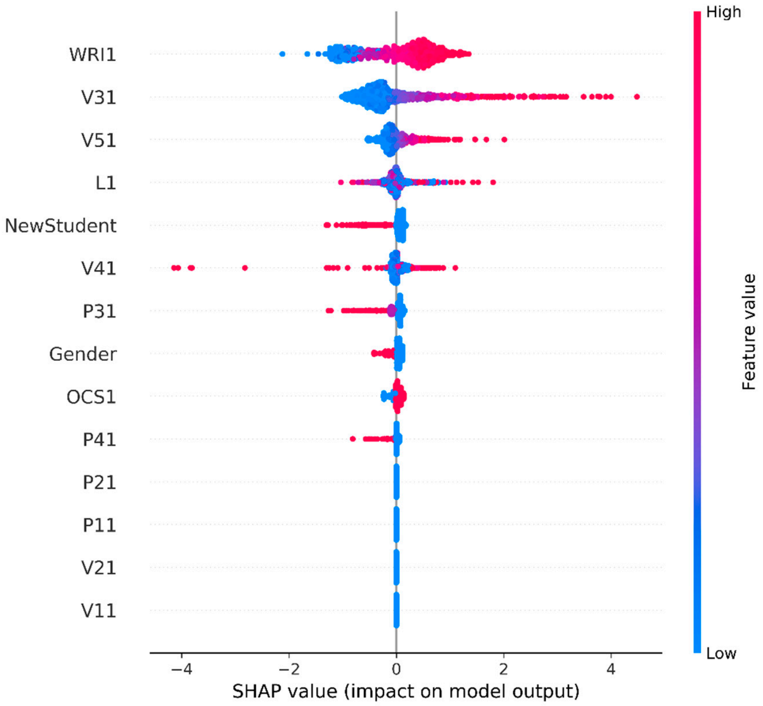

Figure 8. KernelSHAP values of the MLP regressor (D_1 dataset).

7. Conclusions

In the present study, an effort was made to build a highly-accurate semi-supervised regression model based on multi-view learning for the task of predicting student grades in a distance learning course. Additionally, we sought to gain insights and extract meaningful information from the model interpreting the predictions made and providing computed explanations about the predicted grades. The experimental results demonstrate the benefits brought by a natural split of the feature space. Therefore, our work contributes a different perspective to the existing single-view methods by fully exploiting the potential of different feature subsets by extending the COREG framework to the multi-view setting. In addition, it points out the importance of specific attributes that heavily influence the target attribute. Finally, the produced learning model may serve as an early alert tool for educators aiming at providing targeted interventions and support actions to low performers.

Generating synthetic data could be proven a highly favoring technique for mitigating the problem of limited labeled data. A recent demonstrated work has adopted such a strategy for training a boosting variant of the self-training scheme in the context of SSC [56]. In that work, the aspect of Natural Neighbors was preferred applying kNN algorithm as the base classifiers, and their obtained results seem encouraging enough for trying to extend their work also in our case. Another future direction could be applying pre-processing stages that may help us discriminate better the initially gathered data. Combination of semi-supervised Clustering either with conventional learners or ensembles, or even DNNs, as it has been validated in other real-life cases (e.g., geospatial data [57], medical image classification [58]) reducing inherent biases and helping us to uncover better possible underlying data relationships before the learning model could be found quite useful in practice. Another one possible effect of Clustering has been highlighted in Reference [50], where this strategy facilitated the scaling of a time-consuming learner over large volumes of unlabeled examples.

Finally, the strategy of transfer learning has been found great acceptance in the last years over several fields and could be proven beneficial in the case of EDM tasks. The two different aspects of this combination are expressed through either creating pre-trained models based on other learning tasks or enriching the discriminative ability of selected regressors through separate source domains that contain plentiful training data [59,60]. Combination of Active Learning with Semi-supervised learning might find great acceptance especially in cases that limited labeled data are provided, and the provided budget for monetization costs is highly bounded [61]. The modification also of transductive approaches for being considered under inductive learning scenarios seems a brilliant idea that compromises the accuracy of the former category and the generalization ability of the second one. Such a study was presented in Reference [62] and should be studied for SSR tasks.

Author Contributions: The authors contributed equally to the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baker, R.S.J.D.; Yacef, K. The state of educational data mining in 2009: A review and future visions. *JEDM J. Educ. Data Min.* **2009**, *1*, 3–17.
2. Baker, R. Data mining for education. *Int. Encycl. Educ.* **2010**, *7*, 112–118.
3. Costa, E.D.B.; Fonseca, B.; Santana, M.A.; De Araújo, F.F.; Rego, J.B.A. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.* **2017**, *73*, 247–256. [CrossRef]
4. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Fardoun, H.M.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]
5. Kostopoulos, G.; Karlos, S.; Kotsiantis, S.B. Multiview Learning for Early Prognosis of Academic Performance: A Case Study. *IEEE Trans. Learn. Technol.* **2019**, *12*, 212–224. [CrossRef]
6. Shelton, B.E.; Hung, J.-L.; Lowenthal, P.R. Predicting student success by modeling student interaction in asynchronous online courses. *Distance Educ.* **2017**, *38*, 59–69. [CrossRef]
7. Rahman, M.; Watanobe, Y.; Nakamura, K. Source Code Assessment and Classification Based on Estimated Error Probability Using Attentive LSTM Language Model and Its Application in Programming Education. *Appl. Sci.* **2020**, *10*, 2973. [CrossRef]
8. Zhu, X. *Semi-Supervised Learning Literature Survey*; University of Wisconsin-Madison: Madison, WI, USA, 2006.
9. Kostopoulos, G.; Karlos, S.; Kotsiantis, S.; Ragos, O. Semi-supervised regression: A recent review. *J. Intell. Fuzzy Syst.* **2018**, *35*, 1483–1500. [CrossRef]
10. Van Engelen, J.E.; Hoos, H. A survey on semi-supervised learning. *Mach. Learn.* **2019**, *109*, 373–440. [CrossRef]
11. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038. [CrossRef]
12. Xu, C.; Tao, D.; Xu, C. A Survey on Multi-view Learning. *arXiv* **2013**, arXiv:1304.5634.
13. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
14. Karlos, S.; Fazakis, N.; Kalleris, K.; Kanas, V.G.; Kotsiantis, S.B. An incremental self-trained ensemble algorithm. In Proceedings of the IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Rhodes, Greece, 25–27 May 2018; pp. 1–8. [CrossRef]
15. Karlos, S.; Fazakis, N.; Kotsiantis, S.; Sgarbas, K. Self-Trained Stacking Model for Semi-Supervised Learning. *Int. J. Artif. Intell. Tools* **2017**, *26*. [CrossRef]
16. Fu, B.; Wang, Z.; Xu, G.; Cao, L. Multi-label learning based on iterative label propagation over graph. *Pattern Recognit. Lett.* **2014**, *42*, 85–90. [CrossRef]
17. Kang, Z.; Lu, X.; Yi, J.; Xu, Z. Self-weighted multiple kernel learning for graph-based clustering and semi-supervised classification. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2312–2318. [CrossRef]
18. Wang, B.; Tsotsos, J.K. Dynamic label propagation for semi-supervised multi-class multi-label classification. *Pattern Recognit.* **2016**, *52*, 75–84. [CrossRef]
19. Luo, Y.; Ji, R.; Guan, T.; Yu, J.; Liu, P.; Yang, Y. Every node counts: Self-ensembling graph convolutional networks for semi-supervised learning. *Pattern Recognit.* **2020**, *106*, 107451. [CrossRef]

20. Ribeiro, F.D.S.; Calivá, F.; Swainson, M.; Gudmundsson, K.; Leontidis, G.; Kollias, S. Deep Bayesian Self-Training. *Neural Comput. Appl.* **2019**, *32*, 4275–4291. [CrossRef]
21. Iscen, A.; Tolias, G.; Avrithis, Y.; Chum, O. Label Propagation for Deep Semi-Supervised Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5065–5074.
22. Akusok, A.; Gritsenko, A.; Miche, Y.; Björk, K.-M.; Nian, R.; Lauren, P.; Lendasse, A. Adding reliability to ELM forecasts by confidence intervals. *Neurocomputing* **2017**, *219*, 232–241. [CrossRef]
23. Conati, C.; Porayska-Pomsta, K.; Mavrikis, M. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *arXiv* **2018**, arXiv:1807.00154.
24. Liz-Domínguez, M.; Caeiro-Rodríguez, M.; Llamas, M.; Mikic-Fonte, F.A. Systematic Literature Review of Predictive Analysis Tools in Higher Education. *Appl. Sci.* **2019**, *9*, 5569. [CrossRef]
25. Zhou, Z.-H.; Li, M. Semi-Supervised Regression with Co-Training. 2005. Available online: https://dl.acm.org/citation.cfm?id=1642439 (accessed on 31 October 2020).
26. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
27. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electron. J.* **2017**. [CrossRef]
28. Kanakaris, N.; Karacapilidis, N.; Kournetas, G. On the Exploitation of Textual Descriptions for a Better-informed Task Assignment Process. In Proceedings of the 9th International Conference on Operations Research and Enterprise Systems, {ICORES}, Valletta, Malta, 22–24 February 2020; Parlier, G.H., Liberatore, F., Demange, M., Eds.; SCITEPRESS: Setúbal, Portugal, 2020; pp. 304–310.
29. Chatzimparmpas, A.; Martins, R.M.; Jusufi, I.; Kerren, A. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf. Vis.* **2020**, *19*, 207–233. [CrossRef]
30. Lipton, Z.C. The mythos of model interpretability. *Queue* **2018**, *16*, 31–57. [CrossRef]
31. Hosseini, B.; Hammer, B. Interpretable Discriminative Dimensionality Reduction and Feature Selection on the Manifold. *Lect. Notes Comput. Sci.* **2020**, *11906 LNAI*, 310–326. [CrossRef]
32. Plumb, G.; Molitor, D.; Talwalkar, A.S. Model Agnostic Supervised Local Explanations. *Adv. Neural Inf. Process. Syst.* **2018**, 2520–2529.
33. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should {I} Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
34. Tan, S.; Caruana, R.; Hooker, G.; Lou, Y. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society—AIES '18*; ACM Press: New York, NY, USA, 2018; pp. 303–310. [CrossRef]
35. Mollas, I.; Bassiliades, N.; Vlahavas, I.P.; Tsoumakas, G. LionForests: Local interpretation of random forests. In *First International Workshop on New Foundations for Human-Centered AI (NeHuAI 2020)*; Saffioti, A., Serafini, L., Lukowicz, P., Eds.; CEUR: Aachen, Germany, 2020; pp. 17–24.
36. Houidi, S.; Fourer, D.; Auger, F. On the Use of Concentrated Time–Frequency Representations as Input to a Deep Convolutional Neural Network: Application to Non Intrusive Load Monitoring. *Entropy* **2020**, *22*, 911. [CrossRef]
37. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4768–4777.
38. Nuñez, H.; Maldonado, G.; Astudillo, C. Semi-supervised regression based on tree SOMs for predicting students performance. *IET Conf. Publ.* **2018**, *CP745*, 65–71. [CrossRef]
39. Kostopoulos, G.; Kotsiantis, S.; Fazakis, N.; Koutsonikos, G.; Pierrakeas, C. A Semi-Supervised Regression Algorithm for Grade Prediction of Students in Distance Learning Courses. *Int. J. Artif. Intell. Tools* **2019**, *28*, 1940001. [CrossRef]
40. Hady, M.; Schwenker, F. Co-Training by Committee: A Generalized Framework for Semi-Supervised Learning with Committees. *Int. J. Softw. Inform.* **2008**, *2*, 95–124. [CrossRef]
41. Brefeld, U.; Gärtner, T.; Scheffer, T.; Wrobel, S. Efficient co-regularised least squares regression. In Proceedings of the 23rd International Conference on World Wide Web-WWW '14, Seoul, Korea, 7–11 April 2006; pp. 137–144.

42. Liang, R.Z.; Xie, W.; Li, W.; Du, X.; Wang, J.J.Y.; Wang, J. Semi-supervised structured output prediction by local linear regression and sub-gradient descent. *arXiv* **2016**, arXiv:1606.02279.
43. Levatić, J.; Ceci, M.; Kocev, D.; Džeroski, S. Self-training for multi-target regression with tree ensembles. *Knowledge-Based Syst.* **2017**, *123*, 41–60. [CrossRef]
44. Kim, S.W.; Lee, Y.G.; Tama, B.A.; Lee, S. Reliability-Enhanced Camera Lens Module Classification Using Semi-Supervised Regression Method. *Appl. Sci.* **2020**, *10*, 3832. [CrossRef]
45. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Trans. Neural Networks* **2009**, *20*, 542. [CrossRef]
46. Zhou, Z.-H.; Li, M. Semi-supervised learning by disagreement. *Knowl. Inf. Syst.* **2010**, *24*, 415–439. [CrossRef]
47. Barreto, C.A.S.; Gorgônio, A.; Canuto, A.M.P.; João, C.X., Jr. A Distance-Weighted Selection of Unlabelled Instances for Self-training and Co-training Semi-supervised Methods. In *BRACIS*; Springer: Cham, Switzerland, 2020; pp. 352–366. [CrossRef]
48. Liu, Y.; Wang, L.; Mammadov, M. Learning semi-lazy Bayesian network classifier under the c.i.i.d assumption. *Knowledge-Based Syst.* **2020**, *208*, 106422. [CrossRef]
49. Fazakis, N.; Karlos, S.; Kotsiantis, S.; Sgarbas, K. A multi-scheme semi-supervised regression approach. *Pattern Recognit. Lett.* **2019**, *125*, 758–765. [CrossRef]
50. Guo, X.; Uehara, K. Graph-based Semi-Supervised Regression and Its Extensions. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*. [CrossRef]
51. Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. Efficient kNN Classification with Different Numbers of Nearest Neighbors. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 1774–1785. [CrossRef]
52. Karlos, S.; Kanas, V.G.; Aridas, C.; Fazakis, N.; Kotsiantis, S. Combining Active Learning with Self-train algorithm for classification of multimodal problems. In Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 15–17 July 2019; pp. 1–8. [CrossRef]
53. Nigam, K.; Ghani, R. Understanding the Behavior of Co-training. *Softwarepract. Exp.* **2006**, *36*, 835–844.
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Vanderplas, J. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830. [CrossRef]
55. Lundberg, S.M.; Erion, G.; Chen, H.; Degrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
56. Li, J.; Zhu, Q. A boosting Self-Training Framework based on Instance Generation with Natural Neighbors for K Nearest Neighbor. *Appl. Intell.* **2020**, *50*, 3535–3553. [CrossRef]
57. Yao, J.; Qin, S.; Qiao, S.; Che, W.; Chen, Y.; Su, G.; Miao, Q. Assessment of Landslide Susceptibility Combining Deep Learning with Semi-Supervised Learning in Jiaohe County, Jilin Province, China. *Appl. Sci.* **2020**, *10*, 5640. [CrossRef]
58. Peikari, M.; Salama, S.; Nofech-Mozes, S.; Martel, A.L. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. *Sci. Rep.* **2018**, *8*, 1–13. [CrossRef]
59. Tsiakmaki, M.; Kostopoulos, G.; Kotsiantis, S.B.; Ragos, O. Transfer Learning from Deep Neural Networks for Predicting Student Performance. *Appl. Sci.* **2020**, *10*, 2145. [CrossRef]
60. Wang, G.; Zhang, G.; Choi, K.-S.; Lam, K.-M.; Lu, J. Output based transfer learning with least squares support vector machine and its application in bladder cancer prognosis. *Neurocomputing* **2020**, *387*, 279–292. [CrossRef]
61. Karlos, S.; Kostopoulos, G.; Kotsiantis, S.B. A Soft-Voting Ensemble Based Co-Training Scheme Using Static Selection for Binary Classification Problems. *Algorithms* **2020**, *13*, 26. [CrossRef]
62. Yi, Y.; Chen, Y.; Dai, J.; Gui, X.; Chen, C.; Lei, G.; Wang, W. Semi-Supervised Ridge Regression with Adaptive Graph-Based Label Propagation. *Appl. Sci.* **2020**, *8*, 2636. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.