Predicting students' performance in online courses using multiple data sources

Mélina Verger [1]   **Hugo Jair Escalante** [2]

Abstract

Data-driven decision making is serving and transforming education. We approached the problem of predicting students' performance by using multiple data sources which came from online courses, including one we created. Experimental results show preliminary conclusions towards which data are to be considered for the task.

1. Introduction

Predicting students' performance is a multidisciplinary task which has a significant impact to many stakeholders, including students, teachers, and educational institutions (Alamri & Alharbi, 2021). Not only it helps to predict at-risk students to plan interventions for academic support and guidance (Kotsiantis et al., 2004), but it also helps to identify key factors affecting students' performance (Alamri & Alharbi, 2021). It has then become one of the most challenging issues for educational institutions (Rastrollo-Guerrero et al., 2020) regarding the improvement of the teaching-learning process and the more accurate allocation of resources and instructors (Hellas et al., 2018).

During the last decade, such a predictive task has gained more and more attention since online courses along with e-learning resources and educational software had created a new context known as web-based education (Romero & Ventura, 2010) in which learners generate an increasing volume and variety of types of data (Papadogiannis et al., 2020; Romero & Ventura, 2020).

The problem has been approached by considering a variety of data sources, including, academic, behavioral, demographic and personality, where these sources have been used separately and combined. The benefits of predicting students' performance using multiple data sources are twofold. On the one hand, using several data sources enables to discover relationships among different factors and which of these are associated with students' performance, worthwhile at both education and prediction levels. On the other hand, it fills the gap towards the consideration of a more holistic view of student combining non-cognitive and cognitive characteristics of students that few studies address (Hellas et al., 2018). Most of the articles reviewed by (Hellas et al., 2018) have used academic data for prediction. Behavioral data are also common predictor (Alamri & Alharbi, 2021). However, (Villagrá-Arnedo et al., 2016) states that the use of heterogeneous data enriches the final performance of the prediction algorithms.

Despite the increasing volume of educational data, the Educational Data Mining (EDM) field is still a growing research area where the community has pointed out the scarceness of open data as one of its biggest challenges. In most cases, data comes from the research group conducting the studies (Vijayalakshmi & Venkatachalapathy, 2019) (Baker & Yacef, 2009) and is rarely made publicly available for preserving privacy of involved students and teachers. Moreover, most experiments concern a specific context, formed by a population from a single institution and even from a single course (Hellas et al., 2018), which is detrimental to generalizing conclusions with quality confidence because of presumable latent factors (Romero & Ventura, 2007; 2010).

From these remarks, this work explores a generic approach to address students' performance prediction with the aim of answering the following research questions:

1. To what extent using multiple data sources on learners increases prediction performance?
2. Which learners' characteristics are most related to higher academic performance?

In trying to answer these questions, we have performed a study on predicting students' performance as follows. We created an open learning framework to collect learners' data from multiple sources not offered in current open datasets. We studied the collected data with other open datasets focusing on the common features involved. We compared prediction performances using usual data sources but also using the additional ones we collected. We complemented the previous prediction results with feature selection analyses.

The global results show that we obtained encouraging sim-

[1]Paris-Saclay University, Orsay, France. [2]Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico. Correspondence to: Mélina Verger <meverger@gmail.com>, Hugo Jair Escalante <hugojair@inaoep.mx>.

ilar trends and knowledge with different students' populations and courses. We highlight the relevance and importance of academic and behavioral data in contrast to demographic ones. The optimal number of features appears to be 3 and we formulated an hypothetical advantageous relationship between students' who have stronger abilities of memorizing and deducing knowledge from written material and better performance. Additional experiments involving larger datasets are required in order to reach conclusive findings. However, this research work paves the way for the study of new correlations between students' characteristics and their performance.

The remainder of this report is organized as follows. Section II describes the literature review; Section III presents the methodology and the experiments framework; Section IV discusses the results; and Section V provides concluding remarks.

2. Related work

2.1. Students' performance measures

Students' performance could be defined with different measures: exam grade or range, course grade or range, GPA, retention or dropout rate, knowledge gain (Hellas et al., 2018). The choice of a particular metric depends on the educational framework (e.g., in university or in private institution, online or in-person teaching, etc.) and for what it is being predicted. The course grade range (e.g., pass/fail, high-medium-low performance, A-B-C-D-E-F grade) corresponds to the most generic case where instructors are interested in predicting which students will rather belong to a certain class than obtain a precise grade. It is more suitable for adapting their guidance to smaller groups of students. In this work, we focus on pass/fail course grade range which is common with other datasets from online courses framework to make some comparisons.

2.2. Data sources categories

(Hellas et al., 2018) performed a systematic review based on 357 articles and identified five data sources categories: demographic, personality, academic, behavioral, and institutional. Demographic data includes age, gender, academic level, location, health information, family background, etc. Personality data encompasses psychological data (e.g., self-regulation, self-efficacy, big five personality traits (Goldberg, 1990)) and affective data (e.g., happiness, anxiety, depression). Academic data covers course performance and/or past performance, during high school for example. Behavioral data includes time on task, number of interactions with the learning material, etc. Institutional data are about teaching approach, high-school quality, etc

Despite this diversity of data, most of the related studies

have used either one data source, academic (Hellas et al., 2018) or behavioral (Alamri & Alharbi, 2021), or at most two data sources with the combination of demographic and academic data. (Villagrá-Arnedo et al., 2016) highlights an availability-relevance trade-off of educational data which can limit considering more data sources for prediction tasks in the field. Nonetheless, (Villagrá-Arnedo et al., 2016) states that the use of heterogeneous significant data enriches the final performance of the prediction algorithms, deduced from their experiments with either behavioral features, academic features, both of them or the most correlated ones. (Kotsiantis et al., 2004) supports this idea since their highest accuracies were obtained with six machine learning algorithms by adding features progressively. Therefore, this work is intended to be a comprehensive study of all the possible data sources except the institutional one as it is a cross-population experiment without a specific institution. More details of the features involved are provided in subsection 3.2. Moreover, considering several data sources enables to discover relationships among different factors and which of these are associated with students' performance, with a more holistic view of students (Hellas et al., 2018).

2.3. Data sources availability

Most of experiments in EDM are conducted with data provided by a university associated to the research group, such as (Mesarić & Šebalj, 2016; Osmanbegovic & Suljic, 2012; Al-Barrak & Al-Razgan, 2016; Al-Shehri et al., 2017). The access authorization process as well as privacy preserving mechanisms make difficult to publish the data, which is still a current concern discussed in the latest International Conference on Educational Data Mining[1]. Taking this into account with the lack of description on features used (Hellas et al., 2018), it is challenging to select relevant methodology, models, and results among every one of those that are proposed. To make our approach transparent, we have created a learning framework open to anyone and where the anonymized data are publicly available with the consent of the learners (see `https://github.com/melinaverger/ed_project`).

2.4. Explainability purpose

The very active research area on explainable AI (XAI) and in particular some works focusing on predicting students' performance, call to use explainable models to equally consider the correctness of the predicted value and the understandability of the prediction made (Alamri & Alharbi, 2021), contributing to generalize conclusions in spite of limited data availability. In addition, the recent release of the General Data Protection Regulation (GDPR) re-emphasized the

---

[1]`https://educationaldatamining.org/edm2021/`

importance of explainable and trustworthy AI when the decision could affect individuals (Longo et al., 2020). Therefore, among the most used learning algorithms in EDM (e.g., decision tree, logistic regression, naïve bayes, random forest, support vector machine, k-nearest neighbor, feed-forward neural network, deep neural network (Romero & Ventura, 2010; Hellas et al., 2018)), we describe benchmark results with explainable models in the next paragraph.

For a binary classification (e.g., pass or fail target), (Kotsiantis et al., 2004) used combinations of demographic and academic data with a C4.5 decision tree and reported accuracies of 79.22% and 73.99% over 354 and 28 instances respectively, the latter representing a real-condition class. The average over all feature combinations for the 28-instance experiment is 63.43% accuracy, 69.79% sensitivity and 57.67% specificity. For a 3-class classification (e.g., high, medium or low performance targets), (Vijayalakshmi & Venkatachalapathy, 2019) used demographic, academic and behavioral features and reported an accuracy of 69% using a C5.0 decision tree on a dataset with 480 instances. For a 4-class classification, (Lakshmi et al., 2013) used demographic and academic features and compared ID3, C4.5 and CART decision trees algorithms over 120 instances. C4.5 achieved the highest accuracy at 55.83%. Similarly, (Chaudhury et al., 2016) obtained an average over the four classes of 83.25% F-score with C4.5 over 648 instances. In this work, we will continue to use an explainable approach and put our results into perspective for the related combinations.

Finally, this work attempts to address some limitations of current state of the art, such as scarceness of open data, single or few data sources, single-context experiment and non-explainable methods and results, by proposing an approach described in the following section.

3. Methods

Summarizing this work, we created an open learning framework to collect data from multiple sources, and we performed the following: (1) we positioned the collected data with two open datasets regarding the common features involved, (2) we compared prediction performances using usual data sources (3) but also using the additional ones we collected, and (4) we analyzed the previous prediction performances with a feature selection study. This section provides information on how we collected data, how we extracted features and which algorithms we used for the binary classification prediction task and for the feature selection analyses.

3.1. Data collection

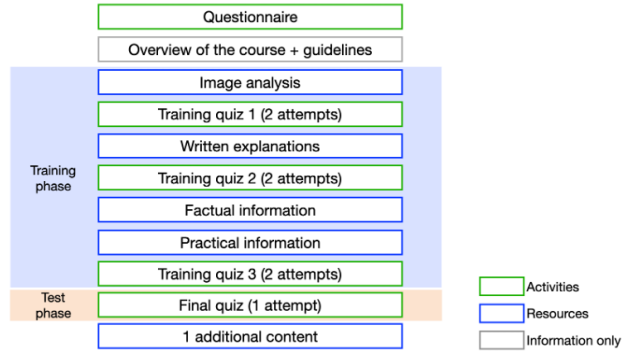As the goal of this work is to exploit multiple data sources not offered in current open data sets, we created an open

Figure 1. Structure of the course. Uniformly with e-learning platforms common vocabulary, we call *activity* any content where participants need to interact with the course (questionnaires, quizzes, etc.) and *resource* any content where participants only need to consult it (mostly learning material).
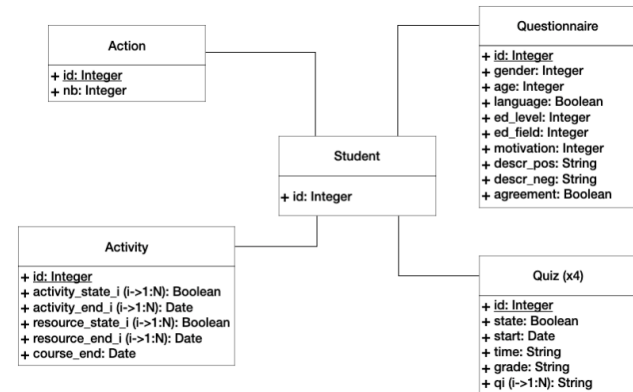

Figure 2. Relational model of anonymized and converted data (73 variables).

course on the e-learning platform The Open University[2] (OU) to collect usual and additional features described in subsection 3.2. This 1-hour general course is an introduction to the organization of the brain, its characteristics, and functionalities, which does not require substantial prior knowledge. It is thus intended for a general audience with any education background. Each participant had been previously informed of the data collection and the use of the data and had explicitly consented to participate in the proposed learning experiment. This complies with the OU privacy policy[3] and GDPR directives[4].

The structure of the course is presented in Figure 1. It

---

[2]https://www.open.ac.uk
[3]https://www.open.ac.uk/about/main/strategy-and-policies/policies-and-statements/website-privacy-ou
[4]https://gdpr-info.eu

was designed as follows. The duration of the course is a trade-off between involving participants in a learning process with enough time and material, capturing their learning process in equal conditions to others with limited duration and demanding volunteer participants' efforts with reasonable duration. In Figure 1, it is observed that each resource concerns either a particular format type or content type of the learning material followed right after by a corresponding quiz. Then, each quiz is composed by half of the questions on memorizing information and the other half on deducing knowledge. It must be noted that academic data in this learning framework, e.g., the training quiz grades, are coherent with the final grade, also derived from a quiz, but consequently specific to this final evaluation material.

Besides, since the OU offers a Moodle-based learning management system (LMS) (Watson & Watson, 2007), the course permits the collection of the raw data whose variables are described in Figure 2, after being anonymized and converted in appropriate formats.

3.2. Feature preprocessing

3.2.1. MINIMAL SET OF FEATURES

With the data previously collected, it is possible to build a minimal set of features of different data sources, demographic, personality, academic and behavioral, mentioned earlier. It is not needed to have institutional data as it is a cross-population experiment. This minimal set was designed to be reproducible in other context and experiment and to make comparisons with other existing open data sets.

For the demographic category, several works (Dutt et al., 2017; Adejo & Connolly, 2018) show that basic demographic information on students have a significative role in characterizing students and predicting their success. Age, gender, level of education and field of education field are then kept.

Moreover, time spent on tasks, number of clicks and whether an activity or resource has been viewed/done are commonly provided and used in similar experiments in the field (Feldman et al., 2015; Ahmad & Shamsuddin, 2010; Popescu, 2009; Graf, 2007). Thus, such behavioral features are also part of the set.

Besides, past academic grades have a predictive importance, reflecting students' consistency in their schooling (Livieris et al., 2018; Adejo & Connolly, 2018). Course performance based on the grades obtained by participant before the final quiz constitute an academic history. Their average is independent of the number of obtained grades and is a reproducible feature that we used.

Regarding personality, few studies consider these kinds of students' characteristics to explore correlation with their

learning performance whereas (Afzaal et al., 2019; Papadogiannis et al., 2020) recent works point out relevant relationships. Since composing a generic set becomes complex when including data for personality recognition, one motivation question and two participants' text descriptions on themselves either positive or negative are the only information asked. The two latter will be processed in future work.

All this information constitutes the minimal set of features detailed in Table 1. Demographic and personality data sources were collected through a questionnaire (see Appendix A) whereas all the other sources were automatically recorded.

3.2.2. ADDITIONAL FEATURES

Additional features were created to explore other hypothetical relationships (see Table 2). These new features average the grades grouped by format type or content type of the learning material and aim at discovering relationships between hypothetic learning preferences and students' performance (Antoniuk, 2019). These are visual or verbal preference regarding the format type and factual or practical regarding the content type. Two other features correspond to a memory and a deduction score. Related analyses are presented in the results section.

To have visual insights on all the features, both minimal and additional, please see Appendix B.

3.3. Supervised learning algorithms

For the sake of explainability and understanding on key factors affecting students' performance to answer the research questions, the requirements of a white-box supervised learning algorithm were met. Features should be understandable and the machine learning process transparent (Loyola-Gonzalez, 2019).

On this point, decision tree and rule-based learning algorithms are supported for the students' performance prediction task in the systematic review of (Alamri & Alharbi, 2021). Decision trees are indeed self-explanatory and deal with any type of features (Breiman et al., 1984). On the other hand, they are sensitive to small variations in the training dataset (Quinlan, 1986).

In this work, the prediction results were obtained using a CART decision tree algorithm[5]. In some of the following experiments, they were also compared with a random forest (RF) algorithm and a support vector machine (SVM) algorithm to evaluate the relative performance of the decision tree (DT). These algorithms' parameters were selected

---

[5]`https://scikit-learn.org/stable/modules/tree.html#tree`

Table 1. Minimal set of features. (Those which are marked with a * will be processed in future work).

Category	Name	Description	Values
Demographic (D)	age	Learner's age	18-99
	gender	Learner's gender	F/M/NA
	ed_level	Learner's level of education (current or last)	1-8[a]
	ed_field	Learner's field of education (current or last)	4 classes[b]
Academic (A)	avrg_grade	Average of the learner's course grades before taking the final exam	0.00-10.00
Behavioral (B)	%_completion	Percentage of course completion	0.00-1.00
	nb_action	Sum of the number of views on the resources and the number of attempts at the activities	0-inf.
	time	Total time learner spent on activities	0-inf.
Personality (P)	motivation	Learner's enrolment motivation	3 classes[c]
	descr_pos*	Learner's description of a positive day	Text
	descr_neg*	Learner's description of a negative day	Text

[a]Number of years after high school graduation/A-level.

[b]Maths, Computer Science, Engineering, Technology... - Health, Biology, Medicine... - Humanities, Languages, Arts, Culture ... - Administration, Law, Management, Commerce...

[c]Very interested in the course subject - Moderately interested but curious - Little interest but will to participate in the educational experience.

Table 2. Additional features.

Category	Name	Description	Values
Learning preferences (L)	visual	Learner's score on visual material	0.00-1.00
	verbal	Learner's score on written material	0.00-1.00
	factual	Learner's score on factual content	0.00-1.00
	practical	Learner's score on practical content	0.00-1.00
	memory	Learner's score on memorizing any kind of information	0.00-1.00
	deduction	Learner's score on deducing any kind of knowledge	0.00-1.00

by performing a grid search for each experiment and by considering the computational resources available.

3.4. Feature selection algorithms

To assess the relevance of the features, we used feature selection algorithms, three of which were forward elimination (FE), backward elimination (BE) and recursive feature elimination (RFE). FE, also called forward selection[6], is a greedy procedure that iteratively finds the best new feature to add to the set of selected features. We start with zero feature and at each iteration we keep adding the feature which best improves our model. The procedure stops when an addition of a new feature does not improve the performance of the model or when the desired number of selected features is reached. BE, also called backward selection, works in the opposite direction to FE. We thus start with all the features and greedily remove the least significant feature at each iteration. Then, RFE[7] is an instance of BE (Guyon et al., 2002) that creates models and determines the worst performing feature at each iteration to find the best performing feature

---

[6]https://scikit-learn.org/stable/modules/feature_selection.html#sequential-feature-selection

[7]https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

subset based on the feature ranking. Finally, these wrapper methods were used with our decision tree models.

4. Results

Three datasets were considered and named respectively D1, D2 and D3. D1 is built from the collected data described in section 3 where 44 students were involved. D2 is the Open University Learning Analytics dataset (Kuzilek et al., 2017) which can be built from several data tables[8]. It has 28,785 students involved and 7 possible courses which could have been taken at defined starting dates in the years 2013 and 2014. D3 is the Canvas Network Person dataset (CanvasNetwork, 2016) comprised of de-identified data from Canvas Network open courses running between January 2014 and September 2015. It contains records from 224,914 students enrolled in 238 possible courses belonging to 10 different disciplines. A summary is displayed in Table 3.

As well as D1, the two datasets D2 and D3 are about varied students' populations. Nonetheless, the two latter were not designed for multiple data sources. After some cleaning and preprocessing to make fair comparisons between the three datasets, 27, 11,420 and 10,005 instances were kept

---

[8]https://analyse.kmi.open.ac.uk/open_dataset#description

Table 3. Comparison of datasets. (D: demographic. A: academic. B: behavioral. P: personality. L: learning preferences).

	D1	D2	D3
# courses	1 (general topic)	7 (3 disciplines)	238 (10 disciplines)
# students	44	28,785	224,914
# training instances	27	11,420	10,005
class-imbalance ratio	59.25%	72.52%	36.78%
data sources	D + B + A + P + L	D + B + A	D + B

from D1, D2 and D3 respectively. It can be noted that the gender attribute was used for data understanding but removed for prediction to be gender inclusive with respect to the predictions made. We recall that the target variable is binary (pass (1) or fail (0)).

4.1. Comparison with open datasets

The first experiment aims at positioning D1 with common educational data represented by D2 and D3. As there is no standard approach yet to collect educational data, students' age, level of education and number of interactions were the only common features found across the three datasets. D3 is indeed the most restrictive. These features represent demographic (D) and behavioral (B) data.

We computed baseline results evaluated with ten-fold cross-validation and we compared them with four tests performing different balancing techniques: upsampling the minority class, downsampling the majority class, up and downsampling both classes and Synthetic Minority Oversampling Technique (SMOTE (Chawla et al., 2002)). The best results are presented in Table 4 (and the others in Appendix C). Across the different balancing technique tests, these techniques especially improved performance with D1 whereas D2 and D3 showed more stable results and always had close performance. At the end, upsampling the minority class and downsampling the majority class presented the best results globally and will be reused for all the other experiments.

Table 4. Results with the upsampled minority class and the downsampled majority class.

	D1	D2	D3
Accuracy (%)	66.66	69.05	69.77
Precision (%)	59.16	69.15	70.06
Recall (%)	67.50	69.05	69.77
F-score (%)	57.99	69.01	69.66

We then display in Figure 3 the predictive performances with error bars obtained on the three datasets with the different models. Not surprisingly, D1 leads to lower performance and higher variability than D2 and D3 since D1 has comparatively much less instances. It corresponds to what we expected but the mean accuracy is encouraging since, with more samples, it is probable to obtain closer results. Furthermore, the DT model seems relevant for predictive

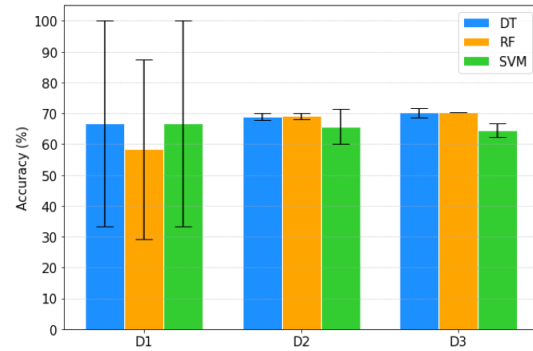performance compared to the other models, in addition to their explainability.

Figure 3. Resulting accuracies with DT, RF and SVM on D1, D2 and D3.

Besides, we also experimented transfer learning between the datasets to see if there is a presumably shared knowledge with open educational data and our collected data. The results are in Figures 22, 23 and 24 in Appendix D. We deduced encouraging results again because in the three graphs, when testing on D1, we obtained about the same performance when we trained our DT model either on D2 or D3 than when we trained the model on D1 (normal case).

4.2. Usual data sources

The second experiment aims at comparing observable trends between the collected data D1 and the open data D2, that represent different populations and courses, regarding the usual demographic (D), academic (A) and behavioral (B) data sources (see Appendix E for details on the features). D3 was excluded since it has too few comparable features. The experiment consists of comparing performance results on different feature combinations. The results are in Figure 4 and comments in terms of performance and feature importance are below. The detailed results are in Table 10 in Appendix F.

Academic data and behavioral data enable to perform the best results with both datasets. They are indeed the features most representative of what is being predicted, students' performance, which supports their common use. When
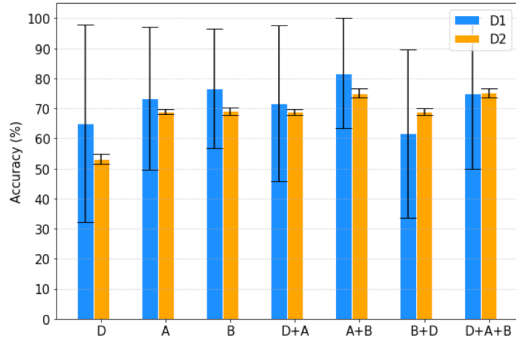
Figure 4. Resulting accuracies with the different usual sources.

combined, results slightly improve, with a stronger emphasis with D2. We also observe the same ratio about 40-60% of prediction importance between academic data and behavioral data within the two datasets. Besides, the time is the main predictor among the behavioral data.

Regarding demographic data, it shows one of the lowest performances when compared to other combinations. In fact, these demographic features are too general to discriminate groups of students (see distributions according the target in Appendix G) and other features such as working conditions or socio-economic features (work and living environments, parental education, financial background, etc.) could be more useful. In addition, about the same performances are obtained when demographic data are combined with other sources like academic or behavioural data, suggesting that it does not bring additional information and even add noise in some cases with lower results. Feature importance confirms that academic features are 99% and 100% important in D2 and D1 respectively, and that behavioral features in particular the time are 98% important too, compared to demographic ones (0-2%). However, these general demographic features are very useful to understand the data and make statistical analyses outside a prediction task so that they are worth collecting every time an educational study is performed.

To conclude this part, we observe similar trends in both datasets representing different students' populations and courses. Demographic data has low prediction importance contrary to academic data and behavioral data. Moreover, a combination of the two latter gives the best results. Thus, using only two of the usual data sources seems to be relevant.

4.3. Additional data sources

The third experiment aims at testing and comparing additional combinations provided by D1. The same approach as the previous subsection is adopted and the results tables and

figures are displayed in Appendix H.

For the usual data sources, we obviously find the same conclusions than the previous experiment including the 40-60% ratio between academic and behavioral data when combined, and the time as a dominant behavioral feature (87% against 13% for the number of interactions since the percentage of completion has a null variance in D1). Besides, demographic data still has a quasi-null or null importance when combined with any type of data, and academic and behavioral data always have high importance (90-100%) when not in the same combination (for instance with personality data). Nonetheless, learning preferences data that are derived from academic data show only 36% of importance for the verbal score which is the most significant learning preferences feature. It suggests that learning preferences could be useful for prediction but not in every case like academic data. Among the learning preferences features, the verbal, the factual, visual and deduction scores have a consistent performance across the several-source combinations with a more significative role of the verbal score. Then, a general performance loss is observed with 3 sources or more especially with 5 sources which give the lowest performances.

To conclude this part, the level of importance either of demographic data, academic data or behavioral data generalizes with additional data sources, except in the case when academic data are combined with learning preferences data as the latter is derived from the former. The time is still the most important predictor among behavioral data and some scores, in particular the verbal one, seem to be significantly important. Finally, testing 1-source, 2-source, 3-source, 4-source and 5-source combinations leads to consider 2-source combinations as the most performing combinations such as academic and behavioral data again.

4.4. Feature selection analyses

In this part, we performed feature selection in order to assess the relevance of the features individually in contrast to combinations previously. All the features in the minimal set and the additional set are considered. As there are many features and comparably few instances, feature selection leads to simplifying the model for better understanding and improving prediction performance. Reducing computational cost and time is not the main purpose but this outcome will be also met. We first used wrapper methods and then filter methods to first find an optimal feature set for the prediction task and then to have insights on the relationships between the features for domain knowledge (usefulness vs. relevance (Guyon & Elisseeff, 2003)). It has to be noted that the feature of percentage of course completion has a null variance and was not taken into account for the following since the preprocessing step kept only records of students who fully
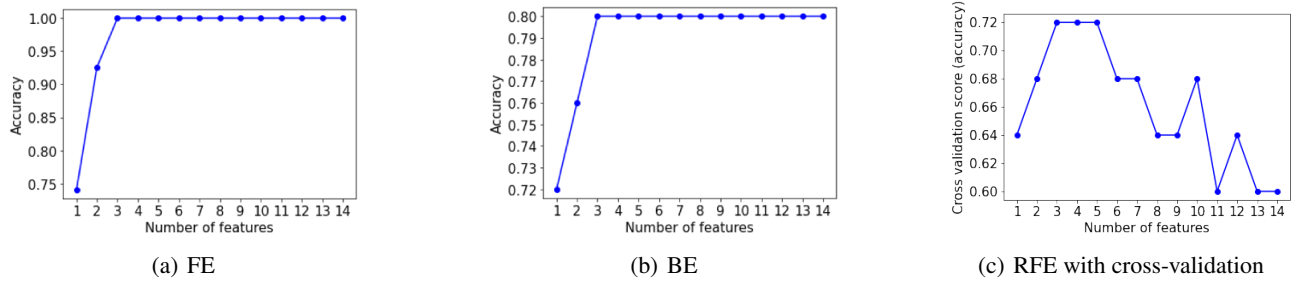
(a) FE   (b) BE   (c) RFE with cross-validation

Figure 5. Feature selection wrapper methods results.

completed the course, including filling the questionnaire, to have all the needed data notably the related demographic one for the study.

Forward elimination (FE), backward elimination (BE) and recursive feature elimination (RFE) with cross-validation are performed to find a smaller set of not redundant or complementary features. According to the experimental results, the optimal number of features appears to be 3, corresponding mostly to a 2- or 3-source combination with one or few features in each source. Taking 3 as the optimal number of features, the selected ones from each of the methods are indeed: time (B), verbal score (L) and deduction score (L) for RFE; field of education (D), time (B) and verbal score (L) for FE; and the same as FE for BE. These results emphasize the predictive power of the time and the verbal score whose importance was demonstrated previously.

Regarding filter methods, variable ranking methods were used such as statistical measures ANOVA (linear correlation coefficients) and Kendall (nonlinear correlation coefficients). The correlation between each feature and the target are computed in Figure 30 and Figure 31. Thus, independently of the prediction model, the verbal score is again significantly relevant for students' performance. Academic data or learning preferences data derived from the former are also well ranked, followed by behavioral data, demographic data and personality data. Among the behavioral features, the time is indeed better ranked than the number of actions and among the demographic data, learner's field of education has a much higher correlation with the target. This can be easily explained by the prior knowledge brought by the field of education that helps the learner to better succeed. Besides, in the population of D1, learner's level of education and learner's age are highly correlated as seen in Figure 32, describing why a good coefficient of one of them implies a low coefficient for the other in Figure 30 and Figure 31.

To conclude this part, this feature selection study, combined with the previous experiments, stresses the usefulness and the relevance of the verbal score and the time. Such role for the verbal score means that the skills of well extract, memorize and deduce knowledge from written learning ma-

terial could be a strong advantage and a major ability related to success. However, the statistical measures expose more general relationships such as students' field of education or average grade, which allows to have a more holistic view of students contrary to the time for instance. Moreover, the wrapper methods point out again that 2 data sources or about 3 features lead to the best results with a minimal and complementary set of features. Academic data and behavioral data are still the preferred sources.

5. Conclusion

In this work, we predicted students' performance using multiple data sources from online courses, including one we created to that purpose. We compared results with different learning algorithms and different datasets. Two research questions were posed: (1) To what extent using multiple data sources on learners increases prediction performance? (2) Which learners' characteristics are most related to higher academic performance? The results showed encouraging similar trends and knowledge with different students' populations and courses. We highlighted the relevance and the importance of academic and behavioral data in contrast to demographic ones and the use of multiple data sources where the optimal number of complementary features, hence complementary data on students', appears to be 3. Eventually, we formulated a hypothetical advantageous relationship between students' who have stronger abilities of memorizing and deducing knowledge from written material and better performance. However, it has to be kept in mind that most of the results are based on a small number of instances stressing the limitations of the study.

Future work will be related to having more participants to the learning experiment to challenge these current results, implementing a personality traits recognition process to enrich personality features, exploiting the explainability of the decision trees, creating new features from existing ones, and approaching the prediction task as a multi-class classification problem for refined results.

Acknowledgements

References

Adejo, O. W. and Connolly, T. Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, 10(1):61–75, 2018. ISSN 17581184. doi: 10.1108/JARHE-09-2017-0113.

Afzaal, S., Siau, N. Z., and Suhali, W. S. H. Evaluating students' personality and learning styles in higher education: Pedagogical considerations. *International Journal of Learning, Teaching and Educational Research*, 18(7):145–164, 2019. ISSN 16942116. doi: 10.26803/ijlter.18.7.10.

Ahmad, N. B. H. and Shamsuddin, S. M. A comparative analysis of mining techniques for automatic detection of student's learning style. *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10*, pp. 877–882, 2010. doi: 10.1109/ISDA.2010.5687150.

Al-Barrak, M. A. and Al-Razgan, M. Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, 6(7):528–533, 2016. ISSN 20103689. doi: 10.7763/ijiet.2016.v6.745.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J., and Olatunji, S. O. Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *EEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2017. ISSN 21693536. doi: 10.1109/ACCESS.2021.3061368.

Alamri, R. and Alharbi, B. Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access*, 9:33132–33143, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3061368.

Antoniuk, A. Learning Styles: Moving Forward from the Myth. *Canadian Journal for New Scholars in Education*, 10(2):85–92, 2019.

Baker, R. S. J. D. and Yacef, K. The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, 1(1):3–16, 2009. ISSN 21572100.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. A. *Classification and Regression Trees*. Taylor & Francis, 1984.

CanvasNetwork. Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Open Dataset. *Harvard Datavers*, 2, 2016. doi: 10.7910/DVN/1XORAL.

Chaudhury, P., Mishra, S., Tripathy, H. K., and Kishore, B. Enhancing the capabilities of student result prediction system. *ACM International Conference Proceeding Series*, 04-05-Marc(March), 2016. doi: 10.1145/2905055.2905150.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16 (Sept. 28):321–357, 2002. ISSN 10769757.

Dutt, A., Ismail, M. A., and Herawan, T. A Systematic Review on Educational Data Mining. *IEEE Access*, 5: 15991–16005, 2017. ISSN 21693536. doi: 10.1109/ACCESS.2017.2654247.

Feldman, J., Monteserin, A., and Amandi, A. Automatic detection of learning styles: state of the art. *Artificial Intelligence Review*, 44(2):157–186, 2015. ISSN 15737462. doi: 10.1007/s10462-014-9422-6.

Goldberg, L. R. An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, 59, 1990.

Graf, S. Adaptivity in Learning Management Systems Focussing on Learning Styles. *PhD Thesis*, 1298561821 (December 2007), 2007.

Guyon, I. and Elisseeff, A. An Introduction to Variable and Feature Selection. *Machine Learning Research*, 3: 1157–1182, 2003.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002. doi: 10.1023/A:1012487302797.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S. N. Predicting academic performance: A systematic literature review. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, pp. 175–199, 2018. ISSN 1942647X. doi: 10.1145/3293881.3295783.

Kotsiantis, S., Pierrakeas, C., and Pintelas, P. Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426, 2004. ISSN 08839514. doi: 10.1080/08839510490442058.

Kuzilek, J., Hlosta, M., and Zdrahal, Z. Open University Learning Analytics dataset. *Sci Data 4*, 170171, 2017. ISSN 21693536. doi: 10.1038/sdata.2017.171.

Lakshmi, T., Martin, A., Begum, R., and Venkatesan, V. An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data. *International Journal of Modern Education and Computer Science*, 5(5):18–27, 2013. ISSN 20750161. doi: 10.5815/ijmecs.2013.05.03.

Livieris, I. E., Drakopoulou, K., and Mikropoulos, T. A. Research on e-Learning and ICT in Education. *Research on e-Learning and ICT in Education*, 2018. ISSN 1436-4522. doi: 10.1007/978-3-319-95059-4.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., and Holzinger, A. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. *Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science, vol 12279. Springer, Cham.*, 2020. doi: https://doi.org/10.1007/978-3-030-57321-8_1.

Loyola-Gonzalez, O. Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2949286.

Mesarić, J. and Šebalj, D. Decision trees for predicting the academic success of students. *Croatian Operational Research Review*, 7(2):367–388, 2016. ISSN 18480225. doi: 10.17535/crorr.2016.0025.

Osmanbegovic, E. and Suljic, M. Data Mining Approach for Predicting Student Performance. *Journal of Economics and Business*, 10, 2012.

Papadogiannis, I., Poulopoulos, V., and Wallace, M. A Critical Review of Data Mining for Education: What has been done, what has been learnt and what remains to be seen. *International Journal of Educational Research Review*, pp. 353–372, 2020. ISSN 2458-9322. doi: 10.24331/ijere.755047.

Popescu, E. Diagnosing students' learning style in an educational hypermedia system. *Cognitive and Emotional Processes in Web-Based Education: Integrating Human Factors and Personalization*, pp. 187–208, 2009. doi: 10.4018/978-1-60566-392-0.ch011.

Quinlan, J. R. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. ISSN 21693536.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., and Durán-Domínguez, A. Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences (Switzerland)*, 10(3), 2020. ISSN 20763417. doi: 10.3390/app10031042.

Romero, C. and Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007. ISSN 09574174. doi: 10.1016/j.eswa.2006.04.005.

Romero, C. and Ventura, S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 40(6):601–618, 2010. ISSN 10946977. doi: 10.1109/TSMCC.2010.2053532.

Romero, C. and Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):1–21, 2020. ISSN 19424795. doi: 10.1002/widm.1355.

Vijayalakshmi, V. and Venkatachalapathy, K. Comparison of Predicting Student's Performance using Machine Learning Algorithms. *International Journal of Intelligent Systems and Applications*, 11(12):34–45, 2019. ISSN 2074904X. doi: 10.5815/ijisa.2019.12.04.

Villagrá-Arnedo, C., Gallego-Duran, F. J., Compan-Rosique, P., Llorens-Largo, F., and Molina-Carmona, R. Predicting academic performance from Behavioural and learning data. *International Journal of Design and Nature and Ecodynamics*, 11(3):239–249, 2016. ISSN 17557445. doi: 10.2495/DNE-V11-N3-239-249.

Watson, W. and Watson, S. L. An argument for clarity: What are learning management systems, what are they not, and what should they become? *TechTrends*, 51: 28–34, 2007.

A. Questionnaire for collecting demographic and personality data

The main course and questionnaire were in French for communication and spread purpose of the learning experiment. We thus report this questionnaire version with translations.

1. Vous êtes (*You are*) :

   • une femme (*a woman*).
   • un homme (*a man*).
   • Ne se prononce pas (*no answer*).

2. Quel âge avez-vous ? *How old are you?*

   • Réponse numérique. *Numerical answer*.

3. Le français est-il votre langue maternelle ? *Is French your native language?*

   • Oui (*yes*).
   • Non (*no*).

4. Quel est votre niveau d'études actuel (ou dernier) ? *What is your current (or last) level of education?*

   • Choix de Bac +1 à Bac +8. *Categorical answer from A-level to PhD*.

5. Dans quelle filière d'études êtes-vous (ou la dernière) ? *What is your current (or last) field of education?*

   • Maths, Informatique, Ingénierie, Technologie... *Maths, Computer Science, Engineering, Technology...*
   • Santé, Biologie, Médecine... *Health, Biology, Medicine...*
   • Lettres, Langues, Arts, Culture... *Humanities, Languages, Arts, Culture...*
   • Administration, Droit, Management, Commerce... *Administration, Law, Management, Commerce...*

6. En apprendre plus sur le cerveau vous intéresse :

   • beaucoup (le sujet m'intéresse). *You are very interested in learning more about the brain (you are interested in the subject)*.
   • moyennement (mais je suis curieux(se)). *You are moderately interested in learning more about the brain (but you are curious)*.
   • peu (mais je souhaite participer à cette expérience pédagogique). *You have little interest in learning more about the brain (but you want to participate in this educational experience)*.

7. Racontez une histoire hypothétique dans laquelle vous décrivez une journée idéale où tout s'est bien passé. Exemple : Aujourd'hui était une super journée pour moi, je me suis réveillé(e) tôt le matin et j'ai pu courir 5 km avant de me rendre au travail. Pendant que je travaillais, mon ami George m'a apporté une tasse de café, c'était une bonne surprise car la machine à café du bureau ne fonctionne pas. J'ai déjeuné avec Lucy, George et Mike, c'était vraiment sympa. Le soir, j'ai pu sortir mon chien pour une petite promenade et j'ai eu une belle vue sur la lune. *Tell a hypothetical story in which you describe an ideal day when everything went well. Example: Today was a great day for me, I woke up early in the morning and was able to run 5 km before going to work. While I was working, my friend George brought me a cup of coffee, which was a nice surprise because the office coffee machine doesn't work. I had lunch with Lucy, George and Mike, which was really nice. In the evening I was able to take my dog out for a little walk and had a nice view of the moon.*

   • Description libre. *Free description*.

8. Racontez une histoire hypothétique dans laquelle vous décrivez une mauvaise journée où vous avez rencontré des difficultés. Exemple : Aujourd'hui n'était pas une bonne journée pour moi, je me sentais très fatigué(e) le matin et il m'était difficile de me concentrer au travail le matin. Les choses se sont améliorées pendant le déjeuner, j'ai pu discuter un peu avec Lucy et elle m'a invité à une fête ce week-end. Cependant, il pleuvait quand j'ai quitté le bureau et j'ai oublié mon parapluie, j'étais vraiment trempé(e) quand je suis arrivé(e) à la maison, à cause du temps je n'ai pas pu sortir mon chien aujourd'hui, j'espère que je pourrai le faire demain. *Tell a hypothetical story in which you describe a bad day when you had difficulties. Example: Today was not a good day for me, I felt very tired in the morning and it was difficult to concentrate at work in the morning. Things improved over lunch, I was able to chat to Lucy a bit and she invited me to a party at the weekend. However, it was raining when I left the office and I forgot my umbrella, I was really soaked when I got home, because of the weather I couldn't take my dog out today, I hope I can do it tomorrow.*

   • Description libre. *Free description*.

9. J'accepte en suivant ce cours de donner la permission d'utiliser mes réponses anonymes pour une étude scientifique sur les modèles d'apprentissage en ligne. Aucune donnée personnelle ne sera stockée ni partagée (plus d'informations ici[9]). *By taking this course I consent to give permission for the usage of my anonymized answers for a scientific study associated with e-learning patterns. Your personal data will not be stored or shared (more information here).*

   • Oui (*yes*).

------

[9] https://sites.google.com/view/braincourse/utilisation-des-donnees

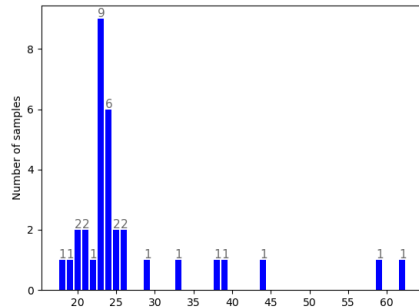# B. Data visualization (collected data)
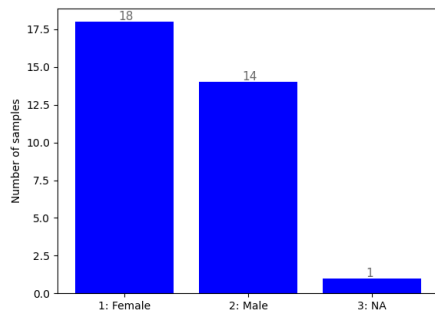

Figure 6. Learners' age.
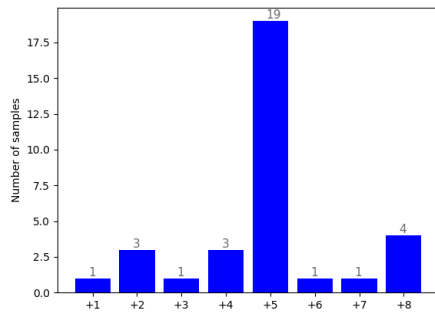

Figure 7. Learners' gender.

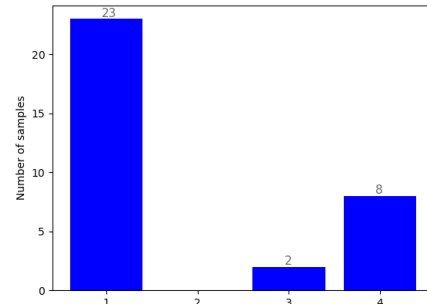*Figure 8.* Learners' level of education.

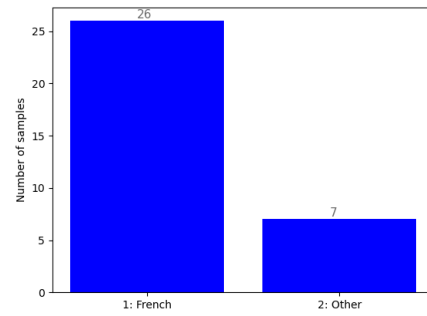
Figure 9. Learners' field of education.

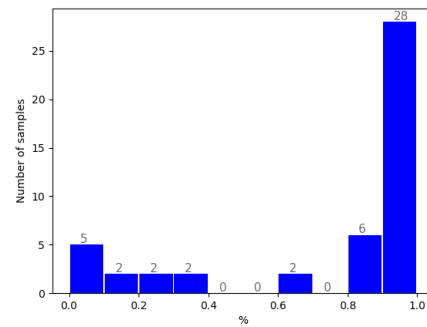*Figure 10.* Learners' native language.

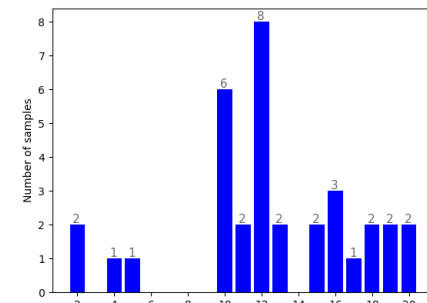*Figure 11.* Learners' percentage of completion.
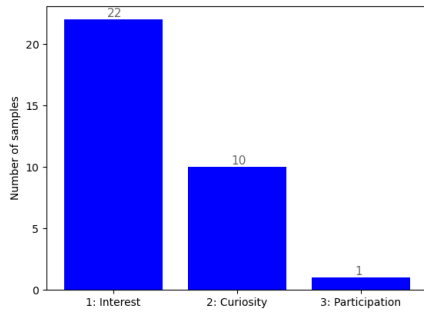

Figure 12. Learners' number of actions.

Figure 13. Learners' motivation.

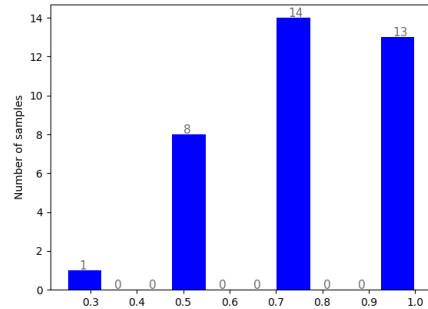*Figure 14.* Learners' visual score.

Figure 15. Learners' verbal score.

Figure 16. Learners' factual score.

Figure 17. Learners' practical score.
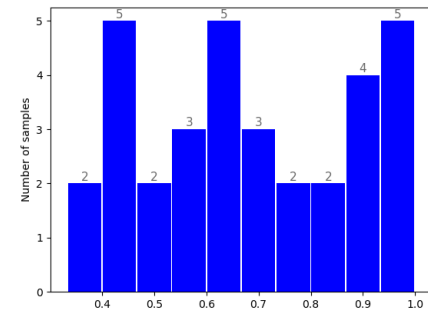
Figure 18. Learners' memory score.
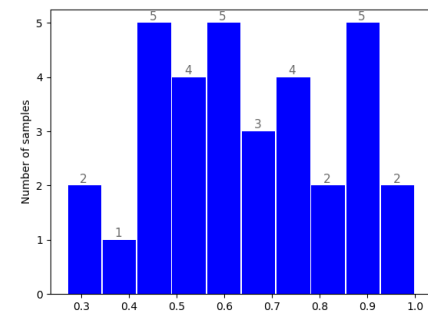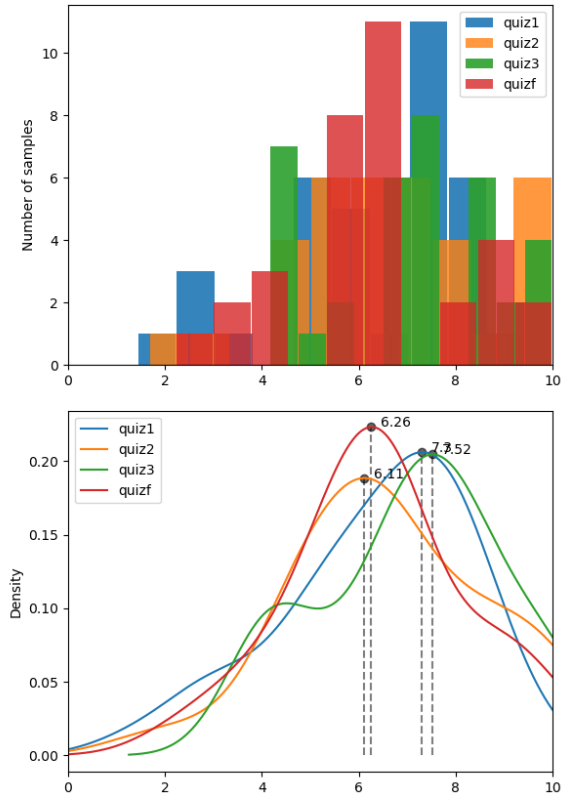

Figure 19. Learners' deduction score.

Figure 20. Learners' quiz grades and their kernel density estimates.
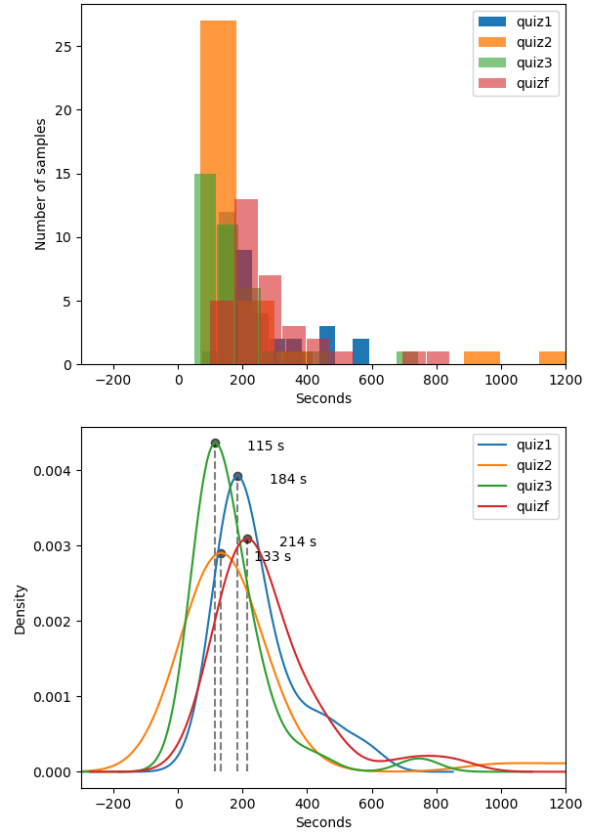

Figure 21. Learners' quiz times and their kernel density estimates.

C. Balancing techniques results

Table 5. Baseline results without balancing technique.

	D1	D2	D3
Accuracy (%)	48.33	77.63	69.34
Precision (%)	42.50	78.25	67.28
Recall (%)	50.00	65.09	67.41
F-score (%)	41.66	65.85	67.28

Table 6. Results with upsamling.

	D1	D2	D3
Accuracy (%)	52.49	69.81	69.72
Precision (%)	47.50	70.09	69.95
Recall (%)	55.00	69.82	69.72
F-score (%)	47.16	69.71	69.63

Table 7. Results with downsamling.

	D1	D2	D3
Accuracy (%)	66.66	69.82	69.49
Precision (%)	53.33	70.18	69.70
Recall (%)	65.00	69.82	69.49
F-score (%)	57.33	69.69	69.41

Table 8. Results with SMOTE.

	D1	D2	D3
Accuracy (%)	59.16	71.19	67.60
Precision (%)	59.16	71.92	67.97
Recall (%)	60.0	71.19	67.60
F-score (%)	55.49	71.02	67.45

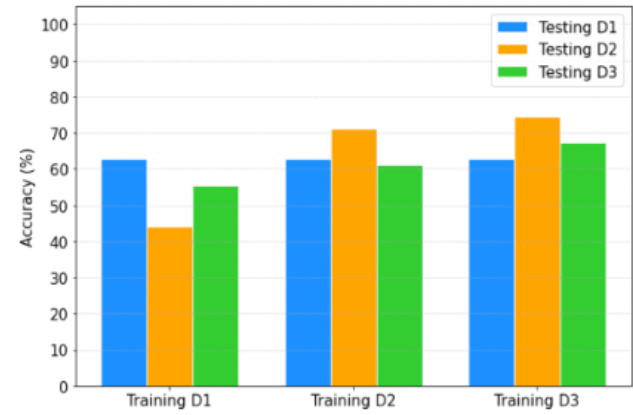## D. Transfer learning results

Figure 22. Results with DT.

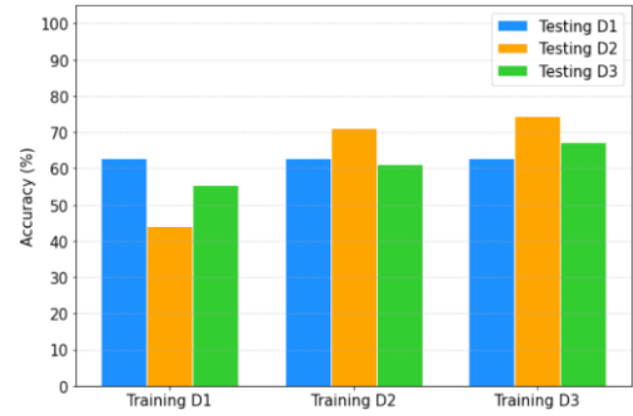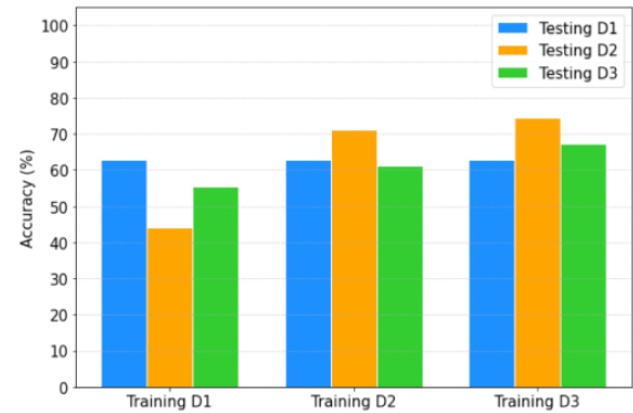*Figure 23.* Results with RF.

Figure 24. Results with SVM.

E. Features involved in the second experiment

Table 9. Features' comparison.

	D1	D2
Demographic (D)	- age: Learner's age - ed_level: Learner's level of education	
	/	- ed_field: Learner's field of education
Academic (A)	- avrg_grade: Average of the learner's course grades before taking the final exam	
Behavioral (B)	- sum_click: Total number of clicks during the courses	- %_completion: Percentage of course completion - nb_action: Total number of views and achievements during the course - time: Total time learner spent on activities

F. Detailed results of the second experiment

Table 10. Results of combinations of usual data sources.

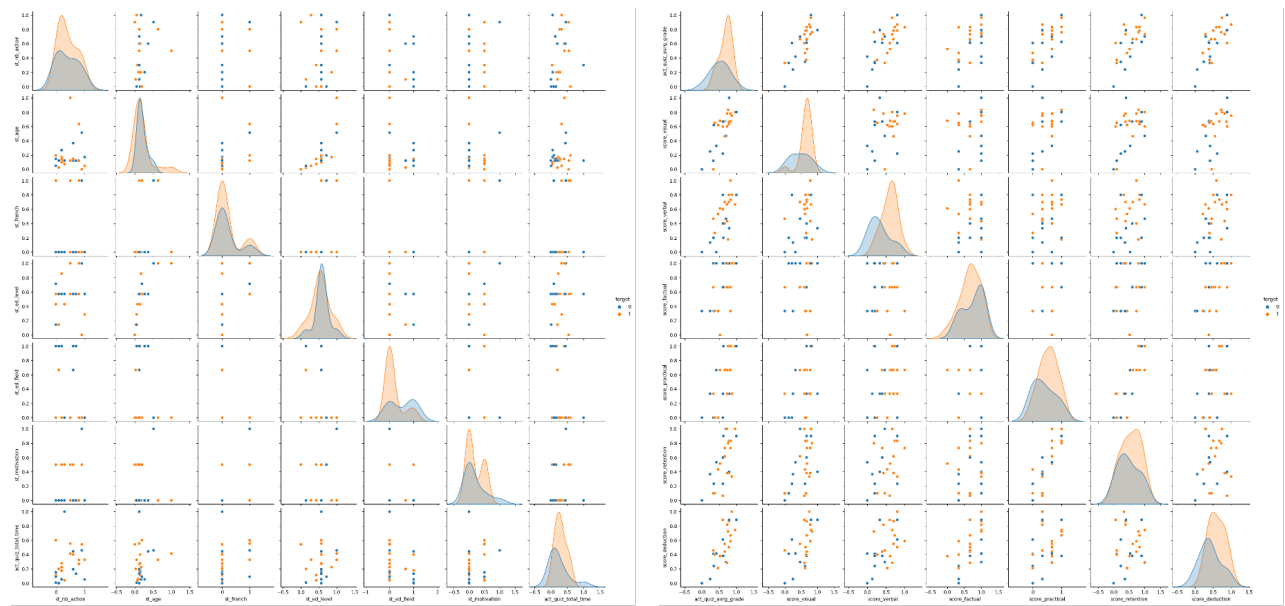|	Accuracy		Precision		Recall		F-score	
	D1	D2	D1	D2	D1	D2	D1	D2
D	66.66	53.17	58.53	53.25	65.00	53.17	59.00	52.86
A	71.66	68.56	68.33	69.21	72.50	68.56	66.49	68.29
B	70.00	69.10	72.50	69.22	72.50	69.10	70.00	69.04
D + A	68.33	68.85	53.33	69.57	62.50	68.85	55.16	69.56
A + B	75.00	75.18	62.50	75.31	75.00	75.18	66.33	75.15
B + D	68.33	69.10	54.16	69.22	65.00	69.09	57.00	69.04
D + A + B	63.33	75.38	55.00	75.43	65.00	75.38	55.66	75.36

G. Scatter matrix plots

Figure 25. Scatter matrix plots. (See the close distributions of demographic data according to the target).
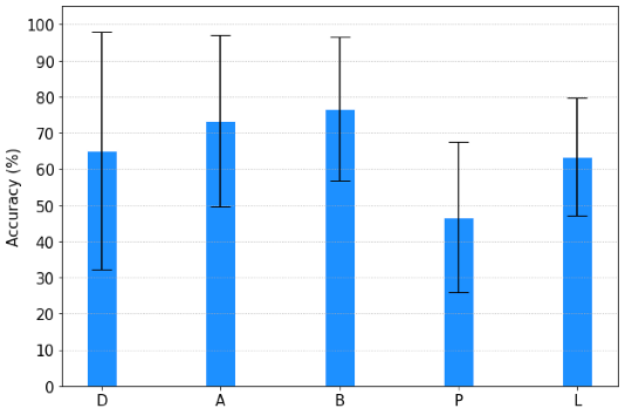
H. Additional sources results

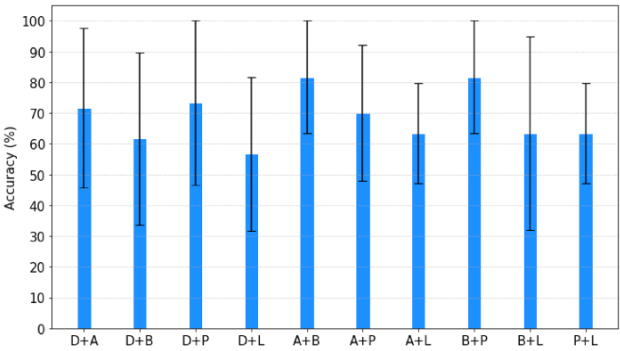

Figure 26. Resulting accuracies with 1-source combinations.

Figure 27. Resulting accuracies with 2-source combinations.

Table 11. Results of 1-source combinations.

	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
D	68.33	55.00	62.50	57.00
A	76.66	72.50	77.50	72.33
B	71.66	65.00	72.50	65.66
P	46.66	23.33	45.00	30.16
L	61.66	54.16	65.00	54.99

Table 12. Results of 2-source combinations.

	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
D + A	70.00	65.00	70.00	64.83
D + B	65.00	55.00	67.50	56.49
D + P	68.33	55.00	62.50	57.00
D + L	56.66	48.33	55.00	48.99
A + B	65.00	54.16	65.00	57.33
A + P	76.66	72.50	77.50	72.33
A + L	65.00	61.66	67.50	60.83
B + P	75.00	70.00	**77.50**	70.00
B + L	66.66	57.50	67.50	58.99
P + L	58.33	49.16	60.00	50.83

Table 13. Results of 3-source combinations.

	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
D + A + B	60.00	51.66	60.00	53.33
D + A + P	70.00	65.00	70.00	64.83
D + A + L	63.33	56.66	65.00	57.50
D + B + P	65.00	55.00	67.50	56.49
D + B + L	55.00	46.66	57.50	48.16
D + P + L	56.66	48.33	55.00	48.99
A + B + P	63.33	52.50	62.50	54.83
A + B + L	56.66	44.16	57.50	47.33
A + P + L	53.33	45.83	57.50	47.49
B + P + L	60.00	49.16	62.50	51.49

Table 14. Results of 4-source combinations.

	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
D + A + B + P	63.33	52.50	62.50	54.83
D + A + B + L	51.66	41.66	52.50	43.99
D + A + P + L	53.33	46.66	55.00	47.49
D + B + P + L	66.66	57.50	67.50	58.99
A + B + P + L	56.66	44.16	57.50	47.33

Table 15. Results of 5-source combinations.

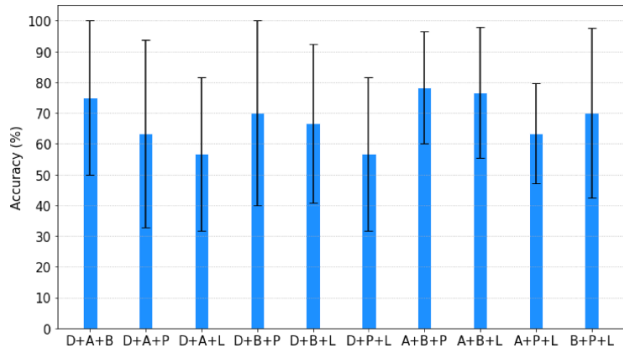|	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
D + A + B + P + L	51.66	41.66	52.50	43.99

Figure 28. Resulting accuracies with 3-source combinations.
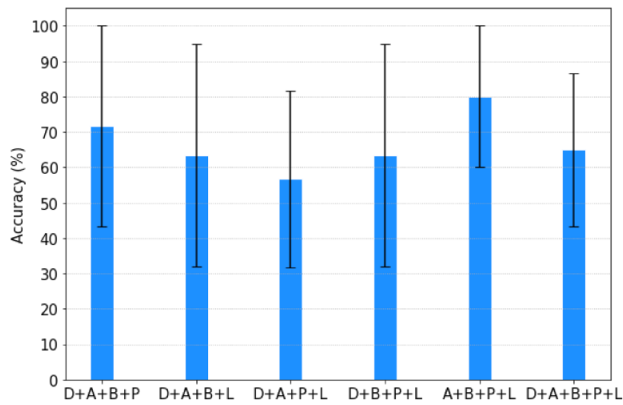

Figure 29. Resulting accuracies with 4- and 5-source combinations.
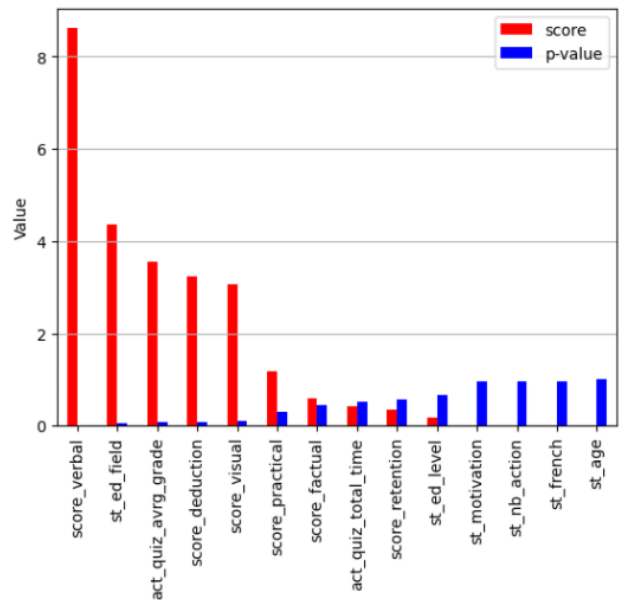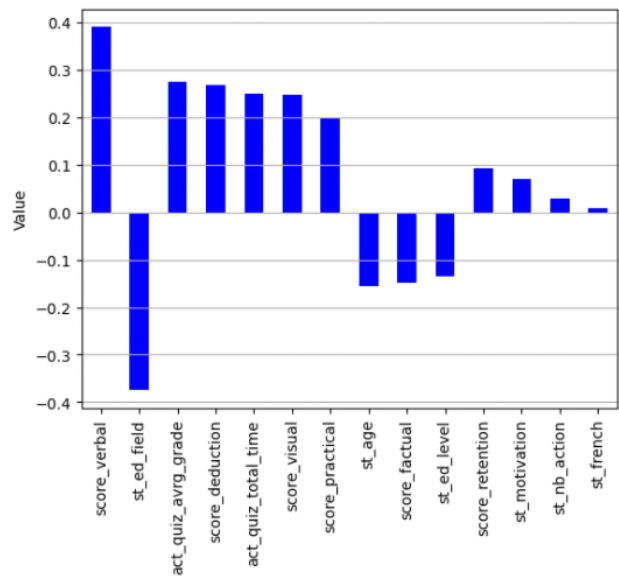
I. Correlation plots

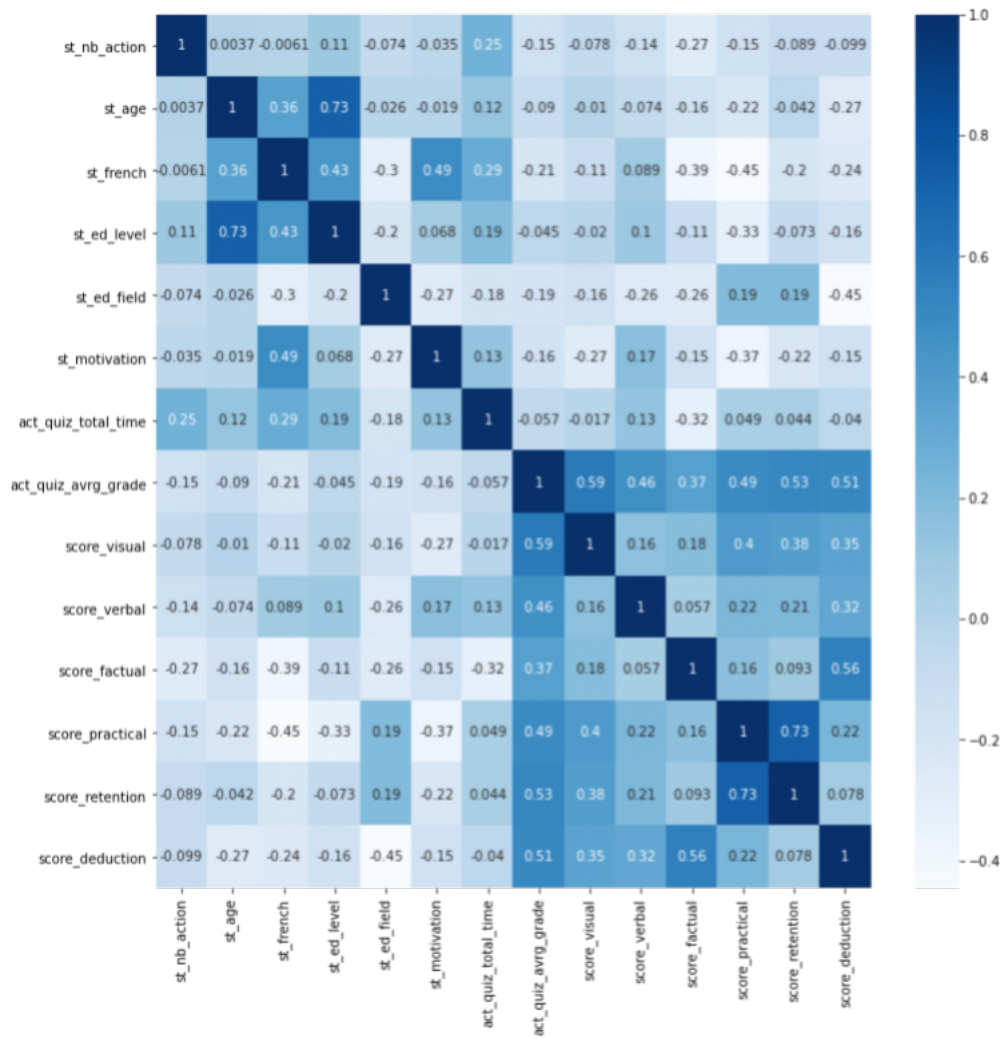*Figure 30.* ANOVA F-values.

Figure 31. Kendall's correlation coefficients.

Figure 32. Correlation matrix between features.