Applying Predictive Analytics Using Clickstream Data for Improving the Students Performance

Nidhi Sharma[1], Ritik Bhardwaj[2]

[1]Research Scholar, CSE Department SURET Jhunjhunu, India
[2] Research Scholar, IISER Bhopal, India

Abstract:
Student performance analysis is an essential aspect of educational institutions. Machine learning (ML) has emerged as a promising tool for analyzing student performance in recent years. To predict the learning abilities of students and prescribe them a personalized learning curriculum, it is necessary to estimate their behavior to learn about their weaknesses and strengths and help institutions improve enrollment and retention. If the teachers can predict in advance and prescribe ways to the at-risk and dropout students, they can plan more effectively to help them. We are describing in this paper various intelligent tutoring systems with Educational Data Mining, Predictive Learning Analytics, prediction of at-risk students at an earlier basis and how this prediction task is done. Predictive Analytics can also offer insights to help students make informed decisions about each student to improve outcomes by understanding what drives each student's behavior and how much the institution can create intensional, specific plans that will positively impact students.

Index Terms: Predicting Student At- Risk, EDM, Learning Analytics, Fully Connected Neural Network (FCNN), Long Short-Term Memory (LSTM), Virtual Learning (VL), Machine Learning (ML), Deep Learning, Recurrent Neural Network (RNN)

I. INTRODUCTION

The vast amount of educational data available provides opportunities to make use of it for many purposes like tapping the learning behaviors of the stakeholders involved, improving their behavior by resolving their issues, and optimizing the learning environment for them[1]. With the ease of data accessibility, several research communities have shown interest in predicting students' patterns and extracting meaningful insights from them. The new emerging communities are not limiting them to only improving the students' performance. Instead, they also show interest in optimizing the learning environment[2].

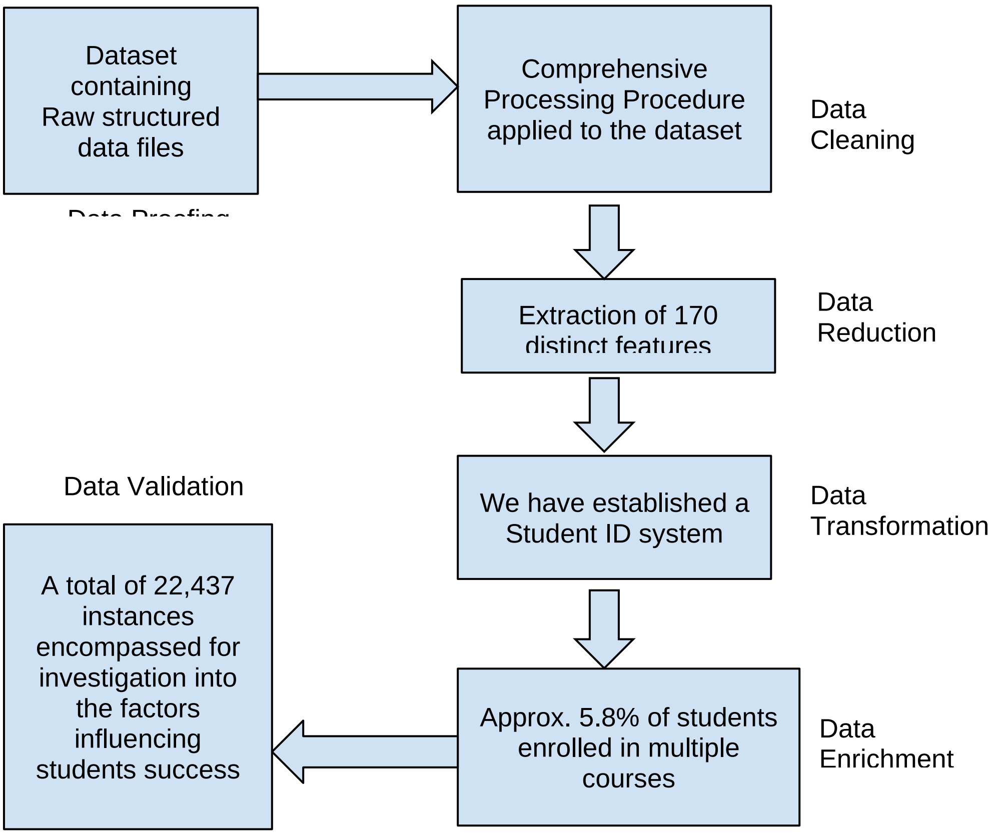
Dataset containing Raw structured data files
Data Proofing
Comprehensive Processing Procedure applied to the dataset
Data Cleaning
Extraction of 170 distinct features
Data Reduction
We have established a Student ID system
Data Transformation
Approx. 5.8% of students enrolled in multiple courses
Data Enrichment
Data Validation
A total of 22,437 instances encompassed for investigation into the factors influencing students success

Figure 1 : Steps performed during the Data Processing

1.1. Data Collection:

The first step is to collect the data. The data collected should be comprehensive and include various attributes that impact student performance, such as socio-economic status, family background, academic history, and personal traits. The data can be collected using multiple methods, including surveys, interviews, and standardized tests.

The present study utilized the OULAD dataset, which consisted of raw structured data files, to investigate the relationship between student interactions with the Virtual Learning Environment (VLE) and their academic outcomes. The log-file data underwent a comprehensive processing procedure, resulting in the extraction of 170 distinct features. These features were derived by carefully analyzing the data tables within the database.

1.2. Data Preprocessing:

Once the data is collected, it must be preprocessed to remove irrelevant or redundant information. The data should also be cleaned to remove any errors or inconsistencies.

In the present study, we have utilized the OULAD dataset, which consists of raw structured data files used to investigate the relationship between student interactions with the Virtual Learning Environment (VLE) and in predicting their academic outcomes. The log-file data underwent a comprehensive processing procedure that resulted in the extraction of 170 distinct features. These features were derived carefully by analyzing the data tables within the database. Notably, the computations were conducted weekly, with each week encompassing a consistent set of students and sharing the same activity features. This ensured a homogenous grouping of students across the weeks, implying that students in a particular week were also present in the preceding weeks.

A student ID system was established to uniquely identify each student within the dataset. However, it was observed that only a tiny fraction, approximately 5.8%, of the dataset represented students who enrolled in multiple courses. Consequently, the focus of the study was directed towards analyzing students' academic performance within a single class, disregarding any repetitions of the same class by individual students. As a result, each student was assigned a unique ID based on their previous ID, the course they undertook, and the specific interval during which the course was offered.

The study's primary objective was to examine instances of 'pass' and 'fail' outcomes. To simplify the classification task, the 'pass' instances were merged with instances indicating a distinction, resulting in a single class for analysis. This consolidation allowed for a more straightforward categorization of student outcomes. Ultimately, the formulated dataset encompassed 22,437 instances, facilitating a comprehensive investigation into the factors influencing student success.

In addition to the VLE interaction data, the study also considered incorporating demographic information. This supplementary dataset contained various attributes crucial to understanding the teenage development of students, including gender, region, age band, and highest education level. To ensure compatibility with the analysis, the demographic data instances were subjected to label encoding. Furthermore, dimensionality reduction techniques were applied to the encoded dataset to mitigate any potential issues arising from high-dimensional data.

1.3. Feature Engineering:

It involves selecting the relevant features from the preprocessed data. The components selected should be those that have a significant impact on student performance. Feature engineering is critical as it determines the effectiveness of the ML model in predicting student performance.

1.4. Model Development:

The next step is to develop the ML model. There are various ML algorithms that can be used, such as linear regression, decision trees, and random forests. The ML algorithm selected should be based on the type of data and the prediction task. The ML model should also be optimized to improve its performance.

1.5. Model Evaluation:

Once the ML model is developed, it must be evaluated to determine its accuracy and effectiveness. The evaluation should be done using a test dataset different from the dataset used to train the model. The evaluation should also use appropriate metrics, such as accuracy, precision, and recall.

II. LITERATURE REVIEW:

A predictive analysis model was developed to identify the non-cognitive skills of the students of the class so that their performance can be improved. Tailored teaching was applied to improve the class performance of the students. There is an increase in the class grade performance of the students after implementing the model. There is a subsequent 9% increase in the student's average performance when the model is applied.[1]

The behavior of students in an online learning environment varies from traditional classroom settings where the learner got the motivation from the computer assisted tools like Excel. This significantly contributed to their Learning, making them solely responsible for their Good or Bad performance.[4]

In the literature, substantial studies emphasize predicting student performance by applying various data analytic techniques and highlighting multiple factors impacting a learner's performance, categorizing different features contributing to low student retention.[5]

In an online educational platform, time is a vital feature for impacting learners and their routines, followed by the incoherent support provided by the instructors. Moreover, effective curriculum content is paramount to driving intrinsic motivation in learners, encouraging them to participate in positive and active participation, ultimately influencing a learner's performance and intention to complete a particular program [6,7].

Another dimension involves predicting students' grades for the next term's potential courses [8,9].

Methods for predicting future course grades should be based on sparse linear models and low-rank matrix factorizations specific to each course or student-course tuple. The procedures were evaluated using a data set from the University of Minnesota. The evaluation showed that the course-specific models outperformed various competing schemes, with the best strategy achieving an RMSE across courses of 0.632 versus 0.661 for the best-competing method.[8]

A systematic performance prediction for students was investigated from machine learning and data mining perspectives. To get a feel for the methods involved, an experiment is conducted on a dataset from the institute and a public dataset. The work provides developments and challenges in the study task of SPP and facilitates the progress of personalized education.[9]

Morsy et al. [10] characterized the course space as a latent vector and proposed regression models for predicting the grade of the following semester.

Similarly, another study predicted course-specific grades for the next semester using Markov models [11]. Marbouti et al. [12] used logistic regression to assess the risk of student failure by including attributes of their attendance, quizzes, and exam behavior. They identified students at risk of doing poorly in multiple weeks of courses, and their predictions improved for the final weeks. In addition, logistic regression was also used as a baseline assessment method to predict at-risk students [13]. Student history, including previous grades in previous courses, assessments, and entrance tests, is also essential in classifying and predicting an individual's performance [14].

The learning analytics community's association with Deep Learning techniques is still in its early stages, and there is little evidence in the existing literature. Deep Learning, which consists of multiple nonlinear layers, enables self-adaptive models through the phenomenon of hierarchical representation, where each layer passes the learned information to subsequent layers in an abstract manner [15].

Corrigan & Smeaton [16] used a variant of recurrent neural networks (RNN) to predict learning success by incorporating student interaction with the VLE. The deep learning approach they used outperformed the traditional baseline approach. Similarly, another study indicated success rate by including student attendance and querying their behavior using log data [17].

They predicted students' grades, including their engagement and interactions with the VLE, by using RNN and LSTM models. Through the sequential model used, they aimed to predict students' grades and identify at-risk students early. The technique used was compared to conventional regression analysis and was more effective than traditional regression approaches in predicting rates early.

Fei and Yeung [18] used a set of features consisting of lectures viewed and downloaded, assessment scores and attempts, forum activity, and the number of comments in forums to predict student performance and rank students at risk. They employed several techniques, including support vector machines, logistic regression, input–output hidden Markov model, RNN, and LSTM, and found that LSTM outperformed the other methods. An exciting dimension in the educational community is the use of virtual reality in the learning context of distance education, which makes the interaction between instructors and learners more convenient and sophisticated [19].

This implicitly helps teachers to adapt teaching methods to different categories of students. In addition, Kohler has provided a theoretical framework for effective teaching to achieve optimal learning outcomes in traditional and blended classroom settings [20].

E-portfolios are another tool that facilitates teachers to personalize their teaching methods and support students' interventions [21].

Overall, we found only a limited number of studies in a VLE that address applying deep learning techniques to assess and understand student behavior rigorously. Our analysis leverages the power of Deep Learning for the early prediction of at-risk student behavior rigorously in a VLE.

III. METHODOLOGY:

The methodology employed in this study encompasses data preprocessing and the implementation of an LSTM-based model for predicting students' outcomes based on their interactions with the Virtual Learning Environment (VLE) and student demographic information.

IV. MODEL ARCHITECTURE:

The LSTM Model, a deep learning architecture, was implemented for predicting student outcomes. This model incorporated an LSTM (Long Short-Term Memory) layer, capable of capturing temporal dependencies within the student interaction sequences with the virtual learning environment (VLE). The LSTM layer was configured with a specified input size, hidden size, number of layers, and number of classes for classification. In addition to the interaction data, the model also considered an additional feature vector of student’s demographic data.

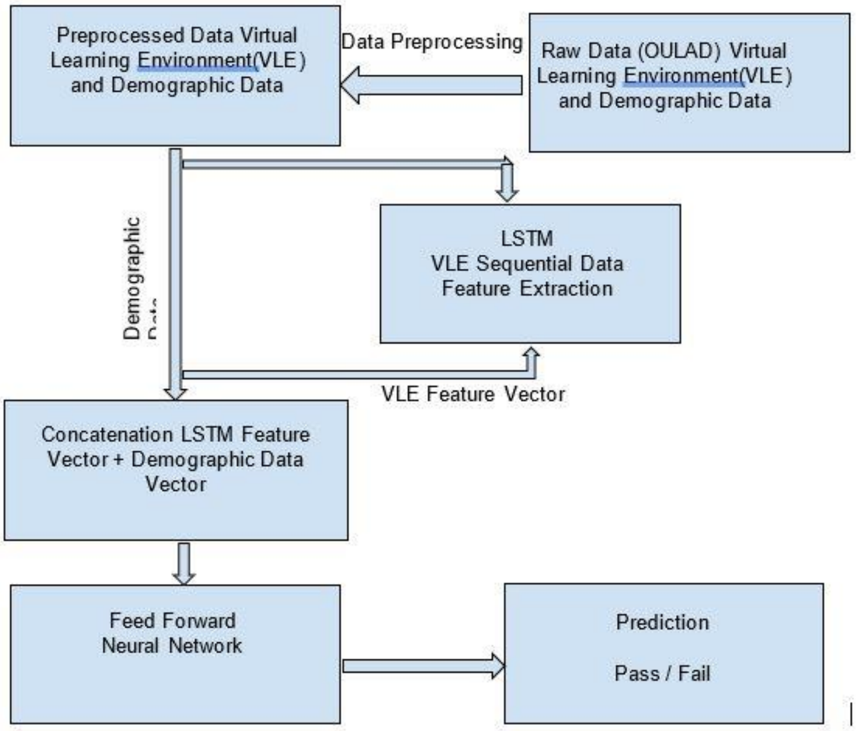
Preprocessed Data Virtual Learning Environment(VLE) and Demographic Data
Data Preprocessing
Raw Data (OULAD) Virtual Learning Environment(VLE) and Demographic Data
Demographic Data
LSTM VLE Sequential Data Feature Extraction
VLE Feature Vector
Concatenation LSTM Feature Vector + Demographic Data Vector
Feed Forward Neural Network
Prediction
Pass / Fail

Figure 2: The Long Short Term Memory(LSTM) Model

The input sequences were first sorted in descending order based on their lengths during the model's forward pass. This sorting allowed efficient processing of variable-length lines within the LSTM. The additional feature vector was aligned with the sorted inputs. Packed padded sequences were then generated to handle variable-length sequences in the LSTM. The LSTM layer was initialized with zero tensors for the hidden state and cell state, and the packed inputs were passed through the LSTM. The output was then unpacked and restored to the original order using the sorted indices.

The final step involved combining the LSTM output (feature vector returned from the virtual learning environment data) with the additional feature vector (demographic data feature vector). The last hidden state of the LSTM output sequence was extracted and concatenated with the additional features. This combined feature vector (VLE obtained feature vector and demographic data feature vector) was fed into a fully connected neural network (FCNN) consisting of multiple layers with ReLU activation functions. The FCNN gradually reduced the dimensionality of the features, ultimately producing a single output value using a sigmoid activation function. This output represented the probability of student success.

In summary, the methodology comprised the preprocessing of the OULAD dataset, including the computation of interaction features and dimensionality reduction of the demographic data. Subsequently, an LSTM Model was constructed, leveraging an LSTM layer to capture temporal dependencies in the student interaction sequences. The model incorporated additional features and utilized a fully connected neural network for classification. This methodology forms the foundation for analyzing student outcomes based on VLE interactions and demographic information.

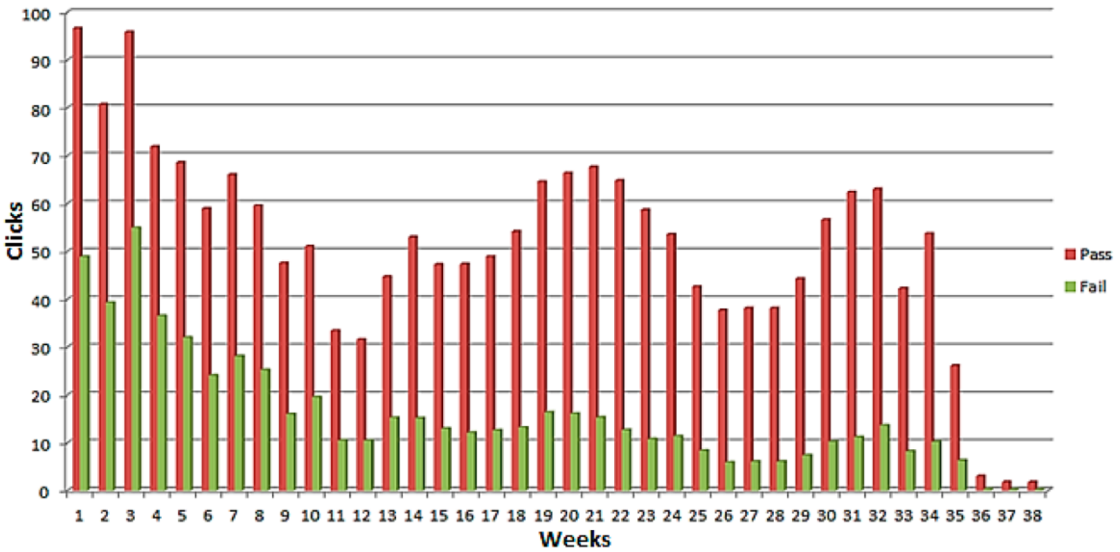

Figure 3 : Depicts the number of clicks for the two classes—pass and fail—where the aggregated activities for each class were normalized according to per student.[3]

The aggregated average clicks per student were processed weekly to visualize the students' weekly interactions. It can be observed that the two classes demarcated in terms of the interaction level in the VLE, with 'pass' instances being more active than ‘fail’ instances.[3]

V. RESULTS AND DISCUSSIONS:

The developed model demonstrates a noteworthy ability to forecast future student outcomes by leveraging a combination of demographic information and data from the Virtual Learning Environment

(VLE). With an accuracy rate of approximately 75%, the model outperforms traditional statistical methods that solely rely on demographic data, yielding an improvement of approximately 6%.

By incorporating both demographic and VLE data, the model benefits from a more comprehensive understanding of the factors influencing student performance. Demographic information, including attributes such as gender, region, age band, and highest education level, provides valuable insights into the individual characteristics and backgrounds of students. On the other hand, VLE data captures students' nuanced interactions and engagement within the virtual learning environment, offering a glimpse into their learning patterns and behaviors.

Integrating these two data sources allows the model to consider various factors impacting student outcomes. By taking into account both the personal attributes of students and their interactions with the VLE, the model gains a more holistic perspective on the complex dynamics influencing academic success.

In terms of accuracy, the model's performance surpasses that of traditional statistical approaches that solely rely on demographic data. These methods typically achieve an accuracy rate of around 69%. Therefore, the developed model demonstrates a clear advantage, exhibiting an improvement of approximately 6% in predictive accuracy.

The enhanced accuracy of the model highlights the value of incorporating VLE data alongside demographic information. The VLE data provides valuable insights into student engagement, participation, and utilization of online learning resources. By capturing these interactions, the model is better equipped to identify patterns and trends contributing to successful academic outcomes.

In summary, the model's ability to predict future student outcomes is significantly enhanced by integrating demographic information with VLE data. Including VLE data offers a more comprehensive understanding of student engagement and learning behaviors, resulting in a predictive accuracy of around 75%. This marks a notable improvement of approximately 6% compared to traditional statistical methods solely relying on demographic data.

VI. References

1. Yi, C.; Kang-Yi, C. Predictive analytics approach to improve and sustain college students' non-cognitive skills and educational outcomes. Sustainability 2018, 10, 4012. [CrossRef]
2. Schumacher, C.; Ifenthaler, D. Features students really expect from learning analytics. Computer Human Behavior 2018, 78, 397–407. [CrossRef]
3. Naif Radi Aljohani, Ayman Fayoumi, and Saeed-Ul Hassan, In Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment.December 2019.
4. Davis, H.C.; Dickens, K.; Leon Urrutia, M.; Vera, S.; del Mar, M.; White, S.MOOCs for Universities and Learners: an Analysis of Motivating Factors. In Proceedings of the 6th International Conference on Computer Supported Education, Barcelona, Spain, 1–3 April 2014.
5. Hone, K.S.; El-Said, G.R. Exploring the factors affecting MOOC retention: A survey study. *Comput. Educ.* **2016**, *98*, 157–168. [CrossRef]
6. Fidalgo-Blanco, Á.; Sein-Echaluce, M.L.; García-Peñalvo, F.J.; Conde, M.Á. Using learning analytics to improve teamwork assessment. *Comput. Hum.Behav.* **2015**, *47*, 149–156. [CrossRef]
7. Khan, I.U.; Hameed, Z.; Yu, Y.; Islam, T.; Sheikh, Z.; Khan, S.U. Predicting the acceptance of MOOCs in a developing country: Application of task-technology fit model, social motivation, and self-determination theory. *Telemat. Inform.* **2018**,*35*, 964–978. [CrossRef]
8. Polyzou, A.; Karypis, G. Grade prediction with course and student specific models. In *Pacific-Asia*

Conference on Knowledge Discovery and Data Mining; Springer: Cham, Germany, 2016; pp. 89–101. 6. Baker, R.S.; Inventado, P.S. Educational data mining and learning analytics. In *Learning Analytics*; Springer: New York, NY, USA, 2014; pp. 61–75.

9. Bydžovská, H.A., Comparative Analysis of Techniques for Predicting Student Performance. In Proceedings of the 9th International Conference on Educational Data Mining 2016, Raleigh, NC, USA, 29 June–2 July 2016.
10. Morsy, S.; Karypis, G. Cumulative Knowledge-based Regression Models for Next-Term Grade Prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining, Houston, TX, USA, 27–29 April 2017.
11. Hu, Q.; Rangwala, H., Course-Specific Markovian Models for Grade Prediction.Advances *in Knowledge Discovery and Data Mining. PAKDD 2018. LectureNotes in Computer Science*; Phung, D., Tseng, V., Webb, G., Ho, B., Ganji, M.,Rashidi, L., Eds.; Springer: Cham, Germany, 2018; Volume 10938, pp. 29–41.
12. 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]
13. Marbouti, M.F.; Diefes-Dux, H.A. Building course-specific regression-based models to identify at-risk students. *Age* **2015**, *26*, 1.
14. Leitner, P.; Khalil, M.; Ebner, M. Learning analytics in higher education—A literature review. In *Learning Analytics: Fundaments, Applications, and Trends*; Springer: Houston, TX, USA, 27–29 April 2017; pp. 1–23.
15. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
16. Corrigan, O.; Smeaton, A.F. A Course Agnostic Approach to Predicting Student Success from VLE Log Data Using Recurrent Neural Networks. In Proceedings of the European Conference on Technology Enhanced Learning, Tallinn, Estonia, 12–15 September 2017; pp. 545–548.
17. Okubo, F.; Yamashita, T.; Shimada, A.; Ogata, H. A Neural Network Approach for Students' Performance Prediction. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13–17 March 2017.
18. Fei, M.; Yeung, D.Y. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic, NJ, USA, 14–17 November 2015.
19. Klampfer, A. Virtual/Augmented Reality in Education Analysis of the Potential Applications in the Teaching/Learning Process. Available online: https://www.researchgate.net/publication/318680101_VirtualAugmented_Reality_in_Education_Analysis_of_the_Potential_Applications_in_the_TeachingLearning_Process (accessed on 28 November 2019).
20. Gettinger, M.; Kohler, K.M. Process-outcome approaches to classroom management and effective teaching. In *Handbook of Classroom Management*; Routledge: Abingdon, UK, 2013; pp. 83–106.
21. Klampfer, A.; Köhler, T. Learners' and teachers' motivation toward using e-portfolios. An empirical investigation. *Int. J. Cont. Eng. Educ. Life-Long Learn.* **2015**, *25*, 189. [CrossRef]