

Data Mining Techniques to Improve Predictions Accuracy of Students Academic Performance: A Case Study with Xorro-Q

Gomathy Ramaswami^{1*}, Teo Susnjak¹, Anuradha Mathrani¹, James Lim² and Pablo Garcia³

¹Institute of Natural and Mathematical Science, Massey University, Auckland, New Zealand

²Department of Civil Engineering, University of Auckland, New Zealand

³ Xorro Solutions Limited, Auckland, New Zealand.

Corresponding Author email: *ramaswamigomathy@gmail.com

Abstract: Recent research in analytics has assisted in making data-driven predictions to help policy makers create a vision for developing strategies for achieving their business objectives. This is especially relevant in educational environments where predictions around students' academic performance can help educators provide suitable pedagogical support to students and align a teaching strategy which encourages and assists students with their learning. This study uses extensive data collected from an educational tool (Xorro-Q) used in an engineering course delivery. Xorro-Q is an audience interaction tool which facilitates synchronous and asynchronous interactions with the speaker (or, teacher in this case) and the audience (or students). Data collected from Xorro-Q comprised student scores from real-time and self-paced activities set by educators over a 12-week semester period along with students' final exam scores and scores from a compulsory prerequisite course. Data mining techniques have been applied to predict the academic performance of students based on data extracted from Xorro-Q. This is done by training the classifier using four different algorithms, namely, Naive Bayes, Logistic regression, K nearest neighbor and Random Forest individually and the performance of each classifier was tested using test data. To increase the accuracy further, the Process Mining was applied to the event logs which were derived from the dataset of Xorro-Q and process models for high-risk, medium-risk and low-risk students were developed. Process conformance checking was applied which takes a model one at a time and the event log as input, and returns a set of differences between the behavior captured in the process model and event logs. The fitness score which was obtained from the conformance checking was applied along with the general features to find out the effectiveness, such as improvement in accuracy. Results show that when general features derived from Xorro-Q are combined with the process mining features, there is an improvement in accuracy of predictions made. Among the four algorithms tested, the study shows Random Forest classifier has better accuracy compared to other three algorithms in a statistically significant way on this dataset. The study has further implications in enhancing value of the role of analytics for predictive modelling by incorporating process mining features in the training set of data.

Keywords: Education tool; Data mining techniques; Multiple classifiers; Accuracy; Conformance checking.

1. Introduction

There is much concern among the educational sector regarding poor academic performance of students leading to them not being able to meet the minimum course requirements. Many research studies have been conducted in recent years to find factors which contribute to low performance of students across different education levels [12]. With rapid advances in information and communication technologies, huge amount of student online data is continually being added to educational databases over their period of study. By leveraging analytics, this dataset can be analyzed to derive data-driven insights on how to resolve these concerns. Analytic techniques used for knowledge discovery with education-related databases are referred to as education data mining (EDM). EDM aids in realization of hidden features and patterns within the massive amounts of raw data to further enable making predictions on final outcomes and behavioral aspects of the subject under study. This includes processes like machine learning, statistical and visualization techniques to examine and present resulting knowledge in a form that can be easily comprehended by humans [14].

EDM is an emerging stream of data mining research. It is defined as the implementation of data mining methods and tools for analyzing educational data. Therefore, in this field of study, researchers analyze academic databases to recognize unique patterns emerging from data and develop data-driven strategies. From the knowledge thus obtained, educational institutions can increase their understanding of student behaviors, thereby enabling them to provide proper support to students resulting in overall improvements in their academic performances and reduction in failure rates. This will further lead to improvement in teaching pedagogies in educational institutions [13].

This study uses extensive data collected from an educational tool (Xorro-Q) used during an engineering course delivery. Xorro-Q is an audience interaction tool which facilitates synchronous and asynchronous interactions with the speaker (or, teacher in this case) and the audience (or students). Data collected from Xorro-Q comprised student scores from real-time and self-paced activities set by educators over a 12-week semester period along with students' final exam scores and scores from a compulsory prerequisite course. The study compares four algorithms, namely, Random Forest (RF), Naive Bayes (NB), K- Nearest Neighbor (k-NN) and Logistic Regression (LR) to gather insights from the student data. Four algorithms have been used to train the model separately and performance of each classifier for predicting the student's academic performance was tested using the test data. Process conformance checking fitness score was also added along with the general features to check whether the integration of process mining features helped in improvement of the accuracy of predictions on students' performances.

2. Research Questions

The paper addressed the following research questions posed in this study

- Can the student's performance be predicted based on the data gathered from the educational tool (Xorro-Q)? If so, which algorithm gives better accuracy and how early can we reliably predict students' outcomes?
- Does the use of Process Mining features help to increase the prediction accuracy?
- If a student got above 70% in the activity score in first two attempts, would it be possible for the student to get the same score in exam?

3. Related Works

EDM has been applied mostly to higher education levels, such as high school and university level for making predictions on student performances [4]. Carlos, Alberto and Romero (2012) [3], used real data of 670 high school students to predict the school dropout rate using some algorithms. Different classification approaches were used for predicting end-of-the-course student performance. Authors propose the use of a genetic programming model for getting highly accurate and all-inclusive classification rules. Their study showed improvement in accuracy by selecting the best attributes, cost sensitive classification and data balancing methods.

Abdous, He and Yen (2012) [1] performed a study of EDM and regression analysis was used to analyze live video streaming to gather insights on students' online behavior and compare it with their performance in the course. A total of 298 students were included in this study, however, no correlation was found between online behavioral aspects like students' questions, chat messages total login times and students' success.

Kaur, Singh and Josan (2015) [6] aimed to identify slow learners' using a dataset of 152 students' academic records. Data were tested with various classification based algorithms like multilayer perception, SMO, Naive Bayes, J 48 Reptree etc. Five classifiers were compared to find the best classification algorithm, and the study identified multilayer perception to be the best classifier with an accuracy of 75 %. In yet another study, authors Kotsiantis and Pintelas (2005) [9] used state-of-the-art regression algorithm to predict students' marks. Six algorithms were tested and their study indicated the most accurate machine learning algorithm for predicting poor student performance to be Naive Bayes classifier.

Dorina (2013) [5] undertaken a study to find whether university level students' performance can be predicted based on their personal and pre-university characteristics. WEKA software was used in this study implementing CRISP –DM model [cross industry standard process for datamining] with a dataset of 10330 students. This study concluded that Decision tree classifier J 48 performed best followed by rule learner JRP and KN classifier but overall accuracy of all classifiers was less than 70%. Also, the classifiers performed differently for different classes of students like excellent, bad etc. Among the factors studied, study stated university admission score and number of failures to be important in judging university performance.

4. Dataset Features and Descriptions

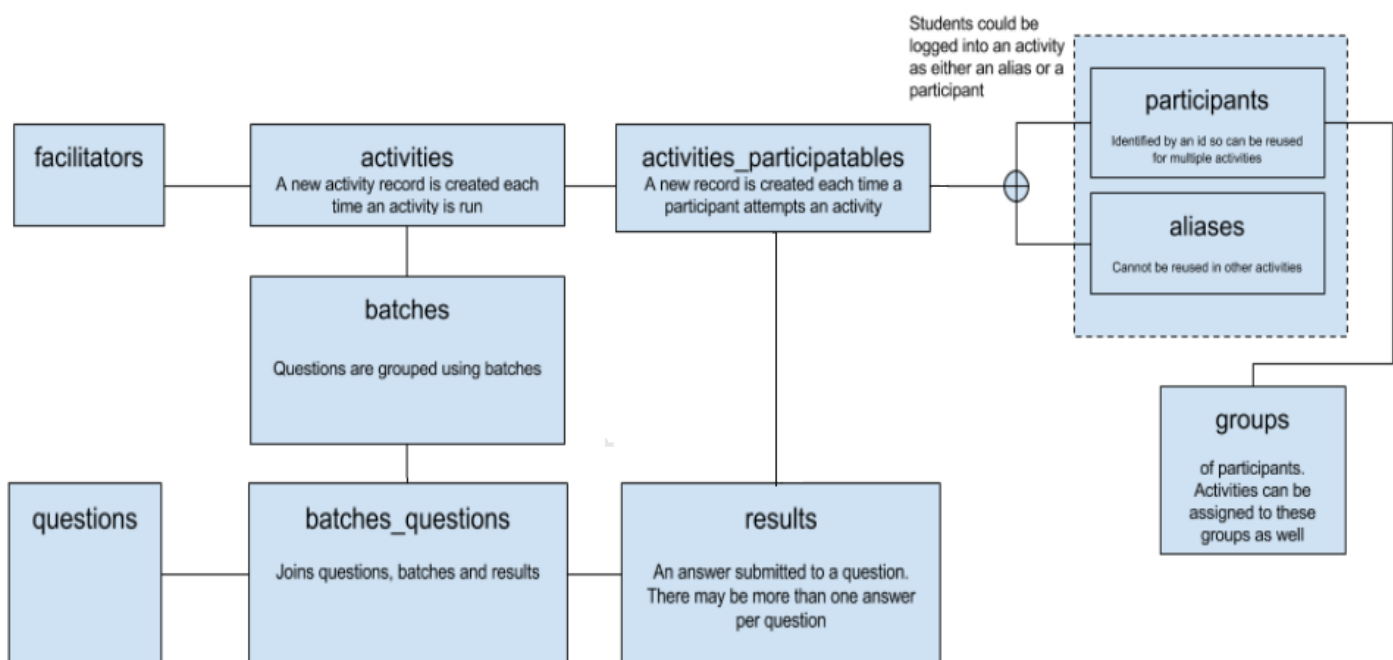


Fig.1: Xorro-Q database design

In this study data obtained from Xorro-Q has been used. Database tables containing information related to the activity name, activity type, maximum score of the activities, activity id, participant id, participant type, total score, number of attempts, number of questions answered correctly, previous course grade, test1 and test2 scores were used for analysis purposes. The database schema is shown in Fig.1. The activities are divided into real-time activities (which happened live in the class) and self-paced activities which could be taken at any time by students over a 12- week semester period and could be attempted as many times as needed by the student to achieve a minimum threshold of 70%. Self-paced activities are further divided into voluntary and mandatory activities. Total 230 students participated here. The target is to predict the students' final performance based on activities performed by the students using the Xorro-Q. The students are categorized as high-risk students (who

have scored less than 55) in the final exam score, medium-risk students (who have scored between 75 and 55) and low-risk students (whose score is above 75). Two datasets were used for the prediction.

4.1. Dataset 1 Features

This dataset contains all the features which have been derived from Xorro-Q database (refer Table 1). Since the students can take many attempts to achieve the minimum threshold of 70%, therefore student's minimum and maximum scores in these activities have also been considered as one of the features for the prediction.

TABLE I: General Features Description Obtained from Xorro-Q database

S.no	Feature Description
1.	Average lowest score taken in the activities
2.	Average highest score taken in the activities
3.	Mean score of the activities
4.	Average number of attempts to do an activity
5.	Total number of answers correct
6.	Previous course grade score
7.	Test1 score
8.	Test2 score
9.	Average of Test1 and Test2

4.2. Dataset 2 Features

A dataset has been generated using event logs of weekly Xorro-Q activities. Event logs are the beginning stage of all process mining techniques that utilize them to discover, verify or extend the model for the process [15]. Prom tool was used for performing process mining. In process mining Conformance techniques was applied and it is a comparison of model with behavior recorded in an event log to identify discrepancies [11]. Conformance checking was performed using model representing high-risk, low-risk and medium-risk student's weekly activities. The event log of the students was replayed using the mentioned model to find the relationship between the event and the model and to analyze the deviation of students from the modelled behavior. Fitness score obtained from conformance testing was combined along with the general features obtained from dataset1 and used for predictions.

4.3. Organizing Data into Sets

To evaluate how well a classifier is performing, we need to test the model on unseen data. So, the dataset has been divided in two sets, training and test set. The training set is used to train the model and then the trained model is used to make predictions on unseen data. Models can be created using many machine learning algorithms but for this study four algorithms LR, RF, NB and k-NN have been chosen. After training the models using these algorithms separately, the trained model is used to predict on the test data to test the performance of the classifier. The students' final outcomes are divided as low-risk, high-risk and medium-risk students to get better accuracy. This has been applied to both datasets. The dataset has been further divided weekly-wise to find how early can prediction on students' performance be made. Moreover, for predicting the final students' performance based on week 3 data, all the week 3, week2 and week1 data are taken into consideration.

4.4. Effectiveness Metrics

To measure how accurate a classifier is F measures is used. F measures is preferred when the dataset is highly imbalanced because accuracy alone might not be sufficient. Since the dataset is highly imbalanced F measures is used and it is the measure of prediction accuracy, which is the harmonic mean of precision and recall.

$$\text{Precision} = \frac{Tp}{Tp+FP} ; \quad \text{Recall} = \frac{Tp}{TP+FN}$$

True Positive(TP): Observation is positive, and is predicted to be positive.

False Positive(FP): Observation is positive, and is predicted to be negative.

False Negative(FN): Observation is positive, but predicted negative.

$$F \text{ measures} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5. Results

Research questions 1: Can the student's performance be predicted based on the data gathered from the educational tool (Xorro-Q)? If so, which algorithm gives better accuracy and how early can we reliably predict students' outcomes?

To answer this question four machine learning algorithms on two datasets have been used. Results shows that the maximum F measures was achieved on week 10 and the F measures increased weekly wise. Also, there was a notable difference in accuracy on week 10 compared to the other weeks because the Test 2 which happened in week 10 was added to the week 10 dataset and Test 2 is linearly related to exam (see Fig.2).

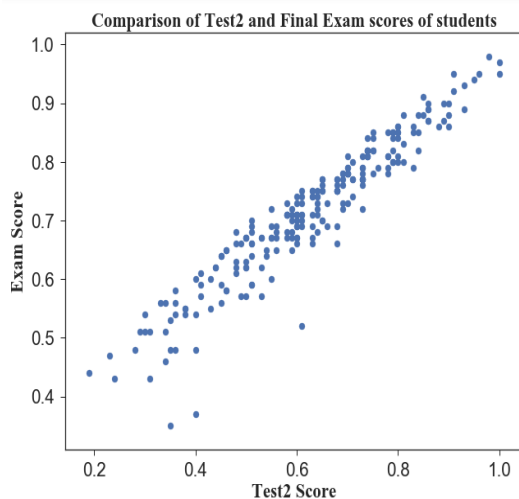


Fig.2: Scatter plot between Test2 and Exam

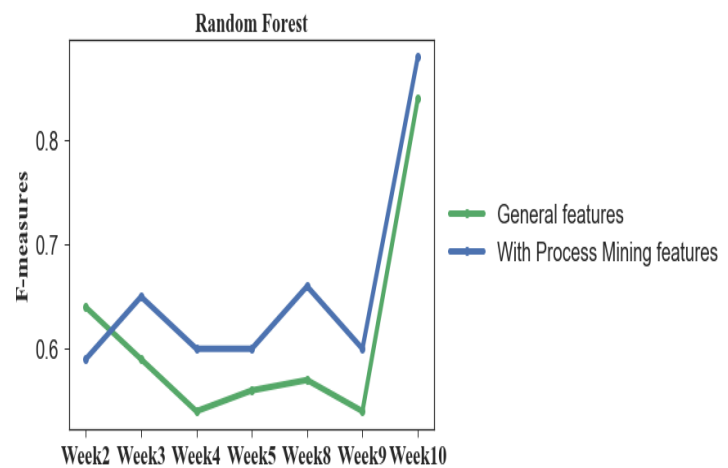


Fig.3: Comparative results of Random Forest classifier

Research question 2: Does process mining features help to increase the prediction accuracy?

Table II shows results of classifiers using general features and process mining features. When process mining features are applied to the general features (refer Fig.3) there is some improvement in accuracy over weeks but it was not a significant improvement. Here too Random forest performed better in predicting the results when compared to the other classifiers.

TABLE II: F-Measures of Classification Algorithm with Standard Features and Process Mining Features

Weeks	LR	LRP	KNN	KNNP	NB	NBP	RF	RFP
Week2	.59	.60	.56	.54	.55	.49	.64	.59
Week3	.62	.68	.53	.48	.58	.55	.59	.65
Week4	.63	.55	.45	.57	.53	.56	.54	.60
Week5	.63	.57	.50	.57	.57	.56	.56	.60
Week8	.51	.63	.57	.54	.59	.57	.57	.66
Week9	.62	.56	.44	.56	.44	.54	.54	.60
Week10	.71	.62	.54	.61	.65	.84	.84	.88

Next, to make comparisons of multiple classifiers and measure the significance of above findings Friedman test was conducted. Friedman test is a non-parametric test used to determine whether the performance of the classifiers is consistent or not. When the p-value is small (<.05), null hypothesis is rejected. Friedman test is conducted to check if there is any difference in performance of classifiers in this study. Null hypothesis is "There is no difference in performance of classifiers". After conducting Friedman test on both datasets, the p-values

obtained are $p=.07$ and $p=.08$; since the p -value is greater than $.05$ null hypothesis is accepted, or there is no difference in performance of multiple classifiers. The test also determines the rank of the algorithm. The best algorithm gets the highest mean rank. Findings show that for dataset 1 the logistic regression outperforms all classifiers, while Random forest gets the highest mean rank when applied to dataset 2.

TABLE III: Results of F-Measures and Rank (mean) on Datasets of Process Mining Features

Weeks	LRP	KNNP	NBP	RFP
Week2	.60	.54	.49	.59
Week3	.68	.48	.55	.65
Week4	.55	.57	.56	.60
Week5	.57	.57	.56	.60
Week8	.63	.54	.57	.66
Week9	.56	.56	.44	.60
Week10	.62	.61	.59	.88
Rank(Mean)	2.14	3.00	3.5	1.28

Research question 3: A student got 70% in the activity score in first two attempts. Would it be possible for the student to get the same score in exam?

Scatter plot was plotted between the activity score and the exam score to answer this question. The scatter plot between the activity score and exam score should have been a 45° line if their scores were same. But such a relationship is not observed (Fig.4). Students who scored above 70% in activities could score as 30% less in exam (see box plot in Fig.5). The distributions are not comparable. Next, to find the difference between the activity score and the exam score, mean absolute percentage difference between the two was calculated as 29.9%. Hence, we cannot conclude that the activity scores and exam scores are same for students who got above 70% in first 2 attempts. It should be noted that the activity score of 70% was considered for first 2 attempts only of the self-paced activities. We can only speculate on the reasons, like students had some assistance during their attempts. We could also argue that the activity score of 70% should be for more than 2 attempts. Our reasons for selecting 2 attempts is that since these activities could be done by students in their own times, it best represented their confidence level, though this is debatable.

Comparison of Activity and Final Exam% of students with Xorro-Q activity score above 70%

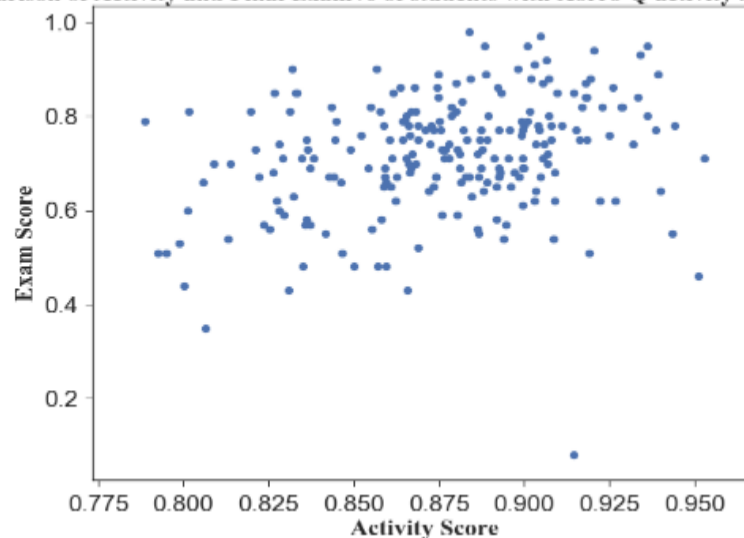


Fig.4: Scatter plot between Activity score and Exam

Box plot comparison of Final Exam score of students with Xorro-Q activity score of above 70%

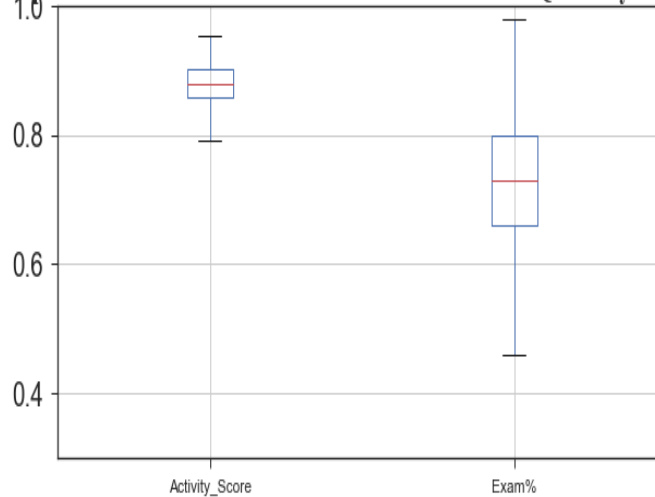


Fig.5: Box plot between Activity score and Exam

6. Conclusions

Predicting students' academic performance using Xorro-Q was a considerable difficult task. This is because the data available was highly imbalanced and selecting the best attributes was challenging. Many experiments were carried out and four classification algorithms were selected. The four different algorithms namely LR, RF, NB and k-NN were applied to the two datasets, one, with the general features, and the other with the process mining features. Accuracy increased when the process mining features were added. While checking the early prediction on students' performance, the accuracy increased over the weeks but better accuracy showed on 10th week over a 12- week semester course. The results also show that a student who got 70% in the activities is likely to get 30% less in exam. Among the four algorithms tested, Random Forest proves to be the best performer on this dataset in a statistically significant way.

The limitation of the study is the missing values and that the dataset is highly imbalanced. The demographic information of the students is also missing. Moreover, the real-time activities and voluntary activities of self-paced were done only for two weeks by the students. There are more chances of increasing the accuracy by adding these missing data. Having said this, this study has shared new insights on enhancing value of the role of analytics for predictive modelling by incorporating Process Mining features in the training set of data.

7. References

- [1] Abdous, M., He, W., & Yen, C.-J. (2012). Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade . Educational Technology and Society.
- [2] Bhegade, K. B., & Swati V. Shinde, P. (2016). Student Performance Prediction System with Educational Data Mining . International Journal of Computer Applications .
- [3] Cano-, C. M.-V., & Ventura, C. R. (2012). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Springer Science+Business Media, 315-330.
- [4] Cristobal, R., Manuel-Ignacio, L., Jose-Maria, L., & Sebastian, V. (2013). Predicting students final performance from participation in on-line discussion forums. Elsevier, 458-472.
- [5] Dorina, K. (2013). Predicting Student Performance by Using Data Mining Methods for Classification .

- [6] Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector . 3rd International Conference on Recent Trends in Computing 2015 (pp. 500 – 508). procedia Computer Science.
- [7] Kohli, K., & Birla, S. (2016). Data Mining on Student Database to Improve Future Performance . International Journal of Computer Applications .
- [8] Kotsiantis S: Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education. Knowledge-Based system, 529-535.
- [9] Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting Students' Marks in Hellenic Open University . Fifth IEEE International Conference on Advanced Learning Technologies . IEEE Computer Society.
- [10] Li, K. F., Rusk, D., & Song, F. (2013). Predicting Student Academic Performance. Seventh International Conference on Complex, Intelligent, and Software Intensive Systems. IEEE Computer Society.
- [11] A.Rozinat, W. v. (2008). Conformance checking of processes based on monitoring. ScienceDirect, 64-95.
- [12] C.Marquez-Vera, C. R., Morales, S., & Ventura. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Journal of Latin-American Learning Technologies, Vol.8 no.1.
- [13] Romero, C. (2010). Educational Data Mining: A Review of the State of the Art. IEEE, Vol.40. No.6.
- [14] Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, Issue 113.
- [15] Munoz-Gama, J. (2014). Conformance checking and diagnosis in process mining. Spain: Universitat Politecnica De Catalunya Barcelonatech.