Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach

Mirka Saarela[*,†] MIRKA.SAARELA@JYU.FI
Bülent Yener[†] YENER@CS.RPI.EDU
Mohammed J. Zaki[†] ZAKI@CS.RPI.EDU
Tommi Kärkkäinen[*] TOMMI.KARKKAINEN@JYU.FI

[*]University of Jyväskyla, Department of Mathematical Information Technology, 40014 Jyväskylä, Finland
[†]Rensselaer Polytechnic Institute, Computer Science Department, 12180 Troy, New York, USA

Abstract

Large-scale educational assessment studies (LSAs) regularly collect massive amounts of very rich cognitive and contextual data of whole student populations. Currently, LSAs are limited to reporting student proficiencies in the form of *plausible values* (PVs). PVs are random draws from the posterior distribution of a student's ability, which is based on the Bayesian approach with the prior distribution modeling the student background within the population and the likelihood test item response using the Rasch model. While PVs have shown to be a reliable estimate for proficiencies of populations, a more comprehensive study of these rich data sets by deploying machine learning algorithms may provide a better understanding of the underlying factors affecting student performance and thus yield to better and more interpretable predictive models. This paper presents such a novel approach to learn directly from LSA data by deploying a combination of both unsupervised and supervised learning feature selection algorithms to predict student performance on math scores. Our technique learns the difficulty level of different math questions and predicts weather or not a student with a particular background profile will be successful in answering correctly.

1. Introduction

Since 2000 triennially, the Organisation for Economic Co-operation and Development (OECD) collects a massive amount of data of stratified samples of 15-year-old students from all over the world for the Programme for International Student Assessment (PISA). The sampled students not only take a cognitive test—in which they have to demonstrate their math, reading and science skills—but also reply to a questionnaire, in which they provide information about their social and economical background, as well as their motivations, behaviors, and attitudes towards various aspects of education. All collected data is publicly available[1] and according to the OECD, of very high quality in terms of degree of validity and reliability (OECD, 2009; 2012). Moreover, these data are comparable throughout different countries so that they provide a very rich database for educational machine learning (ML) and data mining (DM) applications.

The participating countries pay large sums of money (Musik, 2016) primarily with the goal to utilize PISA data and analysis results for research. However, as concluded by Rutkowski et al. (2010), not many researchers work with these freely available and high quality datasets because of the many technical complexities within them. The major difficulty of conducting secondary analysis with PISA data is that many desired properties that describe the students are not originally observed features, but are already pre-processed and made available as derived variables through a combination of different state-of-the-art methodologies. One example is that there are no single performance scores for the cognitive test in PISA datasets. Instead, for each student and each assessment domain—reading, math, and science—five plausible values (PVs) are reported.

The PVs are random draws from the posterior distribution of a student's ability, which is defined as

$$f(\beta \mid x_i, y_i) \propto P(x_i \mid \beta, \delta) f(\beta \mid \lambda, y_i), \qquad (1)$$

MLDEAS workshop papers of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

[1]PISA data can be downloaded from http://www.oecd.org/pisa/pisaproducts/.

where $P(x_i \mid \beta, \delta)$ denotes a Rasch Model (Rasch, 1960) given the student's ability β and the test items' difficulties δ, and $f(\beta \mid \lambda, y_i)$ denotes a population model with the background information of the student encoded in y_i[2]. This population model for a student i is estimated with the latent regression model (Tarpey & Petkova, 2010) $\beta_i = y_i^T \lambda + \epsilon_i$, where $\epsilon_i = \mathcal{N}(0, \sigma^2)$ (Marsman, 2014; OECD, 2014), and with λ denoting the regression coefficients.

PVs have shown to be a reliable estimate for proficiencies of populations (Monseur & Adams, 2008; Wu & Adams, 2002; OECD, 2009) and are used not only in PISA, but also in other LSA studies, such as the National Assessment of Educational Progress[3], the European Survey on Language Competences[4], the Trends in International Mathematics and Science Study, and the Progress in International Reading Literacy Study[5]. However, these estimations are done on normalized data and are based on linear regression (i.e., the λ parameter in f above). Thus, it is worth investigating how deploying a general framework of ML can complement the current state of art by using the raw data which is publicly available.

In this paper, we describe a ML approach that combines unsupervised learning with several supervised learning algorithms and deploys various feature selection algorithms by working directly with raw data. One particular challenge is the sparsity of raw cognitive data due to the design of tests, and missing values in the questionnaire data (Saarela & Kärkkäinen, 2014; 2015a;b; Kärkkäinen & Saarela, 2015; Rutkowski et al., 2010). This work addresses the high sparsity of the cognitive data by clustering the scored cognitive item response data into several difficulty bins and using each bin as a label as we explain later in Section 3.1. Since there were enough data points without missing contextual data from the PISA background questionnaire, we defer to imputation for the future work and focused on complete data. We examined the interaction between different classifier-feature selection algorithms and show that ML is a promising and complementary approach to understand and predict student performance.

The structure of this paper is as follows. In Section 2, we describe the PISA data. After that, our overall method is explained in Section 3, and the experimental results are presented in Section 4. Finally, in Section 5, overall results are summarized and directions for further work are discussed.

[2]In the official PISA literature, it is not explicitly reported which features of the student's background are actually taken into account (OECD, 2014). However, Monseur and Adams (2008) argue that all information from the background questionnaire is utilized.

[3]nces.ed.gov/nationsreportcard/

[4]www.surveylang.org/

[5]See both http://timssandpirls.bc.edu/

Table 1. Item cluster allocation to booklets in PISA 2012. PM denotes cluster of math, PR cluster of reading, and PS cluster of science items.

BOOKLET ID	ITEM CLUSTER			
B1	PM5	PS3	PM6A	PS2
B2	PS3	PR3	PM7A	PR2
B3	PR3	PM6A	PS1	PM3
B4	PM6A	PM7A	PR1	PM4
B5	PM7A	PS1	PM1	PM5
B6	PM1	PM2	PR2	PM6A
B7	PM2	PS2	PM3	PM7A
B8	PS2	PR2	PM4	PS1
B9	PR2	PM3	PM5	PR1
B10	PM3	PM4	PS3	PM1
B11	PM4	PM5	PR3	PM2
B12	PS1	PR1	PM2	PS3
B13	PR1	PM1	PS2	PR3

2. Data

We use the two main student datasets from the latest PISA assessment, which was conducted in 2012 (the 2015 data is not yet public): the *scored cognitive item response* and the *student questionnaire data file*. Both datasets have 485,490 observations (the students who attended the 2012 PISA assessment) and a couple of hundreds of variables.

As explained above, every student that attends the PISA test is assigned only a small fraction of the whole item battery. In PISA 2012, there were 13 main different tests—called *booklets*—and 210 different cognitive test items. Since mathematics was the main assessment domain in PISA 2012, the majority of the items, i.e. 108 of them, are items that test the math proficiency of the students. These cognitive test items were organized into groups—in PISA denoted as *item clusters*—so that each booklet contained four item clusters (this is illustrated in Table 1) and was estimated to be completable in two hours. As can be seen from Table 1, each booklet contained at least one cluster with math items. Our goal in this study is to predict the math performance of the students, which is why we use the sparse $108 \times 485,490$ matrix of the scored math items for building the labels of our classifiers (this will be further explained in Section 3.1).

For the classification features, we are interested in all attributes that are directly concerned with the students' attitudes towards mathematics and that might explain their math performance. In the PISA background questionnaire, there are 53 different math attitudinal statement questions[6], in each of which the student is asked to tick one box of a Likert-scale depending on the degree to which he or she agrees (*totally disagree*, *disagree*, *agree*, or *totally agree*)

[6]Variables ST29Q01–ST46Q09 (position 67–119) in PISA questionnaire data set, see https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf

with the given statement. Examples of such statements include *I will learn many things in mathematics that will help me get a job* and *my parents believe studying mathematics is important*. All 53 questions/statements can be found in Figure 1. We select all students that have non-missing values for all of these questions. Because of the rotated design in PISA, these are a bit less than one third of the students from each country. For example, in the Finnish subset, there are 2,491 (out of 8,829) students, which have non-missing values for all these 53 features, and in the whole PISA data, there are 136,344 (out of 485,490) students with complete values for this feature set.

3. Methodology

3.1. Unsupervised learning from cognitive data for label creation

We define identifying the students that are likely to succeed or fail math items of certain difficulty as a prediction problem. Our goal is to train a supervised learning algorithm that predicts success or failure from the data. However there are several problems with identifying the labels necessary for this approach. First, the plausible values cannot be used, since that would be akin to engineering an already known formula (see Section 1). Second, as discussed in Section 2, the students were administered different cognitive tests and the single items in the tests vary in their difficulty (OECD, 2014), which is why we cannot simply use the total sum of correct items for each student as their label. The raw scored cognitive data has a high percentage of missing data and no aggregated test scores and no item difficulties are available. Besides the PVs, the only available information about the actual performance of each student in the cognitive test is the fact whether he or she was administered an item and—in case the item was administered—the score the student obtained for it. The score values can be either 0 (fail), 1 or 2 (partially or fully correct).

To be able to work with the available data, we designed an algorithm to extract labels from raw data and use these labels to train a predictive model. For every different test/booklet, we summed up the total scores of the included math items. Then, we assigned each math item that was included in the test—a summary of the cluster of different items of the main tests was provided in Table 1—to a bin which we denote as *difficulty level* in such a way that each difficulty level is of same size (i.e., includes the same number of items). We chose the number of difficulty levels for our label matrix Λ to be seven, because the OECD defined seven math proficiency levels (see Figure 15.4 in the PISA 2012 technical report by the OECD (2014)). Hereby, it is assumed that all of the different booklets are consistent with regard to their average difficulty, which is supported by the fact that each test should be fair and solvable within two hours.

We created a binary label for each student and each of the seven difficulty levels, which takes value 1 if the student answered more than half of the questions in that category correctly and 0 otherwise. The labels were stored in the seven-dimensional label matrix Λ. Basically, we consider the student to be able to solve items of a certain difficulty if he or she answered the majority of the items of this difficulty bin in his/her particular test correctly. This matrix is complete, i.e. with no missing values, since each booklet contains items from each category. Depending on the target group we are interested in, we either create our label matrix Λ only for one country (for instance, for Finland the $8,829 \times 7$ matrix) or for a bigger group (for example, for all PISA countries the $485,490 \times 7$ matrix).

3.2. Supervised learning for multi-label prediction

Having the label matrix Λ fixed, we have to decide which kind of classifier should be trained for our data. Many different supervised learning algorithm have been introduced in the ML literature (Kotsiantis et al., 2007). However, the performances of different prediction models can vary depending on the data and their preprocessing. A model that performs perfectly on one dataset might perform very poorly on another dataset. Since we could not know what the best model and preprocessing for our data were, we first compared different approaches for the Finnish subset of PISA (see Section 4) before we selected the best approach to produce the final results.

In Zaki and Meira (2014), classification techniques have been categorized into probabilistic classification, decision tree classifier, linear discriminant analysis (LDA), and support vector machines (SVM). We chose at least one from each of these categories of classifiers with different objectives and compared their performances in terms of their prediction accuracy. Altogether, we compared two probabilistic classifiers (nearest neighbour and naïve bayes), one LDA, one SVM, and one decision tree based classifier (random forest). For each of the different classifiers, the Finnish subset of PISA was randomly divided, so that two thirds of the data was used for training the classifier, and one third was used for testing it.

The most important step for learning from the data is the dimension reduction in the feature space. We were looking for the minimal set of features to represent our data, since redundant or even noisy features lower the accuracy of prediction models, make them less comprehensible, and increase the computational complexity. Generally, dimension reduction methods can be divided into those techniques that extract features and those that select features (Tang et al., 2014). To get the best results, we tested with each classification algorithm two *feature extraction*—i.e., Principal

Component Analysis (PCA) and Isomap—and four *feature selection* methods—i.e., Fisher (Duda et al., 2000), Anova (Elssied et al., 2014), Gini (Hall, 1999), and MRMR (Peng et al., 2005).

3.3. Difficulty levels are predictive

Correct answers for easier questions are predictive for harder ones. With the intention to predict the performance of the students in each difficulty level as accurately as possible, we implemented an additional set of classifiers, which were the same as described above but with the difference that for each classifier, the information if the student mastered the previous difficulty level(s) was iteratively added to the original set of 53 features. That means that for predicting difficulty level λ_6 we had 54 features, for predicting λ_5, we had 55 features, and for predicting λ_1, we had 58 features. The order of the difficulty levels is $\lambda_1 < \lambda_2 \ldots < \lambda_7$, with λ_1 being the easiest and λ_7 being the most difficult one.

4. Results

We tested our algorithmic approaches by using the Finnish subset in PISA only, and then we applied the best approach first, to the Finnish (Section 4.3) and second, to the whole PISA data (Section 4.4). In Table 2, the results of the experiments with the different classifiers and dimension reduction methods are reported. As can be seen from the table, with respect to the classifier, SVM performed overall the best.

Moreover, we made the observation that the prediction accuracy was for all models the best for the highest difficulty level λ_7 and the worst for the second easiest one λ_2. The prediction accuracy for λ_1 went up again, probably because the classifiers had learned that most of the students succeed in the math items of the easiest category.

4.1. Iterative approach

To test our hypothesis that the information whether or not the student had mastered the previous difficulty level can enhance the accuracy of our classifier for the next difficulty level (see Section 3.3), we iteratively added–before predicting the next item difficulty—the previous item difficulty vector(s) as a further feature(s) to the classifiers. Naturally, testing and training data were divided according to the same indices as our original feature and label matrix. With this adjustment, the prediction accuracy improved noticeably (on average $2-5\%$) for difficulty level six to two for all classifiers. For difficulty level seven, the features remained the same and the accuracy of the classifier could not improve. For difficulty level one, the accuracy of the classifier actually dropped slightly. A possible explanation for that fact is, as discussed in Section 4, the general difficulty to predict the performance on the second easiest math difficulty level λ_2 correctly, as well as the observation that the prediction accuracy of the easiest difficulty level λ_1 was very high in the non-iterative approach.

4.2. Feature selection

As pointed out in Section 3.2, to avoid overfitting, we are interested in a prediction model that uses the most important features only. Therefore, we saved from all of our classifiers all features that were selected by the four feature selection algorithms in each iterative step. Then, when building the final prediction model we used for each iterative step only those features that were chosen by the different feature selection algorithms (see Section 4.3). Moreover, for training the prediction model two thirds of the data were used, and for testing it the remaining third of the data was used.

In Figure 1, the histogram of all the selected features for all iterative steps and all 53 initial features is shown. As can be seen from the histogram, the variable *Maths Self-Concept - Get Good Grades* is the most chosen feature by the feature selection algorithms, and therefore the most important variable in our math performance prediction model. Furthermore, it can be seen that, for instance, the feature *Subjective Norms - Parents Like Mathematics* is never chosen by any of the feature selection algorithms and that this feature therefore, seems to be negligible/insignificant when predicting the math performance of Finnish students.

Figure 2 also illustrates the sum of chosen features by the different feature selection algorithms. However, in this figure also the additional features λ_7-λ_2 are included. As can be seen, the information whether a student was able to master the preceding difficulty levels, are important features for the math performance prediction of the next difficulty level. It should be noted that the sums of the lasts six features cannot be fully compared, because λ_7 had the chance to be selected in all of the six last prediction models, while λ_2 could be selected only in the very last prediction models.

4.3. Results for Finland

In Table 3, the final results of the best approach for the Finnish data, i.e. the iterative SVM classifier with only the features that had been chosen at least five times (original features) or at least three times (additional λ features) by the feature selection algorithms, are reported. In each iterative step, only the features that were selected for this step were included. The table shows the accuracy, precision, recall, and f-score, which were computed on the confusion matrix of the test data.

As expected, the accuracy results are better for the higher

Table 2. Comparison of prediction accuracy (Finnish students performance in math items of different difficulty defined in label matrix Λ) with different classifiers and feature selection algorithms. The best accuracies for each level are underlined.

	Predicting success in math items of difficulty level 7						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.936816525	0.935601458	0.935601458	0.930741191	0.933171324	0.940461725	0.934386391
Naïve Bayes	0.749696233	0.933171324	0.917375456	0.764277035	0.776427704	0.940461725	0.767922236
LDA	0.919805589	0.917375456	0.899149453	0.883353584	0.878493317	0.940461725	0.876063183
SVM	0.940461725	0.939246659	0.940461725	0.940461725	0.940461725	0.940461725	0.940461725
Random Forests	0.938031592	0.940461725	0.939246659	0.933171324	0.929526124	0.940461725	0.931956258
	Predicting success in math items of difficulty level 6						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.834750911	0.825030377	0.835965978	0.82746051	0.817739976	0.838396112	0.817739976
Naïve Bayes	0.720534629	0.843256379	0.831105711	0.731470231	0.742405832	0.838396112	0.742405832
LDA	0.808019441	0.809234508	0.833535844	0.784933171	0.795868773	0.838396112	0.795868773
SVM	0.838396112	0.837181045	0.838396112	0.838396112	0.838396112	0.838396112	0.838396112
Random Forests	0.834750911	0.832041312	0.832320778	0.812879708	0.834750911	0.838396112	0.815309842
	Predicting success in math items of difficulty level 5						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.696233293	0.690157959	0.716889429	0.693803159	0.708383961	0.64763062	0.696233293
Naïve Bayes	0.662211422	0.722964763	0.705953827	0.673147023	0.670716889	0.708383961	0.67436209
LDA	0.699878493	0.688942892	0.705953827	0.690157959	0.693803159	0.708383961	0.685297691
SVM	0.722964763	0.716889429	0.710814095	0.721749696	0.722964763	0.713244228	0.719319563
Random Forests	0.722964763	0.701470231	0.714459295	0.720534629	0.704738761	0.713244228	0.722964763
	Predicting success in math items of difficulty level 4						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.614823815	0.611178615	0.575941677	0.592952612	0.599027947	0.585662211	0.605103281
Naïve Bayes	0.648845687	0.626974484	0.640340219	0.668286756	0.660996355	0.619684083	0.659781288
LDA	0.640340219	0.650060753	0.634264885	0.620899149	0.636695018	0.619684083	0.645200486
SVM	0.67800729	0.679222357	0.643985419	0.653705954	0.65127582	0.623329283	0.653705954
Random Forests	0.650060753	0.646415553	0.611178615	0.64763062	0.625759417	0.623329283	0.622114216
	Predicting success in math items of difficulty level 3						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.648845687	0.656136087	0.602673147	0.635479951	0.657351154	0.652490887	0.669501823
Naïve Bayes	0.64763062	0.652490887	0.671931956	0.64763062	0.645200486	0.652490887	0.650060753
LDA	0.637910085	0.631834751	0.662211422	0.637910085	0.643985419	0.659781288	0.65127582
SVM	0.675577157	0.668286756	0.662211422	0.665856622	0.65127582	0.64763062	0.667071689
Random Forests	0.667071689	0.648845687	0.611178615	0.67436209	0.62818955	0.641555286	0.643985419
	Predicting success in math items of difficulty level 2						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.573511543	0.583232078	0.539489672	0.543134872	0.539489672	0.546780073	0.546780073
Naïve Bayes	0.571081409	0.582017011	0.622114216	0.569866343	0.579586877	0.602673147	0.583232078
LDA	0.577156744	0.580801944	0.602673147	0.59781288	0.589307412	0.607533414	0.57472661
SVM	0.59781288	0.59781288	0.605103281	0.596597813	0.599027947	0.605103281	0.599027947
Random Forests	0.572296476	0.599027947	0.545565006	0.591737546	0.571081409	0.603888214	0.567436209
	Predicting success in math items of difficulty level 1						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.733900365	0.738760632	0.720534629	0.732685298	0.713244228	0.769137303	0.733900365
Naïve Bayes	0.606318348	0.753341434	0.769137303	0.617253949	0.616038882	0.769137303	0.618469016
LDA	0.714459295	0.708383961	0.741567436	0.716889429	0.733900365	0.769137303	0.730255164
SVM	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303
Random Forests	0.769137303	0.763061968	0.737545565	0.760631835	0.732685298	0.759416768	0.739975699

difficulty levels (because most students will fail this level) and the lower difficulty levels (because most students will master this level) than for the middle difficulty levels. On the other hand, the precision increased monotonically from the most difficult to the easiest question difficulty level. This was most probably the case, because the classifier had learned that most students fail items of the highest difficulty and hence, simply returned 0 for the majority of the test instances. Since accuracy is not the best measure of performance we focus on the precision for the rest of the discussion.

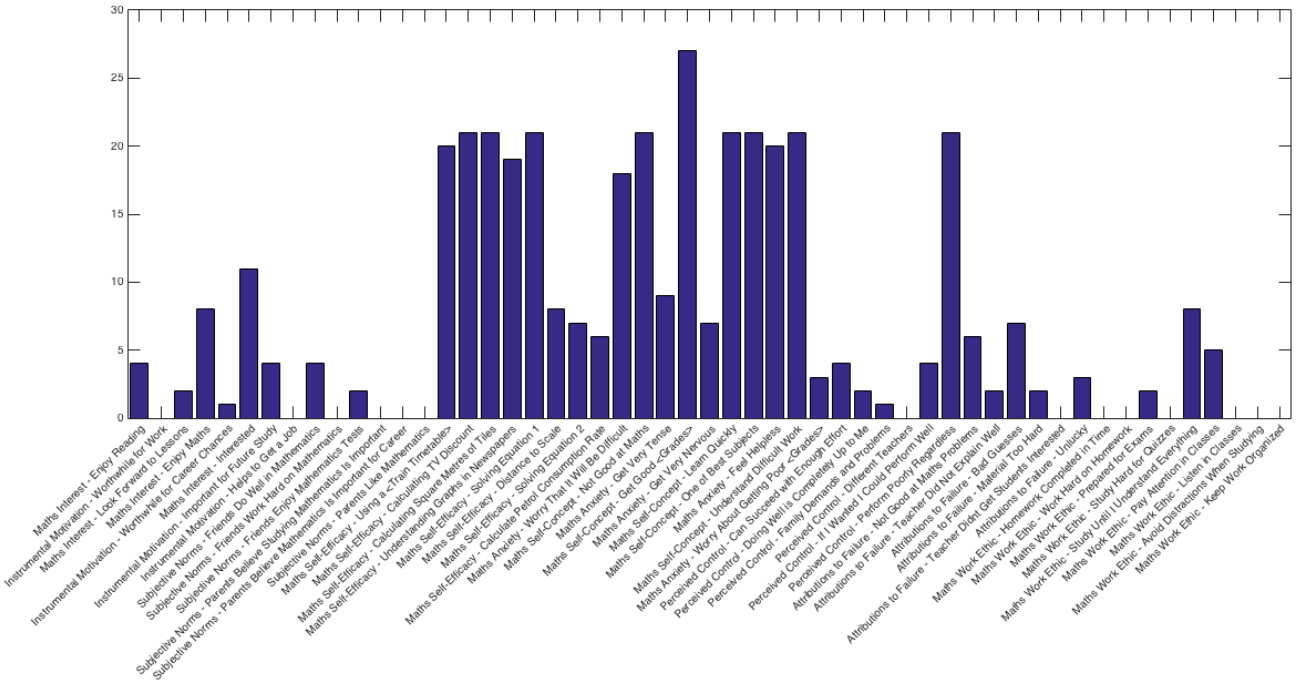

Figure 1. Frequency of selected features of the 53 initial features by the four feature selection algorithms for the Finnish student data. The higher the bar of a feature, the more often this feature was selected, and the more important this feature is for the prediction model.

Table 3. Results of iteratively predicting success in math items of the different difficulty levels for Finnish students with SVM and—for each difficulty level—only the most selected features by the four feature selection algorithms.

Difficulty	**Accuracy**	**Precision**	**Recall**	**F-score**
Level 7	0.9579	0.0312	1.0000	0.0606
Level 6	0.8555	0.1385	0.5294	0.2195
Level 5	0.7427	0.3309	0.6866	0.4466
Level 4	0.7843	0.4029	0.6975	0.5108
Level 3	0.7096	0.5496	0.7791	0.6445
Level 2	0.6757	0.6530	0.7095	0.6801
Level 1	0.7630	0.9493	0.7833	0.8583

4.4. Results for all countries participating in PISA

Table 4 shows the prediction results for all PISA countries (i.e. the 136344×53 feature matrix of all students that had complete values for all 53 features from the background questionnaire and the corresponding 136344×7 label matrix for the same students). However, it should be noticed that the same settings as for Finland were used, that is the classification algorithm and the selected features that were optimized for the Finnish data. For difficulty levels λ_6 and λ_5 the prediction accuracies are actually higher than for the Finnish data. However, this is most likely the case because most of the world's students are not able to solve items of this difficulty level. This assumption is supported by the very low precision values. Moreover, we see again the worst result for predicting λ_2, where the prediction accuracy is only slightly better than guessing.

Table 4. Results of iteratively predicting success in math items of the different difficulty levels for students from all in PISA participating countries with SVM and—for each difficulty level—only the most selected features by the four feature selection algorithms.

Difficulty	**Accuracy**	**Precision**	**Recall**	**F-score**
Level 7	0.9524	0.0003	0.0714	0.0006
Level 6	0.8872	0.0027	0.2414	0.0054
Level 5	0.7723	0.0133	0.3118	0.0255
Level 4	0.6156	0.1627	0.5016	0.2457
Level 3	0.5817	0.7934	0.5918	0.6779
Level 2	0.5350	0.5629	0.5404	0.5514
Level 1	0.6539	0.9591	0.6656	0.7859

5. Discussion and future work

PISA data—as well as LSA data generally—provide an interesting source for educational ML and DM applications, because they are of high quality, internationally comparable, and publicly available. However, the challenges of working with these data are the high sparsity of the raw data and the lack of any readily available and comparable cognitive test results of the students.

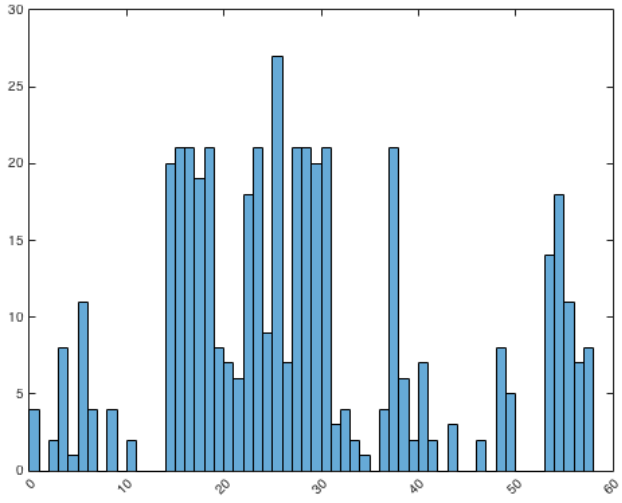

Figure 2. Histogram of selected features of the 53 initial features plus the 6 additional ones for the iterative steps by the four feature selection algorithms for the Finnish student data.

In this paper, we have presented an approach to prepare LSA data for supervised ML approaches. In addition, initial results of using our approach for predicting success in math items of various difficulty, have been presented. Hereby, we have tested different classification and dimension reduction algorithm for the Finnish data, and then applied the best classifier with only the selected features of different feature selection algorithm for the Finnish and for the whole PISA data. The prediction accuracy was further improved by adding for each succeeding difficulty level the information whether the student mastered the preceding difficulty level(s). An analysis of the chosen features by the feature selection algorithm enabled a predictive power ranking of the questions asked in the background questionnaire that actually explained the students' math performance.

The results presented in this paper are only preliminary and we intend to extend and improve our experiments and study in various directions. First of all, the results that were presented here are based on the fully available raw data only. We intend to perform similar experiments for the whole contextual data by first imputing the missing values.

We also intend to compare our approach to the Rasch model and plausible value approach currently used in most LSAs, which has evolved from the psychometric literature. It has been argued that one of the weaknesses of the Rasch model is the fact that all students with the same raw score (i.e., number of correctly solved tasks) obtain the same ability estimate (Embretson & Reise, 2013). It would be interesting to compare this to our approach, where the difficulty level of the solved items is taken into account. As discussed by Baker and Yacef (2010), comparing and integrating machine learning techniques to the ones from the psychometrics literature, is one of the most distinguishing features that separates the educational ML/DM discipline from the traditional ML/DM research area.

References

Baker, R. and Yacef, K. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1, 2010.

Duda, Richard O., Hart, Peter E., and Stork, David G. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000. ISBN 0471056693.

Elssied, Nadir Omer Fadl, Ibrahim, Othman, and Osman, Ahmed Hamza. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3):625–638, 2014.

Embretson, Susan E and Reise, Steven P. *Item Response Theory*. Psychology Press, 2013.

Hall, Mark A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

Kärkkäinen, T. and Saarela, M. Robust Principal Component Analysis of Data with Missing Values. In *Machine Learning and Data Mining in Pattern Recognition*, pp. 140–159. Springer, 2015. ISBN 978-3-319-21023-0. doi: 10.1007/978-3-319-21024-7_10.

Kotsiantis, Sotiris B, Zaharakis, I, and Pintelas, P. *Supervised machine learning: A review of classification techniques*. OS Press, 2007.

Marsman, Maarten. *Plausible Values in Statistical Inference*. Universiteit Twente, 2014.

Monseur, Christian and Adams, Ray. Plausible Values: How to Deal with Their Limitations. *Journal of applied measurement*, 10(3):320–334, 2008.

Musik, Alexander. Philologenverband bezeichnet Pisa-Studie als Geldverschwendung. http://www.deutschlandfunk.de/bildungsforschung-in-der-kritik-philologenverband.680.de.html?dram:article_id=347675, 2016.

OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009. ISBN 9789264056268.

OECD. *PISA 2009 Technical Report*. OECD Publishing, 2012.

OECD. PISA 2012 Technical Report, 2014.

Peng, Hanchuan, Long, Fuhui, and Ding, Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Rasch, Georg. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. 1960.

Rutkowski, Leslie, Gonzalez, Eugenio, Joncas, Marc, and von Davier, Matthias. International Large-Scale Assessment Data Issues in Secondary Analysis and Reporting. *Educational Researcher*, 39(2):142–151, 2010.

Saarela, M. and Kärkkäinen, T. Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 60–68, 2014.

Saarela, M. and Kärkkäinen, T. Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 156–163, 2015a.

Saarela, M. and Kärkkäinen, T. Weighted clustering of sparse educational data. *In 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 337–342, 2015b.

Tang, Jiliang, Alelyani, Salem, and Liu, Huan. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, pp. 37, 2014.

Tarpey, Thaddeus and Petkova, Eva. Latent regression analysis. *Statistical modelling*, 10(2):133–158, 2010.

Wu, M and Adams, RJ. Plausible values: Why they are important. In *11th International Objective Measurement Workshop, New Orleans*, 2002.

Zaki, Mohammed J and Meira Jr, Wagner. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.