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.
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:


Presentations given from this project so far:


ACHIEVEMENTS

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-
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
# List of Tables

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