Computers & Education 131 (2019) 22–32

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Utilizing early engagement and machine learning to predict student outcomes

Cameron C. Gray*, Dave Perkins

School of Computer Science, Bangor University, Gwynedd, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Learning analytics
Student retention*MSC:*
68-U35
68-T10
97-B40

ABSTRACT

Finding a solution to the problem of student retention is an often-required task across Higher Education. Most often managers and academics alike rely on intuition and experience to identify the potential risk students and factors. This paper examines the literature surrounding current methods and measures in use in Learning Analytics. We find that while tools are available, they do not focus on earliest possible identification of struggling students. Our work defines a new descriptive statistic for student attendance and applies modern machine learning tools and techniques to create a predictive model. We demonstrate how students can be identified as early as week 3 (of the Fall semester) with approximately 97% accuracy. We, furthermore, situate this result within an appropriate pedagogical context to support its use as part of a more comprehensive student support mechanism.

1. Introduction

Student retention, the practice of avoiding situations where Higher Education (HE) students do not continue their studies to a successful outcome, is an area of active research and study (McCoy & Byrne, 2017; Paige, Wall, Marren, Dubenion, & Rockwell, 2017). Methods and responses to retention, and student engagement in general have been varied (Li & Carroll, 2017; Webb, Wyness, & Cotton, 2017). Most fit into two broad categories; *reactive* - where tutors and support staff respond to specific cases and causes, or *preventative* - action designed to target larger groups of students to highlight the successes and benefits they all enjoy. The general actions usually aim to raise overall student satisfaction, counteracting any local discontentment (both organizationally - within a department or course, and temporally - a particular point in time).

It has long been held that academic performance is not necessarily the only factor involved in retention problems (Lotkowski et al., Noeth). Reasons can include a complex psycho-social interplay of factors leading to intuitive responses from educators rather than decisions based on data (DeBerard, Spielmans, & Julka, 2004). As such, the results of these responses are varied. Those educators with experience in dealing with both academic and pastoral issues will, undoubtedly, out-perform their less experienced colleagues.

To meet retention (and others, such as pastoral care) goals, educators require the development of rapid decision models, processes, and support to both identify and mediate issues uncovered. Learning Analytics products and tools are progressing to help in this endeavor. However, these products either do not focus on earliest possible identification or require significant human interpretation as to the best course of action.

Our research has been focused on providing the earliest possible identification of students that would benefit from tutor intervention. We hypothesize that by applying machine learning techniques to data already collected about session attendance, we would be able to make such an identification.

* Corresponding author.
E-mail address: c.gray@bangor.ac.uk (C.C. Gray).

https://doi.org/10.1016/j.compedu.2018.12.006
Received 4 July 2018; Received in revised form 15 December 2018; Accepted 17 December 2018
Available online 21 December 2018

Within this paper we present four contributions;

1. evaluation of several candidate machine learning methods using derived metric measuring student attendance/engagement to produce predictions of student outcome.
2. a case study using a full academic year and the trained model at Bangor University, UK.
3. discussion of related issues, including student motivation and potential interventions tutors may wish to undertake.
4. a substantial view on the ethical and data protection issues as part of the overall discussion (see Section 5).

2. Related work

Attendance at timetabled sessions has been found to be a consistent predictor of likely student retention (Fike & Fike, 2008). The same study also correlated attendance and outcome of 'developmental' courses, which usually occur in the first year of a degree program. While the debate over a cohesive model including all plausible causes of student departure rages; there has been broad agreement that engagement with a course usually leads to higher achievement (Seidman, 2005; Shelton, 2003; Tinto, 2010).

Previous studies (Veenstra, 2009; Price, 1993; Martin) have shown that one of the most successful strategies for retaining students is an early intervention (with varying definitions of 'early'). Robbins, Oh, Le, and Button examined the link between types of intervention showing a possible 13% increase in retention when linked to an academic skill or attendance (Robbins, Oh, Le, & Button, 2009). Learning Analytics tools do include methods that can be used to support early interventions, but they are not designed for that purpose.

West et al. conducted a study linking learning analytics specifically to retention indicators and efforts (West et al., 2016). Their findings show that students self-reporting issues is the most common data source when provided with fixed categories. However, when given a free-form answer field, the majority of comments singled out 'class attendance' as the most offered answer. This result was also reported by Anderson, Whittington, and Li (Anderson, Whittington, & Li, 2016) confirming that attendance can be used as a strong indicator of a student's final grade.

Various Learning Analytic predictive models exist in both literature and commercial use. One class attempts to predict students as either passing or failing based on their educational resource usage, such as Virtual Learning Environments (VLEs), Libraries, and other support services (Arnold & Pistilli, 2012; Daniel, 2015; Nguyen, Rienties, & Toetenel, 2017). Usage patterns can vary drastically meaning these systems require time to build up a profile of each new class. There is usually a correlation between reduced engagement with support and teaching resources, but the analytics can only flag a potential issue with any given student. A weak correlation means that the 'usage' is only one factor in the model, and usage alone cannot be used to make a definitive prediction due to the variances observed.

The second class of products/tools and accompanying research uses marks/grades to predict the likelihood of a poor outcome or retention issues (Herodotou et al., 2017; Rovira, Puertas, & Igual, 2017). These models have a high success rate in flagging poor outcomes as they are directly descended from the constituent components of that outcome. There is one drawback to using grades as the predictor variables; at the point grades are awarded, the student cannot do anything to influence them. This deficiency may be acceptable where there are multiple or formative assessments but fails in courses with single and/or major summative assessments.

During the late 2000s and early 2010s, a third meta-class became popular where the previous two styles of analytics were combined with in-person interviews or oral/written depositions (Heaton-Shrestha, May, & Burke, 2009; Tempelaar, Rienties, & Nguyen, 2017). The findings in this work show that both teachers and students are looking for more insight into why any particular flagged event was significant, rather than mere reporting of statistics and lists. Within Learning Analytics this is known as moving from descriptive analytics to insight analytics (Picciano, 2012).

There are several predictive models developed and presented in the literature, discussed later in this section. These works broadly fall into one of two groups: identification of groups or sub-populations; and exploration of the efficacy of including other data-sets. Most of the work on learning analytics deals with either the US or Australian systems, although we also examine cases from UK HE institutions as they are particularly relevant.

There have been several studies into the statistical worth of 'performance indicators' and 'metrics' to describe students (Ball & Wilkinson, 1994; Draper & Gittoes, 2004; Richards, 2011). While the relative merit conclusions differ slightly, all three of these studies support the idea that metrics need to be individualized. The level of resolution could vary, from specific metrics for the institution based on the rationale of deploying analytics to different factors being considered for each student. This view is supported by Gašević et al. cautioning against a one-size-fits-all analytics approach (Gašević, Dawson, Rogers, & Gasevic, 2016).

Baker et al. studies which factors are most telling of success in online courses and how early these can be used (Baker, Lindrum, Lindrum, & Perkowski, 2015). As the courses are entirely computerized, the factors are entirely based on activity as a surrogate for engagement/attention. They report 65%–70% accuracy based on metrics obtained in the first week of the course. This level of accuracy is not sufficient to be considered predictive, which the study admits, but the authors do feel this is strong enough to support intervention with a student. Ye and Biswas conducted a similar study into Massive Online Open Course (MOOC) data (Ye & Biswas, 2014). Their study found that better prediction rates were found when using timely data, in this case a behaviour, within the previous week.

JISC, a UK not-for-profit technical and policy support organization for post-compulsory education, has conducted a thorough review of the state of learning analytics in HE (Sclater, Peasgood, & Mullan, 2017). As part of this report, the authors present 11 case studies. Each focuses on a different aspect of learning analytics. These include 'Early Alert' at the University of New England - where the system developed has the same mission as our work. Their approach, however, correlated emoticon and text input to a 'wellbeing status'. Where this dipped or remained low, students are gradually offered more and more support. This project caused a decrease in the attrition/dropout rate from 18% to 12% (Davis, 2015).

Nottingham Trent University created the 'NTU Student Dashboard' as a pilot project in deploying learning analytics. The system is considered predictive as it correlates low engagement with the high risk of a poor outcome. The model derives an engagement metric from multiple sources including VLE interaction, library usage, attendance checks and submissions of assessments. Interestingly, retention is not a particular concern in this institution, but they acknowledge that the dashboard will assist in this regard. The main goal is to foster and improve the relationship between the student, the institution and their tutor. Therefore the system focuses on a positive engagement metric, rather than the negative risk factor for withdrawal. This project found that 27% of first-year undergraduates had changed their behaviour based on the information presented.

Learning analytics is not to be seen as a 'silver bullet'. While analytics can assist educators, they cannot provide every answer. This is the same conclusion Pardo and an international team arrived at. Their study looked at the usefulness of predictive models in assisting educators. They found that provision of a learning analytics platform is not the catalyst some believed it would be. The educators required more support with identifying the groups or clusters within their cohorts to design appropriate interventions (Pardo et al., 2016).

Evaluation of these efforts, in real-world terms, is difficult. By the very nature of deploying analytics and educational interventions, the resulting data-set is altered. This means that producing a control or objective group becomes an ethical dilemma for the educators and researchers (Slade & Prinsloo, 2013). Scientific rigor demands an impartial evaluation. However, this would mean knowingly and deliberately withholding intervention from some students that may be in dire need of it. Gašević et al. remind us that while the technology and science are a fundamental necessity, the entire endeavor relates to improving learning experiences (Gašević, Dawson, & Siemens, 2015).

3. Applying machine learning to engagement

The primary goal of this work is to make early identification of students that would benefit from intervention possible. While some of the work undertaken could further the search, the authors are not intending to locate the 'best' predictors for retention nor engagement. This work is specifically looking for the earliest possible, accurate identification of students that may be at risk of low achievement and/or a retention problem.

Therefore, we conducted a set of feature selection experiments to ascertain at what point the attendance becomes a reliable predictor of students' academic outcome for the academic year. These experiments were conducted using the WEKA (Frank et al.) workbench, classifiers, and tools.

3.1. Data collection and structure

Bangor, as with other UK HE institutions, sets out an Engagement Monitoring policy. Under this policy, all students' attendance will be monitored through checks of their student ID in most, if not all, sessions. Staff are equipped with bar-code scanners which record the unique number of each ID card along with a time stamp. At the end of each monitored session, this record is uploaded to a central database. Each time stamp is then matched with the timetabled session in which it was taken. The database also allows for meetings with their assigned tutor to be recorded, along with other custom events.

As part of the student's registration, they agree to various processing of this data - including for analytics purposes. In Bangor's policy documents (and through various engagement activities), the purpose of these analytics is covered at length. The aim is to enhance the experience of every student, not to simply 'spy' or 'check up on them'. The students have responded positively to this approach and readily provide permission to use their data in this and other analytics systems.

Within the database, at any point in time, there would be k observations for each student. Each observation is coded as either one for an attendance and zero otherwise. These observations are recorded in the set $z = \langle z_1, z_2, \ldots, z_k \rangle$. Initially, a ratio of sessions attended was considered - termed Engagement Ratio or ER - defined mathematically as Equation (1).

$$ER_k = 1/k \sum_{j=1}^{k} z_j \tag{1}$$

Ultimately this proved to be less effective for prediction of the student outcome. See Section 3.6 for details of the experiment that lead to abandoning this metric.

The second metric, termed the Bangor Engagement Metric (BEM), combined both attendance and non-attendance into a single reading. In the absence of definitive guidance from past work, we crafted the metric independently and based on an idealized view of attendance. It is formed, for any point in time, as the number of sessions attended less the number of sessions missed. Using the same set of k observations in set z, the metric is formally defined as Equation (2).

$$BEM_k = \sum_{j=1}^{k} (-1)^{(1-z_j)} \tag{2}$$

As the structure and timetable for each program variant is different, it is not possible to compare the raw set z. To provide a common number of results, and frame of reference, the items in the set are summed by academic week. Each student now has a consistent 12 ER and BEM values for each complete semester. Examination attendance is also monitored using the same process, adding an additional six readings for a year. This totals 30 BEM readings for a complete academic year.

The data-set includes three biographical variables; the numeric codes for the student's program, the school, and the year of study. These elements are handled as categorical data in the data-set forcing a set of values. In addition, each student is assigned one of five Academic Standing codes signifying the completion state of that academic year/level. The Academic Standing code is the class label that the machine-learning model needs to match/predict. These codes are;

- PA - pass.
- FN - fail (cannot progress).
- FC - conditional fail (requires supplementary assessment).
- RY - repeat the year/level.
- RS - repeat a single semester.

3.2. Sources of (potential) external effects

While attendance is often a student choice there are several factors that are beyond their control. These range from the basic and somewhat expected, such as sickness, to more impacting work and family commitments. There may also be demographic considerations, such as religious events, that can also influence the student's attendance.

Micere's meta-study of US college entrants, their characteristics and demographics (Micere, 2013) shows a significant demographic shift in completion rates. However, the raw data on engagement shows only Black males having a significant drop in traditional involvement with their studies. A meta-analysis of over 21,000 US college students (Credé, Roch, & Kieszczynka, 2010), has shown a weak relationship between student characteristics and their attendance. In this case, characteristics means personality traits or qualities such as diligence rather than any demographic or social division. The student population can also be divided by study mode, full-time or part-time. Micere's study did examine this factor, and the impact of part-time employment and was rejected as a significant factor.

When a student is completing their studies part-time, their overall attendance value will be lower proportionate to the number of credits they are completing. Their resulting BEM value will be expected to be lower by the same proportion, allowing for class schedule differences. As a result of these findings, this study has chosen not to place any specific corrections or biases into the construction of the Bangor Engagement Metric. If this information were included in the model, it can be a bias that will cause specific over-fitting to the data. By not including these specifics, the algorithms (and resulting model) must generalise the patterns of attendance for all circumstances.

3.3. Initial data exploration

Our initial data-set contained the weekly values of all undergraduate students' BEM, program, and year of study for the 2016/17 academic year at Bangor University. These were the most recent historical and complete results. As a result, we can match predicted outcomes with the ground truth for each student. This set has; $N = 4970$ instances, $n = 32$ features and $C = 5$ classes.

The first exploration of this data found that while the skew and mean of each weekly set of metrics changed marginally, it remained a normal distribution. Fig. 1 shows the population using 20 bins calculated from the maximum range of the Engagement Metric.

The Heat-map also shows that as the academic year progresses, the range of engagement widens. This result confirms those found by other studies (Massingham & Herrington, 2006) and anecdotal evidence from colleagues.

Subsequent explorations examined time series plots of individual student data. These were split by year/level and department due to the large amount of over-plotting when visualizing the entire set. The observations made hold true for all divisions in the data-set. Fig. 2 shows the results for the 2016/17 freshman (first year) cohort for Computer Science.

The stand-out observation is that allowing for a certain amount of variance; students tend to remain on their starting trajectory throughout an academic year. Some of these students were already the beneficiaries of pastoral care. This observation would call into question the effectiveness or timeliness of staff efforts.

This figure also clearly shows the inevitable outliers in any cohort. The topmost line (green/solid) could be assumed to be the most diligent student who attends all sessions and will therefore pass. The bottom-most (again a greed/solid in this example/plot) would fit the profile of a naturally talented student that feels that they do not need to attend sessions but is also expected to pass.

3.4. Feature selection

Armed with previous findings, a new set of experiments was devised. They aimed to determine when these trajectories could be separated into the final academic standings. This is a feature selection task, a process where the metrics are evaluated for their worth in a classification problem. The process aims to keep metrics that do describe differences accurately, and therefore drop those that are irrelevant or redundant.

Floating Search methods have already been shown to provide superior results in less computational time. Therefore the Sequential Forward Selection (SFS) method was chosen (Kittler, 1986; Whitney, 1971). This method operates by selecting the strongest (usually defined as the highest accuracy) individual feature, then pairing this with all others to find the strongest pair, and so on.

The initial experiment utilized SFS and the Nearest Neighbor (1-NN) classifier, an approach followed by Kudo and Sklansky (Kudo

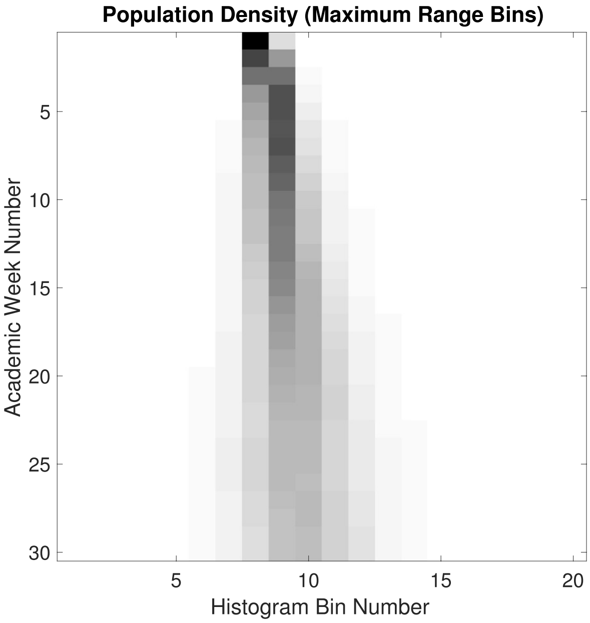

Fig. 1. Heat-map showing relative population density at each academic week. Each week has 20 bins calculated from the maximal range of the Bangor Engagement Metric. The darker a cell the more students occupy that position.

& Sklansky, 2000). Due to the PA/pass class drastically outnumbering other labels, the F-Measure statistic (Joshi, 2002) is used to measure performance. This experiment produces a ranked list of features that should be considered. The top four features (in order) were found to be School, Week 4 BEM, Week 5 BEM, and Week 3 BEM. A check of these results was made, utilizing the 1-NN classifier and the selected features. Table 1 shows the results of this experiment. The headline result is a LOOCV accuracy of 80.10%, and an weighted F-Measure value of 0.788. However, these results also show very low F-Measure values for the least represented classes.

Encouraged by this result, a further attempt increased the accuracy, by using a C4.5 Pruned Tree classifier (Quinlan, 1993) and the same protocol and features, to 84.85%. However, this choice actually weakened the predictive power for students that would need to complete supplementary work, with this class' F-Measure dropping to 0.023.

As the algorithmic results pointed toward early weeks from the fall semester, the next set of experiments tracked the accuracy when progressively adding weekly data. These experiments used the same C4.5/LOOCV combination; the results are shown in Table 2. The best overall accuracy achieved was 86.20%, which includes all of the weeks of the fall semester. This does not meet the stated goal of early identification of students. The second best result, 86.10%, or just 0.10%/five students less, required only the first three weeks' values without the school and year of study being included. Using this model, tutors would be in a position to potentially make appropriately targeted interventions from academic week 4. Each experiment was conducted both including and excluding two additional features; the students school/department and year of study. The rationale for year of study was to examine if cohortal effects were sufficiently large to influence the model. The school/department was included as the mode of study within each discipline is likely to be different. Without an extra discriminator, the model was in danger of being over-generalized to ignore these differences.

These experiments provided sufficient evidence that there is sufficient predictive power by using only the first three weeks BEM values to provide a suitable model. It also shows that including discriminator features, such as year, school, or program increases some F-Measures. A further experiment is needed to determine which of these three discriminators is most powerful. This will result in a selected feature set of:

- Week 1 BEM Value
- Week 2 BEM Value
- Week 3 BEM Value
- School, Program, or Year
- Academic Standing (Class Label)

3.5. Classifier selection

To this point, the success metric had been the overall classification accuracy, the number of students correctly matched to their actual outcomes. However, the authors quickly realized that accurate classification into all five classes, while ideal, is not strictly required. Identification of potentially at-risk students is not contingent on which failure mode they may achieve, just that it is not expected to be a pass. With this extra constraint, it would be reasonable to reduce the problem to a two-class classification problem

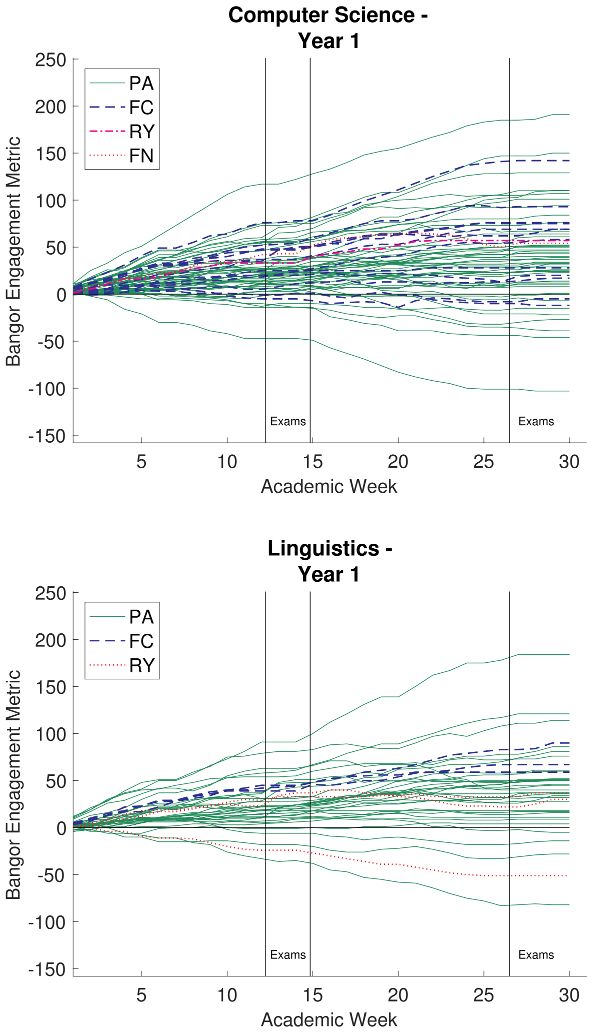

Fig. 2. Time-series plots showing the 2016/17 1st Year Cohort in Computer Science and Linguistics at Bangor University. Each line represents one student, and is colored by their final academic standing at the end of the year. Academic standings in the legend are; PA/pass, FN/fail, FC/supplementary work required, RY/repeat year, and RS/repeat semester. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Experimental results using the 1-NN classifier and the top four algorithmically selected features. The last row shows the weighted average, based on relative class size, for the entire classification problem. (TP = True Positive, FP = False Positive, Prec. = Precision, AUC = Area under ROC Curve).

Class	TP Rate	FP Rate	Prec.	Recall	F-Measure	AUC
Pass	0.914	0.764	0.878	0.914	0.896	0.628
Supplementary	0.118	0.062	0.166	0.118	0.138	0.565
Repeat Year	0.051	0.013	0.058	0.051	0.054	0.632
Fail	0.161	0.022	0.198	0.161	0.178	0.668
Repeat Semester	0	0	0	0	0	0.699

and re-categorize the student outcomes. However, it is still useful information for tutors to understand the severity of the potential outcome by mode.

Confusion Matrices are the tool for reporting summary output in classification tasks, a simple table showing true/actual labels as rows and predicted labels as the columns. Each cell contains a count of instances/objects that fall at that intersection. The perfect outcome would be correct counts on the main diagonal showing that the true and predicted labels match. All counts elsewhere in the matrix indicate a classification error. There are two separate regions; above the diagonal showing false negatives, Type I errors, and below showing false positives known as Type II errors. Table 3 shows an example confusion matrix from the C4.5 Feature Selection experiment.

Table 2
Results from testing successive week feature sets, using C4.5 Trees with Leave One Out Cross Validation. Rows are arranged by accuracy.

Weeks	School/Year Inc.	Accuracy %	Per-Class F-Measure				
			PA	FC	RY	FN	RS
1–12	Y	86.20	0.935	0.202	0.068	0.190	0.000
1–3	N	86.10	0.926	0.049	0.000	0.231	0.000
1–4	N	86.04	0.929	0.040	0.065	0.149	0.000
1–4	Y	85.77	0.929	0.151	0.019	0.178	0.000
1	N	85.75	0.923	0.000	0.000	0.000	0.000
1	Y	85.75	0.923	0.000	0.000	0.000	0.000
1, 2	Y	85.75	0.923	0.000	0.000	0.000	0.000
1–5	N	85.73	0.927	0.058	0.065	0.245	0.000
1–5	Y	85.71	0.929	0.141	0.075	0.211	0.000
1–6	Y	85.69	0.930	0.135	0.038	0.200	0.000
1–6	N	85.69	0.927	0.075	0.041	0.216	0.000
1–3	Y	85.65	0.928	0.139	0.000	0.110	0.000
1, 2	N	85.59	0.924	0.000	0.000	0.056	0.000
1–12	N	84.35	0.923	0.103	0.017	0.183	0.000

The shaded cells in the table show classification results where students are identified as having a positive outcome when in fact they would achieve a negative one. By re-framing the training of classifiers to minimize Type II errors, these inaccurate and misleading errors will be minimized as well. This adjustment was made as it is better to intervene with a student that may well succeed on their own than to mis-classify a weak student as passing. Instead of minimizing the raw Type II count, maximizing the F-Measure score of all classes will tend toward the 'perfect' classifier.

One final set experiments evaluated suitable classifiers on the initial data-set. Each experiment ran a combination of classifier algorithm, protocol (Resubstituion, or Leave-One-Out), and a cohort discriminator. This discriminator was either the Academic School or Degree Program. The use of the School and Year was examined in a previous experiment, but ultimately the Year did not provide any additional predictive power.

Selecting the final combination, became a multivariate optimization problem. The F-Measure for all classes needed to be maximized, along with the final accuracy rate; while minimizing the difference between the different protocols. These goals are set to create the best classifier, while resisting over-fitting on the data set. The results of this benchmark can be found in Table A.4.

All of the classifiers benchmarked have good performance with the majority class (PA). This is unsurprising as a pass outcome applies to 85.75% of the instances. Neither are any of the classifiers able to differentiate the 2 instances for the RS (Repeat Semester) class. The selection criteria must therefore revolve around the accurate prediction of the other failure modes. The F-Measure provides a surrogate for individual accuracy, therefore the sum of the failure mode F-Measures can be used as a comparative metric, in the range [0…5]. The top six classifiers (ranked by this metric) achieve between 36% and 47% of the maximum score (1.878/5 to 2.342/5).

The top three classifiers all use the degree program as the cohort discriminator. These three also have larger differences between the resubstitution and the LOO CV protocols. This leads to the conclusion that degree program is less appropriate when determining the patterns within a cohort, and those of the preceding or successive cohorts. Excluding program combinations, leave Random Tree and Random Forest using the School discriminator. The classification of failure modes is within 1% of each other, whereas Random

Table 3
Confusion Matrix from a C4.5 Feature Selection Experiment. Shaded cells represent 'problematic' classifications where a poor outcome would be missed.

↓ Actual / Predicted →	PA	FC	RY	FN	RS
Pass (PA)	4218	23	2	19	0
Resit[1] (FC)	425	28	2	12	0
Repeat Year (RY)	68	3	0	7	0
Fail (FN)	117	10	3	31	0
Repeat Semester (RS)	2	0	0	0	0

[1] A student achieving an FC/Resit, or more formally Conditional Fail, status would need to undertake supplementary assessments to pass their courses. In the U.S. system, this would be equivalent of Summer school/courses.

Tree has an 11% edge with the pass class. On that basis, the Random Tree with School will be used as the classifier of choice.

We believe that classifiers with a stochastic element are able to overcome local maxima/minima during training. This ability allows the classifier to provide a more rounded approach to the data-set. When examining the classification regions and trees produced by other candidate algorithms; the areas, and numbers of instances they represent, become to small to avoid misclassification.

3.6. Alternative metric experiment

As previously noted, utilizing a percentage of sessions attended is a plausible alternative to the BEM. Using the same classifier and parameters, we conducted a companion experiment but using the proportion data-set instead. The results were gathered using both the resubstitution and Leave-One-Out CV protocols from the first three teaching weeks. When using resubstitution, the classifier achieves 98.28% accuracy with only three students mischaracterized under the 'on-mission' metric. However, when using n-Fold Cross Validation, it fairs 12% worse (85.90%) with 290 students misclassified. We can conclude from these results, that the classifier is prone to over-fitting when using the proportion. It also achieves 7.66% less in overall accuracy. This would lead to more potentially unnecessary student interventions. While this is preferred to letting a poor outcome continue unimpeded, it may cause undue stress for passing students that did not require intervention.

4. Unseen case study results

So far, the results have been testing against the same data-set as the classifiers have been trained on. The model and practices devised over previous experiments were applied to a new, previously unseen, data-set. This data is from the 2015/16 academic year, predating any work on learning analytics at Bangor University. As a result, this data could not be influenced by any factor, intentional or not. This data-set comprises $N = 4877$ instances/students. We utilize the same $n = 4$ features (school and Week 1–3 values of the BEM) with the same $C = 5$ classes. The experiment trained the model using data from the 2016/17 academic year, and then tested using the 2015/16 set.

The exact accuracy of model on the unseen data dropped to 84.79% (4135/4877 instances), a difference of 8.77% from the train/test on the same set. This is understandable as there will be cohort effects, as well a graduating class introducing different patterns that will no longer be present. These effects can also be caused by changes in lower education, filtering their way through the system, so that the incoming first-year cohort do not act in the same way as their predecessors.

However, when the model is evaluated against its primary goal, the correct identification of a potential poor outcome rises to 97.33%. This equates to 130 students across the institution that would not be identified, but the model also flags 583 (11.95%) students that would have passed unaided as potentially requiring intervention. This was deemed to present too high a potential risk to otherwise capable students. As a result we sought an alternative classifier. Our classifier evaluation had previously discounted the RandomForest classifier as too unstable with changing data; however, it offers a significantly lower Type II error rate as well a marginal improvement in overall accuracy.

A new model, using the RandomForest classifier but keeping all other parameters the same, was created from the 2016/17 data and again tested on 2015/16 results. This model yielded significantly better results. The overall accuracy improved to 88.05% (+3.26%/159 instances), while the mis-classification rate for otherwise passing students dropped to 8.65% (−3.3%/164 instances). Crucially, the model only marks an additional two students as passing when they would not, bringing the total to 132 not identified. This constitutes a 0.041% drop in effective accuracy.

5. Discussion

Our results show that we have been able to produce a model to identify potential targets for interventions from an early juncture. The model has been engineered to minimize collateral damage, but will still misrepresent some students. Our guiding principle can be seen as a rewording of Blackstone's Formulation (originally applying to criminal law) (Blackstone, 1830). The wording becomes; it is better to intervene with 10 passing students, than to miss a single failing one. As with all predictive models, it is only as good as the information used to train it. There will be students that are mischaracterized, so it is imperative that tutors use this model as an aid instead of a fait accompli.

As the model is based on attendance, some may argue that any model may be inappropriate and that the raw values for attendance should be provided instead. The main rationale for exploring a model is that records show there are students that have low to no attendance but are still perfectly capable and do achieve. Similarly, there are students that have perfect attendance and are not able to achieve. A model employing machine learning techniques should be able to account for these counter-patterns. A suitably advanced model will then naturally reduce collateral damage to either engagement or confidence with students that are naturally talented.

We have noted that within our results, including the unseen data-set that student trajectories appear fixed. There is a slight amount of variance from week to week, but no radical changes in behaviour. As this data-set is from within the institution, we are able to correlate dates with other information - such as tutor visits. This implies that either existing interventions merely prevent a situation getting worse, or outright fail to achieve the intended goal.

The authors have taken a global view of achievement and engagement rather than seeking to identify students that are at risk within one module or subject area. Therefore, the results of this study need to be viewed in a similar context. There will be situations where early engagement is an inappropriate measure on a module level. An individual lecturer could set different criteria for their particular learning design (e.g. missing week 6 would be the worst time to become disengaged due to the material/skill presented that week). Similarly, this work does not attempt to examine suitability nor competancy of a student for future academic programs.

The model produced would be most applicable to Directors of Student Engagement, Senior Tutors, Directors of Teaching and Learning or another departmental role most responsible for overarching pastoral and academic care. As the model seeks to identify potential failing of an entire year, it is less useful to individual lecturers within a single module/course. While the previously identified roles will be primarily interested in the findings, it would fall to individual pastoral tutors to actually perform any intervention.

When dealing with achievement, welfare, and confidence of a student there are serious ethical considerations. Interventions, however well-intentioned, will affect a student's mindset. How large that effect, and whether it is positive or negative will depend on the skill and care of the educator involved. Practitioners will need to adopt a new approach when dealing with students identified by any analytics. We would advocate adopting a similar set of principles as in the modern form of the Hippocratic Oath for clinicians (Lasagna). In the modern version, practitioners acknowledge that they are not treating a disease or set of symptoms but a sick human being. This is something educators may end up losing sight of, when algorithms make the identifications instead of their own intuition.

These effects can not only be triggered by an intervention, but also just from being identified as potentially benefiting from assistance. Some students could see this as a oblique method of assessing their performance and become withdrawn (Lonn, Aguilar, & Teasley, 2015). It is with this in mind that any analytics need to be used openly and transparently. We do not recommend that students are 'kept in the dark' on how and why they have been selected. This position is also advocated by several ethical investigations (Drachsler & Greller, 2016; Pardo & Siemens, 2014). Some even go as far as labeling analytics as 'harmful' when not used in this fashion (Dringus, 2012).

Very little practical guidance is provided for educators on specifically how to react to analytics flagging one of their students. The guidance available encourages them to ground the intervention in solid pedagogical terms (Siemens, 2012; Wise, 2014). One thing is clear, that educators and students must discuss any issues present frankly and honestly if the student is to fulfill their potential.

These interventions may be as simple as a hallway conversation inquiring whether the student is having issues (Yeager & Walton, 2011). For some students, knowing that their tutors have noticed something amiss is enough to effect a change in behaviour (Woolf et al., 2009). In other cases, usually where the causes for disengagement are more varied, the intervention will need to be more complex and involve support staff/services as well as the academics. Current pedagogical research suggests that while the mechanics of the intervention are important, the relationship between student and their support is critically important (Klem & Connell, 2004).

With the European Union's General Data Protection Regulations (Hoel & Chen, 2016) looming large on the horizon, institutions (including Bangor) will need to disclose the manner in which student data is processed. At Bangor, we have taken the view that any analytics that can aid staff must also be visible by the student concerned. We have also developed a position on seeking permission from students to have their data included. We seek blanket permission for student data to be included in our model, but allow students to opt-out of having the resulting analytics used. In this case neither the student nor staff are shown the analytics results. This respects the students' rights under current (and near future) UK and EU law but does not compromise the availability of complete data to the model. Each institution will need to work with their student bodies and any relevant authorities to form their own implementation policies before deploying their solution.

6. Conclusions

As part of this work, we have defined a descriptive metric - the Bangor Engagement Metric - which appears to hold significant predictive power. Initially this metric has been used to enhance Bangor University's existing student information systems. We have now shown how machine learning can be brought to bear on the problem of student retention. Through a set of experiments, we have selected a combination of classifier and measurements to best meet the mission of early identification. This combination has achieved an accuracy in excess of 97%. It minimizes both the number of students needlessly intervened with and the number of students incorrectly predicted as having a positive outcome.

However, as with all machine learning applications the model will not remain static. Different cohorts of students will progress through their courses, requiring different patterns to be identified. This means that the training of this model will be ongoing, including new data and removing aging data. We also recognize that models are only as good as their training, meaning that while this model works for Bangor's students it will not be identical in other institutions. However; the method, classifier and features selected would be transferable. Therefore, there is a need to conduct a longitudinal study, both within Bangor and comparing like-for-like groups in other institutions.

Appendix A. Classifier Benchmark Results

The delta Δ metrics are between the two protocols, not other combinations. This table is sorted by the Δ Accuracy metric, then by the ΔΣ F-Fail (F-Measure for all classes but PA) metric.

Table A.4
Full Results of Classifier Evaluation.

Classifier	Discrim.	Protocol	Acc. %	Per-Class F-Measures					Σ F-Fail*	ΔΣ F-Fail*	Δ F-PA	Δ Acc.
				PA	FC	RY	FN	RS				
Random Tree	Program	Resub.	96.01	0.978	0.817	0.881	0.87	0	2.568	2.342	0.09	17.12
Random Tree	Program	LOO CV	78.89	0.888	0.134	0.041	0.051	0	0.226	2.342	0.09	17.12
RandomTree Regres.	Program	Resub.	96.02	0.978	0.817	0.881	0.87	0	2.568	2.346	0.079	16.99
RandomTree Regres.	Program	LOO CV	79.03	0.899	0.13	0.016	0.076	0	0.222	2.346	0.079	16.99
Random Forest	Program	Resub.	96.01	0.965	0.697	0.746	0.75	0	2.193	1.934	0.064	14.76
Random Forest	Program	LOO CV	81.25	0.901	0.146	0.047	0.066	0	0.259	1.934	0.064	14.76
RandomTree Regres.	School	Resub.	93.6	0.965	0.693	0.736	0.739	0	2.168	1.863	0.059	13.38
RandomTree Regres.	School	LOO CV	80.22	0.906	0.117	0.067	0.121	0	0.305	1.863	0.059	13.38
Random Tree	School	Resub.	93.56	0.965	0.688	0.736	0.739	0	2.163	1.897	0.064	12.74
Random Tree	School	LOO CV	80.82	0.901	0.124	0.029	0.113	0	0.266	1.897	0.064	12.74
Random Forest	School	Resub.	93.6	0.965	0.697	0.746	0.75	0	2.193	1.878	0.057	11.37
Random Forest	School	LOO CV	82.23	0.908	0.127	0.016	0.172	0	0.315	1.878	0.057	11.37
RotationForest	Program	Resub.	88.31	0.936	0.256	0.323	0.437	0	1.016	0.736	0.012	2.74
RotationForest	Program	10-Fold CV†	85.57	0.924	0.093	0.024	0.163	0	0.28	0.736	0.012	2.74
RotationForest	School	Resub.	88.07	0.935	0.246	0.283	0.386	0	0.915	0.693	0.011	2.48
RotationForest	School	LOO CV	85.59	0.924	0.043	0	0.179	0	0.222	0.693	0.011	2.48
Multi-layer Perceptron	School	Resub.	86.28	0.929	0.102	0.049	0.341	0	0.492	0.25	0.004	0.75
Multi-layer Perceptron	School	10-Fold CV†	85.53	0.925	0.08	0	0.162	0	0.242	0.25	0.004	0.75
Self-Organising Map	Program	Resub.	86.03	0.925	0.033	0.074	0.048	0	0.155	0.155	0.003	0.64
Self-Organising Map	Program	LOO CV	85.39	0.922	0	0	0	0	0	0.155	0.003	0.64
Nave Bayes	Program	Resub.	85.51	0.925	0.048	0	0.264	0	0.312	0.098	0.004	0.78
Nave Bayes	Program	LOO CV	84.73	0.921	0.024	0	0.19	0	0.214	0.098	0.004	0.78
Multi-layer Perceptron	Program	Resub.	85.75	0.923	0	0	0	0	0	0.096	0.002	0.01
Multi-layer Perceptron	Program	10-Fold CV†	85.74	0.925	0.029	0	0.067	0	0.096	0.096	0.002	0.01
DecisionTable	Program	Resub.	86	0.925	0	0	0.136	0	0.136	0.023	0.001	0.16
DecisionTable	Program	LOO CV	85.84	0.924	0	0	0.113	0	0.113	0.023	0.001	0.16
DecisionTable	School	Resub.	86	0.925	0	0	0.136	0	0.136	0.023	0.001	0.16
DecisionTable	School	LOO CV	85.84	0.924	0	0	0.113	0	0.113	0.023	0.001	0.16
Nave Bayes	School	Resub.	84.95	0.922	0.02	0	0.208	0	0.228	0.017	0.001	0.14
Nave Bayes	School	LOO CV	84.81	0.921	0.012	0	0.199	0	0.211	0.017	0.001	0.14
Self-Organising Map	School	Resub.	85.75	0.923	0	0	0	0	0	0	0	0.02
Self-Organising Map	School	LOO CV	85.73	0.923	0	0	0	0	0	0	0	0.02
C4.5 Tree	Program	Resub.	85.75	0.923	0	0	0	0	0	0	0	0
C4.5 Tree	Program	LOO CV	85.75	0.923	0	0	0	0	0	0	0	0
C4.5 Tree	School	Resub.	85.75	0.923	0	0	0	0	0	0	0	0
C4.5 Tree	School	LOO CV	85.75	0.923	0	0	0	0	0	0	0	0

Key: Acc. - Accuracy, Discrim. - Discriminator, Regres. - Regression, Resub. - Resubstituion.
* F-Fail represents the F-Measure scores for all classes except PA (Pass).
†10-Fold CV Experiments were substituted for Leave One Out CV due to time constraints.

References

Anderson, T., Whittington, C., & Li, X. J. (2016). Classes to passes: Is class attendance a determinant of grades in undergraduate engineering subjects? *AAEE2016 CONFERENCE coffs Harbour, Australia*.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at purdue: Using learning analytics to increase student success. *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing early at-risk factors in higher education e-learning courses. *Proceedings of the 8th international conference on educational data mining (EDM)* (pp. 150–155). . http://www.educationaldatamining.org/EDM2015/proceedings/edm2015_proceedings.pdf.

Ball, R., & Wilkinson, R. (1994). The use and abuse of performance indicators in UK higher education. *Higher Education, 27*(4), 417–427.

Blackstone, W. (1830). *Commentaries on the laws of England, Vol. 4*. Collins & Hannay.

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272–295.

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46*(5), 904–920.

Davis, D. (2015). *Altis consulting: He information management specialists presentation.* UK Learning Analytics Networkhttps://analytics.jiscinvolve.org/wp/files/2015/05/Jisc-LA-Network-Davis.pdf.

DeBerard, M. S., Spielmans, G., & Julka, D. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal, 38*(1), 66–80.

Drachsler, H., & Greller, W. (2016). Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 89–98). ACM.

Draper, D., & Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society: Series a, 167*(3), 449–474.

Dringus, L. P. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks, 16*(3), 87–100.

Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community College Review, 36*(2), 68–88.

E. Frank, M. A. H. Hall, I. H. Witten, The WEKA workbench. Online Appendix, data mining: Practical machine learning tools and Techniques. Fourth Edition.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.

Gašević, D., Dawson, S., & Siemens, G. (2015). Lets not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

Heaton-Shrestha, C., May, S., & Burke, L. (2009). Student retention in higher education: What role for virtual learning environments? *Journal of Further and Higher Education, 33*(1), 83–92.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., Hlosta, M., & Naydenova, G. (2017). Implementing predictive learning analytics on a large scale: The teacher's perspective. *Proceedings of the seventh international learning analytics & knowledge conference, LAK '17* (pp. 267–271). New York, NY, USA: ACM. https://doi.org/10.1145/3027385.3027397http://doi.acm.org/10.1145/3027385.3027397.

Hoel, T., & Chen, W. (2016). Implications of the european data protection regulations for learning analytics design. *Workshop paper accepted for presentation at the international workshop on learning analytics and educational data mining (LAEDM 2016) in conjunction with the international conference on collaboration technologies (CollabTech 2016), Kanazawa, Japan-September* (pp. 14–16). .

Joshi, M. V. (2002). On evaluating performance of classifiers for rare classes. *EEE international conference on data mining, 2002. Proceedings., 2002* (pp. 641–644). I. https://doi.org/10.1109/ICDM.2002.1184018.

Kittler, J. (1986). *Feature selection and extraction, Handbook of pattern recognition and image processing.* 59–83.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262–273.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition, 33*(1), 25–41.

L. Lasagna, A modern hippocratic oath, Tufts University School of Medicine.

Li, I. W., & Carroll, D. (2017). *Factors influencing university student satisfaction, dropout and academic performance: An Australian higher education equity perspective.* Perth, Western Australia: National Centre for Student Equity in Higher Education, Curtin University.

Lonn, S., Aguilar, S. J., & Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Computers in Human Behavior, 47*, 90–97.

V. A. Lotkowski, S. B. Robbins, R. J. Noeth, reportThe role of academic and non-academic factors in improving college retention, ACT Policy Report.

S. A. Martin, Early intervention program and college partnerships, ERIC Digest.

Massingham, P., & Herrington, T. (2006). Does attendance matter? An examination of student attitudes, participation, performance and attendance. *Journal of University Teaching and Learning Practice, 3*(2), 82–103.

McCoy, S., & Byrne, D. (2017). Student retention in higher education. *Economic insights on higher education policy in Ireland* (pp. 111–141). Springer.

Micere, K. (2013). Getting them enrolled is only half the battle: College success as a function of race or ethnicity, gender, and class. *American Journal of Orthopsychiatry, 83*(2pt3), 310–322. https://doi.org/10.1111/ajop.12033.

Nguyen, Q., Rienties, B., & Toetenel, L. (2017). Unravelling the dynamics of instructional practice: A longitudinal study on learning design and vle activities. *Proceedings of the seventh international learning analytics & knowledge conference, LAK '17* (pp. 168–177). New York, NY, USA: ACM. https://doi.org/10.1145/3027385.3027409http://doi.acm.org/10.1145/3027385.3027409.

Paige, S. M., Wall, A. A., Marren, J. J., Dubenion, B., & Rockwell, A. (2017). *The learning community experience in higher education: High-impact practice for student retention.* Taylor & Francis.

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., & Gašević, D. (2016). Generating actionable predictive models of academic performance. *Proceedings of the sixth international conference on learning analytics & knowledge, LAK '16* (pp. 474–478). New York, NY, USA: ACM. https://doi.org/10.1145/2883851.2883870http://doi.acm.org/10.1145/2883851.2883870.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology, 45*(3), 438–450.

Picciano, A. G. (2012). The evolution of big data and learning analytics in american higher education. *Journal of Asynchronous Learning Networks, 16*(3), 9–20.

Price, L. A. (1993). *Characteristics of early student dropouts at allegany community college and recommendations for early intervention.* Allegany Community Collegehttp://files.eric.ed.gov/fulltext/ED361051.pdf.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Richards, G. (2011). Measuring engagement: Learning analytics in online learning. http://www.academia.edu/download/37420980/Kazan2011-Measuring_Engagement_vf2.docx.

Robbins, S. B., Oh, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and retention as mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *Journal of Applied Psychology, 94*(5), 1163.

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS One, 12*(2), 1–21. https://doi.org/10.1371/journal.pone.0171207https://doi.org/10.1371/journal.pone.0171207.

Sclater, N., Peasgood, A., & Mullan, J. (2017). *Learning Analytics in Higher Education.* Tech. rep., JISC.

Seidman, A. (2005). *College student retention: Formula for student success.* ACE/Praeger series on higher education, Praeger Publishershttps://books.google.co.uk/books?id=ckk5B_ADM_YC.

Shelton, E. N. (2003). Faculty support and student retention. *Journal of Nursing Education, 42*(2), 68–76.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 4–8). ACM.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1510–1529.

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2017). Towards actionable learning analytics using dispositions. *IEEE Transactions on Learning Technologies, 10*(1), 6–16.

Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. *Higher education: Handbook of theory and research* (pp. 51–89). Springer.

Veenstra, C. P. (2009). A strategy for improving freshman college retention. *Journal for Quality and Participation, 31*(4), 19.

Webb, O., Wyness, L., & Cotton, D. (2017). *Enhancing access, retention, attainment and progression in higher education: A review of the literature showing demonstrable impact.* Higher Education Academy.

West, D., Huijser, H., Heath, D., Lizzio, A., Toohey, D., Miles, C., et al. (2016). Higher education teachers experiences with learning analytics in relation to student retention. *Australasian Journal of Educational Technology, 32*(5), 48–60.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers C, 20*(9), 1100–1103. https://doi.org/10.1109/T-C.1971.223410.

Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 203–211). ACM.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. *International Journal of Learning Technology, 4*(3–4), 129–164.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: Theyre not magic. *Review of Educational Research, 81*(2), 267–301.

Ye, C., & Biswas, G. (2014). Early prediction of student dropout and performance in moocs using higher granularity temporal information. *Journal of Learning Analytics, 1*(3), 169–172. http://learning-analytics.info/journals/index.php/JLA/article/view/4212/4429.