Transforming Learning Data into a Machine Learning Model to Help STEM Students Transition to University

Neeraj Katiyar, Armin Yazdani, Janette Barrington, Kira Smith, Valerie Bourassa, Hilary Sweatman and Marcy Slapcoff

May 19, 2023

International Society of the Learning Sciences

Transforming Learning Data into a Machine Learning Model to Help STEM students Transition to University

Neeraj Katiyar, Armin Yazdani, Janette Barrington, Kira Smith, Valerie Bourassa, Hilary Sweatman, Marcy Slapcoff
neeraj.katiyar, armin.yazdani, janette.barrington, kira.smith, valerie.bourassa, hilary.sweatman, marcy.slapcoff{@mcgill.ca}
Office of Science Education, McGill University

Abstract: Previous research has applied Machine Learning (ML) to predict student success in higher education using entry data and cumulative GPA scores. Our research aims to add student learning and performance data in specific STEM courses to the ML modelling process. In the initial phase, the data included self-report scores on inventories that assess students’ learning strategies, metacognitive awareness, mindset, and misconceptions about how the brain works, as well as learning analytics and course grades. This data is collected as part of an orientation program that aims to develop students’ self-regulated learning capabilities. Our goal is to provide evidence to inform this program, use the results to predict student success and challenges in first-year STEM courses, and inform proactive help for students’ transition to university. This paper provides a step-by-step introduction to the methodology used to build a prototype of the ML model underpinning this research and future directions.

Keywords: Machine Learning, Students' Performance Prediction, STEM Education

Introduction

Learning is known to be challenging and stressful for incoming undergraduate STEM students especially given the sheer volume of content to be mastered and the limited capacity of the human brain [1]. The McGill Office of Science Education supports students as they transition to university-level learning with an orientation program (SciLearn) first implemented in Fall 2020. SciLearn uses insights from the learning sciences, specifically from neuroscience, to help students gain awareness about how their brain works and how to become self-regulated learners (Yazdani et al, manuscript in preparation). For example, students learn how their current study behaviors (e.g., deliberate practice and peer collaboration) and lifestyle habits (e.g., exercise, relaxation, and sleep) may impact their future academic performance. Inspired by a citizen science framework, the program engages students as active participants and collaborators in using a wide range of self-collected data to understand how they can evolve as learners. The program is designed and facilitated by neuroscientists and education specialists and comprises a series of synchronous and asynchronous workshops (also referred as SciLearn lab), peer collaboration sessions, and special events. Since its inception, a significant amount of student learning data has been generated.

In the initial phase of our research, we have tracked students as they progressed through the SciLearn program in two large introductory science courses and collected self-report scores on inventories that assess their learning strategies, metacognitive awareness, mindset, and misconceptions about how the brain works. This baseline data was collected right before completion of the program together with demographic data, a measure of student progression in the course learning management system, and course grades. Our data inventory so far includes over 400 students in an introductory organic chemistry course and over 500 students in an introductory psychology course. Early results lay a strong foundation for further research to improve the training and testing accuracy of a machine learning predictive model.

Machine Learning (ML) is a promising tool for analyzing complex patterns and recent research shows its potential in helping students become self-regulated learners [2;8]. Previous research has applied ML to predict student performance in higher education using entry data and cumulative GPA [3]. This study aims to add the additional features described above to the ML modelling process towards providing evidence to inform the SciLearn program, using the results to predict student success and challenges in first-year STEM courses, and helping instructors and educational specialists identify early markers of risk and intervene where necessary to prevent students from having poor learning outcomes. The research question guiding our work is: *How can we leverage attributes that are highly correlated with students’ academic success to help early undergraduates become self-regulated learners and educators identify risks?*

Related Work

Much research has been done in the area of educational mentor support where a predictive model is built to forecast student performance to identify at-risk students as well as contributing features in the course. This research is taking place because of the complex nature of learning, e.g., performance depends on many characteristics related to the learner. Possible characteristics include the student's recent academic assessments, demographics, psychological profile, culture, educational background [3], and engagement with course content. Demographic factors consist of family background, gender, disability and age, and all of these are considered important attributes [4]. If we look at academic progress, the student's grade is among the most important attributes that can be used to assess performance in a specific course [5], whereas academic potential can be evaluated by the student's GPA especially during transition to university education. Our research introduces a few new attributes that focus on using descriptive features from the learning sciences related to study behaviours as well as content engagement and their mutual effect on performance. Different ML and data mining techniques used to predict student performance also relate to our work, such as:Support Vector Machine (SVM); Logistic Regression(LR); Decision Tree (DT); Random Forest (RF); and Bayesian Network (BN). Following (see Table 1) shows a summary of a few relevant research papers.

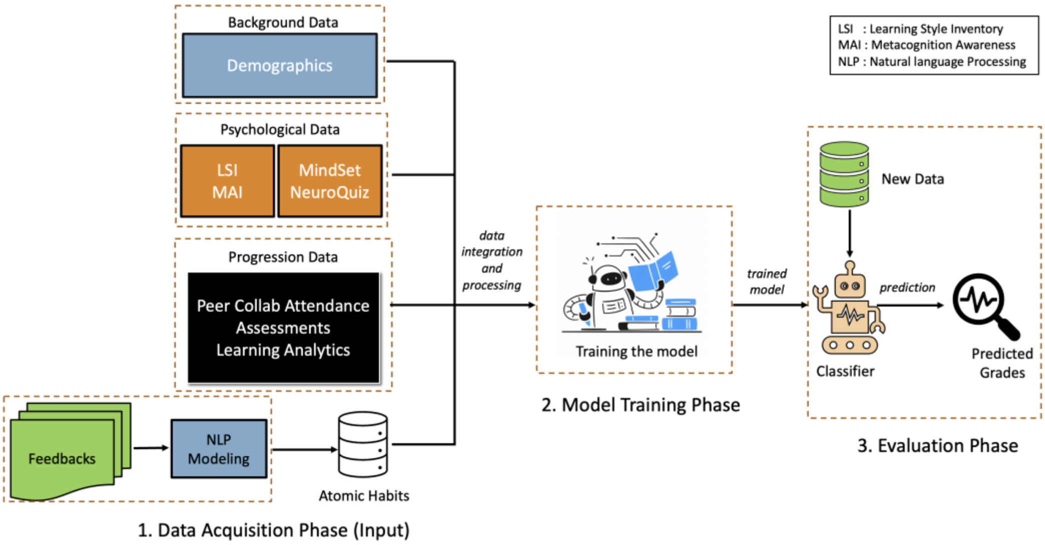
Table 1:
Relevant research papers

Paper	Attributes	ML Model	Best Model
Alharbi et al, 2016 [7]	Student demographics, general performance, students' modules	LR, DT and Ensemble approach	No overall winners
Gray et al, 2014 [6]	Aptitude, Personality, 914 Motivation Learning strategies	DT, LR,SVM	SVM
Guleria et al, 2014 [5]	Class Performance, Attendance, Assignment, Lab Work, Sessional Performance	DT	DT

Machine Learning Methodology

As mentioned, our research focused on first and second term undergraduate students registered in two STEM courses. With the newly added features, study behaviour and content engaggement, discussed above in the related work section, we are building a protoype ML model from two perspectives of patterns: classification vs. time series. In this paper we have discussed the classification approach as the later is still in progress. Our methodology involved the normalization of data, correlation of study features, and generation of predictive models (classification and time series) with predicted grades.

Figure 1
System model for predictive modeling



Experiment System Model

Above figure 1 depicts the main steps and components of the proposed experimental ML model for predicting student success and challenges in first-year STEM courses. The system architecture is divided into three phases *Data Acquisition*, *Modeling*, and *Evaluation.* Next, we briefly discuss each phase.

Data Acquisition:

This phase involves collection of raw data from various sources. We grouped the data into four categories with related attributes and sources (see Table 2). Primarily, the data were collected through surveys using McGill's learning management system portal. This first stage of our research uses data from over 400 students enrolled in either Organic Chemistry (Fall 2021) and (Winter 2022).

Table 2:
Data Attributes and Source

Data Category	Attributes Collected	Data Source
Demographics (Background)	Background data: Educational background, gender, first-generation, disability etc.	Survey
Learning Inventory Data (Psychological profile)	Learning Strategies Inventory (LSI), Metacognitive Awareness Inventory (MAI), Mindset score, Neuromyths score.	Survey
Progression Data	SciLearn Peer collaboration and lab attendance, grades on assignments(assesments), learning analytics (content engaggement) progression data	McGill Learning Management System
Atomic Habits Adopted	Atomic Habits (a series of evidence-based habits from neuroscience: napping, scheduling free time, notes taking, avoiding multitasking and peer learning)	Feedback text McGill Learning Management System (myCourse portal)

Model Training:

This is known as the Machine Learning stage and involves data cleaning (removal of balnk values and noisy data), normalization (scailing the feature values in a specific range to better analyze the patterns), one-hot encoding (converting texts into numbers in order to convert the data for training compatible), feature engineering (addition or substraction of features based on correlation), exploration analysis (in depth understanding of data distribution), and modeling. Next section will discuss these operations in detail.

Model Performance Assesment (Evaluation):

Once our model was trained, we proceeded to assess its performance based on appropriate metrics discussed with domain experts and suitable for classification models. The domain experts are the course instructors, education developers, and neuroscientists. A number of performance metrics have been proposed in different application scenarios. For example, accuracy is typically used to measure the percentage of correctly classified test instances. It is so far the primary metric for assessing classifier performance [13] and [14]; along with precision and recall metrics which are widely used in information retrieval [15]; Following (see Table 3) shows the metrics used to evaluate the performance of our model.

Table 3:
Metrics to assess the performance of ML model

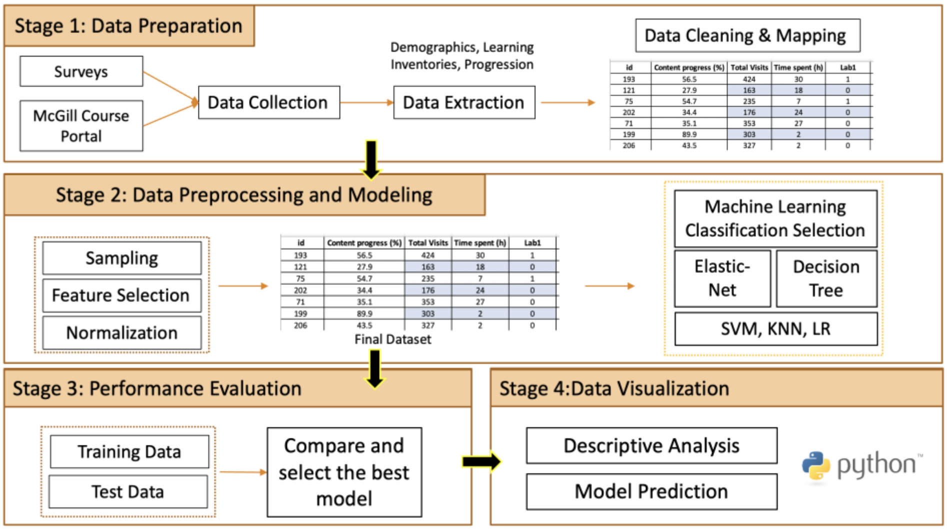
Metrics	Description
Accuracy	How many times the ML model was correct overall
Recall	Model's ability to detect positive samples.
Precision	How good the model is at predicting a specific category

In the following section author discusses the technical architecture of the model which combines the *Modeling* and *Evaluation* phases from the system model (see Figure 1).

Technical Architecture:

The technical implementation of our ML model is divided into three stages as shown below (see figure 2). Further section discuss each stage in detail.

Figure 2:
Technical model for predictive modeling


Stage 1: Data Preparation
Surveys
McGill Course Portal
Data Collection
Demographics, Learning Inventories, Progression
Data Extraction
Data Cleaning & Mapping
Stage 2: Data Preprocessing and Modeling
Sampling
Feature Selection
Normalization
Final Dataset
Machine Learning Classification Selection
Elastic-Net
Decision Tree
SVM, KNN, LR
Stage 3: Performance Evaluation
Training Data
Test Data
Compare and select the best model
Stage 4:Data Visualization
Descriptive Analysis
Model Prediction
python™

id	Content progress (%)	Total Visits	Time spent (h)	Lab1
193	56.5	424	30	1
121	27.9	163	18	0
75	54.7	235	7	1
202	34.4	176	24	0
71	35.1	353	27	0
199	89.9	303	2	0
206	43.5	327	2	0

id	Content progress (%)	Total Visits	Time spent (h)	Lab1
193	56.5	424	30	1
121	27.9	163	18	0
75	54.7	235	7	1
202	34.4	176	24	0
71	35.1	353	27	0
199	89.9	303	2	0
206	43.5	327	2	0

Stage 1: Data Preparation

Once the data is collected from different sources, the next steps include cleaning the data (e.g., null value removal), mapping features to numerical values to make it compatible with the ML model, and data labeling.

Stage 2: Data Processing and Modeling

Once the data is labeled, the next step is to sample and extract the relevant features from the cleaned dataset. We conducted a Pearson correlation analysis on the selected features to understand the important attributes. The next step is to normalize the data to ensure the feature values fall within a confined range. This step is helpful during the data visualization step if the value range of two features is extremely high. Now the data is ready to feed into the ML model for training.

Classification is one of the most popular techniques used in predicting students academic performance. There are many classification methods used for this prediction. Among those we used in the current study are Multinomial LR, DT, SVM), Elastic-net Classifier(EC) and RF. Following is a brief description of these models:

- *Decision tree* are often used due to its clarity and simplicity in discovering and predicting data. Many scholars found that decision trees can be easily understood since it's based on IF-THEN rules [11].
- *Support vector machine (SVM)* is good for handling a small dataset and has a greater generalization ability compared with other methods [6].
- *Elastic-net Classifier* are extremely scalable and require several linear attributes to learn a problem. We found five articles that have applied the Naïve Bayes method in predicting the student's academic performance.
- *Random Forest* stores and classifies classes based on certain measure of similarity such as distance function and is ensemble classification technique [12].
- *Multinomial Logistic Regression:* Modified version of logistic regression that predicts a multinomial probability (i.e. more than two classes) for each input example.

Stage 3 & 4: Performance Evaluation and Visualization

As discussed by Muraina in [9], a 80%-20% train-test split is feasible and effective to capture the possible patterns especially with less data. Hence, we have used 80% of our data for training with random sampling while 20% is used to evaluate the model's performance. As referred in the (see Table 3). We have used accuracy, precision, and recall as the base for the evaluation of our models. Once the model is trained, descriptive analysis [10] has been used for indepth understanding of patterns, inferences and correlations as discussed in the next section.

Results and Key Observations

Next, in this section author further discusses the statistical analysis and experimented ML model's performance.

Correlation analysis:

For statistical analysis we have used pearson correlation coefficient score (a value which reflects the corrospondance of one variable with other, ranges from 0-1) to understand the relationship between input features and the target category (grades). The value closer to 1 reflect high correspondence with the target feature. The analysis have shown a positive correlations of five main attributes; Scilearn lab attendance, internal assesment (recent grades), background, learning profile, and the adoption of atomic habits. While learning profile (LSI, Mindset and MAI attributes), and progression data (assesments and SciLearn lab attendance) have shown the strongest correlation with grades as shown below (see Table 5).

Table 5
High impacting features with the grades

Data Category	Features	Correlation score with Grade
Learning Inventory (Psychological profile)	LSI	0.81
	Mindset	0.85
	MCAI Knowledge	0.76
	MCAI Regulation	0.72
Progression Data	Assesments	0.89
	Scilearn Lab attendance	0.72

Performance Analysis:

Five classification models have been created and tested using five ML techniques, MLR, RF, SVM, EC and DT. Results (Table 4) demonstrate the accuracy and performance measures for each model. As shown, the MLR model has the lowest accuracy index, equal to 58.2, with the highest error of 41.8. While DT, RF and EC models perform average with the mean accuracy of ~63%. So far, the most accurate model is built on the SVM (with high correlated data), which has an accuracy of 78% with an error index of 22%. The interesting and positive observation is; when the SVM model is trained on all the features, the accuracy drops to 62.3% from 78%, which is a validation of our correlation analysis. Five main features influence the classification decision based on our correlation exploratory analysis discussed in the above section: SciLearn lab attendance, assessments, background, learning profile and specific atomic habits referenced before (see Table 2). Therefore, in the next phase of research with new cohort data, our focus will be more on these attributes, and we will also include learning analytics (see Table 2) to capture the predictions from a time series correlation perspective.

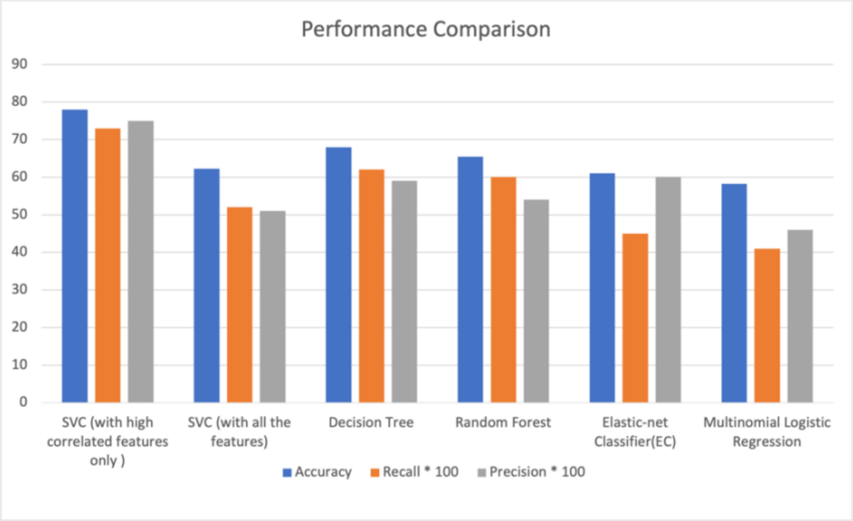
Table 4
Model Comparison

Model	Accuracy	Recall	Precision
Support Vector Classifier	78%	0.73	0.75
Decision Tree	68%	0.62	0.59
Random Forest	65.5%	0.60	0.54
Elastic-net Classifier(EC)	61%	0.45	0.60
Multinomial Logistic Regression	58.2%	0.41	0.46

As referenced in the next section, limitation, below; the low number of data points could potentially introduce biases in the prediction (low recall and precision of the high accuracy) as evident from EC model (see Table 4) which also presents an opportunity for performance improvements with more datasets. Hence, we are continuing our analysis with more recent cohort student data (Fall 2022 and Winter 2023 for introductory psychology course). Also, looking at the modelling from a time series perspective with more time-dependent features such as learning analytics and SciLearn peer collaboration attendance referenced before (see Table 2).

Figure 3

Graphical representation and comparison of model's perfromance



Limitations

This was a pilot study hence we acknowledge current data restrictions and future research is planned to increase the number of students enrolled to train our models and increase reliability. We also rely on self-reported data that can be problematic. Although the SciLearn program is incentivized with a small bonus grade, participants are self-selected and therefore not fully representive of McGill's first-year cohort. The courses in which data were collected are also taught by instructors known to be excellent teachers, as our sample grows other features related to course design (e.g., type of assessments) could affect the ML modeling process.

Conclusion and Future Work

Student performance prediction and feature analysis are essential, especially for incoming undergraduates transitioning to university. This kind of analysis can identify students at risk and help educators improve their performance. Students can also enhance their own learning and become self-regulated learners by understanding how their current study and life habits predict their future performance. This research aimed to build ML models using students' demographic background, learning attributes, early academic performance, and learning analytics. Our initial analysis has shown a high correlation among early learning attributes and grades as discussed in the results section. Also, our initial classifier has shown an accuracy of 78%, which gives us a promising direction for future analysis with more time-dependent features. These early results also validate the correlation between metacognitive awareness and learning strategies and lay a strong foundation for further research with a larger dataset to improve our ML model training and testing accuracy. One additional avenue we are exploring is to look at student attributes and habits as they evolve using time series modelling. We aim to identify promising features early in the academic journey which correlate highly with a student's path towards success and to provide timely guidance, when required.

References

[1] Wieman C., (2007). Why Not Try a Scientific Approach to Science Education?. Change: The Magazine of Higher Learning. 39. 9-15. 10.3200/CHNG.39.5.9-15.

[2] Moon-Heum Cho & Jin Soung Yoo (2017). Exploring online students' self-regulated learning with self-reporte reported surveys and log files: a data mining approach. Interactive Learning Environments, 25(8), 970-982.

[3] K. P. Shaleena and S. Paul (2015), "Data mining techniques for predicting student performance," in ICETECH 2015 - IEEE International Conference on Engineering and Technology, no. March, pp. 0–2.

[4] A. M. Shahiri., Wahidah H., (2015) "A Review on Predicting Student's Performance Using Data Mining Techniques," in *Procedia Computer Science*.

[5] P. Guleria, N.,Thakur N., and Sood M. (2015), "Predicting student performance using decision tree classifiers and information gain," *Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC*, 126–129.

[6] Gray, C. McGuinness, et al, “An application of classification models to predict learner progression in tertiary education,” in *Souvenir of the 2014 IEEE International Advance Computing Conference, IACC*

[7] Z. . Alharbi, J. et al (2016), “Using data mining techniques to predict students at risk of poor performance,” *Proc. 2016 SAI Comput. Conf. SAI 2016*, pp. 523–531.

[8] Francisco, M., Amado,C. (2021). Perusall’s Machine Learning Towards Self-regulated Learning. International Conference on Innovative Technologies and Learning.

[9] Muraina, Ismail, (2022). “Ideal Dataset Splitting Ratios In Machine Learning Algorithms: General Concerns for Data Scientists & Data Analyst” -7th International Mardin Artuklu Scientific Research Conference

[10] Khadem Charvadeh, (2020). Data Visualization and Descriptive Analysis for Understanding Epidemiological Characteristics of COVID-19: 18. 10.6339/JDS.202007_18(3).0018.

[11] M. M. Quadri, et al (2010) "Drop out feature of student data for academic performance using decision tree techniques," Global Journal of Computer Science and Technology.

[12] Sun, Wang, H., Xue, B., Jin, Y., Yen, G. G., & Zhang, M. (2019) Surrogate-assisted evolutionary deep learning using an end-to-end random forest-based performance predictor. IEEE Transactions on Evolutionary Computation, 2019. 24(2): p. 350-364.

[13] Ben-David, A. (2007). A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence, 20(7), 875–885. doi:10.1016/j.engappai.2007.01.001

[14] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 17(3), 299–310.

[15] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Moderninformationretrieval (Vol. 463). New York: ACM press.

Acknowledgment

The Office of Science Education is grateful for the continued support of the DGDM Family Foundation, which enables us to promote student learning through the SciLearn program.