Article

Predicting Student Performance Using Clickstream Data and Machine Learning

Yutong Liu [1,*], Si Fan [2], Shuxiang Xu [1], Atul Sajjanhar [3], Soonja Yeom [4] and Yuchen Wei [1]

1 School of Information and Communication Technology, University of Tasmania, Launceston, TAS 7250, Australia
2 School of Education, University of Tasmania, Launceston, TAS 7250, Australia
3 School of Information Technology, Deakin University, Burwood, VIC 3125, Australia
4 School of Information and Communication Technology, University of Tasmania, Hobart, TAS 7000, Australia
* Correspondence: yutong.liu@utas.edu.au

Abstract: Student performance predictive analysis has played a vital role in education in recent years. It allows for the understanding students' learning behaviours, the identification of at-risk students, and the development of insights into teaching and learning improvement. Recently, many researchers have used data collected from Learning Management Systems to predict student performance. This study investigates the potential of clickstream data for this purpose. A total of 5341 sample students and their click behaviour data from the OULAD (Open University Learning Analytics Dataset) are used. The raw clickstream data are transformed, integrating the time and activity dimensions of students' click actions. Two feature sets are extracted, indicating the number of clicks on 12 learning sites based on weekly and monthly time intervals. For both feature sets, the experiments are performed to compare deep learning algorithms (including LSTM and 1D-CNN) with traditional machine learning approaches. It is found that the LSTM algorithm outperformed other approaches on a range of evaluation metrics, with up to 90.25% accuracy. Four out of twelve learning sites (content, subpage, homepage, quiz) are identified as critical in influencing student performance in the course. The insights from these critical learning sites can inform the design of future courses and teaching interventions to support at-risk students.

Keywords: Learning Analytics; Educational Data Mining; student performance prediction; clickstream data

1. Introduction

Learning Analytics (LA) is a rapidly growing research field. The most widely used definition for LA is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (p. iii) [1]. As the definition implies, LA is connected to computer-supported learning environments and educational data that are collected from Virtual Learning Environments (VLEs), such as Learning Management Systems (LMSs) [2]. The past two decades have seen an increased adoption of LMSs, leading to the availability of large educational data sets. With appropriate LA and data mining techniques, analysis of these data sets can provide a better understanding and insights into learners' learning processes and experiences.

Current research in LA has explored various methodologies, including descriptive, diagnostic, predictive and prescriptive analytics [3]. Predictive analytics in education involves inferring uncertain future events or outcomes related to learning or teaching [3]. Some tasks predict an aspect of teaching, such as course registration, student retention, or the impact of a given instructional strategy on learners [3]. Other tasks focus on learning and learners' perspectives, such as predicting academic success, course grades or skill acquisition [3]. Student performance prediction is one of the significant areas because

it can be used to enhance student academic performance and reduce student attrition, for example, by using early warning systems to support at-risk students [4]. This research is an exploration of LA for the purpose of student performance prediction.

The data used in student performance prediction tasks rely on the course design and the data generated from LMSs. Commonly used data include demographics, academic background, and learning behaviour data [5–7]. Some studies only focus on behaviour data (e.g., video-viewing data) to perform these tasks [8]. Others use mixed-category data (e.g., both demographics and behaviour data) to predict student performance [9]. However, due to a variety of course design, models built for specific courses using specific course-related data can be difficult to reuse in other courses. As a category of data reflecting student behaviour, clickstream data indicates the path a student takes in navigating one or more learning sites in LMSs. The strengths of using clickstream data include its ease of access, regardless of course conditions, such as course structure, assessments or learning activities. One limitation of this data category is its less explicit connections with students' learning behaviours compared to other data, such as discussion board data. This may explain why it has received insufficient attention from educators and researchers to date. Despite this limitation, studies have used clickstream data and observed connections between click actions and students' learning behaviours [10], confirming its value and potential in student performance prediction tasks. In addition to the introduction (Section 1), Section 2 of this paper presents a literature review on Learning Analytics and Educational Data Mining, student performance prediction, and clickstream data. The research methods, including the research aim and objectives, data sets, and experiments, are discussed in Section 3. The implementation and results of the experiments are presented in Section 4. The findings are discussed in Section 5, and the conclusion is provided in Section 6.

2. Literature Review

2.1. Learning Analytics and Educational Data Mining

This research concerns two closely related research fields, Learning Analytics (LA) and Education Data Mining (EDM). LA is an increasingly explored area in education [11]. The data used in LA are considerably diverse, from log and survey data to eye tracking, automated online dialogue and "Internet-of-Things" data [2]. Generally, LA aims to generate insights from educational data to improve learning and teaching [12,13]. These insights help educational institutions to enhance education-related policies, management strategies, and learning systems or environments [14]. Educational Data Mining (EDM) is a closely related field to LA [15]. EDM "is an emerging multidisciplinary research area, in which methods and techniques for exploring data originating from various educational information systems have been developed" (p. 3) [16]. EDM has a stronger emphasis on the technical element of data mining and analysis, but shares the overarching goal with LA of generating insights to support learning and teaching improvement.

LA and EDM can be used to support students, teachers, researchers, and educational institutions in a number of ways. LA and EDM can provide insights into students' learning experiences, processes and performance [17]. For teachers, it can lead to enhanced course planning and material evaluation tools [18]. For students, it allows those who encountered learning difficulties to acquire timely interventions from their instructors [19]. For educational researchers, it enables them to better understand learners' behaviours and the impact of the learning environment on student learning [17]. For institutions, it can help improve student engagement and potentially achieve higher retention rates [19]. Predictive analysis is a key focus in LA and EDM research. A systematic review identified four main areas of focus in LA and EDM: (a) computer-supported prediction analytics, (b) computer-supported behaviours analytics, (c) computer-supported learning analytics, and (d) learning visualisation analytics [20]. Ref. [21] pointed out that for the studies they investigated from 2000 to 2017, the biggest proportion of research was in the area of computer-supported prediction analytics (63.25%). Classification and regression are popular prediction methods in EDM, involving using variables to predict students' academic performance or success

(e.g., whether drop-out students or academic grades). Early warning systems are one type of such applications [4,22]. Another method, called Knowledge Inference or Latent Knowledge Estimation, is associated with using EDM techniques to "assess the skills of students based on their responses to a problem-solving exercise" (p. 184) [21]. For example, natural language processing technologies (e.g., Transformers) have the potential to be used to measure human language knowledge acquisition [23]. This research sits the classification method, one of the commonly seen methods in predictive analysis [20].

2.2. Student Performance Prediction

With the rapid expansion in the volume of educational data, methodological approaches in LA and EDM research have continued to grow and mature, along with sophisticated data analysis techniques [17,24]. Today, LA and EDM drive an advanced method of prediction that enhances traditional techniques in student performance prediction. The term 'student performance' is used differently across different studies. There are generally two types: performance at the program level [25–27] and at the course level [28]. For the program level, prediction tasks can be related to identify the probability of student dropouts or graduates from a degree program (e.g., a bachelor's degree). For the course level, student performance is defined as students' learning outcomes such as assignments or assessments in their courses after a study period [29]. Its prediction tasks can focus on predicting students' final scores or grades, or their pass or fail status at the end of the course [4,7]. At-risk students are those who are more likely to fail the course [4]. The aim of this type of prediction task is to identify at-risk students and help them achieve their academic goals. This research focuses on the course-level performance prediction, and investigates student learning behaviours on various learning tasks.

Predictive analysis at the course level can provide valuable insights for enhancing teaching and learning. It can help instructors understand student behaviour (e.g., the total time in learning content viewing) [12] and design instruction accordingly [17]. It can also give students who encounter learning difficulties the opportunities to receive timely interventions from their instructors [19] and adjust their learning strategies. Furthermore, the integration of predictive analysis has been shown to enhance educational strategies from an institution perspective. For example, [30] investigated the adoption levels of blended learning systems using online behaviour data found that the use of predictive analysis can support the strategic developments of LMS. Previous research has employed feature importance analysis in performance prediction tasks to determine the influence of student-related features on academic performance [31,32]. Identifying these important features can provide valuable insights for improving teaching and learning [33].

A number of approaches and techniques are adopted in student performance prediction tasks. From the data usage perspective, comparative analysis of the trained datasets (e.g., generating multiple feature sets through feature extraction) to validate predictive models has less attention. A systematic literature review reported that nearly 92% of studies used only one dataset in model training, while 8% used multiple datasets to search for optimal predictive models [34]. Moreover, in classification tasks for predicting student performance, several data mining algorithms have proven particularly popular, including Decision Tree, K-Nearest Neighbor, Support Vector Machines, Naïve Bayes, Random Forest, Boosted Trees, Adaptive Boosting and Gradient Boosting [34,35]. In addition to achieving high performance in predictive modeling tasks, the popularity of certain algorithms may also be related to their ability to generate explainable output. This means that these algorithms are able to provide clear and intuitive explanations of their predictions, which can be useful for interpreting the results of the model. Explainable models are significant in solving problems in education, including predicting student performance [36]. Explainable models such as tree-based algorithms (53%) and rule-based algorithms (33%) are more commonly used than deep learning algorithms (6%) [36]. Deep learning models are instances of "black-box" models, meaning that the way models work cannot be explained or understood by human. However, recent studies have revealed a shift in focus from explainable models

or "white-box" models to more complex "black-box" models that are capable of solving challenging problems [37].

2.3. Clickstream Data

Researchers devote themselves to investigating student performance prediction using features extracted from the LMS usage data. LMSs produce various data based on their functionalities and the course design. Some researchers use data regarding students' personal and social background, previous academic achievements (e.g., transcripts, admission data), or student self-reporting (e.g., interview or survey data) [38–40]. These data indicate individuals' information or learning environments, which could impact students' academic performance. Other data categories, such as event-stream data, are also popular due to their direct link to students' learning behaviours in activities or tasks, such as discussions/forums, quizzes, learning material access, video viewing, and assignments/homework.

Recent literature shows the usage of clickstream data in predictive analysis. For instance, in business analysis, clickstream data are analysed to inform webpage design and evaluate the effectiveness of marketing programs [41]. In education, clickstream indicates the path(s) a student takes through one or more learning sites. Clickstream data have been used in research to understand student learning behaviours in online courses [42], and how their learning behaviours impact their academic performance [7]. For example, visitation of some pages or sites may be consistent with students' navigation habits in learning [43]. Therefore, those pages can be used to display the most important materials or notifications when considering course presentation design.

Clickstream data in educational research has value in understanding the teaching and learning process. However, like any other data, clickstream data have limitations. It shows non-continuous events in behaviour patterns, resulting in sparse data. Each click action could be the start point or endpoint of each fragment in learning, so the mid-process could be missed. For example, a click on a URL indicates that a student has requested the URL directory path (e.g., a sample URL https://xxxxxx/homepage) but the request itself is not semantically meaningful in educational contexts [42]. Therefore, the link between a click action on a particular site or page (e.g., homepage in the case of the sample URL) and the corresponding learning behaviour is implicit.

Despite the limitations, research has confirmed that clickstream data are reliable and offer valid and nuanced information about students' actual learning processes [42]. For example, clickstream data contain subtle information with time-stamped "footprints" on individual learning behavioural pathways [42,44]. In prediction tasks, these "footprints" can indicate learning efforts, and are more reliable measures compared to conventional methods of self-reporting [42,45]. For instance, recent research has used clickstream data to examine student engagement with videos in learning (by clicking on pausing or changing playback speed on videos) [46], as well as students' effort regulation and time management behaviours [42]. In recognition of its strengths and potential in student performance prediction tasks, clickstream data are chosen for the purpose of this research.

3. Methods

3.1. Research Aim and Objectives

In this research, we aim to examine how machine learning and clickstream data can be used to predict student performance in online learning and teaching. This aim will be achieved by addressing the following Research Objectives (RO):

- RO1: Feature extraction: to extract feature sets from the original datasets, that can be effectively used for student performance prediction
- RO2: Feature selection: to investigate the impact of different features on prediction outcomes with the aim to identify the important features
- RO3: Model evaluation: to compare the performance of different models when using different feature sets, with and without a feature selection method, as well as multiple

machine learning (including deep learning) algorithms, to find the most optimal model for predicting student performance.

To achieve these objectives, this research adopted a commonly used prediction method in data science, including data preparation, feature extraction (RO1), and model training with/without a feature selection method (RO2) and model evaluation (RO3).

3.2. Data Sets

This research utilised open-source clickstream data OULAD (Open University Learning Analytics Dataset) from a distance-learning university—Open University [47]. According to the description of OULAD, the data adhered to the Data Protection Policy and Policy on Ethical Use of Student Data for Learning Analytics [47]. The data were collected anonymously, and all the students consented to their data being used for academic research. The Open University courses were represented in a VLE with typical online course structures [47]. Each course (called a Module) has multiple module presentations. Each module presentation consists of a few formative assessments and a final exam. The original data reflected four aspects of students: registrations, assessments, VLE interactions, and demographics [47]. Only two of these aspects were used due to their alignment with the aim of this research. One is demographics, containing the classification label (the final result of the course); its dataset named studentInfo [47]. Another is VLE interactions, containing students' clickstream data; its dataset named studentVLE [47]. Details of studentVLE are shown in Table 1.

After assessing the OULAD data, it was determined that a single course's data was used for this research to gain deep insights into teaching and learning for that specific course. To select a course with the largest dataset, the number of non-withdrawn students was compared among seven courses from the raw datasets. As a result, the course *BBB* was selected as it had the most number of non-withdrawn students. This was a course in the Science, Technology, Engineering and Mathematics (STEM) discipline [47].

Table 1. Details of the studentVLE dataset [47].

#	Columns	Description	Data Type
1	code_module	the module identification code	nominal
2	code_presentation	the presentation identification code	nominal
3	id_site	the VLE material identification number	numerical
4	id_student	the unique student identification number	numerical
5	date	the day of student's interaction with the material	numerical
6	sum_click	the number of times the student interacted with the material	numerical

3.3. Experiments

This session covered data preparation, followed by feature extraction (RO1), model training (RO2), and model evaluation (RO3) methods. A feature is a measurable piece of data that can be used for analysis or prediction.

Data preparation. For the purpose of simplicity, the original labels in the studentVLE dataset with three classes *pass*, *fail* and *distinction* were simplified to two classes *pass* and *fail*. Specifically, the label *distinction* was merged with *pass*. As a result, for the total sample of 5521 students, 68% (3754) were labelled as *pass*, and 32% (1767) students were labelled as *fail*. The course *BBB* had 12 activity categories (denoted by Act in this research), including: *forum* (Act1), *content* (Act2), *subpage* (Act3), *homepage* (Act4), *quiz* (Act5), *resource* (Act6), *url* (Act7), *collaborate* (Act8), *questionnaire* (Act9), *onlineclass* (Act10), *glossary* (Act11), and *sharedsubpage* (Act12) (According to OULAD, id_site are categorised into 12 activities, named forumng (Act1), oucontent (Act2), subpage (Act3), homepage (Act4), quiz (Act5), resource (Act6), url (Act7), oucollaborate (Act8), questionnaire (Act9), ouelluminate (Act10), glossary (Act11) and sharedsubpage (Act12). The activity category ouelluminate (Act10) indicates an online classroom for live lecture broadcasting or real-time tutorials).

Before extracting feature sets, the data were cleaned. When merging studentInfo and studentVLE, it was found that 180 students had no recorded click behaviours in studentVLE. These 180 samples were discarded, leaving 5341 students' data for the research.

Feature extraction. Two feature sets representing click counts for activity categories were extracted from OULAD. These feature sets were extracted by (1) aggregating time-based click count and (2) transforming the data structure. As the first step of the feature extraction, the click counts were summed with two levels of granularity sizes - week and month, indicating the click count every week and month. To do this, the whole course period (T) was divided into two parts. The first course period was T0 (e.g., *week0* or *month0*), indicating the time period before the course officially commences. The aggregation method was to sum up all the click counts in T0. The second course period started from the date when the course officially commences to the end of the course (T1, . . ., Tn) (e.g, from *week1* to *week39*, from *month1* to *month9*). As the second step of the feature extraction, the pieces of data were further transformed into the structure of balanced panel data. A balanced panel data refers to a structure where each panel member (i.e., student) is observed in regular time intervals. As a result, two feature sets, named WEEK and MONTH, were generated as the structure of Figure 1. The columns indicate activity categories, the rows indicate time-based observations of students' click counts. The total of 5341 sample students was extended to 213,640 observations in WEEK, and 53,410 observations in MONTH. The rationale of the strategy is to avoid high-dimensional feature sets while keeping both the time and activity dimensions.
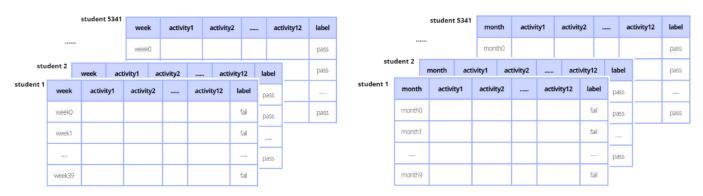

Figure 1. Generated two feature sets WEEK (**left**) and MONTH (**right**) for training models. WEEK contains 213,640 rows (that is 5341 students × 40 weeks) × 13 columns (that is 12 features + 1 label). MONTH contains 53,410 rows (that is 5341 students × 10 months) × 13 columns (that is 12 features + 1 label).

Model training methods. In this research, machine learning, including traditional machine learning (LR, k-NN, RF, GBT) and deep learning (1D-CNN, LSTM) methods, were used for model training. For traditional machine learning, LR (Logistic Regression) was selected as the baseline algorithm because of its effectiveness with clickstream data in previous research [14]. LR is a transformation of linear regression that can be seen as a simple version of the regression model for binary classification [48]. It is based on the estimation mechanism of probability with the 0 or 1 output of the model [48]. k-NN (k-Nearest Neighbors) is another traditional machine learning method used for classification. It is a simple, similarity-based algorithm that classifies by comparing the similarities between testing and training samples [49]. The choices of RF (Random Forest) and GBT (Gradient Boosting Trees) [5], were motivated by their boosted capacities. RF is an ensemble classifier that improves the performance of classic tree-based algorithms using the Bagging method [50]. It creates an uncorrelated forest of trees and combine them in parallel [51]. The prediction by the committee of RF is more accurate than that of any single tree [51]. GBT, or GBDT (Gradient Boosting Decision Trees), uses the Boosting method to sequentially combine weak learners (typically shallow decision trees) to allow each new tree to correct the previous errors [52]. This method reduces Bias, which is one of the

components of accuracy. Some recent studies have found GBT to perform better than RF in clickstream data classification tasks [5,53].

For deep learning methods, CNN (Convolutional Neural Network) is designed for image-related tasks. An innovative study used an enhanced 2D-CNN (Two-dimensional Convolutional Neural Network) model on a set of temporal educational data by transforming the data into a colour-image-like structure [54]. This current research adopted 1D-CNN (One-dimensional Convolutional Neural Network) models, using hyperparameters *kernel* and *stride* to control the feature-extracting window size and the window slice step, to examine its capacity on temporal clickstream data. LSTM (Long Short-Term Memory) is a Recurrent Neural Network (RNN) designed for dealing with sequential data (e.g., time-series data) by 'memorising' earlier inputs. In complex settings, such as multidimensional data with inter-sequential or inter-temporal interactions, LSTM is more powerful than traditional models [55]. LSTM were selected for its capacity to handle temporal data and its outstanding performance in tabular clickstream data in prediction tasks [5]. This research used a stacked LSTM architecture (two LSTM layers, followed by one Fully Connected Layer).

M1 and M2 models. Two models were trained with WEEK and MONTH feature sets when using LR, k-NN, RF and GBT. One model did not use feature selection (denoted by M1 models in this research). Another model used Information Gain feature selection filter method (denoted by M2 models in this research). For 1D-CNN and LSTM, only one model was trained for each feature set (WEEK and MONTH) without using the feature selection method (i.e., M1 models). This is because 1D-CNN and LSTM have the capability to weight features within the neural network mechanism.

Cross-validation. To maximise the data used for training and obtain solid results, this research adopted 10-fold cross-validation. Specifically, for each fold, 90% of the data were used as a train set, and 10% were used as a test set.

Model evaluation metrics. The performance of the models were evaluated with accuracy, F1-score and AUC.

4. Results

The results are presented in three main aspects: model training implementation, model performance (i.e., model evaluation results) and feature importance. The first two aspects show how the models were trained, how well they perform, and which model is the best. The third aspect, feature importance, examines the dominant features of the best model to gain insights into teaching and learning.

4.1. Model Training Implementation

This session demonstrated the implementation process of M1 models (no feature selection) and M2 models (using Information Gain for feature selection).

M1 Models. LR M1 models adopted normalised input data (0, 1) and were trained using the default hyperparameters of the Logistic Regression operator in RapidMiner. k-NN M1 models also adopted normalised input data (0, 1) and used an optimal k value of 7, and a distance measure hyperparameter of MixedMeasures. RF M1 Models adopted original input data (i.e., without normalisation) and involved tuning two hyperparameters: *number_of_trees* is set to 350, and *maximal_depth* is set to 3. GBT M1 models adopted original input data and involved tuning three hyperparameters: *number_of_trees* is set to 61, *maximal_depth* is set to 3 and *learning_rate* is set to 0.1. 1D-CNN models, structured with an input layer, a 1D-CNN layer, a pooling layer, a fully connected layer and an output layer, used a *batch* size of 128, a *epochs* of 400, a *learning rate* of 0.001, and *dropout rate* of 0.4. The 1D-CNN models using input data with a shape of (1, 13) (for both WEEK and MONTH feature sets) have 897 trainable parameters. LSTM models, structured with an input layer, two LSTM hidden layers, a fully connected layer and an output layer, adopted an *optimisation function* of *Adam stochastic gradient descent*, a *loss function* of *categorical binary cross-entropy*, a *learning rate* of 0.0001, a *dropout rate* of 0.2, a *batch* size of 128, and an *epochs*

of 700. The LSTM models using input data with a shape of (40, 13) (i.e., WEEK feature set) have 32 hidden units in first LSTM layer and 8 hidden units in the second LSTM layer, with a total of 7209 trainable parameters. The LSTM models using input data with a shape of (10, 13) (i.e., MONTH feature set) have 32 hidden units in first LSTM layer and 16 hidden units in second LSTM layer, with a total of 9041 trainable parameters.

M2 models. The model training process involved (1) weighting features by Information Gain and acquiring weight scores of features, and (2) finding the best threshold of weight scores that enables a selection of the best subset of features. The implementation of this process involved two parts. The first part was obtaining weight scores for all click behaviour features using the operator *Weight by Information Gain* in RapidMiner. The results of this process were shown in Table 2. The second part of the implementation process involved training models to find the best threshold with the best model performance. For LR and k-NN using both WEEK and MONTH feature sets, the best threshold was 0.7. This indicated only one feature, *homepage* (Act4), was involved in the model. For RF and GBT using WEEK, the best threshold was 0.3. This indicated two features, *homepage* (Act4) and *forum* (Act1), were involved in the model. For RF and GBT using MONTH, the best threshold was 0.3, indicating that four features were involved in the model. These were *homepage* (Act4), *forum* (Act1), *subpage* (Act3) and *resource* (Act6).

Table 2. Weight scores of feature weighting by Information Gain for WEEK and MONTH feature sets (normalised in between 0 and 1).

Feature Name	Activity Category	Weight Scores in WEEK	Weight Scores in MONTH
Act1	forum	0.60	0.63
Act2	content	0.28	0.34
Act3	subpage	0.28	0.54
Act4	homepage	1.00	1.00
Act5	quiz	0.12	0.32
Act6	resource	0.22	0.43
Act7	url	0.13	0.19
Act8	collaborate	0.04	0.09
Act9	questionnaire	0.03	0.06
Act10	onlineclass	0.00	0.01
Act11	glossary	0.03	0.06
Act12	sharedsubpage	0.00	0.00

4.2. Model Performance

The model performance was evaluated using accuracy (Table 3), F1-score (Table 4) and AUC (Table 5). As 10-fold cross-validation was used in model training, the variances were also demonstrated in the model performance results. Overall, the results showed LSTM ›1D-CNN ›GBT, RF ›k-NN, LR. The best model was the LSTM (M1) model using feature set WEEK, with an accuracy of 89.25% ($\pm$0.97%), F1-score of 92.71% ($\pm$0.62%) and AUC of 0.913 ($\pm$0.014). Also, the variances of the model performance were relatively low. The second-best model was the LSTM (M1) model using feature set MONTH, with 88.67% ($\pm$1.27%) accuracy, 92.37% ($\pm$0.81%) F1-score and 0.906 ($\pm$0.013) AUC. However, the rest of the models did not perform well. The models k-NN, RF, GBT, 1D-CNN with WEEK and MONTH showed low AUC scores, between 0.50 to 0.78 (Table 5). This range of AUC indicated that the model's ability is equivalent to a random guess or slightly better. Also, their corresponding accuracy ranges from 66% to 76% with F1-score ranging from 72% to 85%. Due to the imbalanced data (approximately 68% Pass, 32% Fail), this performance range is not ideal in practice.

Table 3. Model performance—accuracy (variances are demonstrated within the parentheses).

Algorithms	M1 Using Feature Set WEEK	M2 Using Feature Set WEEK	M1 Using Feature Set MONTH	M2 Using Feature Set MONTH
LR	70.24% (±0.02%)	70.25% (±0.00%) [1]	70.24% (±0.03%)	70.25% (±0.00%) [2]
k-NN	66.29% (±0.30%)	66.10% (±0.32%) [3]	76.46% (±0.48%)	76.46% (±0.49%) [4]
RF	70.25% (±0.00%)	70.25% (±0.00%) [5]	73.82% (±3.83%)	76.39% (±0.47%) [6]
GBT	70.25% (±0.00%)	69.02% (±1.99%) [7]	76.47% (±0.49%)	76.47% (±0.49%) [8]
1D-CNN	70.25% (±0.27%)	n/a	77.55% (±0.88%)	n/a
LSTM	89.25% (±0.97%) *	n/a	88.67% (±1.27%) **	n/a

Features used in M2 models: [1] Act4; [2] Act4; [3] Act4; [4] Act4; [5] Act4, Act1; [6] Act4, Act1, Act3, Act6; [7] Act4, Act1; [8] Act4, Act1, Act3, Act6. * The best performance: LSTM algorithm with M1 using feature set WEEK. ** The second-best performance: LSTM algorithm with M1 using feature set MONTH.

Table 4. Model performance—F1-score (variances are demonstrated within the parentheses).

Algorithms	M1 Using Feature Set WEEK	M2 Using Feature Set WEEK	M1 Using Feature Set MONTH	M2 Using Feature Set MONTH
LR	82.51% (±0.01%)	82.53% (±0.00%) [1]	82.51% (±0.02%)	82.53% (±0.00%) [2]
k-NN	72.94% (±0.25%)	72.67% (±0.27%) [3]	84.06% (±0.37%)	84.06% (±0.38%) [4]
RF	82.52% (±0.00%)	82.52% (±0.00%) [5]	83.88% (±1.65%)	83.96% (±0.37%) [6]
GBT	82.52% (±0.00%)	79.58% (±4.75%) [7]	84.04% (±0.38%)	84.04% (±0.38%) [8]
1D-CNN	82.52% (±0.18%)	n/a	85.24% (±0.82%)	n/a
LSTM	92.71% (±0.62%) *	n/a	92.37% (±0.81%) **	n/a

Features used in M2 models: [1] Act4; [2] Act4; [3] Act4; [4] Act4; [5] Act4, Act1; [6] Act4, Act1, Act3, Act6; [7] Act4, Act1; [8] Act4, Act1, Act3, Act6. * The best performance: LSTM algorithm with M1 using feature set WEEK. ** The second-best performance: LSTM algorithm with M1 using feature set MONTH.

Table 5. Model performance—AUC (variances are demonstrated within the parentheses).

Algorithms	M1 Using Feature Set WEEK	M2 Using Feature Set WEEK	M1 Using Feature Set MONTH	M2 Using Feature Set MONTH
LR	0.597 (±0.007)	0.607 (±0.006) [1]	0.482 (±0.008)	0.500 (±0.000) [2]
k-NN	0.670 (±0.006)	0.666 (±0.007) [3]	0.671 (±0.008)	0.670 (±0.010) [4]
RF	0.690 (±0.004)	0.674 (±0.004) [5]	0.751 (±0.008)	0.734 (±0.009) [6]
GBT	0.698 (±0.004)	0.690 (±0.004) [7]	0.763 (±0.007)	0.763 (±0.007) [8]
1D-CNN	0.720 (±0.005)	n/a	0.786 (±0.006)	n/a
LSTM	0.913 (±0.014) *	n/a	0.906 (±0.013) **	n/a

Features used in M2 models: [1] Act4; [2] Act4; [3] Act4; [4] Act4; [5] Act4, Act1; [6] Act4, Act1, Act3, Act6; [7] Act4, Act1; [8] Act4, Act1, Act3, Act6. * The best performance: LSTM algorithm with M1 using feature set WEEK. ** The second-best performance: LSTM algorithm with M1 using feature set MONTH.

4.3. Feature Importance

The important features of the best model, the LSTM (M1) model using WEEK feature set, were examined by observing how the model's performance changes when each feature in the model is discarded. As a result, the model performance showed a dropped accuracy each time a feature was removed. The accuracy decreased significantly when the features Act2 (*content*), Act3 (*subpage*), Act4 (*homepage*) and Act5 (*quiz*) were removed. Therefore, these four features can be considered dominant. A comparison of the accuracy after each feature were removed was shown in Table 6.

Table 6. Feature importance analysis result.

Removed Feature	Activity Category	Model's Accuracy	Dropped Accuracy
Act1	forum	89.22% (±0.79%)	0.03%
Act2 *	content	89.03% (±0.68%)	0.22%
Act3 *	subpage	88.90% (±0.78%)	0.35%
Act4 *	homepage	88.90% (±1.27%)	0.35%
Act5 *	quiz	89.07% (±0.93%)	0.18%
Act6	resource	89.18% (±0.87%)	0.07%
Act7	url	89.12% (±0.89%)	0.13%
Act8	collaborate	89.16% (±0.97%)	0.09%
Act9	questionnaire	89.23% (±0.89%)	0.02%
Act10	onlineclass	89.16% (±0.83%)	0.09%
Act11	glossary	89.22% (±0.77%)	0.03%
Act12	sharedsubpage	89.22% (±0.69%)	0.03%

* The first dominant features are Act3 (*subpage*) and Act4 (*homepage*), the second dominant feature is Act2 (*content*), and the third dominant feature is Act5 (*quiz*)

5. Discussion

This research trained multiple predictive models using machine learning methods and clickstream data, achieving up to 90.25% (89.25% + 0.97%) accuracy. The results provide insights into effective ways to extract features, train and evaluate predictive models in student performance prediction tasks using students' clickstream data. From a data science perspective, this research contributes three major findings through answering its three Research Objectives: (1) Feature extraction, (2) Feature selection, and (3) Model evaluation. In addition, although this research was conducted based on a data science approach rather than with a pedagogical focus, the identified important features from the best model can inform future course design and teaching interventions.

5.1. Research Objective 1: Feature Extraction

In this research case, the best practice was to adopt the weekly view (i.e., the WEEK feature set, aggregating click counts based on weekly time intervals), even though in some cases, it did not perform as well as the monthly view (i.e., the MONTH feature set, aggregating click counts based on weekly time intervals). Panel data appeared to be unsuitable for all traditional machine learning (LR, k-NN, RF, GBT) and 1D-CNN, so the weekly and monthly views were not worth comparing for these algorithms. However, for LSTM, the weekly view showed the best model with 89.25% accuracy, significantly higher than the monthly view at 88.67%. Therefore, although traditional machine learning models learned features better from the monthly view, their accuracy were still lower than the model using the weekly view with the LSTM algorithm. This result is consistent with some current studies that use a weekly view to represent clickstream data and effectively conduct their prediction tasks [14]. The feature sets used in this research involve both time and activity dimensions, which can be seem as an effective way to conduct feature extraction. As suggested by previous research, using features from multiple dimensions is more likely to achieve better predictive results [56].

5.2. Research Objective 2: Feature Selection

In data science, feature selection is a powerful technique that reduces the number of features and makes the model easier to learn patterns [57]. In this research, features were selected based on their correlation with the label (in M2 models). However, the traditional machine learning models (LR, k-NN, RF, GBT) did not boost the results despite using feature selection. One potential reason is that the dataset structure may not be suitable for the chosen machine learning algorithms. Another possible reason is that removing the lower correlated features could lead to a loss of valuable information for prediction. In contrast, the LSTM (M1) model using feature set WEEK did not use the feature selection method but still achieved the highest accuracy. This finding suggests that feature selection

can be optional in student performance prediction with clickstream data, at least when the feature set is not high-dimensional. For high-dimensional feature sets, feature selection is commonly used for training prediction models [49]. As WEEK and MONTH feature sets only have the total of 12 features, they are not considered high-dimensional. Therefore, feature selection may not be necessary in this case.

5.3. Research Objective 3: Model Evaluation

The experiment results show the varying capacities of different machine learning algorithms when dealing with clickstream data with a panel data structure. First, k-NN is often used for course-specific prediction using event-stream data [9]. It did not produce good results when applied to clickstream data in this research case. This finding is consistent with previous clickstream data studies that used k-NN for student performance prediction and obtained poor performance [18]. Second, the 1D-CNN models had higher accuracy than the traditional machine learning algorithms, but it was still significantly lower than the best LSTM model in this study. This may be because 1D-CNN is not as well-suited for dealing with sequential data. Third, LSTM is effective for learning features from panel data with a matrix structure that indicates student click behaviours. This may be because LSTM can handle sequential observations in each panel member. Similar findings have been reported in other papers. For example, a study concludes that LSTM is a outstanding candidate for training customer prediction models using a form of panel data [55]. Also, LSTM can achieve the same effects as traditional machine learning with complex feature selection processes, without requiring manual feature selection. In this study, LSTM generated better results than the traditional machine learning with large feature selection workloads. Therefore, when using LSTM to predict student course results, feature selection may not be necessary.

Deep learning can be used in areas where analytical-result-driven decisions can be explained and used to affect individuals [36]. This research used LSTM to train models, then conducted feature importance analysis to generate explainable results (see the following section Implications for Learning and Teaching). Despite the perceived "black-box" nature of deep learning LSTM, this research demonstrated how it can provide explainable achievements for student performance prediction and support decision-making in teaching and learning.

5.4. Implications for Learning and Teaching

In student performance prediction tasks, feature importance analysis can reveal the factors that influence students' performance. Based on the feature importance analysis of this research, the students' click behaviours on the course homepage and subpages are the most important to the prediction tasks. The second-most important activities are the course content and quizzes. The rest of the activity categories demonstrated minor importance, meaning that students' click behaviour patterns on those sites have a minor impact on predicting students' academic results. These findings can inform teaching and learning in online courses with a similar structure, with the goal of improving the learning environment or facilitating teaching intervention practices. First, the identified dominant activity categories can be used to guide course design. Taking the findings in this study as an example, students' click behaviours on the course homepage and its subpages may reflect students' habits of interacting with the course learning environment, for example, student starting their learning by accessing the course's homepage. Therefore, key course information can be placed on the homepage and subpages to align with students' behaviour habits. Additionally, as the content and quizzes are identified as critical, educators or instructors can encourage students to engage with the learning content or materials and quiz activities.

6. Conclusions

This research investigates the potential of using clickstream data to predict student performance. Student performance prediction is a sub-topic of LA and EDM. The ability to predict student performance can be beneficial in identifying at-risk students and providing them with learning support. This research aims to build a student performance prediction model using clickstream data. In the experiments, multiple predictive models were trained and analysed, using aggregated click data (number of clicks in a weekly and monthly basis), machine learning algorithms (LR, k-NN, RF, GBT, 1D-CNN and LSTM) and a feature selection method. This research found that weekly-based click count aggregation in the form of panel data, together with the LSTM model, is the best practice for this student performance prediction case. Feature selection is optional in this case. Moreover, by analysing important features from the best model, this research found that clicks on the homepage, subpages, content and quizzes are significant in predicting student performance. Based on these findings, educators can consider improving the online learning environment by utilising the advantage of students visiting homepage and subpages of the course. Teaching intervention practices to help at-risk students are suggested to provide student support around learning materials and quizzes.

Abbreviations

The following abbreviations are used in this manuscript:

LA	Learning Analytics
EDM	Educational Data Mining
LMS	Learning Management System
OULAD	Open University Learning Analytics Dataset
VLEs	Virtual Learning Environments
LR	Logistic Regression
k-NN	k-Nearest Neighbors
RF	Random Forest
GBT	Gradient Boosting Trees
CNN	Convolutional Neural Network
1D-CNN	One-dimensional Convolutional Neural Network
2D-CNN	Two-dimensional Convolutional Neural Network
LSTM	Long Short-Term Memory

References

1. Siemens, G. Message from the LAK 2011 General & Program Chairs. In Proceedings of the LAK11: 1st International Conference on Learning Analytics and Knowledge, Banff, AB, Canada, 27 February–1 March 2011.
2. Nistor, N.; Hernández-Garcíac, Á. What types of data are used in learning analytics? An overview of six cases. *Comput. Hum. Behav.* **2018**, *89*, 335–338. [CrossRef]
3. Society for Learning Analytics Research (SoLAR). Available online: https://www.solaresearch.org/about/what-is-learning-analytics (accessed on 30 August 2022).

4.   Akçapınar, G.; Altun, A.; Aşkar, P. Using learning analytics to develop early-warning system for at-risk students. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 40. [CrossRef]

5.   Chen, F.; Cui, Y. Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *J. Learn. Anal.* **2020**, *7*, 1–17. [CrossRef]

6.   Imran, M.; Latif, S.; Mehmood, D.; Shah, M.S. Student Academic Performance Prediction using Supervised Learning Techniques. *Int. J. Emerg. Technol. Learn.* **2019**, *14*, 92–104. [CrossRef]

7.   Yang, Y.; Hooshyar, D.; Pedaste, M.; Wang, M.; Huang, Y.M.; Lim, H. Prediction of students' procrastination behaviour through their submission behavioural pattern in online learning. *J. Ambient. Intell. Humaniz. Comput.* **2020**, 1–18. [CrossRef]

8.   Brinton, C.G.; Chiang, M. MOOC performance prediction via clickstream data and social learning networks. In Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM), Kowloon, Hong Kong, 26 April–1 May 2015; pp. 2299–2307.

9.   Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **2016**, *103*, 1–15. [CrossRef]

10. Rodriguez, F.; Lee, H.R.; Rutherford, T.; Fischer, C.; Potma, E.; Warschauer, M. Using clickstream data mining techniques to understand and support first-generation college students in an online chemistry course. In Proceedings of the LAK21: 11th International Conference on Learning Analytics and Knowledge, Irvine, CA, USA, 12–16 April 2021; pp. 313–322.

11. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]

12. Oliva-Cordova, L.M.; Garcia-Cabot, A.; Amado-Salvatierra, H.R. Learning analytics to support teaching skills: A systematic literature review. *IEEE Access* **2021**, *9*, 58351–58363. [CrossRef]

13. Viberg, O.; Hatakka, M.; Bälter, O.; Mavroudi, A. The current landscape of learning analytics in higher education. *Comput. Hum. Behav.* **2018**, *89*, 98–110. [CrossRef]

14. Aljohani, N.R.; Fayoumi, A.; Hassan, S.U. Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability* **2019**, *11*, 7238. [CrossRef]

15. Romero, C.; Ventura, S. Data mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [CrossRef]

16. Calders, T.; Pechenizkiy, M. Introduction to the special section on educational data mining. *ACM Sigkdd Explor. Newsl.* **2012**, *13*, 3–6. [CrossRef]

17. Akram, A.; Fu, C.; Li, Y.; Javed, M.Y.; Lin, R.; Jiang, Y.; Tang, Y. Predicting students' academic procrastination in blended learning course using homework submission data. *IEEE Access* **2019**, *7*, 102487–102498. [CrossRef]

18. Tomasevic, N.; Gvozdenovic, N.; Vranes, S. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Comput. Educ.* **2020**, *143*, 103676. [CrossRef]

19. Mangaroska, K.; Giannakos, M. Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Trans. Learn. Technol.* **2018**, *12*, 516–534. [CrossRef]

20. Aldowah, H.; Al-Samarraie, H.; Fauzy, W.M. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telemat. Inform.* **2019**, *37*, 13–49. [CrossRef]

21. Aleem, A.; Gore, M.M. Educational data mining methods: A survey. In Proceedings of the 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, India, 10–12 April 2020; pp. 182–188.

22. Cano, A.; Leonard, J.D. Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Trans. Learn. Technol.* **2019**, *12*, 198–211. [CrossRef]

23. Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Dis-Cover AI Minds to Preserve Human Knowledge. *Future Internet* **2021**, *14*, 10. [CrossRef]

24. Dutt, A.; Ismail, M.A.; Herawan, T. A systematic review on educational data mining. *IEEE Access* **2017**, *5*, 15991–16005. [CrossRef]

25. Burgos, C.; Campanario, M.L.; de la Peña, D.; Lara, J.A.; Lizcano, D.; Martínez, M.A. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Comput. Electr. Eng.* **2018**, *66*, 541–556. [CrossRef]

26. Kemper, L.; Vorhoff, G.; Wigger, B.U. Predicting student dropout: A machine learning approach. *Eur. J. High. Educ.* **2020**, *10*, 28–47. [CrossRef]

27. Xu, J.; Moon, K.H.; Van Der Schaar, M. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 742–753. [CrossRef]

28. Marbouti, F.; Diefes-Dux, H.A.; Strobel, J. Building course-specific regression-based models to identify at-risk students. In Proceedings of the 2015 ASEE Annual Conference & Exposition, Seattle, WA, USA, 14–17 June 2015; pp. 26–304.

29. Lemay, D.J.; Doleck, T. Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. *Interact. Learn. Environ.* **2020**, 1782–1793. [CrossRef]

30. Park, Y.; Yu, J.H.; Jo, I.H. Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *Internet High. Educ.* **2016**, *29*, 1–11. [CrossRef]

31. Waheed, H.; Hassan, S.U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* **2020**, *104*, 106189. [CrossRef]

32. Behr, A.; Giese, M.; Theune, K. Early prediction of university dropouts—A random forest approach. *Jahrbücher Für Natl. Und Stat.* **2020**, *240*, 743–789. [CrossRef]

33. Helal, S.; Li, J.; Liu, L.; Ebrahimie, E.; Dawson, S.; Murray, D.J. Identifying key factors of student academic performance by subgroup discovery. *Int. J. Data Sci. Anal.* **2019**, *7*, 227–245. [CrossRef]

34. Namoun, A.; Alshanqiti, A. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Appl. Sci.* **2020**, *11*, 237. [CrossRef]

35. López Zambrano, J.; Lara Torralbo, J.A.; Romero Morales, C. Early prediction of student learning performance through data mining: A systematic review. *Psicothema* **2021**, *33*, 456–465.

36. Alamri, R.; Alharbi, B. Explainable student performance prediction models: A systematic review. *IEEE Access* **2021**, *9*, 33132–33143. [CrossRef]

37. Minar, M.R.; Naher, J. Recent advances in deep learning: An overview. *arXiv* **2018**, arXiv:1807.08169.

38. Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access* **2020**, *8*, 55462–55470. [CrossRef]

39. Nahar, K.; Shova, B.I.; Ria, T.; Rashid, H.B.; Islam, A. Mining educational data to predict students performance. *Educ. Inf. Technol.* **2021**, *26*, 6051–6067. [CrossRef]

40. Zollanvari, A.; Kizilirmak, R.C.; Kho, Y.H.; Hernández-Torrano, D. Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access* **2017**, *5*, 23792–23802. [CrossRef]

41. Filvà, D.A.; Forment, M.A.; García-Peñalvo, F.J.; Escudero, D.F.; Casañ, M.J. Clickstream for learning analytics to assess students' behavior with Scratch. *Future Gener. Comput. Syst.* **2019**, *93*, 673–686. [CrossRef]

42. Li, Q.; Baker, R.; Warschauer, M. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *Internet High. Educ.* **2020**, *45*, 100727. [CrossRef]

43. Broadbent, J.; Poon, W.L. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet High. Educ.* **2015**, *27*, 1–13.

44. Jiang, T.; Chi, Y.; Gao, H. A clickstream data analysis of Chinese academic library OPAC users' information behavior. *Libr. Inf. Sci. Res.* **2017**, *39*, 213–223. [CrossRef]

45. Gasevic, D.; Jovanovic, J.; Pardo, A.; Dawson, S. Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *J. Learn. Anal.* **2017**, *4*, 113–128. [CrossRef]

46. Seo, K.; Dodson, S.; Harandi, N.M.; Roberson, N.; Fels, S.; Roll, I. Active learning with online video: The impact of learning context on engagement. *Comput. Educ.* **2021**, *165*, 104132. [CrossRef]

47. Kuzilek, J.; Hlosta, M.; Zdrahal, Z. Open university learning analytics dataset. *Sci. Data* **2017**, *4*, 170171. [CrossRef] [PubMed]

48. Zou, X.; Hu, Y.; Tian, Z.; Shen, K. Logistic regression model optimization and case analysis. In Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 19–20 October 2019; pp. 135–139.

49. Zhou, Q.; Quan, W.; Zhong, Y.; Xiao, W.; Mou, C.; Wang, Y. Predicting high-risk students using Internet access logs. *Knowl. Inf. Syst.* **2018**, *55*, 393–413. [CrossRef]

50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

51. Bader-El-Den, M.; Teitei, E.; Perry, T. Biased Random Forest For Dealing With the Class Imbalance Problem. *IEEE Trans. Neural Networks Learn. Syst.* **2019**, *30*, 2163–2172. [CrossRef] [PubMed]

52. Gupta, A.; Gusain, K.; Popli, B. Verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets. In Proceedings of the 2016 11th International Conference on Industrial and Information Systems (ICIIS), Roorkee, India, 3–4 December 2016; pp. 457–462.

53. Zhu, G.; Wu, Z.; Wang, Y.; Cao, S.; Cao, J. Online purchase decisions for tourism e-commerce. *Electron. Commer. Res. Appl.* **2019**, *38*, 100887. [CrossRef]

54. Vo, C.; Nguyen, H.P. An enhanced CNN model on temporal educational data for program-level student classification. In *Asian Conference on Intelligent Information and Database Systems*; Springer: Cham, Switzerland, 2020, pp. 442–454.

55. Sarkar, M.; De Bruyn, A. LSTM response models for direct marketing analytics: Replacing feature engineering with deep learning. *J. Interact. Mark.* **2021**, *53*, 80–95. [CrossRef]

56. Hung, J.L.; Rice, K.; Kepka, J.; Yang, J. Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content. *Inf. Discov. Deliv.* **2020**, *48*, 199–212. [CrossRef]

57. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.