

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

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

A thesis presented in a partial fulfilment of the
requirements for
Master of Information Science (IT)
At
Massey University
Auckland
New Zealand
in 2018

Gomathy Suganya
(Supervisor: Dr. Teo Susnjak)
(Co-Supervisor: Dr. Anuradha Mathrani)

Abstract

Recent research in analytics has assisted policy makers capitalize on their ever-increasing data repositories and make data-driven predictions to create a vision for developing strategies to achieve their business targets. This is especially relevant in educational environments where data mining techniques can be applied to make predictions around students' academic performance. This can help educators align a teaching strategy which encourages and assists students with their learning. Suitable pedagogical support can be provided to enhance the overall student learning experience.

This study is in the educational domain where student-related course data has been used to extract insights on student performances over the study period. Extensive data collected from an educational tool (Xorro-Q) used in an engineering course delivery has aided this investigation. 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. Popular 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 neighbour and Random Forest. Process mining techniques have been applied along with the general features to find out the effectiveness, such as improvement in accuracy of predictions. 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.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Teo Susnjak and Co-supervisor Dr. Anuradha Mathrani for providing valuable guidance, comments and suggestions throughout the course of my project. But for their help I would not have completed my project successfully.

I would like to acknowledge my client Pablo Garcia for providing me the data and for the countless hours he has spent in discussing with me the way forward and the potential outcomes of this project.

I wish to express my sincere thanks to Dr. James Lim for clarifying my doubts and giving valuable suggestions for correction modification and improvement of the project.

Finally I am thankful and fortunate enough to get unconditional love and constant encouragement from my friend Rahila Umer and my little son Sai Dhanwin which helped me in successfully completing the project on time.

PUBLICATIONS AND PRESENTATIONS

Publications generated from this project so far:

Suganya, G., Susnjak, T., Mathrani, A., Lim, J., Garcia, P. (2017). Data Mining Techniques to Improve Predictions Accuracy of students' Academic Performance: A Case Study with Xorro-Q. *Proceedings of the 11th International conference on Data Mining. Computers, Communication and Industrial Applications.*

Presentations given from this project so far:

1. Suganya, G., Susnjak, T., Mathrani, A., Lim, J., Garcia, P. (2017). Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q. *Oral presentation at the 11th International conference on Data Mining. Computers, Communication and Industrial Applications, Kuala Lumpur, Malaysia.*

2. Suganya, G., Susnjak, T., Mathrani, A., Lim, J., Garcia, P. (2017). Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q. *Poster session presented at the 5th INMS Postgraduate students at Massey University, Auckland, New Zealand, 26th October 2017.*

3. Suganya, G. (2017). *Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q.* Seminar delivered on 25th August 2017 to academic and industry personnel involved in this project.

ACHIEVEMENTS

Award won Suganya, G., Susnjak, T., Mathrani, A., Lim, J., Garcia, P. (2017). Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q. Won the best session paper in *11th International Conference on Data Mining, Computers, Communication and Industrial Applications*.



Contents

1	Introduction	1
1.1	Background	1
1.1.1	This study’s context	3
1.1.2	Scope and study objectives	4
1.1.3	Research Questions	4
1.1.4	Research contributions	5
1.1.5	Thesis outline	5
2	Educational Data Mining Research	7
2.1	Introduction	7
2.1.1	Different classes of Educational environments	7
2.1.2	Types of data used in EDM research	9
2.1.3	Goals of EDM research	10
2.1.4	Benefits and success factors of education data mining	10
2.1.5	Main applications of EDM	11
2.1.6	EDM methods	11
2.1.7	Commonly used data mining techniques in EDM	12
2.1.8	Logistic regression	16
2.1.9	Related study	21
3	Process mining research	25
3.1	Introduction	25
3.2	Event logs	27
3.3	Process discovery	28
3.4	Conformance checking	31
3.5	Enhancement	34
3.6	The PROM Framework	35
3.7	Goals of process mining in Educational domain	36
3.8	Related works	38
3.8.1	Application of Process mining techniques on educational datamin- ing	38

4	Data sources	41
4.1	Dataset 1-Description.	41
4.1.1	Dataset features- Xorro-Q activities	47
4.2	Dataset2 features - Process Mining	48
4.2.1	Process discovery using Inductive miner	49
4.2.2	Conformance Checking	50
4.2.3	Dataset2 features	51
5	Research approach and methodology	53
5.1	Preparing data for mining	53
5.2	Data pre-processing	54
5.2.1	Data Cleaning	55
5.2.2	Data integration	56
5.2.3	Data Transformation	56
5.2.4	Data reduction	56
5.3	Datasets Construction	57
5.3.1	Numerical data with categorical variable	57
5.4	Dataset Attributes	58
5.5	Machine Learning Training Procedures	61
5.5.1	K-fold cross validation	63
5.5.2	Hold-out methods	63
6	Results	71
6.1	Results	71
7	Discussions and Future Scope	87
7.1	Discussions	87
7.2	Study limitations and future work	87
8	Appendix	97

List of Figures

2.1	Different types of traditional and computer-based educational environments and systems	8
2.2	Randomly selecting features	16
2.3	Logistic regression curve	17
2.4	One Vs All method	18
2.5	kNN classifier	20
3.1	Positioning of main classification of process mining	26
3.2	Various process mining techniques in terms of input or output	28
3.3	Various process patterns	30
3.4	Process model in Petri net notation	30
3.5	Rediscovering process model	31
3.6	The four quality dimensions:fitness,simplicity, generalization,precision	33
3.7	Aligning traces with the model	33
3.8	Event log and process model aligning	34
3.9	ProM framework	36
3.10	ProM framework overview architecture	37
4.1	Example of a multiple-choice type	42
4.2	Example of multiple-choice type with more than one answer	43
4.3	Example of a numeric type	43
4.4	Database schema of Xorro-Q	43
4.5	Database schema of Xorro-Q with the attributes	44
4.6	Activities over weeks	47
4.7	Process model of a Low risk students	50
4.8	Replay results	50
4.9	Alignment legend	51
5.1	Various process mining stages	53
5.2	Various steps in data pre-processing	55
5.3	Grouping Students	57
5.4	Xorro-Q activities scores by students over weeks	58

5.5	Comparison of various scores	59
5.6	Distribution of scores	60
5.7	Comparison of Test2 score with final Exam score	61
5.8	Comparison of prior course grade with Exam score.	61
5.9	Comparison of Test1 score with Exam score	62
5.10	K-fold cross validation	64
5.11	Hold-out method	65
5.12	Bagging model	66
5.13	Voting	67
6.1	Confusion matrix of a kNN classifier	73
6.2	Confusion matrix of a NB classifier	74
6.3	Confusion matrix of a LR classifier	74
6.4	Confusion matrix of a RF classifier	75
6.5	Comparative results of a various classifier	75
6.6	Comparative results of LR classifier	77
6.7	Comparative results of Knn classifier	77
6.8	Comparative results of NB classifier	77
6.9	Comparative results of RF classifier	77
6.10	Confusion matrix of RF classifier	79
6.11	Box plot comparison of Xorro-Q activities with num of attempts for the year 2016	80
6.12	Box plot for difficult activities of num of attempts for the year 2016 .	81
6.13	Box plot comparison of Xorro-Q activities with num of attempts for the year 2017	82
6.14	Box plot for difficult activities with num of attempts for the year 2017	82
6.15	Box plot comparison of answers attempted for the year 2016	84
6.16	Box plot comparison of answers attempted for the year 2017	85
6.17	Box plot comparison for difficult activities answers attempted for the year 2016	86
6.18	Box plot comparison of difficult activities answers attempted for the year 2017	86
8.1	Box plot comparison of Xorro-Q activities with num of attempts . . .	97
8.2	Box plot comparison of difficult activities with num of attempts . . .	98
8.3	Box plot comparison of Xorro-Q activities with answers attempted . .	99
8.4	attributes	100
8.5	attributes	100

List of Tables

3.1	Sample event log	28
4.1	Entity relationship	45
4.2	General features description obtained from Xorro-Q database	48
4.3	Event log generated from process mining	49
4.4	Various fitness scores obtained from conformance checking	51
5.1	Confusion matrix for a multiple classes	68
6.1	F-measures of classification algorithm with standard features and process mining features	77
6.2	Results of F-measures and rank(mean) on datasets of process mining features	78
6.3	Number of students on every category	79

