International Journal of Web-Based Learning and Teaching Technologies
Volume 16 • Issue 2 • March-April 2021

A Withdrawal Prediction Model of At-Risk Learners Based on Behavioural Indicators

Fedia Hlioui, Multimedia Information System and Advanced Computing Laboratory, University of Sfax, Sfax, Tunisia
Nadia Aloui, CCSE Department SWE, Jeddah University, Saudi Arabia & University of Sfax, Tunisia & ISIMS, Sfax, Tunisia
Faiez Gargouri, Multimedia Information System and Advanced Computing Laboratory, University of Sfax, Sfax, Tunisia

ABSTRACT

Nowadays, the virtual learning environment has become an ideal tool for professional self-development and bringing courses for various learner audiences across the world. There is currently an increasing interest in researching the topic of learner dropout and low completion in distance learning, with one of the main concerns being elevated rates of occurrence. Therefore, the early prediction of learner withdrawal has become a major challenge, as well as identifying the factors, which contribute to this increasingly occurring phenomenon. In that regard, this manuscript presents a framework for withdrawal prediction model for the data from The Open University, one of the largest distance learning institutions. For that purpose, we start by pre-processing the dataset and tackling the challenge of discretization process and unbalanced data. Secondly, this paper identifies the semantical issues of raw data by introducing new behavioural indicators. Finally, we reckon on machine learning algorithms for withdrawal prediction model to understand the lack of learners' commitment at an early stage.

KEYWORDS

Behavioural Indicators, Discretization, Learner At-Risk, OULAD, Unbalanced Data, Withdrawal Prediction Model

1. INTRODUCTION

With an increasing interest in open educational resources, the web-based learning has become a commonplace in higher education institutions and organisations. There is plethora of different terms used in literature to describe the online learning delivery platforms like Virtual Learning Environment (VLE), or Learning Management Systems (LMS), or Massive Open Online Courses (MOOC). In the remaining of this paper, we will use the term VLE for designating all the E-learning environments. These modern trends of these platforms is credited to their ability to provide an open, online, high quality and low-cost educational content on a large scale more efficiently (Almatrafi & Johri, 2018). The VLEs attract not only the educational and pedagogical communities, but also scientists from various disciplines such as Philosophy, Educational science, Machine Learning, Statistics, and Computers sciences. Despite the potential and high associated with the VLE, retention rate over

DOI: 10.4018/IJWLTT.2021030103

all are typically very low. Studies on distance learning assert that the percentage of learners who completed the course is only 22% (Reich, 2014), or even 7% reported by (Jiang & Kotzias, 2016). Such percentages seriously doubt the reliability and the efficacy of the VLE (Kloft, et al., 2014).

This in turn can motivate researchers and scientists to analyse and exploit the reasons of the high withdrawals; hence, dropout prediction can be an important aspect in such environment. An early prediction can help the different stakeholders for several reasons. Teachers can anticipate possible issues with learners and adapt their courses or leaning strategies to improve engagement. In addition, instructors or courses' designers can use the predictive model to make decisions about the curricular design and to personalize interventions. Furthermore, learners can receive information about their learning progress, which allow them to reflect on how they are doing and improve their performance.

Rigorous efforts were put for modelling advanced tools for monitoring the learners' progress. These tools often use the VLE learners' characteristics as an input and the predicting the learners' course withdrawal as an output. In this context, various Machine Learning techniques have been successfully applied to obtain statistically high dropout prediction accuracy. They mainly focused on gathering the learners' characteristics and on applying several techniques in order to process this form of data.

In the VLE context, many datasets served to build models aimed at predicting the aforementioned outcomes, like the KDD Cup (Kuzilek, et al., 2017), OULAD (Kuzilek, et al., 2017), the Student Academic Performance dataset (Bharara, et al., 2018), etc. The existing methods for data preprocessing mainly employed the intrinsic statistical characteristics of the data features (the learners' descriptors), in order to prepare an effective data, before the training step. Nevertheless, these features enclosed a semantical characteristic that has an important impact on the extracted knowledge' quality. The abstraction of the learners' descriptors is performed through the establishment of the indicators. In the work done by (Popescu, 2009), the behavioral indicators referred to the relative frequency of these learner actions, the amount of time spent on a specific action type and the order of performing these actions. According to (Bousbia, et al., 2013), an indicator describes the learners' navigation behaviors based on their low-level navigation traces (links followed, clicks, etc.). Based on the above-mentioned definitions, we infer that, the model of behavioral indicators is seen as a meta-knowledge of the traces' observations. Many behavioral indicators are proposed in literature like navigation type (Bousbia, et al., 2013), disorientation (Adda, et al., 2016), concentration rate (Ammor, et al., 2013), collaborative level (Bouzayane & Saad, 2017), contribution rate (Wong, et al., 2015) and effort level (Papanikolaou, 2015). These indicators offered a considerable representative power; they provided new semantically coherent features' set, which is efficient not only for optimizing the predictability of learners' dropout but also for understanding the impact of the indicators on the learners' commitment to the course completion.

For this sake, we propose our own framework for a withdrawal prediction model for the OULAD dataset (Open University Learning Analytics Dataset) of the Open University, one of the largest distance learning institutions in United Kington (Kuzilek, et al., 2017). We start by pre-processing the dataset and tackling the challenge of discretization process and unbalanced data. Secondly, we address the semantical issues of raw data by introducing new indicators. Notably, we focus on the behavioural indicators covering the learners' interactions during the learning sessions. Finally, we reckon on Machine Learning algorithms for ensuring both efficient and semantically coherent predictive model.

The structure of this paper is organized as follows. We begin by presenting the existing works, which tackle with the problem of learners' dropout in VLE. In the third section, we describe the different facets of OULAD dataset. The subsequent section involves our steps for preparing and preprocessing the raw data to be eligible for a datamining task. In the fifth section, we propose new demographic, behavioural and performance indicators. Thereafter, we proceed to experimentally assess the obtained data for predicting the learners' dropout. Finally, we recall the main contributions and identify some future research in short terms and others in the long terms.

2. LITERATURE REVIEW

Dropouts are expensive for educational institutes as well as for society (Barbu, et al., 2017). It is plain to see that researchers along with educational technologies proposed innovative dropout prediction solutions. Considerable efforts have been conducted to collect the learners' characteristics, in order to understand the learners' withdrawals on distance learning. Some research works were experimented by using public platforms like Coursera in (Maldonado-Mahauad, et al., 2018), Edx in (Liang, et al., 2016) and Moodle in (Romero, et al., 2013), or private platforms like in (Zhang, et al., 2017) and (El Haddioui & Khaldi, 2017). Other studies exploited freely datasets like the KDD cup dataset in (Cao & Zhang, 2015), OULAD in (Kuzilek, et al., 2017), the Assessments Benchmark dataset (Piech, et al., 2015), etc. The collected data can cover the discussion forums, the navigation path, the assessments scores, the clickstreams, etc. Other researchers like (Kardan & Conati, 2013) adopted an innovative technology for collecting the learners' traces, which is an eye tracking for monitoring the learner's eye gaze. In (Shareghi Najar, et al., 2015), the authors proved that this process often contain noise, which can be caused by different factors such as participant's head movements, wearing contact lenses and glasses, or inaccurate calibration. Moreover, it is very likely that the collected data may contain privacy information about the learner (May, et al., 2016).

The content of these datasets depends on the specificities of the VLE, the accessibility of the VLE (protected or open source) and on the application's context and aim. Each proposed dataset emphasizes on one information side of learner's individual differences and ignores the other sides. For example, the KDD Cup dataset (Cao & Zhang, 2015) does not contain the demographic and historical data from past courses. Other datasets like the Students' Academic Performance dataset (Bharara, et al., 2018) and the Assessment Benchmark dataset (Piech, et al., 2015), does not cover the behavioral aspect of learners. In Coursera platform, some researchers investigated only in the discussion forums (Wen et al., 2014) (Wang, et al., 2015), but others dealt only with the learners' clickstreams in videos (Shridharan, et al., 2018). However, the experiments made in Coursera platform neglected the demographic and performance features. Moreover, it remains not open-access for scientists due to many privacy issues. VLEs are often declined to publish the data due to confidentiality and privacy concerns (Dalipi, et al., 2018). It was proved that it is not always straightforward or simple to promise absolute privacy, confidentiality and anonymity when using an open VLE (May, et al., 2016). Wherefore, we opt for the Open University Learning Analytics Dataset (OULAD), which the privacy levels are clearly identified, and their protection measures allow us to set rules and policies in terms of learner tracking (Kuzilek, et al., 2017). This dataset is freely available[1], anonymised and certified by the Open Data Institute[2]. It includes both learner demographic data and interaction data with the university's VLE.Previous work with OULAD used different machine learning and pattern recognition techniques to understand the causes of learners' withdrawals. It is a challenging task due to a wide variety of features that are employed, each of which is designed for a slightly different problem specification. In the work done by (Haiyang, et al., 2018), only three kinds of learning activities types (forum, OUcontent and resource) were considered for modeling the dropout problem. According to (Wolff et al., 2014), the four VLE activities (forum, OUcontent, resource, Subpage) and demographic data contributed the most for determining the learners' withdrawal. Nevertheless, other studies like in (Hussain, et al., 2018) highlighted that the significant learning activities' types for predicting low-engagement in an OULAD Dataset were "OUcontent", "forumng", "subpage", and "homepage". In contrast to these findings, the authors (Heuer & Breiter, 2018) combined all activity types into a single formula, i.e. daily activity. In this work, the semantics behind the learning activities are neglected. Therefore, these research works did not exploit all the facets of the OULAD dataset.

From other side, all the studies cited previously viewed the dropout prediction as a binary classification problem to make distinction between withdrawal and retention of learners. In terms of the machine learning and patterns recognition techniques used, some studies adapted one (Kuzilek, et al., 2018) or more supervised classification algorithms (Hussain, et al., 2018) (Heuer & Breiter,

2018). While, these studies did not take into consideration the unbalanced data issue. It corresponds to the case where the prediction classes are not equivalently distributed all over the instances (the learners, in our case). This leads to over-fitted models, which usually tend to predict correctly the majority class and to predict wrongly the minority class.

Moreover, these works are lacking an important dimension of data understandability. Actually, a central challenge to any study includes significant aggregation of raw data sets, often requiring advanced methods that scale to large data sets. The semantics behind the raw attributes are not straightforward. For example, the number of clicks on a certain space in the VLE's corresponding web platform does not translate the level of preservation or the autonomy of a learner. This problem is solved by the use of indicators. The indicators describing the learners' individual differences may help not only the instructors in their monitoring and accompanying tasks, but also the courses' designers for the modeling and adaptation of the learning materials. Additionally, they can help developers to evaluate the VLE effectively and expand system function for future development trend. Despite these metrics showed a powerful representative dimension on various datasets like that of Coursera (Ramesh, et al., 2015) or Moodle (Djouad & Mille, 2018) and, they were not employed on OULAD database yet.

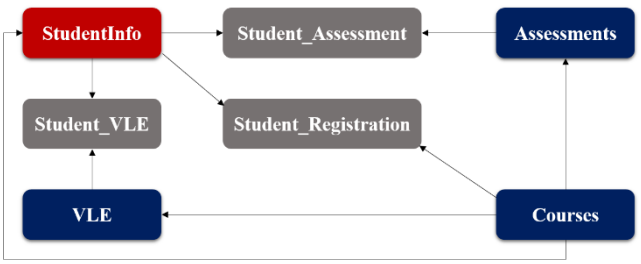
3. OULAD DATASET DESCRIPTION

The following article's contributions are focusing on OULAD dataset, which is a freely available as a set of CSV files, in the web under a CC-BY license (Kuzilek, et al., 2017). This database was collected at the Open University, which is the one of the largest distance learning university in United Kingdom. It represents all the facets of the learners' characteristics. It contains information about 22 courses delivered between 2013 and 2014, which come from two main discipline: Social sciences and Science, Technology, Engineering and Mathematics (STEM). Figure 1 highlights the structure of the OULAD database. Each course is called a module. Modules can be presented multiple times during one or two years. To distinguish between different presentations of a module, each presentation is named by the year and month it starts.

OULAD dataset is obtained by joining seven different tables where all of them are joined to form a central composite table "Student-Info":

- **Table "Student-Info":** Contains learner demographic features with the final_result as final score for achieving the course, imd_band stands for Index of multiple deprivation; UK quality studies containing seven forms including health, education skills and disability for each learner (Gamie, et al., 2019).
- **Table "Course":** Covers all available modules and their presentation. Modules can be presented multiple times during one or two years. To distinguish between different presentations of a module, each presentation is named by the year and month it starts.

Figure 1. The structure of OULAD database



- **Table "Assessment":** Three types of assessments are available: Tutor Marked Assessment, Computer Marked Assessment, and Final Exam. Some module contains a mix of two or more types of assessments. Each assessment has a specific weight and a cut-off day of submission. The number of assessments and their weights vary from module to module.
- **Table "VLE":** Includes all the materials and course contents for each Module-presentation, uploaded by teacher. There are twenty types of learning resources available for each course, such as reading pdf files, access to homepage and glossary, participating in forums and collaborative platforms, undertaken quizzes and questionnaires, etc. Each learning activity has a specific role for the module material and is identified with a label, for example, Forummng, OUcontent, Dataplus.
- **Table "Student_Registration":** Contains the date of enrolment/ un-enrollment of learners' in a certain module-presentation.
- **Table "Student_Registration":** Contains the date of enrolment/ un-enrollment of learners' in a certain module-presentation.
- **Table "Student_Assessment":** Refers to the assessment results for each learner.
- **Table "Student_VLE":** Captures the daily information relating to learner behaviour within an online course, in addition to the number of clicks learners made while studying the course material in the VLE.

Table 1 presents the dictionary of OULAD database, which contains the description of each field in each table.

4. METHODOLOGY OF RAW DATA PREPROCESSING

To overcome the previously stated issues, we propose a new framework for mining OULAD in order to design a new withdrawal predictive model. This framework can help the instructors to interfere with at risk learners at the appropriate time with the right tools for a constructive intervention. Moreover, it can support the courses' designers for adapting the learning content according to the learners' needs or the developers for evaluating the VLE efficiency.

Therefore, this framework (see. Figure 2) encloses four main phases: Data preprocessing, Indicators extraction, K-means-based Data Discretizing, withdrawal prediction. In the present section, we detail the steps of the first phase. They focus on three main axis: (i) data format mapping, (ii) rules-based data discretizing and (iii) data balancing.

4.1. Data Format Mapping

As a first step, we start by mapping the relational database into a tabular structure (dataset). This latter is useful for a data mining procedure. Therefore, we use the foreign keys in the seven tables (see Table 1) in order to merge them into a global table. In our explorative study, we focus on data from only one module-presentation named module "DDD" and presentation "2013B" (i.e. this means that the course started in February 2013). The course belongs on STEM subject with 1303 enrollments. This dataset includes 536,837 interactions (i.e. learners' clickstreams) within the VLE and 173,912 submitted assessments. Aiming at representing each learner by a row in the dataset, we proceed to aggregate the rows that correspond to the same learner's ID.

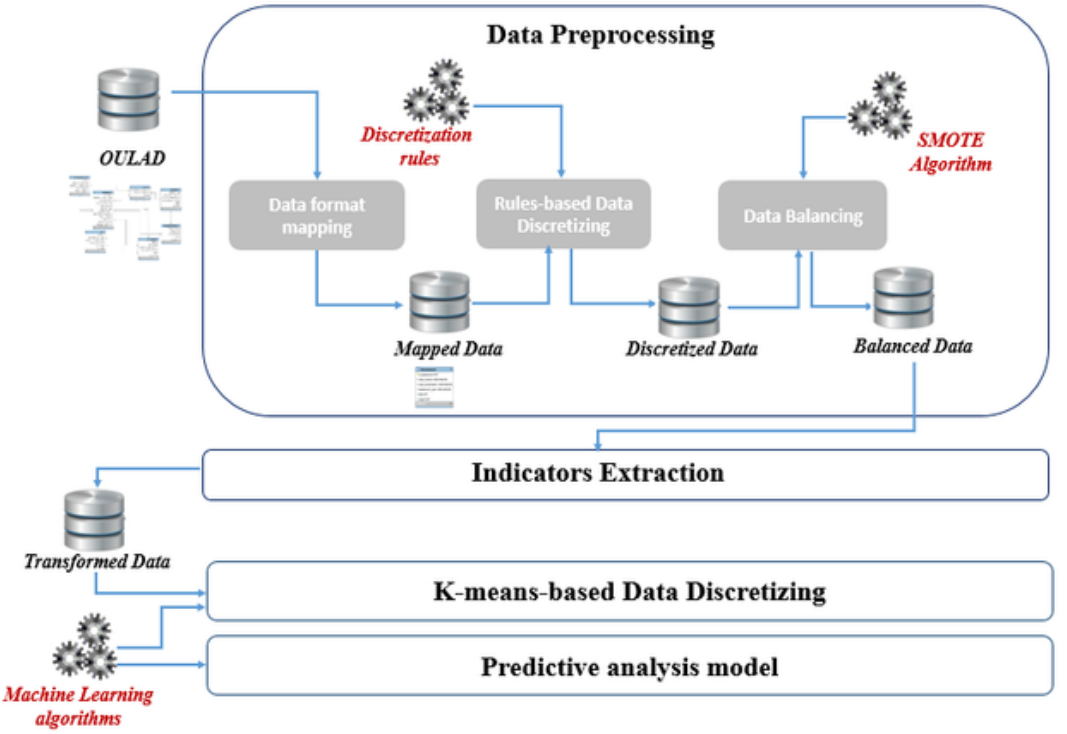
The learner's corresponding rows code the information about their interactions (number of clicks) with a particular site, which represents a certain activity type, at different timestamps (throughout the learning journey). For each activity type, we propose to insert new columns that represent their corresponding total number of clicks. As the dataset contains twenty unique activity types, we suggest classifying them into four families:

Table 1. The Dictionary of OULAD database

Table	Attribute	Description
Student_Info	*id_student*	The unique learner identification number.
	gender	The learner's gender.
	region	The geographic region, where the learner lived.
	highest_education	The highest learner education level on entry to the course.
	imd_band	The IMD band of the place where the learner lived during the course.
	age_band	A band of learner's age.
	num_of_prev_attempts	The number of how many times the learner has attempted this course.
	studied_credit	The number of credits for the modules, which the learner is currently studying.
	disability	The disability of the learner.
	final_result	The learner's final result in the module-presentation.
VLE	*id_site*	The identification number of the learning material.
	activity_type	The role associated with the learning material.
	week_from	The week from which the learning material is planned.
	week_to	The week until which the learning material is planned.
StudentVLE	*date*	The day of learner's interaction with the learning material.
	sum_click	The number of times the learner interacted with the learning material.
Assessment	*id_assessment*	The assessment identification number.
	assessment_type	The type of assessment.
	date	The cut-off day of the assessment.
	weight	The weight of the assessment. The weight of the exam is equal to 100%; the sum of all other assessments is also 100%.
Student_Assessment	*date_submitted*	The day of assessment submission.
	is_banked	The status of an assessment, who has been transferred from a previous Module-presentation
	score	The learner's score in the assessment.
Course	*code_module*	The module identification number.
	code_presentation	The presentation identification number.
	length	The duration of the course in days from module start date to module end date.
Student_Registration	*date_registration*	The day of learner's registration for the module-presentation.
	Date_unregistration	The day of learner withdrawns from the module-presentation.

- Collaboration activities includes Forum interaction (*Forumng),* Wiki (*Ouwiki),* Collaborative tasks (*Oucollaborate)* and Elluminate tasks (*Ouelluminate).*
- Course structure activities includes browsing the type of activities like *glossary*, *homepage, dataplus*, etc.

Figure 2. A withdrawal predictive model for OULAD



- Course content activities includes browsing the type of activities like *Resource*, *Url*, Content (*Oucontent*), *Page, Subpage*, etc.
- Evaluation activities include *quizzes*, *questionnaires* and *external quizzes*.

In fact, each learner is attached to their interactions (total number of clicks) respecting the four categories.

Subsequently, we proceed to compute the learners' scores per assessment. The original dataset explicitly includes both the score and the weight of each assessment. Therefore, we compute the weighted average score for each learners.

Finally, we obtain a dataset composed of 27 features and 1303 instances. In the remaining of this paper, we will use the terms features, attributes and descriptors for designating the same entities. We also use the term instances for designating the records (i.e. the learners in the dataset).

4.2. Rules-Based Data Discretization

Performing a discretization of some numerical values is necessary to enlarge the comprehensibility and interpretation. Discretization divides the numerical data into categorical classes that are more user-friendly for the instructor than precise magnitudes and ranges (Romero, et al., 2008). The process of transforming the continuous/numerical values into a finite set of intervals is called the discretization. It is useful in many datamining techniques because it reduces the impact of data noise. Even some techniques imperatively require discrete data in order to produce their output.

For OULAD dataset, the learner age value is divided with three interval (young age if the value is lower than 35 years; middle age if the value is higher or equal to 35 years and lower than 55 years; senior age if the value is higher or equal to 55 years).

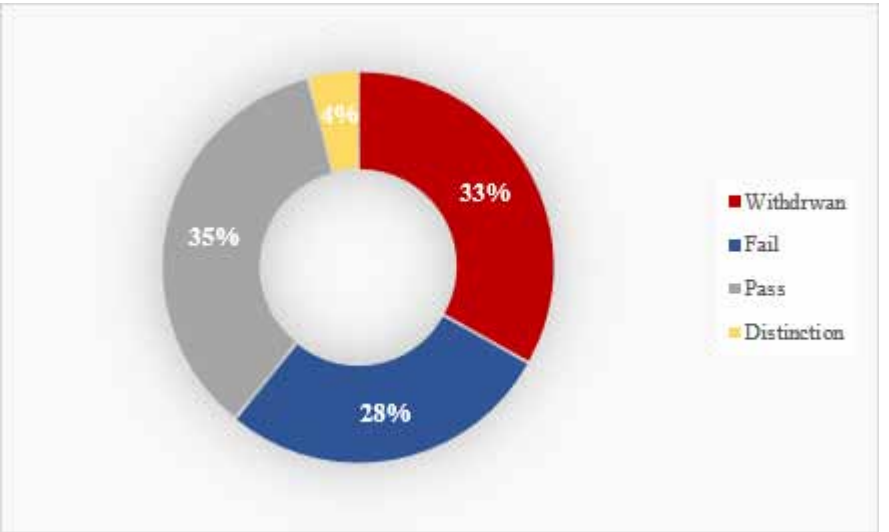
In this study, the target class is the final result (final_result attribute from the learner-info table) achieved by the learner. As mentioned previously, we deal with the dropout prediction as a binary classification problem. We divide the target class into two values: withdrawal and completion. A learners who completed the course, has "Distinction" or "Pass" or "Fail" as a final result value. Otherwise, learner who did not finish the course has "Withdrawn" as a final result value.

4.3. Data Balancing

As we mentioned previously, we resort to Machine Learning algorithms for the withdrawal prediction. Most of learning systems usually assume that training sets used for learning step are balanced. Many researchers like (Batista, et al., 2004) proved that this is not always the case in real world data where one class represents a circumscribed concept, while the other class represents the counterpart of that concept. This is known as the class unbalance problem and is often reported as an obstacle to the induction of good classifiers by Machine Learning models (Batista, et al., 2004). When a model is trained on an unbalanced dataset, it tends to reveal a strong bias to the majority class, since classic Machine Learning algorithms intend to maximize the overall prediction accuracy. Inductive classifiers are designed to minimize errors over the training instances, while Machine Learning algorithms, can neglect classes containing few instances (i.e. Minority class) (Bekkar & Alitouche, 2013). It convenes to note that, in our selection of OULAD, the learners' success rate is equal to 37%. Those percentages are highlighted in Figure 3. As we can see, the data labels are very unbalanced. To obtain a generalized predictive model, the used data should present approximately equivalent distribution of the class labels. Otherwise, the obtained model misleads the predictions.

In recent years, there have been several attempts at dealing with the class unbalance problem in the field of Data Mining and Knowledge Discovery in Databases, to which Machine Learning is a substantial contributor. In the context of our study, we used the Synthetic Minority Over-sampling method (SMOTE), which is an advanced method of over-sampling, proposed by (Chawla, et al., 2002). It is available in Weka[3]environment as a supervised data filter. The principle of this method is to generate new observations in the minority class by interpolating the existing ones (Bekkar & Alitouche, 2013). Recent studies like (Fernandez, et al., 2018) and (Dimic, et al., 2019) proved that the simplicity of SMOTE algorithm and the ability to apply it to different problems make this method a standard in the process of learning from unbalanced data.

Figure 3. Distribution of learners' result in the course "DDD2013B"



5. PROPOSED INDICATORS

As we mentioned previously that, it is necessary to get through more transformation to extract a meta-knowledge (i.e. indicator) of the observations traces. The semantic aspect behind the indicators has an important impact on the extracted knowledge' quality from the raw data. The withdrawal predictive model using such type of information becomes more understandable for the educational institution. Throughout our study, we have classified the indicators into three class: demographic, behavioural and performance indicators.

5.1. Demographic Indicators

As the name implies the demographic indicators explicit information about the learners' social / personal background. Undeniably, they have an important role on the learning experience in the context of VLE (Qiu, et al., 2016). In our context, we consider the following features:

- *Gender*: the learner's gender;
- *Age*: the learner's age, which is divided into three classes: "young", "middle" and "senior";
- *Highest education level:* his learner education degree before registering for the course. It is divided into four level: "A Level or Equivalent"," HE Qualification"," Lower Than A Level", "No Formal quals" and "Post Graduate Qualification";
- *Region*: the geographic region, where the learner lived;
- *Number of previous attempts*: the number of how many times the learner has attempted the course;
- *Disability*: the disability, whether the learner has declared.

5.2. Behavioural Indicators

In this section, we identify the semantical issues of raw data by introducing a meta-model of the learners' traces observations. We introduce four behavioural indicators, such as the perseverance, the autonomy, the commitment, and the motivation.

5.2.1 Autonomy Indicator

Many researchers have tried to define the learner autonomy, resulting in inter-related definitions. The learning autonomy is defined in (Tran & Duong, 2018) as "a self-management involving decision-making abilities that a learner needs to possess". That is, autonomy can be understood as the learner capacity to control their own learning and manage it in a self-reliant way by creating a learning plan, by finding resources that support study and by self-evaluating" (Fotiadou, et al., 2017). By analogy to this context, we refer to the learner autonomy by the navigation frequency within the VLE; it is calculated by the number of visited learning materials, as follows:

$$Autonomy_{L} = \sum_{i=1}^{k} nbConsultation\_L\left(site_{k}\right)$$

with:

- k: the number of pages visited by the learner L, during the learning course.
- nbConsultations_L ($site_k$): the consultations' number of a given learning material ($site_k$) made by the learner L.

5.2.2. Perseverance Indicator

Researchers have used various terms to define the perseverance including "retention", "persistence", and "retention". Those terms and their definitions are various. According to (Farrington, et al., 2012), the perseverance denote the behaviour of being engaged, focused and persistent in pursuit of learning goals within the MOOC setting. In the work done by (Ammor, et al., 2013), the perseverance is defined by the time taken on studying during the learning session. Other scholars defined the concept as the persistence with the various activities inside the course that ultimately results in achievement of a larger goal (Warriem, et al., 2016). From these definitions, we combined the temporal engagement aspect with a specific type of learning activities for describing the learner's perseverance. It is calculated by the number of assessments properly submitted by the learner before the cut-off day of submission:

$$Perseverance_{L} = \sum_{i=1}^{n} submittedAssessnment \Big/ total\ of\ Assessments$$

with:

- n: the number of assessments properly submitted by the learner L before the cut-off day of submission.
- *Total of Assessments*: the total number of assessments belongs the course. In our case, this course provides seven tutor-marked assessments, six computer-marked assessments and one exam.

5.2.3. Commitment Indicator

Commitment (a.k.a Engagement) refers to the learner's psychological investment, his willingness, and his involvement in the educational activities (York, et al., 2015). This concept cannot be easily determined, but rather it can be inferred by interpreting behavioral patterns, which indicate learner's level and type of involvement (Ramesh, et al., 2013). In (Beer, et al., 2010), the authors described the learning commitment as a blend of a distinct elements' number including active learning, collaborative learning, formative communication with academic staff, participation in challenging academic activities and involvement in enriching educational experiences. We follow this intuition, we distinguish between different types of commitment, and how they relate to different activities patterns. We divided the commitment aspect into four categories according to our previously categorized learning activities (See. Section 3.a). Therefore, we distinguish:

- *Collaborative commitment:* this concept is defined as an attitude that expresses the willingness of a transmitter to share his/her own knowledge with the other actors in the VLE (Bouzayane & Saad, 2017). It include the voluntary participation in the discussion forums and the collaborative platforms.
- *Course structure commitment:* many researchers proved that the course structure has a big impact to estimate the learner engagement within the VLE. For instance, Alshabandar et al. (Alshabandar, et al., 2018) pointed out that the learner who access to the home page of a course, in current weeks they will continue to interact with the course in the next week. Otherwise, if the learner fail to frequently access within the home page of the course, the probability of the learner dropout is increased (Wang & Chen, 2016).
- *Course content commitment:* this indicator refer to the learner engagement with the learning content delivered which can be a lectures notes, books, lecture slides and other courses materials delivered via hypertext markup language (HTML), PDFs, and lecture format.

- *Evaluation activities commitment:* Many researchers proved the benefits of tests, quizzes and assessments to both educators and learners. Quizzes are meant to track, report, and evaluate learning progress and outcomes.

Each commitment indicator is calculated by the sum of clicks (interaction) made by the learner in each site according to the type of commitment activity:

$$Commitment_{L-C} = \sum_{i=1}^{n}\sum_{j=1}^{m} sumclick(site_{i,j})$$

with:

- *n*: total number of learning activity types belonging a given commitment category C; For example, the collaborative category has exactly four activity type: *Forumng, Ouwiki, Oucollaborate* and *Ouelluminate.*
- *m*: total number of sites per learning activity type *i*;
- *sum_click(sitei,j)*: sum of clicks made by the learner *L* (attribute *sum_clicks* from table *Student_VLE*) on the site *j,* which is belonging to a learning activity type *i* and a given engagement category C.

5.2.4. Motivation Indicator

The detection of unmotivated learners during the course could bring the teacher to choose the right moment and tools for a constructive intervention, e.g. through creating weekly exercises, or initialize further investigations on the reasons of low activity (Dyckhoff, et al., 2012). Thus, in order to identify the motivation indicator, we calculate for each learner, the sum of all his or her interactions (sum-clicks) made in the VLE (sitej):

$$Motivation_{L} = \sum_{j=1}^{m} sumclick\left(site_{j}\right)$$

with:

- *m*: total number of sites visited by the learner L;
- *sum_click(sitej)*: sum of clicks made by the learner *L* (attribute *sum_clicks* from table *Student_VLE*) on the site *j.*

A learner is defined to be "motivated", if the sum of all their interactions is higher or equal than the average of the learners' interactions. Otherwise, the learner is considered as "unmotivated".

5.3. Performance Indicator

It is calculated by the sum of the assessments' scores obtained by the learner L. At this level, it is convenient to distinguish between the evaluation activities commitment and the performance indicator. The former tracks the learner's browsing actions, while the latter tracks the learner's score on the course's scored activities.

In our case, this course provides seven tutor-marked assessments, six computer-marked assessments and one exam. The number of assessments and their weights vary from module to module:

$$Performance_{L} = \sum_{i=1}^{p} weight_{i} \times score$$

with:

- p: total number of assessments
- *Weight i:* the weight of the assessment *i*

6. K-MEANS-BASED DATA DISCRETIZING

Firstly, we compute the values of the indicators according to the previously stated formulas. Indeed, the obtained values are numeric and they are defined over different intervals (e.g. the perseverance indicator ranges from 0 to 0.725 and the autonomy indicator ranges from 0 to 2683). A numerical indicator's value remains meaningless according to the instructors. Accordingly, we proceed to the discretization step. It means transforming a range of numeric (quantitative) values into a finite set of non-numeric (qualitative) values.

Several discretization methods exist in the literature. Some solutions define thresholds (e.g. the average) to cut the feature's interval of values into a finite number of subintervals (bins). These methods usually aim at obtaining subintervals with the same width or subintervals with the same count of instances (i.e. learners). Other solutions use the clustering methods (e.g. hierarchical algorithm) for grouping the extremely similar values into the same label. Although these methods are more intelligent, they depend on the number of possible subintervals.

Therefore, we proceed to discretizing all the numeric values of the indicators by adopting K-means (MacQueen et al., 1967). The k-means clustering method remains one of the most popular clustering method and the most influential data mining algorithms in the research community. This method, which are unsupervised and static, have been widely used in the literature for the discretising process: See for example (Tahir, et al., 2016) and (Hmamouche, et al., 2015). This method is available in Weka environment.

The K-means-based data discretizing process (See. Figure 4) follows these three steps:

1. Apply K-means on the indicator's corresponding features, in the dataset. Here, set the number of clusters to the value 2 (bins = 2);
2. Compute the "within" score between the obtained subintervals as the average of the variance between each subinterval's members;
3. Repeat steps 1 and 2 for different bins (from 2 to 10).

Normally, the lower is the within score, the better is the number of bins. However, choosing a very high number of bins may negatively affect the prediction process; this is technically called the "overfitting". Consequently, we opt for smaller values in order to ensure a generalized predictive model. To ensure a minimized loss of information through the discretization, we use the elbow method in order to produce the optimal number of bins for each indicator. We present in Figure 4, the variation of the within score of all the indicators throughout different counts of bins.

Considering the curves that correspond to Autonomy indicator, Evaluation activities commitment indicator and Performance indicator, we remark that the within score decrease is very dramatic from bins=2 to bins=3. Contrariwise, the decrease of the within score between bins=3 and bins=4 is relatively small. Therefore, we chose 3 as the input of k-means algorithm for these indicators. As for Perseverance and Collaborative indicators, we remark that their corresponding curves' elbows are present in two values (3 and 5 and respectively 3 and 6). In this case, we solve this issue by

Figure 4. The variation of the within score of all the indicators throughout different counts of bins

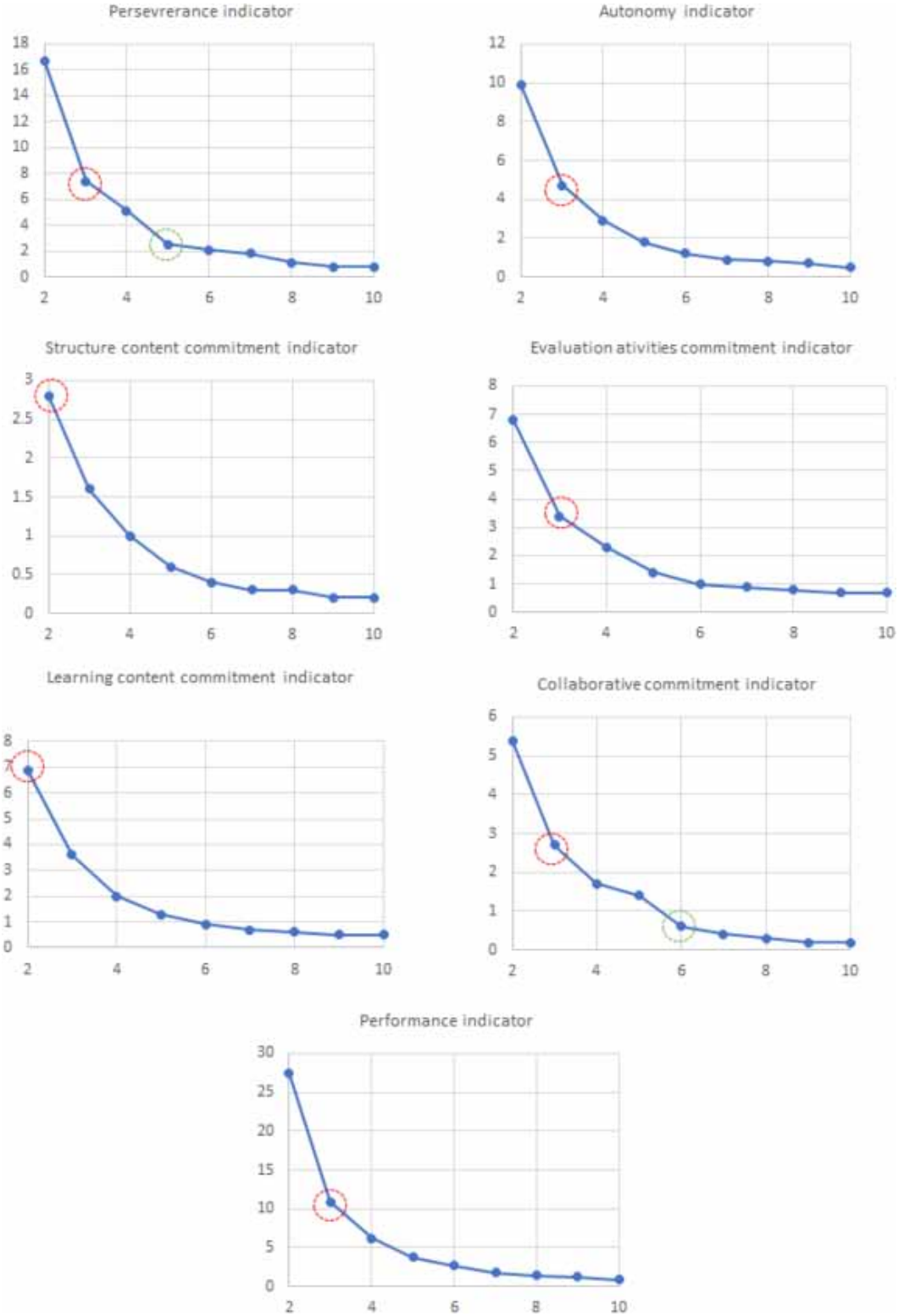


Table 2. The features' categorization

Category	Features
Demography	Age; Region; Highest education level; Gender; Number of previous attempts; disability.
Behaviour	Autonomy; Perseverance; Collaborative commitment; Course structure commitment; Course content commitment; Evaluation activities commitment; Motivation.
Performance	Performance
Class	withdrawal /completion

adopting our main assumption: select the simplest representation. Hence, we choose bins=3 for both indicators. For the curves that correspond to Structure content commitment indicator and Learning content commitment indicator, we remark that large numbers of bins do not obviously enhance the within score. For these cases, we opt for the smallest count of subintervals (bins = 2). Following this demarche, we call the obtained bins' sets {low, medium, high} for 3-binned indicators and {low, high} for 2-binned indicators.

7. PREDICTIVE ANALYSIS

Once the pre-processing, indicators' computing and preparing steps were performed, we aim at finding the classifier that ensures the best withdrawal prediction model. The obtained dataset is composed by 15 features (see Table 2) and 1303 rows. These rows divide the learners into two classes (withdrawal and completion).

For mining OULAD dataset, we perform a set of experiments while using different predictive models, accordingly different algorithms. We used Weka for creating the following models:

- Decision tree: we applied J48 (Quinlan, 1993) and Random forest (Breiman, 2001) for learning two different decision tree models.
- Support Vector Machine (SVM) (Suykens & Vandewalle, 1999): we set the cost to 1, gamma to 0 and loss to 0.1.
- Artificial Neural Network: we used the Multilayer Perceptron (Ruck, et al., 1990) with an automatic set of hidden layers. We set the learning rate to 0.3 and the momentum to 0.2.
- Bayesian classifier: we used the Tree Augmented Naive Bayes (TAN) algorithm combined with K2 score (Friedman, et al., 1997).

As the predictive models are obtained by the supervised algorithms (i.e. supervised machine learning), we propose to assess their reliability by adopting a 5-folded cross validation routine. Firstly, the dataset is equally partitioned into five equal or nearly equal folds. On these partitioned folds, training and testing phases is done in five iterations such that in each iteration, we leave one fold for testing phase and train the model on the remaining four folds. The accuracy obtained in each iteration is then averaged to get the model accuracy. An important thing to note is that dataset is commonly stratified before being split into five folds. Stratification is the process of rearranging data in such a way that each fold is a good representative of the whole (Yadav & Shukla, 2016).

Additionally, the predictive rate does not give a faithful assessment of the predictive model. Thus, we propose to use the F-measure for evaluating the obtained models. The F-measure is defined as the weighted harmonic mean of the Precision and Recall. It is computed as follows:

$$\text{F-measure} = 2 \times \frac{Recall \times Precsion}{Recall + Precision}$$

where:

$$Recall = \frac{TP}{TP + FN}$$

and:

$$Precision = \frac{TP}{TP + FP}$$

where:

TP (True Positives) = the number of learners with label completion who were predicted as the label completion

FP (False Positives) = the number of learners with label withdrawal who were predicted as the label completion

FN (False Negatives) = the number of learners with label completion who were predicted as the label withdrawal

We enhance the experiments in order to spotify the advantages of our methodological choices (i.e. the data balancing impact, the data discretization impact, and the behavioural indicators' extraction motivation). To assess its usefulness, we propose to conduct a comparative study. We choose the original pre-processed dataset (dataset without indicators: 26 features), the non-discretized dataset (dataset with continuous values with indicators: 15 features) as the baseline data and the discretized dataset (dataset with discrete indicators: 15 features). We report, in Table 3, the obtained F-measure of all the predictive algorithms applied on the three above-mentioned datasets. Note that we considered the cases of balanced and unbalanced data.

Firstly, we observe that the data balancing has a remarkable role on enhancing the predictive performance of the models. For the balanced data sets' experimentations, the F-measure ranges from 0.655 to as high as 0.853 (see the fourth column of Table 3). However, all classifier models based on unbalanced data obtain the lowest F-measure values, ranging over values 0.595 to 0.796 (see the first column of Table 3). Therefore, the data-balancing step shows a crucial role for guiding all the predictive models towards a reliable module. The data balancing was remarkable especially for the dataset with discrete indicators, where the classifiers achieve the highest F-measure, yielding values of 0.836 and 0.853, respectively.

Secondly, we remark that the introduction of a meta-model of the learners' features enhance the performance of all the considered classifiers. For the experimentations made on the unbalanced datasets, the results show that the J48 achieve an average score F-measure 0.796, whereas the Random Forest, MLP Classifier, Bayesian Classifier, and SVM Classifier achieve 0.792, 0.762, 0.785, and 0.595, respectively (see the second column of Table 3). Although the Random Forest Classifier picks the highest F-measure for the balanced dataset, with average value of 0.842, (see the fifth column of Table 3). Nevertheless, the impact of the datasets' balancing and behavioural indicators' extraction processes', we note that the SVM Classifier remains the worst result. Therefore, we conclude that the use of the raw features misleads the prediction. Not only a higher number of attributes complicates

Table 3. F-measure of compared predictive models

ML Algorithms	Dataset					
	Unbalanced Dataset			Balanced Data		
	Dataset Without Indicators	Dataset With Numeric Indicators	Dataset With Discrete Indicators	Dataset Without Indicators	Dataset With Numeric Indicators	Dataset With Discrete Indicators
Decision tree (J48)	0.794	0.796	0.789	0.823	0.836	0.844
Random Forest	0.775	0.792	0.754	0.824	0.842	0.853
Bayesian classifier (TAN)	0.798	0.785	0.791	0.829	0.829	0.836
SVM classifier	0.595	0.595	0.796	0.659	0.655	0.841
MLP classifier	0.762	0.786	0.755	0.801	0.822	0.836

the model creation, but also decreases the performance. The indicators' extraction are important step on pre-processing. On the one hand, they summarize the original features into smaller, yet representative, set of attribute. On the other hand, they provide an overview on the impact of each learner's behavioural aspect on the withdrawal/completion.

Thirdly, we observe that the discretization has not remarkably enhance the F-measure for all the classifiers (see the last two columns of Table 3). Nevertheless, all the predictive models obtain a viable performance values, ranging over values 0.836 to 0.853. We note that this procedure has a significant impact especially for the SVM classifier, yielding a value of 0.841. Conversely, the Bayesian and MLP classifiers with a value of 0.836 get the lowest range of F-measure. Although the discretization slightly enhances the withdrawal predictability, it eases the interpretation of the obtained predictive model.

To sum up, each experimentation we made on OULAD datasets turned out to be very positive. We conclude that the decision trees are the best models for predicting the learners' completion. The empirical results show that the highest F-Measure is acquired by the Random Forest followed by J48, with values of 0.853, 0.844 respectively. These predictive models are known for their user-friendly graphical representation; the nodes contain tests about the features of the dataset, and the terminal nodes reflect the decision outcomes (completion and withdrawal).

8. CONCLUSION AND PERSPECTIVES

An early prediction of learners at-risk does not only support learners, instructors, course's designers and pedagogues, but also researchers and developers in design interventions to hearten course completion before a learner falls too far behind. This paper described a Data Mining framework for withdrawal prediction of at-risk learners and illustrated the effectiveness of the method by applying the methodology to the Open University Learning Analytic Dataset. We started by formatting the data in a suitable form for the mining step. Secondly, we identified the semantical issues of raw data by introducing a meta-model of the learners' traces observations. We introduced four behavioural indicators, such as the perseverance, the autonomy, the commitment, and the motivation. Moreover, we considered the demographic and the performance features, which cover all the learners' individual differences. Then, we employed the SMOTE algorithm to tackle the unbalanced data issue. Additionally,

we proceeded for the discretization process for all the numeric values by applying a set of rules or by adopting the K-means algorithm. Finally, we have applied several machine-learning algorithms for the withdrawal prediction model, such as Decision trees, Random forest, SVM, Bayesian classifier and MLP classifier. Experimental results showed that the data balancing and discretization processes had been greatly improved the withdrawal predictability task, but especially the decision tree exhibit better predictive performance in terms of the F-measure value compared to the other tested models.

While the results in this paper are promising and interesting, there are still some important future directions from an education research perspective. Firstly, in order to generalize on the result obtained, more executions must backed up by using more data from other online learning courses from a variety of academic disciplines. Extensions are also possible with other types of Machine Learning techniques, which can be examined individually or in combined form, where for combined form, different level of fusions (such as classifier, feature, or decision) can be applied. Furthermore, we plan to tackle the features' selection challenge for compressing the data processing scale, where the redundant and irrelevant features will be removed. The feature selection technique can pre-process Machine Learning algorithms, and good feature selection results can improve learning accuracy, reduce learning time, and simplify the complexity of the predictive models. Further research can be conducted by the development of a real-time learning analytics dashboard, which helps the course instructors to acquire up-to-date predictions about learners' commitment and to make accurate decisions about struggling learners'. On the other hand, this real-time feedback allows learners to study more strategically as they can easily visualize their progress and their knowledge gaps and know where they should spend their time to improve their content mastery.

REFERENCES

Adda, S. A., Bousbia, N., & Balla, A. (2016). A Semantic Analysis of the Learner's Disorientation. *International Journal of Emerging Technologies in Learning*, *11*(06), 10–18. doi:10.3991/ijet.v11i06.5234

Almatrafi, O., & Johri, A. (2018). Systematic Review of Discussion Forums in Massive Open Online Courses. *IEEE Transactions on Learning Technologies*, *7*, 124809–124827.

Alshabandar, R., Hussain, A., Keight, R., Laws, A., & Baker, T. (2018). *The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses*. Congress on Evolutionary Computation. doi:10.1109/CEC.2018.8477770

Ammor, F.-Z., Bouzidi, D., & Elomri, A. (2013). Construction of deduction system of learning profile from performance indicators. *International Journal of Information and Education Technology (IJIET)*, *3*(2), 129–134. doi:10.7763/IJIET.2013.V3.249

Barbu, M., Vilanova, R., Lopez Vicario, J., Pereira, M. J., Alves, P., Podpdora, M., Prada Medrano, M., Torrebruno, A., Marin, S., & Tocu, R. (2017). *Data mining tool for academic data exploitation: literature review and first architecture proposal*. Projecto SPEET-Student Profile for Enhancing Engineering Tutoring.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, *6*(1), 20–29. doi:10.1145/1007730.1007735

Beer, C., Clark, K., & Jones, D. (2010). Indicators of engagement. Curriculum, technology & transformation for an unknown future. Proceedings ascilite, 75-86.

Bekkar, M., & Alitouche, T. A., (2013). Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process, 3*(4), 15.

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, *23*(2), 957–984. doi:10.1007/s10639-017-9645-7

Bousbia, N., Gheffar, A., & Balla, A. 2013. Adaptation based on navigation type and learning style. In *International Conference on Web-Based Learning*, (pp. 23-31). Academic Press.

Bouzayane, S., & Saad, I. (2017). A preference ordered classification to leader learners identification in a MOOC. *Journal of Decision Systems*, *26*(2), 189–202.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cao, L., & Zhang, C. (2015). *KDD Cup 2015—Predicting dropouts in MOOC*. Academic Press.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. *IEEE Global Engineering Education Conference*, 1007-1014. doi:10.1109/EDUCON.2018.8363340

Dimic, G., Rancic, D., Macek, N., Spalevic, P., & Drasute, V. (2019). Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique. *Information Discovery and Delivery*, *47*(2), 76–83.

Djouad, T., & Mille, A. (2018). Observing and Understanding an On-Line Learning Activity: A Model-Based Approach for Activity Indicator Engineering. Technology. *Knowledge and Learning*, *23*(1), 41–64.

Dyckhoff, A. L., Zielke, D., Bultmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning analytics toolkit for teachers. *Educational Technology & Society, 15*(3), 58-76.

El Haddioui, I., & Khaldi, M. (2017). Study of learner behavior and learning styles on the adaptive learning management system manhali: Results and analysis according to gender and academic performance. *Journal of Software*, *12*(4), 212–227. doi:10.17706/jsw.12.4.212-226

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance—A Critical Literature Review*. ERIC.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. doi:10.1613/jair.1.11192

Friedman, N. G., Fotiadou, A., Angelaki, C., & Mavroidis, I. (2017). Learner Autonomy as a Factor of the Learning Process in Distance Education. *European Journal of Open, Distance and E-learning*, *20*(1), 96–111. doi:10.1515/eurodl-2017-0006

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*(2-3), 131–163. doi:10.1023/A:1007465528199

Gamie, E. A., El-Seoud, M., & Salama, M. A. (2019). A layered-analysis of the features in higher education data set. *Proceedings of the 8th International Conference on Software and Information Engineering*, 237-242. doi:10.1145/3328833.3328850

Haiyang, L., Wang, Z., Benachour, P., & Tubman, P. (2018). A Time Series Classification Method for Behaviour-Based Dropout Prediction. *2018 IEEE 18th International Conference on Advanced Learning Technologies*, 191-195.

He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Heuer, H., & Breiter, A. (2018). Student Success Prediction and the Trade-Off between Big Data and Data Minimization. In DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik. Bonn: Gesellschaft für Informatik e.V.

Hmamouche, Y., Ernst, C., & Casali, A. (2015). Automatic KDD Data Preparation Using Multi-criteria Features. *The Fifth International Conference on Advances in Information Mining and Management*.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational Intelligence and Neuroscience*, *2018*, 21. doi:10.1155/2018/6347186 PMID:30369946

Jiang, S., & Kotzias, D. (2016). *Assessing the Use of Social Media in Massive Open Online Courses*. arXiv preprint arXiv:1608.05668

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(1). Advance online publication. doi:10.19173/irrodl.v15i1.1651

Kardan, S., & Conati, C. (2013). Comparing and combining eye gaze and interface actions for determining user learning with an interactive simulation. *International Conference on User Modeling, Adaptation, and Personalization*, 215-227. doi:10.1007/978-3-642-38844-6_18

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*(1), 170–171. doi:10.1038/sdata.2017.171 PMID:29182599

Kuzilek, J., Vaclavek, J., Fuglik, V., & Zdrahal, Z. (2018). Student Drop-out Modelling Using Virtual Learning Environment Behaviour Data. *European Conference on Technology Enhanced Learning*, 166-171. doi:10.1007/978-3-319-98572-5_13

Li, W., Brooks, C., & Schaub, F. (2019). The Impact of Student Opt-Out on Educational Predictive Models. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 411-420. doi:10.1145/3303772.3303809

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., & Wu, Z. (2016). Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. 2016 international joint conference on neural networks, 3130-3137. doi:10.1109/IJCNN.2016.7727598

Liang, J., Li, C., & Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. *Computer Science & Education ICCSE, 2016 11th International Conference on*, 52-57.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281-297.

Maldonado-Mahauad, J., Pérez-Sanagustin, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Delgado-Kloos, C. (2018). Predicting Learners' Success in a Self-paced MOOC Through Sequence Patterns of Self-regulated Learning. *European Conference on Technology Enhanced Learning*, 355-369. doi:10.1007/978-3-319-98572-5_27

May, M., Iksal, S., & Usener, C. A. (2016). The Side Effect of Learning Analytics: An Empirical Study on ELearning Technologies and User Privacy. *International Conference on Computer Supported Education*, 279-295.

Papanikolaou, K. A. (2015). Constructing interpretative views of learners' interaction behavior in an open learner model. *IEEE Transactions on Learning Technologies*, *8*(2), 201–214. doi:10.1109/TLT.2014.2363663

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 505–513.

Popescu, E. (2009). Learning styles and behavioral differences in web-based learning settings. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, 446-450. doi:10.1109/ICALT.2009.156

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in MOOCs. *Proceedings of the ninth ACM international conference on web search and data mining*, 93-102. doi:10.1145/2835776.2835842

Quinlan, J. R. (1993). *C 4.5: Programs for machine learning. The Morgan Kaufmann Series in Machine Learning*. Morgan Kaufmann.

Ramesh, A., Goldwasser, D., Huang, B., Hal Daumé, I. I., & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS Workshop on Data Driven Education*, 21, 62.

Ramesh, A., Kumar, S. H., Foulds, J., & Getoor, L. (2015). Weakly supervised models of aspect-sentiment for online course discussion forums. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 74-83. doi:10.3115/v1/P15-1008

Reich, J. (2014). *MOOC completion and retention in the context of student intent*. EDUCAUSE Review Online.

Romero, C., Lopez, M.-I., Luna, J.-M., & Ventura, S., (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458-472.

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, *1*(4), 296–298. doi:10.1109/72.80266 PMID:18282850

Shareghi Najar, A., Mitrovic, A., & Neshatian, K. (2015). Eye tracking and studying examples: How novices and advanced learners study SQL examples. *CIT. Journal of Computing and Information Technology*, *23*(2), 171–190. doi:10.2498/cit.1002627

Shridharan, M., Willingham, A., Spencer, J., Yang, T. Y., & Brinton, C. (2018). Predictive learning analytics for video-watching behavior in MOOCs. *52nd Annual Conference on Information Sciences and Systems*, 1-6. doi:10.1109/CISS.2018.8362323

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300. doi:10.1023/A:1018628609742

Tahir, H. M., Said, A. M., Osman, N. H., Zakaria, N. H., Sabri, P. N. A. M., & Katuk, N. (2016). Oving Kmeans clustering using discretization technique in network intrusion detection system. *3rd International Conference on Computer and Information Sciences*, 248-252.

Tan, M., & Shao, P. (2015). Prediction of student dropout in e-Learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning*, *10*(1), 11–17. doi:10.3991/ijet.v10i1.4189

Thai-Nghe, N., Busche, A., & Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. *Ninth International Conference on Intelligent Systems Design and Applications*, 878-883. doi:10.1109/ISDA.2009.15

Tran, T. Q., & Duong, T. M. (2018). EFL learners' perceptions of factors influencing learner autonomy development. *Kasetsart Journal of Social Sciences*. Advance online publication. doi:10.1016/j.kjss.2018.02.009

Wang, F., & Chen, L. (2016). A Nonlinear State Space Model for Identifying At-Risk Students in Open Online Courses. *EDM*, *16*, 527–532.

Wang, W., Yu, H., & Miao, C. (2017). Deep Model for Dropout Prediction in MOOCs. *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, 26-32. doi:10.1145/3126973.3126990

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). *Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains*. International Educational Data Mining Society.

Warriem, J. M., Murthy, S., & Iyer, S. (2016). Shifting the focus from learner completion to learner perseverance: Evidence from a teacher professional development MOOC. *Proceedings of the 24th International Conference on Computers in Education*.

Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Educational Data Mining*.

Wolff, A., Zdrahal, Z., Herrmannova, D., Kuzilek, J., & Hlosta, M. (2014). *Developing predictive models for early detection of at-risk students on distance learning modules*. Academic Press.

Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). An analysis of MOOC discussion forum interactions from the most active users. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 452-457. doi:10.1007/978-3-319-16268-3_58

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing*, 78-83.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and Measuring Academic Success. *Practical Assessment, Research & Evaluation, 20*.

Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., & Chen, H. (2017). Student performance prediction via online learning behavior analytics. *2017 International Symposium on Educational Technology*, 153-157. doi:10.1109/ISET.2017.43

ENDNOTES

[1] https://analyse.kmi.open.ac.uk/open_dataset
[2] https://theodi.org/
[3] https://www.weka.fr/

Fedia Hlioui is a PhD Student enrolled at the final year at Computer science at the faculty of economics and management at the Sfax University. She is also a member at the Multimedia InfoRmation system and Advanced Computing Laboratory at the Sfax University. She obtained the Master degree in 2013. She published four papers in international conferences. Her research interest focuses on different information systems' fields, such as, E-learning, Learning Analytics, Educational Data Mining research, machine learning.

Nadia Aloui is an Assistant Professor of computer sciences at the Higher Institute of Pratical Science and technologies. Nadia Aloui obtained her Master's degree in 2008 from the national school of computer science (ENSI) and Montreal university, and a PhD degree in 2014 from the University of Sfax-Tunisia. She started teaching since 2006 and she is a researcher at the MIRACL Laboratory at the Higher Institute of Computer Science and Multimedia of Sfax. She published many papers in journals and conferences as well as a chapter books in the research field of e-learning and ubiquitous learning. She is member of various international conferences. Her research interest focuses on different computer science fields, such as, mobile learning, context-aware learning, Ontology engineering and knowledge management.

Faïez Gargouri is Professor of computer sciences at the Higher Institute of computer science and multimedia at the University of Sfax, Tunisia (www.isimsf.rnu.tn), where he is the Head (since August 2014 and from 2007 to 2011). He is also the Director of the Multimedia, InfoRmation Systems and Advanced Computing Laboratory (www.miracl.rnu.tn) since 2016 (and also from 2011 to 2014). Faïez Gargouri obtained his Master's degree in Computer Science from the Paris 6 University (1990) and a PhD from the Paris 5 University (1995). In 2002, he obtained an Habilitation Universitaire en Informatique from the Faculté des Sciences de Tunis (Tunisia). His research interest focuses on different information systems' fields, such as, Design, Quality Measurement, Verification, Data Warehousing, Multimedia, Knowledge Management, Ontology. He published more than 300 papers in journals and conferences as well as books (pedagogical and conference proceedings). He is member of the Scientific and Steering committees of various international conferences and journals.