IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2017, accepted August 12, 2017, date of publication August 25, 2017, date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2740980

Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors

AMIN ZOLLANVARI[1], (Member, IEEE), REFIK CAGLAR KIZILIRMAK[1], (Member, IEEE), YAU HEE KHO[2], (Senior Member, IEEE), AND DANIEL HERNÁNDEZ-TORRANO[3]

[1]Department of Electrical and Computer Engineering, Nazarbayev University, Astana 010000, Kazakhstan
[2]School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand
[3]Graduate School of Education, Nazarbayev University, Astana 010000, Kazakhstan

Corresponding author: Amin Zollanvari (amin.zollanvari@nu.edu.kz)

ABSTRACT Predicting students' grades has emerged as a major area of investigation in education due to the desire to identify the underlying factors that influence academic performance. Because of limited success in predicting the grade point average (GPA), most of the prior research has focused on predicting grades in a specific set of classes based on students' prior performances. The issues associated with data-driven models of GPA prediction are further amplified by a small sample size and a relatively large dimensionality of observations in an experiment. In this paper, we utilize the state-of-the-art machine learning techniques to construct and validate a predictive model of GPA solely based on a set of self-regulatory learning behaviors determined in a relatively small-sample experiment. We quantify the predictability of each constituents of the constructed model and discuss its relevance. Ultimately, the goal of grade prediction in similar experiments is to use the constructed models for the design of intervention strategies aimed at helping students at risk of academic failure. In this regard, we lay the mathematical groundwork for defining and detecting *probably helpful interventions* using a probabilistic predictive model of GPA. We demonstrate the application of this framework by defining basic interventions and detecting those interventions that are probably helpful to students with a low GPA. The use of self-regulatory behaviors is warranted, because the proposed interventions can be easily practiced by students.

INDEX TERMS GPA, prediction, classification, intervention.

I. INTRODUCTION

Accurate prediction of students' academic performance in the context of higher education has become a fundamental instrument used to design effective strategies for student recruitment, admission, retention, and individualized educational support throughout a student's studies. However, numerous factors affect the student's performance. Identifying those factors and using them as predictors of success remains a complex problem.

Predictors of academic performance can be organized into three domains. First, traditional predictors of academic performance in university students include standardized intelligence and achievement test scores and academic performance in high school [1]–[3]. Second, student demographic characteristics such as age, gender, and socio-economic status are also among those factors that contribute to student success [4]. Third, during the last decades, increasing attention has been paid to the predictive ability of psychological factors on academic performance, including personality, motivation, and others [5]. In this context, the study and identification of cognitive and metacognitive self-regulatory learning strategies such as rehearsal, elaboration, concentration, help seeking, and time/study management, has proven particularly important in predicting academic performance and improving student learning [2], [6].

In parallel with the growing number of literature on academic performance predictors, research activities have progressively incorporated computational approaches using educational data. Different techniques and models have been applied, such as neural networks [7]–[11], Bayesian networks [12]–[16], rule-based systems [17]–[19], regression [20]–[24], and correlation analysis [25], [26].

These methods, which identify the major factors that influence and predict the overall student performance, differ in accuracy, complexity, and sample size requirements.

As a measure of academic performance, one may use either class-specific grades or grade point average (GPA). Although GPA seems to be a more reliable measure of academic performance, its prediction depends heavily on the reliability and uniformity of its constituents, i.e., the class-specific grades that form GPA [27]. In other words, large differences in grading standards coupled with the amount of elective courses that students can take complicates further the already difficult problem of GPA prediction.

The issues associated with data-driven models of GPA prediction are further amplified by a small sample size and a relatively large dimensionality of observations in an experiment. Several recent works have discussed the effect of having a small sample size in educational studies. For example, data collected in a few review studies [28]–[30] shows that 82 out of 185 studies evaluating a mathematics program have a sample participant size of between 30 and 200. Although a sample size of 200 might be sufficient to study the effect of one single factor, that small of a sample size would not be generally adequate when using traditional statistical methods to study the multivariate effect of, say, 50 factors.

Vapnik, one of the pioneers in machine learning, refers to situations wherein the ratio of sample size (n) to the number of variables in the study (p) is less than 20, i.e., $n/p < 20$ (see [31, p. 11]). A small-sample setting is not the place to rely on intuition nor classical statistical techniques [32]–[34]. Primarily, the classical notion of statistical consistency, which guarantees the performance of many classical statistical techniques, falters because this notion guarantees the performance of a method in scenarios where the number of observations unboundedly increases ($n \to \infty$) for a fixed p. In a finite sample regime, this implies that in order to expect an acceptable performance from a statistical technique, we need to have many more sample points than variables. In other words, one needs to either construct his own statistical learning methods that generalize classical statistical consistency (e.g., Girko G-estimation [35]), or use simplified assumptions on the model (e.g., assumption of model sparsity, variable independence, or first-order dependence tree).

In this work, we are primarily interested in predicting a student's GPA through the multivariate effect of self-regulatory learning behaviors. For this purpose, we conducted a questionnaire survey with 20 questions that reflect various learning strategies of students. Based on the relatively small number of collected samples, we constructed a predictive model of GPA using a maximum-weight first-order dependence tree (MWDT) structure. In this structure, the assumption is that each variable directly depends on only one other variable, commonly referred to as the *parent*. Such simplified assumptions are warranted in small sample situations because the number of sample points is very limited compared to the complexity of a joint distribution and estimating the full joint distribution from data leads to inaccurate results. When the joint distribution is approximated by the first-order tree dependence, we only need to estimate a series of second-order distributions.

Chow and Liu [36] laid down the groundwork for approximating and estimating the joint probability distribution of several variables based on the first-order dependence tree structure. Thus far, this model and its variant form, known as the "Tree Augmented Naive Bayes (TAN)" [37], have been used for classification purposes in various applications, including the classification of hand-printed numerals [36], software fault prediction [38], clinical decision support [39], and various speech and image processing applications [40]–[43]. Several empirical studies have shown that, in many settings, MWDT classifiers can outperform other well-known classifiers such as naive Bayes, which is constructed based on an assumption of variable independence [36], [37], [44]. At the same time, the graphical representation provided by this model shows the interaction and any existing synergy among predictors. In order to validate our predictive model and at the same time remove the effect of the so-called selection bias in small-sample, we utilize a cross-validation procedure *external* to feature selection. This procedure simulates real scenarios where independent data would need to be classified after the classifier is built. We further quantify the predictability of each variable in the constructed model of GPA prediction. Last, we lay the mathematical groundwork for defining interventions and detecting those interventions that are probably helpful to students with a low GPA.

This paper is organized as follows. Section II describes the maximum-weight dependence trees. Section III introduces the feature selection and model assessment. Section IV presents the questionnaire and data collection procedure. Section V presents the constructed data-driven graphical model for GPA prediction. Section VI discusses efficacy of each factor in predicting GPA. In Section VII, we formulate the concept of an intervention using our constructed model and detect those interventions that are probably helpful to students with low GPA. Finally, concluding remarks are presented in Section VIII.

II. MAXIMUM-WEIGHT DEPENDENCE TREES (MWDT)

First-order dependence tree of maximum weight is proposed initially by Chow and Liu [36] and further discussed and extended in [37]. To understand the working principle of first-order dependence tree, consider a vector of random variables $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ with the probability distribution $p(\mathbf{x})$. In this framework, $p(\mathbf{x})$, is approximated by a tree-dependence distribution $p_{\text{Tree}}(\mathbf{x})$ that is obtained by product of $p-1$ pairwise conditional probability distributions

$$p_{\text{Tree}}(\mathbf{x}) = \prod_{i=1}^{p} p(x_i|x_{m_i}), \tag{1}$$

where x_{m_i} is the "parent" of x_i, $m_1 \triangleq 0$, $(m_2, ..., m_p)$ is a permutation of an unknown subset of integers $\{2, 3, \ldots, p\}$, and $p(x_1|x_0) \triangleq p(x_1)$. We assume that $x_1, x_2, \ldots, x_p$ are ranked in

such a way that $m_i < i$, $i = 2, 3, \ldots, p$. A tree is then a graph that is uniquely defined by a p-tuple $\mathbf{m} = (m_1, m_2, ..., m_p)$ where the i-th element of $\mathbf{m}$ shows the parent of variable x_i. To construct the graph, we can assign a node to variable x_i and an edge from x_i to x_{m_i}.

Chow and Liu [36] considered the information theoretic distance measure, the Kullback-Leibler cross-entropy for discrete variables, to assess the goodness of approximating $p(\mathbf{x})$ by $p_{\text{Tree}}(\mathbf{x})$. In this regard, they attempted to find the tree dependence structure τ such that

$$\mathcal{D}_{KL}\big(p(\mathbf{x})||p_{\tau}(\mathbf{x})\big) \leq \mathcal{D}_{KL}\big(p(\mathbf{x})||p_{\text{Tree}}(\mathbf{x})\big), \quad \forall \text{Tree} \in T_p, \tag{2}$$

where T_p is the set of all possible tree dependence structures of p nodes such that $m_i < i$, $i = 2, 3, \ldots, p$, and

$$\mathcal{D}_{KL}\big(p(\mathbf{x})||p_{\text{Tree}}(\mathbf{x})\big) = \sum_{\mathbf{x}} p(\mathbf{x}) \frac{p(\mathbf{x})}{p_{\text{Tree}}(\mathbf{x})}. \tag{3}$$

In [36], it is proved that τ, the solution of the minimization problem (2), is the tree that has the maximum weight among all trees in T_p where the weight of each branch is determined by the mutual information of variables attached to that branch; to wit,

$$\begin{aligned} \text{MWDT} \triangleq \tau &= \underset{\text{Tree}\in T_p}{\text{argmin}}\ \mathcal{D}_{KL}\big(p(\mathbf{x})||p_{\text{Tree}}(\mathbf{x})\big) \\ &= \underset{\text{Tree}\in T_p}{\text{argmax}} \sum_{i=1}^{p} I(x_i, x_{m_i}), \end{aligned} \tag{4}$$

where $I(x_i, x_{m_i})$ is the mutual information between x_i and x_{m_i} defined as

$$I(x_i, x_{m_i}) = \sum_{x_i, x_{m_i}} p(x_i, x_{m_i}) \log \left(\frac{p(x_i, x_{m_i})}{p(x_i)p(x_{m_i})} \right). \tag{5}$$

In a two-class classification problem, we can define an auxiliary "class" random variable C with states "H" (high GPA) and "L" (low GPA) to measure the amount of information between variables given the class variable. In this case, the MWDT becomes the tree with the maximum $\sum_{i=1}^{p} I(x_i, x_{m_i}|C)$, where the class variable is constantly one of the two parents of every other variable, resulting in a tree-augmented structure. Figure 1 provides an example of a tree-augmented structure (when the same tree structure is trained on both classes) approximating a joint distribution of six variables conditional on the class variable. For an observation of unknown class then, one assigns a label H to the observation, if

$$\prod_{i=1}^{p} p(x_i|x_{m_i}, C = H) > \prod_{i=1}^{p} p(x_i|x_{m_i}, C = L). \tag{6}$$

III. FEATURE SELECTION AND MODEL ASSESSMENT

Perhaps the most important aspect of any classifier is its generalization error, defined as the probability of misclassification, since it quantifies the predictive capacity of the classifier. If samples are large, then part of the data can be held out for error estimation. However, when the number of available sample points is comparable in magnitude to the number of potential variables that can be used in the classifier, both the classification and error estimation rules are applied to the same set of training data—a situation that we faced in the current study. At the same time, it is the general consensus that the performance of a constructed classifier does not keep improving as more features (variables) are added to the model. This phenomenon is known as the *curse of dimensionality*, or the *peaking phenomenon* [45], [46]. This phenomenon enforces the procedure of the feature selection to be applied.

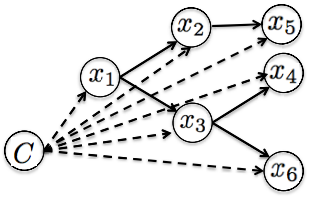

FIGURE 1. An example of a tree-augmented structure encoding $p_{\text{Tree}}(\mathbf{x}) = p(x_1|C) \times p(x_2|x_1, C) \times p(x_3|x_1, C) \times p(x_4|x_3, C) \times p(x_5|x_2, C) \times p(x_6|x_3, C)$ for approximating $p(\mathbf{x}) = p(x_1, x_2, x_3, x_4, x_5, x_6|C)$. Each node indicates a variable and an edge from one node (parent) to another (child) represents the conditional probability distribution of the child node given the state of its parent(s).

Once a feature selection procedure is applied, it is essential to evaluate the performance of each feature subset. In this regard, cross-validation (CV) is a common assessment strategy. However, to avoid bias selection, which results in an optimistic prediction error of the final constructed classifier [47], it is essential to apply the cross-validation procedure external to the feature-selection step [48]. In this work, we have applied this procedure to obtain a realistic view of prediction error (see Fig. 2 for a schematic description of the procedure).

IV. QUESTIONNAIRE, DATA COLLECTION, AND PREPROCESSING

Goldman and Slaughter [27] argued that since students in a specific department mostly take classes within their own major fields, the GPA level of students enrolled in different departments can have a different scale. In other words, some departments may give higher grades than others do even for the same level of performance—a scenario that can potentially undermines the validity of the GPA metric. In order to avoid this problem, we prepared and distributed a 20-question survey among second- to fourth-year students enrolled only in the electrical engineering program at Nazarbayev University. The survey was conducted for one week using Qualtrics, where 82 students responded. The collected data and the distribution of responses are presented in the Supplementary Table S1 and Table S2, respectively.

For the survey, we adopted and customized eighteen questions from [49] that in our opinion deal most directly with

TABLE 1. Students were asked to indicate their extent of agreement with each question. Q_1-Q_{18} choices: I) = Always, II) Most of the time, III) Sometimes, IV) Rarely. Q_{19} choices: I) 0:00-6:00, II) 6:01-12:00, III) 12:01-18:00, IV) 18:01-23:59.

	Question
Q_1	Do you allow time for exercise and socializing with friends?
Q_2	Do you get at least 6 hours of sleep each night?
Q_3	Do you study at least 2 hours per every hour of class?
Q_4	Do you have an area where you always go to study?
Q_5	Is your study area free of noise and distractions and comfortable?
Q_6	Can you study for at least half an hour without getting up, taking snack or phone breaks?
Q_7	Do you use your time between classes to study?
Q_8	Do you start reviewing for major exams at least 3 days in advance?
Q_9	How often do you know what types of questions will be on the test?
Q_{10}	Are you able to finish your tests in the allowed period of time?
Q_{11}	Do you do your homework problems and assignments without looking at the solutions?
Q_{12}	Do you ask questions in class when you don't understand a concept?
Q_{13}	Are you able to take notes in class, keep up with the instructor and understand the concepts at the same time?
Q_{14}	Do you review your notes after each class, preferably right after class?
Q_{15}	Do you make notes and highlight them as you read class materials at home?
Q_{16}	Can you read and learn at the rate of 12-15 pages per hour for history-type material?
Q_{17}	Can you concentrate and understand the material you read without re-reading a second or third time?
Q_{18}	Do you adjust your reading styles when you are reading for literature, social science, or science classes?
Q_{19}	In which of the following time intervals do you usually study?
Q_{20}	What is your cumulative GPA out of 4.00?

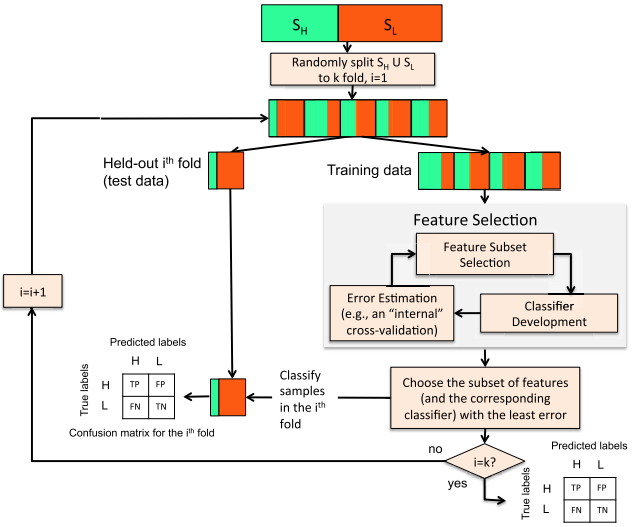

FIGURE 2. A schematic diagram of the cross-validation procedure external to the wrapper feature selection. S_H and S_L denote the full training data from the high and low GPA groups, respectively. We have used 10-fold cross-validation both external and internal to the feature selection process. Without having an external cross-validation the process of feature selection results in selection bias, which results in an optimistic generalization error of the constructed classifier. TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

students' personal learning strategies. These questions cover different aspects of self-regulatory learning behaviors such as time management, study environment, test taking/preparation skills, note-taking skills, reading and writing skills (see Table 1).Questions 19 and 20 were added to collect information on students' study time and their cumulative GPA, respectively. The students were asked to indicate their extent of agreement with each question: I) always; II) most of the time; III) sometimes; IV) rarely. For only Q_{19}, the available choices were the following: I) 0:00-6:00, II) 6:01-12:00, III) 12:01-18:00, or IV) 18:01-23:59.

In our study, the entire set of questions in the questionnaire, i.e., Q_1 to Q_{19}, form the set of potential random variables that can be used in the structure of the final constructed predictive model (through feature-selection process); to wit, each x_i in (1) is one of the questions among Q_1 to Q_{19}. This set of questions as a group take on values in a finite state of $b = 4^{19} > 2\times10^{11}$ possible states. At the same time, an upper bound on the number of states with an available measurement in our dataset is only 82 (responders) $\times$ 4 (options) $=$ 328. Comparing the potential number of possible states to the observed states implies the sparsity of the search space. To reduce the potential dimensionality of the search space, we combine the categories "always" and "most of the time" to a single group "always or most of the time" (hereafter, denoted by **A**), and similarly, the remaining response alternatives to a single category "sometimes or rarely" (hereafter, denoted by **B**). The same procedure has been applied to Q_{19} to combine study hours 0:00-6:00 with 6:01-12:00 (to form group **A**), and combine hours 12:01-18:00 with 18:01-23:59 (to form group **B**).

In order to convert the problem into a classification problem and at the same time avoid overfitting, we dichotomized the GPA response variable based on a threshold that was determined independently from the set of collected data. According to the latest statistics (as of 2013) available from [50], the average GPA of students across many public and private four-year colleges and universities in the United States is 3.15. We have set this value as our threshold and classified any GPA value above or below that as a *high* or *low* GPA, respectively.

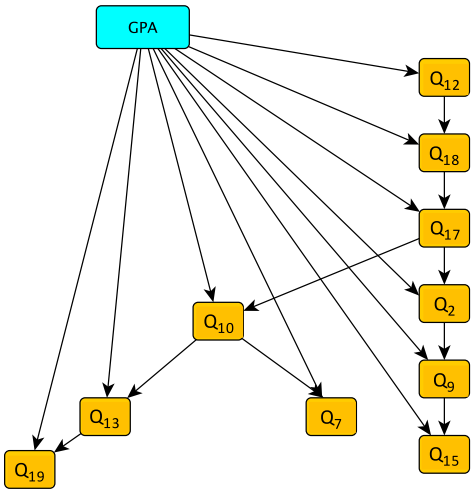

FIGURE 3. The network discriminating a low GPA from a high GPA based on self-regulatory learning behaviors. Hereafter, we refer to this network as the "GPA network".

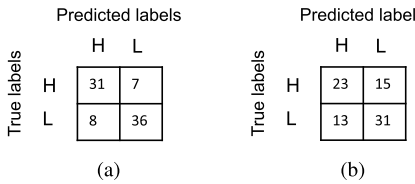

FIGURE 4. Confusion matrices of the constructed MWDT classifier in Fig. 3 as measured by: (a) resubstitution; (b) "external" cross-validation.

V. MODEL CONSTRUCTION AND VALIDATION

We applied an exhaustive search using a wrapper approach for feature subset selection [51]. In this regard, the predictive ability of all $2^{19} - 1 = 524,288$ possible feature subsets is evaluated using the MWDT structure and a 10-fold cross-validation. The final constructed MWDT structure is the classifier with the highest accuracy among all subsets of features. This classifier, which is depicted in Figure 3, is fully characterized by the conditional probabilities between nodes presented in the Supplementary Table S3. Out of 19 potential variables that were part of our survey (see Table 1), the constructed classifier employs 10 variables identified by Q_2, Q_7, Q_9, Q_{10}, Q_{12}, Q_{13}, Q_{15}, Q_{17}, Q_{18}, and Q_{19}. Denoting true positive, true negative, false positive, and false negative, by TP, TN, FP, and FN, respectively, the confusion matrix obtained by resubstitution accuracy estimator is presented in Figure 4a. In this case, the accuracy of the model measured by $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$ is 82%. However, this measure of accuracy is an over optimistic estimator. To have a realistic view of the predictive accuracy of the model, we applied the external cross-validation procedure detailed in Section III. Figure 4b presents the confusion matrix obtained by applying the external cross-validation procedure. Accordingly, the model has an accuracy of 65.85% with a sensitivity $\frac{\text{TP}}{\text{TP+FN}}$ of 63.9% and a specificity $\frac{\text{TN}}{\text{FP+TN}}$ of 67.4%.

VI. EFFICACY OF EACH VARIABLE IN DISCRIMINATING A HIGH AND LOW GPA

In this section, we seek to rank the factors that are parts of our "GPA network" based on their predictability. To formalize the idea, let Q_i, $\psi_{\mathcal{F}'|F}$, Ψ, and $\hat{\varepsilon}(\psi_{\mathcal{F}'|F})$, denote a variable, the classifier constructed with the set of variables $\mathcal{F}'$ out of initial set of variables $\mathcal{F}$ in the dataset, the classification rule Ψ (here wrapper feature selection with MWDT) applied to an initial set of variables $\mathcal{F}$, and the accuracy estimate of the classifier $\psi_{\mathcal{F}'|\mathcal{F}}$, respectively. In other words, $\psi_{\mathcal{F}'|\mathcal{F}}$ is a result of applying Ψ to the dataset. Note that when $\mathcal{F}'$ contains only one single variable, e.g., $\mathcal{F}' = \{Q_i\}$, the constructed MWDT network on the dataset with variables $\mathcal{F}$ reduces to discrete histogram rule [52]—there is no other variable to be used in conjunction with Q_i in the tree structure. The coefficient of determination (CoD) has been extended to and used in classification [53], [54] to find and rank variable(s):

$$\text{CoD} = \frac{\hat{\varepsilon}(\psi_{Q_i|\mathcal{F}}) - \hat{\varepsilon}(\psi_{0|\mathcal{F}})}{\hat{\varepsilon}(\psi_{Q_i|\mathcal{F}})}, \tag{7}$$

where $\hat{\varepsilon}(\psi_{Q_i|\mathcal{F}})$ is the accuracy of histogram rule constructed using Q_i and $\hat{\varepsilon}(\psi_{0|\mathcal{F}})$ is the accuracy of classification with no variable. Note that without any predictor, the classifier reduces to the majority vote; thus, in our case, $\hat{\varepsilon}(\psi_{0|\mathcal{F}}) = 44/82 = 53.65\%$. The CoD as defined in (7), measures the relative increase in the classification accuracy from using the variable Q_i in comparison with classifying the target in the absence of any predictor. Therefore, the higher the CoD for a variable, the greater the predictive capacity of the variable.

As indicated in [53], the CoD in the optimal sense (i.e., in sense where we have the full knowledge of class conditional densities) is a measure between 0 and 1 indicating the usefulness of a variable or a set of variables in filtering or classification. Nevertheless, estimating the CoD in practice even for variables that are part of the constructed classifier (here, the GPA network) may simply result in negative values including even values less than -1. Having a negative CoD is eventually in contrary to the fact that the variable has been part of the final constructed classifier. In [53] and [55] the negative values of estimated CoD from data has been simply defined to be 0 but then the question which remains is what would be the contribution of those variables that are part of the classifier and have a CoD of zero. As described in [53], having a negative CoD in practice is due to estimation in finite sample setting. For a very large sample size (compared to dimensionality of observation), we may not expect such a behavior because we can estimate the full distribution with a reasonable accuracy. Nevertheless, an inherent problem with CoD whether estimated or determined in an optimal sense is that it does not take into account the multivariate relationship of a variable of interest with other contributing variables used in classification rule Ψ.

To resolve the aforementioned issues associated with the use of CoD in ranking and quantifying usefulness of

TABLE 2. The efficacy of each variable in discriminating a high and low GPA. The values of each metric is multiplied by 100 to report in percentage. As seen CoD can have negative values. The overall ranking is based on CoP. However, both CoD and CoP can have potentially many ties in ranks due to use of an error-counting scheme (cross-validation). The values of SDP are used as tie-breaker in ranks determined by CoP. Although these metrics have been computed for all variables in Table 1, we only report the results for variables that are part of the GPA network.

Variable	CoD,	Rank	CoP,	Rank	SDP	Overall Rank
Q_2	-7.30%,	6	1.85%,	5	7.72%	9
Q_7	16.98%,	1	18.52%,	1	32.97%	1
Q_9	12.00%,	2	12.96%,	2	24.10%	2
Q_{10}	$\approx 0\%$,	5	$\approx 0\%$,	6	8.76%	10
Q_{12}	-9.98% ,	7	11.11%,	3	6.65%	7
Q_{13}	$\approx 0\%$,	5	12.96%,	2	10.90%	4
Q_{15}	12.00%,	2	12.96%,	2	20.53%	3
Q_{17}	6.38%,	4	7.41%,	4	13.33%	8
Q_{18}	-15.77%,	8	11.11%,	3	9.88%	6
Q_{19}	10.20%,	3	11.11%,	3	26.78%	5

variables, we reform the CoD and introduce the following new metric to which we refer as the Coefficient of Predictability (CoP):

$$\text{CoP} = \frac{\hat{\varepsilon}(\psi_{\mathcal{F}'|\mathcal{F}}) - \hat{\varepsilon}(\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i})}{\hat{\varepsilon}(\psi_{\mathcal{F}'|\mathcal{F}})}. \tag{8}$$

CoP measures the relative increase in the classification accuracy by using the full set $\mathcal{F}$ in comparison with the set $\mathcal{F}-Q_i$ obtained by omitting Q_i from $\mathcal{F}$. In order to find $\psi_{\mathcal{F}'|\mathcal{F}}$, we apply the classification rule Ψ to the set of variables $\mathcal{F}$, which generally results in a classifier constructed on a lower dimensional space $\mathcal{F}'$. Instead, to construct $\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i}$, we apply the same classification rule Ψ to the set of variables $\mathcal{F}-Q_i$, which results in a classifier constructed generally on a different set of variables than $\mathcal{F}'$, namely, $\mathcal{F}'_i$. In order to avoid overfitting, we assume we are blind to the set of variables selected as part of the GPA network. In other words, we assume $\mathcal{F}$ is the full set of variables in the original dataset. That is to say, in our case,

$$\mathcal{F}' = \{Q_2, Q_7, Q_9, Q_{10}, Q_{12}, Q_{13}, Q_{15}, Q_{17}, Q_{18}, Q_{19}\}, \tag{9}$$

$$\mathcal{F} = \bigcup_{i=1}^{19} \{Q_i\}, \tag{10}$$

and each time we remove a variable from $\mathcal{F}$, we apply the full external cross-validation (as described in Section III) to assess the performance of a newly constructed classifier $\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i}$. At the same time, applying an exhaustive wrapper feature selection procedure, as we have done in Section III, guarantees that

$$\hat{\varepsilon}(\psi_{\mathcal{F}'|\mathcal{F}}) \geq \hat{\varepsilon}(\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i}), \quad \forall i, \tag{11}$$

which leads to $\text{CoP} \geq 0$. The inequality (11) holds because the initial feature set $\mathcal{F}$ to construct $\psi_{\mathcal{F}'|\mathcal{F}}$ contains one more feature than the initial set of features $\mathcal{F} - Q_i$ that leads to $\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i}$. Therefore, using an exhaustive search leads to a an equal or smaller error of $\psi_{\mathcal{F}'|\mathcal{F}}$ than $\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i}$. The accuracy of the GPA network depicted in Fig. 3, i.e., $\hat{\varepsilon}(\psi_{\mathcal{F}'|\mathcal{F}}) = 65.85\%$, is also an upper bound on $\hat{\varepsilon}(\psi_{\mathcal{F}'_i|\mathcal{F}-Q_i})$, $\forall i$, and as a result we have $0 \leq \text{CoP} \leq 1$.

In order to calculate CoD and CoP, we employ cross-validation procedure, which is essentially an error-counting scheme and can potentially result in (relatively many) ties in ranks. To resolve this issue, we employ a Bayesian inferential scheme as a tie-breaker. In this regard, we define the Sum of absolute Difference between Posterior (SDP) probabilities of class variable given an evidence (the state of variable Q_i) as follows:

$$\text{SDP} = \frac{1}{2} \sum_{q_i \in \{\mathbf{A},\mathbf{B}\}} |\text{P}(\text{GPA} = \text{H}|Q_i = q_i) - \text{P}(\text{GPA} = \text{L}|Q_i = q_i)|, \tag{12}$$

where **A** and **B** are the newly formed categories of answers as described in Section IV. In order to find the posterior probabilities we adapt the bucket elimination algorithm [56]. Interpretation of SDP in terms of predictive capacity of each factor is not straightforward. Instead, SDP gives a relative difference in posterior probabilities of class variable conditioned on the state of each variable, which in turn accounts for the efficacy of each variable in discriminating a high and low GPA. Nevertheless, we are primarily interested in the predictive capacity of each factor and, in this regard, we utilize CoP as the main ranking metric and SDP as the tie-breaker.

Table 2 presents the efficacy of each variable in discriminating a high and low GPA. The 10 variables (Q_2, Q_7, Q_9, Q_{10}, Q_{12}, Q_{13}, Q_{15}, Q_{17}, Q_{18}, and Q_{19}) that contribute to the model in Fig. 3 can be associated with five self-regulatory learning behaviors from the literature [57]. The first self-regulatory learning behavior with a substantial predictive capacity on GPA is goal setting and planning, represented by Q_7, which is defined as student's ability to set goals, plan and organize their own learning. The second self-regulatory learning behavior with a discriminant capacity from the questionnaire relates to information seeking, or the ability to seek relevant information from nonsocial sources—like teachers or other students—when undertaking an assignment

or test. This learning strategy is represented by the variables Q_9, Q_{13}, and Q_{12}.

Thirdly, keeping records and monitoring also contribute to the model. This refers to student-initiated efforts to take notes and keep records about important information, events or results and is represented by variable Q_{15}. Fourthly, the ability to rearrange instructional materials and strategies to improve learning, labeled organization and transformation, also demonstrates predictive capacity on students' GPA. This learning strategy is represented by the variables Q_{18}, Q_2, and Q_{10}. Finally, rehearsing or student-initiated efforts to memorize material by overt or covert practice, represented by variable Q_{17}, also evidenced predictive capacity to discriminate student GPA. These results are in line with previous studies which show that students who exhibit higher academic achievements use more of these self-regulatory learning behaviors than those who exhibit lower academic achievements [58].

Interestingly, none of the variables related to the ability of students to select and arrange the physical setting to make learning easier (Q_4, Q_5) are included in our predictive model. This means that in our study these variables have no or negligible influence on predicting GPA when compared to the multivariate influence of other variables that are part of the network. Nonetheless, this observation does not align well with previous research, which evidenced the importance of the learning environment in the academic performance of university students (e.g., [59]), and further research needs to be conducted to verify and quantify the effect of these variables on GPA, if any.

VII. INTERVENTION STRATEGY

Although Table 2 provides an insight to the efficacy of each factor in discriminating a high and a low GPA, it does not provide us with an intervention strategy. An advantage of using the MWDT structure, which is essentially a Bayesian network suited for learning in small-sample situations, is its translatability into intervention strategies. In order to mathematically formalize an intervention strategy, we define an *observed* set of evidence e_O to be a set of answers given by a student to a set of questions from the GPA model,

$$e_O = \big\{(Q_i, q_i) | Q_i \in \mathcal{F}', q_i \in \{\mathbf{A}, \mathbf{B}\}\big\}, \tag{13}$$

where $\mathcal{F}'$ is the set of all variables in the classifier (in our case obtained from (9)), and $\mathbf{A}$ and $\mathbf{B}$ are the reformed categories of answers as described in Section IV. Similarly, let e_T be a set of *testable* evidence based on the set of questions in e_O that might help the student to improve the GPA,

$$e_T = \{(Q_i, p_i) | (Q_i, q_i) \in e_O, p_i \in \{\mathbf{A}, \mathbf{B}\}\}. \tag{14}$$

Assuming the student has a low GPA given e_O (otherwise, we assume there is no need for intervention), we define the Fold change in the Posterior Probability (FPP) of a high GPA given e_T compared to a low GPA given e_O as follows:

$$\mathrm{FPP}_{e_T|e_O} = \frac{\mathrm{P}(\mathrm{GPA} = \mathrm{H}|e_T)}{\mathrm{P}(\mathrm{GPA} = \mathrm{L}|e_O)}. \tag{15}$$

Definition 1: Given two sets of e_T and e_O and assuming the student has a low GPA given e_O, an intervention is defined to be the relative complement of e_T in e_O defined as

$$\mathrm{Int}_{e_T|e_O} \triangleq e_T \setminus e_O = \{(Q_i, q_i \to p_i) \mid (Q_i, q_i) \in e_O, (Q_i, p_i) \in e_T, q_i \neq p_i\}. \tag{16}$$

□

To mathematically characterize whether an intervention defined in (16) is helpful in terms of improving a student's GPA, we define the concept of a *helpful intervention*. Nevertheless, due to the probabilistic nature of our framework, we refer to such an intervention as "probably" a helpful one.

Definition 2: We refer to an intervention $\mathrm{Int}_{e_T|e_O}$ as *Probably a Helpful Intervention* (PHI), if, $\mathrm{FPP}_{e_T|e_O} > 1$. □

Definition 3: We define an intervention $\mathrm{Int}_{e_T|e_O}$ to be of the $m \lhd n$ order (m from n) if $|\mathrm{Int}_{e_T|e_O}| = m$ and $|e_O| = n$, where $|S|$ denotes the cardinality of set S and m and n are two integers such that $m \leq n$. In other words, e_O contains n tuples and e_T differs from e_O in m tuples. Note that definition (14), implies that $|e_O| = |e_T|$ so the cardinality of e_T is implicit in this definition. □

Example: As an example, suppose the set of observed evidence for a student is

$$e_O = \{(Q_2, \mathbf{A}), (Q_7, \mathbf{A}), (Q_9, \mathbf{B}), (Q_{17}, \mathbf{B})\}, \tag{17}$$

and the set of testable evidence is

$$e_T = \{(Q_2, \mathbf{A}), (Q_7, \mathbf{B}), (Q_9, \mathbf{B}), (Q_{17}, \mathbf{A})\}. \tag{18}$$

In this case,

$$\mathrm{Int}_{e_T|e_O} = \{(Q_7, \mathbf{A} \to \mathbf{B}), (Q_{17}, \mathbf{B} \to \mathbf{A})\}, \tag{19}$$

which is an intervention of the order $2 \lhd 4$. □

Note that depending on the set of observed evidence, testable evidence, and a predictive network of GPA, there could be potentially many interventions. Note that the cardinality of set $\mathcal{F}'$ (the number of variables in the GPA network) restricts the order of interventions, i.e., possible values of m and n in the following lemma (see Definition 3).

Lemma 1: For a GPA network of p variables (excluding the class variable) with each variable having two possible states, there are $\frac{2^{2p+2}-3\times 2^{p+1}+2}{3}$ possible $m \lhd n$ order interventions where $1 \leq m \leq n \leq p$.

Proof: For a fixed e_O with cardinality $n \leq p$, there are $\binom{n}{m}$ interventions of order $m \lhd n$ where $1 \leq m \leq n$. Therefore, for this fixed e_O, there exist $\sum_{m=1}^{n}\binom{n}{m} = 2^n - 1$ potential interventions of order $m \lhd n$. Note that there are 2^n possible set of observed evidence e_O of cardinality n. This means that there are in total $(2^n - 1) \times 2^n$ potential interventions of order $m \lhd n$ for all possible e_O of cardinality n. Since n can be any integer from 1 to p, there are in total $\sum_{n=1}^{p}(2^{2n} - 2^n)$ potential interventions and the result follows. □

In our case where the GPA network includes 10 variables (see Fig. 3), Lemma 1 suggests the existence of 1,396,054 possible interventions. A detailed characterization and description of these many interventions is not simply

TABLE 3. The set of probably helpful interventions (PHIs) of order $1 \lhd 1$ determined from the GPA network in Fig. 3. Note that the results of even rows can be obtained from the results of odd rows (see Lemma 2 and its proof). Note that from Definition 1, intervention $\text{Int}_{e_T|e_O}$ is defined only when a student has a low GPA given e_O.

	e_O	e_T	$\text{Int}_{e_T\|e_O}$	$\text{FPP}_{e_T\|e_O}$	is it a PHI?
1	$\{(Q_2, \mathbf{A})\}$	$\{(Q_2, \mathbf{B})\}$	$\{(Q_2, \mathbf{A} \to \mathbf{B})\}$	0.522/0.555=0.940	no
2	$\{(Q_2, \mathbf{B})\}$	$\{(Q_2, \mathbf{A})\}$	$\{(Q_2, \mathbf{B} \to \mathbf{A})\}$	(1-0.555)/(1-0.522)=0.930	no
3	$\{(Q_7, \mathbf{A})\}$	$\{(Q_7, \mathbf{B})\}$	$\{(Q_7, \mathbf{A} \to \mathbf{B})\}$	0.383/0.286=1.339	yes
4	$\{(Q_7, \mathbf{B})\}$	$\{(Q_7, \mathbf{A})\}$	$\{(Q_7, \mathbf{B} \to \mathbf{A})\}$	1.157	yes
5	$\{(Q_9, \mathbf{A})\}$	$\{(Q_9, \mathbf{B})\}$	$\{(Q_9, \mathbf{A} \to \mathbf{B})\}$	0.407/0.351=1.159	yes
6	$\{(Q_9, \mathbf{B})\}$	$\{(Q_9, \mathbf{A})\}$	$\{(Q_9, \mathbf{B} \to \mathbf{A})\}$	1.094	yes
7	$\{(Q_{10}, \mathbf{A})\}$	$\{(Q_{10}, \mathbf{B})\}$	$\{(Q_{10}, \mathbf{A} \to \mathbf{B})\}$	0.443/0.530=0.835	no
8	$\{(Q_{10}, \mathbf{B})\}$	$\{(Q_{10}, \mathbf{A})\}$	$\{(Q_{10}, \mathbf{B} \to \mathbf{A})\}$	0.843	no
9	$\{(Q_{12}, \mathbf{A})\}$	$\{(Q_{12}, \mathbf{B})\}$	$\{(Q_{12}, \mathbf{A} \to \mathbf{B})\}$	0.461/0.528=0.873	no
10	$\{(Q_{12}, \mathbf{B})\}$	$\{(Q_{12}, \mathbf{A})\}$	$\{(Q_{12}, \mathbf{B} \to \mathbf{A})\}$	0.875	no
11	$\{(Q_{13}, \mathbf{A})\}$	$\{(Q_{13}, \mathbf{B})\}$	$\{(Q_{13}, \mathbf{A} \to \mathbf{B})\}$	0.483/0.592=0.815	no
12	$\{(Q_{13}, \mathbf{B})\}$	$\{(Q_{13}, \mathbf{A})\}$	$\{(Q_{13}, \mathbf{B} \to \mathbf{A})\}$	0.789	no
13	$\{(Q_{15}, \mathbf{A})\}$	$\{(Q_{15}, \mathbf{B})\}$	$\{(Q_{15}, \mathbf{A} \to \mathbf{B})\}$	0.368/0.426=0.863	no
14	$\{(Q_{15}, \mathbf{B})\}$	$\{(Q_{15}, \mathbf{A})\}$	$\{(Q_{15}, \mathbf{B} \to \mathbf{A})\}$	0.908	no
15	$\{(Q_{17}, \mathbf{A})\}$	$\{(Q_{17}, \mathbf{B})\}$	$\{(Q_{17}, \mathbf{A} \to \mathbf{B})\}$	0.406/0.460=0.882	no
16	$\{(Q_{17}, \mathbf{B})\}$	$\{(Q_{17}, \mathbf{A})\}$	$\{(Q_{17}, \mathbf{B} \to \mathbf{A})\}$	0.909	no
17	$\{(Q_{18}, \mathbf{A})\}$	$\{(Q_{18}, \mathbf{B})\}$	$\{(Q_{18}, \mathbf{A} \to \mathbf{B})\}$	0.414/0.486=0.851	no
18	$\{(Q_{18}, \mathbf{B})\}$	$\{(Q_{18}, \mathbf{A})\}$	$\{(Q_{18}, \mathbf{B} \to \mathbf{A})\}$	0.877	no
19	$\{(Q_{19}, \mathbf{A})\}$	$\{(Q_{19}, \mathbf{B})\}$	$\{(Q_{19}, \mathbf{A} \to \mathbf{B})\}$	0.423/0.308=1.373	yes
20	$\{(Q_{19}, \mathbf{B})\}$	$\{(Q_{19}, \mathbf{A})\}$	$\{(Q_{19}, \mathbf{B} \to \mathbf{A})\}$	1.199	yes

feasible. Nevertheless, in order to demonstrate the application of the aforementioned framework to define and detect the set of probably helpful interventions, we only consider the set of interventions of order $1 \lhd 1$. Lemma 2 characterizes an interesting property of interventions of order $1 \lhd 1$.

Lemma 2: Assume each variable in the GPA network has only two possible states **A** and **B**. For $1 \leq i \leq p$, an intervention $\{(Q_i, \mathbf{A} \to \mathbf{B})\}$ of order $1 \lhd 1$ is a PHI if and only if $\{(Q_i, \mathbf{B} \to \mathbf{A})\}$ is a PHI.

Proof: If $\{(Q_i, \mathbf{A}, \mathbf{B})\}$ is a PHI, then

$$\text{FPP}_{e_T|e_O} = \frac{\text{P}\big(\text{GPA} = \text{H} \mid \{(Q_i, \mathbf{B})\}\big)}{\text{P}\big(\text{GPA} = \text{L} \mid \{(Q_i, \mathbf{A})\}\big)} > 1. \tag{20}$$

At the same time, note that we have,

$$\text{P}\big(\text{GPA} = \text{H} \mid \{(Q_i, \mathbf{B})\}\big) + \text{P}\big(\text{GPA} = \text{L} \mid \{(Q_i, \mathbf{B})\}\big) = 1. \tag{21}$$

Using (21) in (20) yields,

$$\text{FPP}_{e_O|e_T} = \frac{\text{P}\big(\text{GPA} = \text{H} \mid \{(Q_i, \mathbf{A})\}\big)}{\text{P}\big(\text{GPA} = \text{L} \mid \{(Q_i, \mathbf{B})\}\big)} > 1. \tag{22}$$

Similarly, we can prove that if the left side of (20) is less than 1, then the left side of (22) is less than 1 and the result follows. □

Table 3 presents the set of PHIs of order $1 \lhd 1$ determined from the GPA network in Fig. 3. These results suggest three areas to consider when designing an intervention to improve the academic performance of students in higher education. First, educators should pay attention to the way students use the time between classes (Q_7). Studying between classes seems to help students to achieve high academic performance. This is probably due to the fact that planning and using time effectively help students to achieve their goals and to close the gap between goal setting and its realization, as has been suggested in other studies [60], [61]. Some students, on the other hand, do not seem to benefit from this behavior, i.e., studying between classes does not increase those students' chances of obtaining a high GPA, so a more appropriate strategy for them might be to recommend that they rest and devote their time to other nonacademic activities between classes.

Second, the ability and disposition to learn about the types of questions that will be on a test (Q_9) also affect a university student's academic performance and, therefore, should be considered when designing an educational intervention in this direction. The contribution of this behavior to academic performance appears to be determined by the student's perception of whether seeking information about the characteristics of the test will preserve his/her sense of self-esteem. Thus, if a student believes that seeking information about the types of questions that will be included on the test will preserve or improve his/her sense of self-esteem, students will be more likely to engage in this behavior; on the other hand, if he/she considers that this behavior will have negative consequences for their self-esteem, then the student will be less likely to engage in seeking for this type of information [62]. Accordingly, encouraging students to seek information about tests and assignments in class and facilitating the access of relevant information about the format, number of questions, duration, assessment criteria, and scoring of a test (i.e., the value of particular questions or sections) seem to be successful intervention strategies that will increase students' academic achievement.

Third, the results evidenced that the time of day when students actually study should be also subject of consideration in an intervention targeted to improve increase students academic achievement in higher education. More specifically, students should have opportunities to explore their preferences associated with morning or evening activities (i.e., chorotype) and learn what time of the day they perform better to adjust, whenever possible, their study patterns to these dispositions, since synchronizing chorotype and study behaviors has shown to be positively related to student performance [63].

VIII. CONCLUSION

Predictive mathematical modeling of a student's performance is an imperative step in moving education in the direction of a predictive science. There is a large body of work on predicting students' academic performance as measured either by GPA or by class-specific grades based on a student's prior performance. Nevertheless, the problem of GPA prediction has proven to be a far more complicated process given the inherent limitations associated with the GPA metric. These problems are further amplified by a small experimental sample size, which is commonplace in educational studies.

The underlying hypothesis of this investigation is that a set of cognitive and metacognitive self-regulatory behaviors can influence the academic performance of students as measured by GPA. In this regard, we studied the possibility of predicting GPA solely based on students' self-regulatory behaviors rather than using their prior performance as in many previous studies—for instance, see [7], [11], [16], [64], [65], to just cite a few articles. Using a relatively small-sample cohort, we constructed a data-driven model for classifying low and high GPA students. Although the constructed model differentiates the set of training data with an accuracy of 82%, a more realistic view of the accuracy of the model, i.e., 65.85%, is achieved by a cross-validation procedure conducted externally to the feature-selection. This type of cross-validation simulates real scenarios where independent data would need to be classified after the classifier is built. The obtained accuracy is well above the baseline accuracy of 53.6% that one would obtain when always predicting a test sample by the majority class. We believe that a higher accuracy in predicting GPA is achieved by inclusion of extraneous variables that were not part of the designed questionnaire. Developing a broader questionnaire that includes a more comprehensive set of self-regulatory factors, conducting new data collection experiments, and constructing new predictive models need further research and we have left them for future investigations.

Ultimately, the main purpose of identifying the set of self-regulatory learning behaviors with a predictive capacity is to help students who demonstrate poor performance improve their performance. To achieve this ultimate goal, we laid the theoretical groundwork upon which we formulated the concept of helpful interventions and used our constructed model to nominate basic intervention strategies. The focus of our study is essentially warranted because it can be potentially used to nominate intervention targets and help to match students with the prevention and intervention strategies most likely to work for them. This line of research has also important implications on how teachers can interact with students based on their personality and behavior. In the future, we also investigate the possibility of combining students' self-regulatory behaviors with their prior performance (e.g., high-school grades) to improve the accuracy of constructed predictive models. Such investigations may also enable the possibility of proposing intervention strategies based on a joint combination of self-regulatory factors and students' prior performance.

ACKNOWLEDGEMENT

The authors wish to thank the final year undergraduate students Kamilla Aliakhmet and Diana Zhussip for assisting in data collection.

REFERENCES

[1] T. Chamorro-Premuzic and A. Furnham, "Personality, intelligence and approaches to learning as predictors of academic performance," *Pers. Individual Differences*, vol. 44, no. 7, pp. 1596–1603, May 2008.
[2] A. L. Dent and A. C. Koenka, "The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis," *Edu. Psychol. Rev.*, vol. 28, no. 3, pp. 1–50, Sep. 2016.
[3] A. Furnham, T. Chamorro-Premuzic, and F. McDougall, "Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance," *Learn. Individual Differences*, vol. 14, no. 1, pp. 47–64, 2002.
[4] D. Voyer and S. D. Voyer, "Gender differences in scholastic achievement: A meta-analysis," *Psychol. Bull.*, vol. 140, no. 4, p. 1174, 2014.
[5] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of University students' academic performance: A systematic review and meta-analysis," *Psychol. Bull.*, vol. 138, no. 2, p. 353, 2012.
[6] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Edu. Psychol.*, vol. 25, no. 1, pp. 3–17, 1990.
[7] V. O. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba, "Predicting students' academic performance using artificial neural network: A case study of an engineering course," *Pacific J. Sci. Technol.*, vol. 9, no. 1, pp. 72–79, 2008.
[8] T. Wang and A. Mitrovic, "Using neural networks to predict student's performance," in *Proc. Int. Conf. Comput. Edu.*, Dec. 2002, pp. 969–973.
[9] T. D. Gedeon and S. Turner, "Explaining student grades predicted by a neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 1. Nagoya, Japan, Oct. 1993, pp. 609–612.
[10] M. D. Calvo-Flores, E. G. Galindo, M. C. P. Jiménez, and O. P. Piñeiro, "Predicting studentsŠ marks from Moodle logs using neural network models," *Current Develop. Technol.-Assisted Edu.*, vol. 1, no. 2, pp. 586–590, 2006.
[11] L. V. Fausett and W. Elwasif, "Predicting performance from test scores using backpropagation and counterpropagation," in *Proc. IEEE Int. Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, vol. 5. Jun. 1994, pp. 3398–3402.
[12] N. T. N. Hien and P. Haddawy, "A decision support system for evaluating international student applications," in *Proc. 37th Annu. Frontiers Edu. Conf.-Global Eng., Knowl. Without Borders, Opportunities Without Passports*, Oct. 2007, pp. F2A-1–F2A-6.
[13] Z. Pardos, N. Heffernan, C. Ruiz, and J. Beck, "The composition effect: Conjuntive or compensatory? An analysis of multi-skill math questions in ITS," in *Proc. 1st Int. Conf. Edu. Data Mining*, Montreal, QC, Canada, 2008, pp. 147–156.
[14] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, "The effect of model granularity on student performance prediction using Bayesian networks," in *Proc. Int. Conf. User Modeling*, 2007, pp. 435–439.

[15] R. Stevens, A. Soller, A. Giordani, L. Gerosa, M. Cooper, and C. Cox, "Developing a framework for integrating prior problem solving and knowledge sharing histories of a group to predict future group performance," in *Proc. Int. Conf. Collaborative Comput., Netw., Appl. Worksharing*, 2005, p. 9.

[16] E. Ayers and B. W. Junker, "Do skills combine additively to predict task difficulty in eighth grade mathematics," in *Proc. Edu. Data Mining, Papers AAAI Workshop*, Menlo Park, CA, USA, 2006, pp. 14–20.

[17] P. García, A. Amandi, S. Schiaffino, and M. Campo, "Evaluating Bayesian networks' precision for detecting students' learning styles," *Comput. Edu.*, vol. 49, no. 3, pp. 794–808, Nov. 2007.

[18] C.-C. Chan, "A framework for assessing usage of Web-based e-learning systems," in *Proc. 2nd Int. Conf. Innov. Comput., Inf. Control (ICICIC)*, Sep. 2007, p. 147.

[19] C.-M. Chen, M.-C. Chen, and Y.-L. Li, "Mining key formative assessment rules based on learner profiles for Web-based learning systems," in *Proc. 7th IEEE Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2007, pp. 584–588.

[20] S. B. Kotsiantis and P. E. Pintelas, "Predicting students marks in hellenic open University," in *Proc. 5th IEEE Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2005, pp. 664–668.

[21] N. O. Anozie and B. W. Junker, "Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system," in *Proc. Edu. Data Mining, Papers From AAAI Workshop*, Menlo Park, CA, USA, 2006, pp. 1–6.

[22] P. Golding and O. Donaldson, "Predicting academic performance," in *Proc. Frontiers Edu. 36th Annu. Conf.*, Oct. 2006, pp. 21–26.

[23] D. Martinez, "Predicting student outcomes using discriminant function analysis," 2001.

[24] N. Myller, J. Suhonen, and E. Sutinen, "Using data mining for improving Web-based course design," in *Proc. Int. Conf. Comput. Edu.*, Dec. 2002, pp. 959–963.

[25] D. Pritchard and R. Warnakulasooriya, "Data from a Web-based homework tutor can predict student's final exam score," in *Proc. World Conf. Edu. Multimedia, Hypermedia Telecommun.*, 2005, pp. 2523–2529.

[26] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[27] R. D. Goldman and R. E. Slaughter, "Why college grade point average is difficult to predict," *J. Edu. Psychol.*, vol. 68, no. 1, pp. 9–14, 1976.

[28] R. Slavin and D. Smith, "The relationship between sample sizes and effect sizes in systematic reviews in education," *Edu. Eval. Policy Anal.*, vol. 31, no. 4, pp. 500–506, 2009.

[29] R. Slavin and C. Lake, "Effective programs in elementary mathematics: A best-evidence synthesis," *Rev. Edu. Res.*, vol. 78, no. 3, pp. 427–515, 2008.

[30] R. Slavin, C. Lake, and C. Groff, "Effective programs in middle and high school mathematics: A best-evidence synthesis," *Rev. Edu. Res.*, vol. 79, no. 2, pp. 839–911, 2009.

[31] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[32] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, "The illusion of distribution-free small-sample classification in genomics," *Curr. Genomics.*, vol. 12, no. 5, pp. 333–341, 2011.

[33] A. Zollanvari, "High-dimensional statistical learning: Roots, justifications, and potential machineries," *Cancer Informat.*, vol. 5, pp. 109–121, Apr. 2015.

[34] R. Clarke *et al.*, "The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data," *Nature Rev. Cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[35] V. L. Girko, *Statistical Analysis of Observations of Increasing Dimension*, Dordrecht, The Netherlands: Kluwer, 1995.

[36] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[37] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2, pp. 131–163, Nov. 1997.

[38] K. Dejaeger, T. Verbraken, and B. Baesens, "Toward comprehensible software fault prediction models using Bayesian network classifiers," *IEEE Trans. Softw. Eng.*, vol. 39, no. 2, pp. 237–257, Feb. 2013.

[39] M. B. Sesen, A. E. Nicholson, R. Banares-Alcantara, and M. Brady, "Bayesian networks for clinical decision support in lung cancer care," *PLoS ONE*, vol. 8, no. 12, p. e82349, 2013.

[40] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[41] V. Y. F. Tan, S. Sanghavi, J. W. Fisher, and A. S. Willsky, "Learning graphical models for hypothesis testing and classification," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5481–5495, Nov. 2010.

[42] S. Tschiatschek and F. Pernkopf, "On Bayesian network classifiers with reduced precision parameters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 774–785, Apr. 2015.

[43] R. Mottaghi, S. Fidler, A. Yuille, R. Urtasun, and D. Parikh, "Human-machine CRFs for identifying bottlenecks in scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 74–87, Jan. 2016.

[44] M. G. Madden, "On the classification performance of TAN and general Bayesian networks," *Knowl.-Based Syst.*, vol. 22, no. 7, pp. 489–495, Oct. 2009.

[45] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY, USA: Wiley, 2004.

[46] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate Gaussian data," *Pattern Recognit.*, vol. 10, nos. 5–6, pp. 365–374, 1978.

[47] G. J. McLachlan, "The bias of the apparent error rate in discriminant analysis," *Biometrika*, vol. 63, no. 2, pp. 239–244, 1976.

[48] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6562–6566, 2002.

[49] V. N. Gordon, P. Royster, and T. L. Minnich, "University survey: A guidebook and readings for new students," Ohio State Univ., Columbus, OH, USA, Tech. Rep., 2000.

[50] (2017). *Grade Inflation at American Colleges and Universities*. [Online]. Available: http://www.gradeinflation.com/

[51] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[52] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.

[53] E. R. Dougherty, S. Kim, and Y. Chen, "Coefficient of determination in nonlinear signal processing," *Signal Process.*, vol. 80, no. 10, pp. 2219–2235, Oct. 2000.

[54] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[55] U. M. Braga-Neto and E. R. Dougherty, "Bolstered error estimation," *Pattern Recognit.*, vol. 37, no. 6, pp. 1267–1281, Jun. 2004.

[56] R. Dechter, "Bucket elimination: A unifying framework for probabilistic inference," in *Proc. 12th Conf. Uncertainty Artif. Intell.*, Portland, OR, USA, 1996, pp. 211–219.

[57] B. J. Zimmerman and M. M. Pons, "Development of a structured interview for assessing student use of self-regulated learning strategies," *Amer. Edu. Res. J.*, vol. 23, no. 4, pp. 614–628, 1986.

[58] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," *Theory Pract.*, vol. 41, no. 2, pp. 64–70, 2002.

[59] A. Lizzio, K. Wilson, and R. Simons, "University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice," *Stud. Higher Edu.*, vol. 27, no. 1, pp. 27–52, 2002.

[60] H. Aarts, A. P. Dijksterhuis, and C. Midden, "To plan or not to plan? Goal achievement or interrupting the performance of mundane behaviors," *Eur. J. Social Psychol.*, vol. 29, no. 8, pp. 971–979, 1999.

[61] P. M. Gollwitzer, G. Oettingen, K. Vohs, and R. Baumeister, "Planning promotes goal striving," in *Handbook of Self-Regulation: Research, Theory*, vol. 2. New York, NY, USA: Guilford Press, 2011, pp. 162–185.

[62] S. A. Karabenick and J. R. Knapp, "Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students," *J. Edu. Psychol.*, vol. 83, no. 2, p. 221, 1991.

[63] D. Goldstein, C. S. Hahn, L. Hasher, U. J. Wiprzycka, and P. D. Zelazo, "Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: Is there a synchrony effect?" *Pers. Individual Differences*, vol. 42, no. 3, pp. 431–440, Feb. 2007.

[64] C. S. Davis, C. L. Akers, C. J. Green, and R. E. Zartman, "Variables that influences student performance in an introductory soils class," *J. Natural Resour. Life Sci.*, vol. 35, no. 1, pp. 127–131, 2006.

[65] J. D. Vitale, S. P. Wanger, and D. C. Adams, "Explaining student performance in an undergraduate agricultural economics classroom," *NACTA J.*, vol. 54, no. 1, pp. 2–9, Mar. 2010.

AMIN ZOLLANVARI (M'10) received the B.Sc. and M.Sc. degrees in electrical engineering from Shiraz University, Iran, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, in 2010. He held a postdoctoral position at the Harvard Medical School and the Brigham and Women's Hospital, Boston, MA, from 2010 to 2012, and then joined the Department of Statistics, Texas A&M University, as an Assistant Research Scientist, from 2012 to 2014. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan. His research interest includes signal processing, statistical and machine learning, and bioinformatics.

REFIK CAGLAR KIZILIRMAK was born in Izmir, Turkey, in 1981. He received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bilkent University in 2004 and 2006, respectively, and the Ph.D. degree in electrical and electronics engineering from Keio University, Yokohama, Japan, in 2010. He has been involved in projects related to defense and security as a Technical Leader in Turkey. He is currently with the School of Engineering, Nazarbayev University, Astana, Kazakhstan. His research interests are in the field of communication theory and signal processing. He was a recipient of the IEEE VTC Young Researcher Encouragement Award in 2008.

YAU HEE KHO (S'06–M'08–SM'12) was born in Kuching, Malaysia, in 1974. He received the B.Eng. (Hons.) and Ph.D. degrees from the University of Canterbury, Christchurch, New Zealand, in 1997 and 2008, respectively. From 1998 to 2002, he was an Electrical Design Engineer in Singapore. He joined the Swinburne University of Technology (Sarawak Campus) as a Senior Lecturer from 2009 to 2013. He was an Assistant Professor with Nazarbayev University from 2013 to 2017. He is currently with the Victoria University of Wellington, New Zealand. His research interests include wireless communications, signal processing, and electronics and engineering education. He is a Chartered Engineer with the Engineering Council, U.K., and a Senior Fellow of the Higher Education Academy, U.K. He is also a member of the Institution of Engineering and Technology, U.K. He received the Professional Certificate in learning and teaching (Higher Education) from Swinburne University of Technology, Melbourne, Australia, in 2013.

DANIEL HERNÁNDEZ-TORRANO received the Ph.D. degree in educational psychology from the University of Murcia, Spain. He has held research positions at the University College London, U.K., the Universidade do Minho, Portugal, and the University of Connecticut, USA. He is currently an Assistant Professor with the Graduate School of Education, Nazarbayev University. His main areas of interest are academic excellence, talent development, gifted education, social and emotional factors influencing learning, and young people's wellbeing.