

AUS Repository

A Methodology of Rule Discovery from Large-scale Multi-tier Noisy Educational Data

Item Type	Thesis
Authors	Yousuf, Tasneem
Download date	2026-06-08 17:59:39
Link to Item	http://hdl.handle.net/11073/9357

A METHODOLOGY OF RULE DISCOVERY FROM LARGE-SCALE MULTI-
TIER NOISY EDUCATIONAL DATA

by

Tasneem Yousuf

A Thesis presented to the Faculty of the
American University of Sharjah
College of Engineering
In Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in
Computer Engineering

Sharjah, United Arab Emirates

May 2018

Approval Signatures

We, the undersigned, approve the Master's Thesis of Tasneem Yousuf.

Thesis Title: A Methodology of Rule Discovery from Large-scale Multi-tier Noisy Educational Data

Signature

Date of Signature

(dd/mm/yyyy)

Dr. Imran A. Zualkernan

Associate Professor, Department of Computer Science and Engineering

Thesis Advisor

Dr. Michel Pasquier

Associate Professor, Department of Computer Science and Engineering

Thesis Committee Member

Dr. Hazim El-Baz

Associate Professor, Department of Industrial Engineering

Thesis Committee Member

Dr. Fadi Aloul

Head, Department of Computer Science and Engineering

Dr. Ghaleb Hussein

Associate Dean for Graduate Affairs and Research

College of Engineering

Dr. Richard T. Schoephoerster

Dean, College of Engineering

Dr. Mohamed El-Tarhuni

Vice Provost for Graduate Studies

Acknowledgements

I would like to express my deepest gratitude for my advisor Dr. Imran Zualkernan for his guidance, motivation, and time throughout my thesis. His active participation in all the stages of this thesis and immense knowledge persuaded me to think analytically and be enthusiastic about this research. I would also like to thank my thesis committee members Dr. Michel Pasquier and Dr. Hazim El-Baz for their valuable comments and suggestions. I am also very grateful to Dr. Assim Sagahyroon and Ms. Salwa Mohamed for granting me Graduate Teaching Assistantship in American University of Sharjah.

In addition, I would like to show my gratitude for my husband Yousuf Mushtaq Ali for his support and understanding throughout the course of this research and for always being there for me. I would like to thank my parents Muhammad Ismail and Mehjabeen Ismail for their kind advices and for instilling in me the qualities of hard work and perseverance. I would also like to appreciate and thank my siblings Zehra Ismail, Munira Ali Asghar for being the people I could always look upto and Ruqaiya Ismail, Hamza Ismail, and Sakina Muhammad for always encouraging me and lending a listening ear to me.

Dedication

To my dear daughter Maria for adding to the joys of my life...

Mamma loves you loads...

Abstract

Recent availability of very large amounts of educational data in digital format often leads to data overload where it is difficult to determine important trends and patterns beyond those provided by traditional statistical techniques. Therefore, educational data mining (EDM) has emerged. Association mining is a type of EDM technique which is well-known for discovering relationships from data with high scale and velocity, but low variety and veracity. This analysis can be performed at the micro-level (e.g., for teachers), meso-level (e.g., for cohorts of schools), or at macro-levels (e.g., at region, province, or country level). This thesis proposes a methodology for the application of association mining to multi-tier sparse and error-ridden educational data. The methodology uses rule templates and is organized around the four analytical dimensions of people, process, environment, and outcomes. The methodology defines Extract Transform and Load (ETL) processes for this type of data and shows how data from lower levels is aggregated to baskets at higher levels. The proposed methodology was applied to data collected from a large-scale continuous professional development (CPD) process for 2,613 teachers in a developing country. The methodology was used to mine interesting rules which were evaluated using the objective metrics of Support, Confidence, and Lift to determine the quality of rules. The Confidence for each level was set to be at least 0.85. The results are that micro-level analysis (n = 2613 teachers) yielded little or no rules with a very low mean Support of 0.00345 (sd. = 0.00214) and mean Lift 6.98 (sd. = 4.63). The situation remained somewhat the same at the meso-level (n = 1391 schools) with a mean Support of 0.0059 (sd. = 0.00051) and mean Lift of 5.46 (sd. = 3.23). The results were significantly better at the macro level (n = 59 clusters) with a mean Support of 0.089 (sd. = 0.021) and mean Lift of 5.925 (sd. = 2.5). The mined rules discovered several anomalies and fidelity violations in the CPD process at various levels. The methodology was also useful in identifying small groups of teachers (6-8 teachers), schools (8-10 schools), and clusters (4-7 clusters) with common characteristics that can be further administered to help improve the CPD process.

Keywords: *educational analytics, association mining, rule discovery, Apriori, market basket analysis, developing countries*

Table of Contents

Abstract	6
List of Tables	12
List of Figures	15
List of Abbreviations	17
Chapter 1 . Introduction	18
1.1. Background	18
1.2. Problem Statement	19
1.3. Motivation and Goals	19
1.3.1. Students performance	19
1.3.2. School characteristics	20
1.3.3. Teacher characteristics	20
1.3.4. Mentoring and assessment processes	20
1.4. Expected Nature of Large-Scale Educational Data	20
1.5. Appropriate Data Description Techniques	21
1.5.1. Clustering techniques	22
1.5.2. Associative methods	23
1.6. Thesis Statement	24
1.7. Significance of the Research	24
1.8. Organization of the Thesis	25
Chapter 2 . Background and Literature Review	26
2.1. Educational Data Mining	26
2.2. Apriori Algorithm Applications in EDM	26
2.2.1. Course recommendation	26
2.2.2. Studying students' behavior and its effect	29
2.2.3. Teachers' evaluation	30
2.2.4. Predicting students at risk	30
2.2.5. Profiling high-achieving students	31
2.2.6. Book recommendation	31
2.2.7. Rare itemset mining	31
2.3. Summary and Relevance	32
Chapter 3 . Proposed Methodology	34

3.1.	Data Profiling.....	34
3.2.	Identifying Key Stakeholders	34
3.2.1.	Macro-level.....	35
3.2.2.	Meso-level.....	36
3.2.3.	Micro-level.....	36
3.3.	Extract, Transform, and Load (ETL)	37
3.3.1.	Data cleansing.....	37
3.3.2.	Attribute selection.....	37
3.3.3.	Attribute discretization.....	38
3.3.4.	Projections.....	39
3.4.	Formulation of Education Baskets.....	39
3.4.1.	Approach.....	39
3.4.2.	Items.....	39
3.4.3.	Itemset.....	40
3.4.4.	Transactions.....	40
3.4.5.	Rules.....	40
3.4.6.	Educational baskets and rules.....	41
3.5.	Application of Apriori Algorithm.....	42
3.5.1.	Templates for rule generation.....	43
3.6.	Evaluation Approach	44
3.6.1.	Objective evaluation metrics.....	44
3.6.2.	Subjective evaluation metrics.....	45
3.6.3.	Visualization plots.....	45
3.7.	Summary	47
Chapter 4 . Case Study: Applying the Methodology		49
4.1.	Case Study	49
4.2.	The CPD Framework	50
4.2.1.	Objective.....	50
4.2.2.	Organizations and personnel.....	50
4.2.3.	Mentoring areas.....	51
4.2.4.	Educational levels.....	51
4.3.	Formation of Education Baskets in the CPD Framework.....	52
4.4.	Variables in the Educational Data.....	53

4.3.1.	Process variables.....	53
4.3.2.	Environment variables.	54
4.3.3.	People variables.	56
4.3.4.	Outcome variables.	56
4.5.	Formulation of Rule Templates in the CPD Framework.....	62
4.5.1.	Rule templates at the micro-level.	62
4.5.2.	Rule templates at the meso-level.	63
4.5.3.	Rule templates at the macro-level.....	64
4.6.	Algorithm for Rule Discovery on Educational Data.....	65
4.7.	Summary.....	66
Chapter 5 . ETL.....		67
5.1.	Data Extraction	67
5.2.	Data Transformation	70
5.2.1.	Data cleansing.....	70
5.2.2.	Attributes consistency.....	72
5.2.3.	Data derivation.....	72
5.2.4.	Reshaping data.....	73
5.2.5.	Attribute discretization.....	74
5.2.6.	Attribute selection.....	74
5.3.	Data Load.....	75
5.4.	Summary	75
Chapter 6 . Micro-level Educational Analytics.....		76
6.1.	Teacher Outcome Analysis.....	76
6.1.1.	Educational basket for teacher outcome analysis.	76
6.1.2.	Experiment 1 – Teacher outcome = Good.	78
6.1.3.	Experiment 2 – Teacher outcome = Bad.	80
6.1.4.	What distinguished good and bad teachers.....	84
6.1.5.	Conclusion.	84
6.2.	Subjects Outcome Analysis	85
6.2.1.	Educational basket for subjects outcome analysis.	86
6.2.2.	Experiment 1 – English outcome = Good.....	86
6.2.3.	Experiment 2 – English outcome = Bad.	89
6.2.4.	Comparison of Good vs. Bad English outcome.....	91

6.2.5.	Experiment 3 – General Knowledge (GK) outcome = Good.	91
6.2.6.	Experiment 4 – General Knowledge (GK) outcome = Bad.	93
6.2.7.	Experiment 5 – Religion outcome = Good.	94
6.2.8.	Experiment 6 – Religion outcome = Bad.	97
6.2.9.	Experiment 7 – Mathematics outcome = Good.	98
6.2.10.	Experiment 8 – Mathematics outcome = Bad.	102
6.2.11.	Experiment 9 – National language (NL) outcome = Good.	102
6.2.12.	Experiment 10 – National language (NL) outcome = Bad.	104
6.2.13.	Experiment 11 – Science outcome = Good.	105
6.2.14.	Experiment 12 – Science outcome = Bad.	106
6.2.15.	Experiment 13 – Social Studies (SS) outcome = Good.	106
6.2.16.	Experiment 14 – Social Studies (SS) outcome = Bad.	107
6.2.17.	Conclusion.	107
6.3.	Teacher Training Analysis.	109
6.3.1.	Educational basket for teacher training analysis.	109
6.3.2.	Experiment 1 – Recommended teacher training.	111
6.3.3.	Conclusion.	115
6.4.	Summary.	115
Chapter 7 . Meso-level Educational Analytics.		116
7.1.	School Outcome Analysis.	116
7.1.1.	Educational basket for school outcome analysis.	116
7.1.2.	Experiment 1 – School outcome = Good.	118
7.1.3.	Experiment 2 – School outcome = Bad.	121
7.1.4.	What distinguished Good from Bad schools?	125
7.1.5.	Conclusion.	126
7.2.	School Size Analysis.	126
7.2.1.	Educational basket for school size analysis.	127
7.2.2.	Experiment 1 – School size.	127
7.2.3.	Comparison of different school sizes.	131
7.2.4.	Conclusion.	132
7.3.	Summary.	132
Chapter 8 . Macro-level Educational Analytics.		133
8.1.	Cluster Outcome Analysis.	133

8.1.1.	Educational basket for cluster outcome analysis.	133
8.1.2.	Experiment 1 – Cluster outcome = Good.	135
8.1.3.	Experiment 2 – Cluster outcome = Bad.	137
8.1.4.	What distinguished Good from Bad clusters?	140
8.1.5.	Conclusion.	140
8.2.	Summary	141
Chapter 9 . Discussion, Conclusion, and Future Directions		142
9.1.	Quality of Rules	143
9.2.	Limitations	145
9.2.1.	Educational data characteristic limitations.	145
9.2.2.	Formulation of rule templates.	145
9.2.3.	Application of constraints.	145
9.2.4.	Analysis of rules.	146
9.3.	Future Directions	146
References.....		148
Appendix A: Apriori Algorithm		154
Appendix B: ETL for “Teacher”		155
Appendix C: Rules Generation for Teacher Outcome Analysis.....		157
Vita.....		159

List of Tables

Table 1.1: Different levels of educational analytics	18
Table 1.2: The expected 4 V's of large-scale education data	21
Table 2.1: Prior work in EDM using Apriori algorithm	33
Table 3.1: Example transactions	40
Table 3.2: Example educational transactions.....	42
Table 3.3: Example of support measure	43
Table 3.4: Interesting itemsets from Fig. 3.11	47
Table 3.5: Summary of the rule mining process	48
Table 4.1: Data aggregation techniques used for different variables.....	53
Table 4.2: Process variables for the CPD process in the educational data	54
Table 4.3: Environment variables for the CPD process in the educational data	55
Table 4.4: Outcome variables for the CPD process in the educational data.....	56
Table 4.5: People variables for the CPD process in the educational data.....	57
Table 4.6: Variables at different levels of educational analytics	60
Table 4.7: Sample micro-level educational basket created from educational data generated by CPD framework.....	61
Table 4.8: Example rules generated using template # 1 – micro-level analysis	62
Table 4.9: Example rules generated using template # 2 – micro-level analysis	63
Table 4.10: Example rules generated using template # 1 – meso-level analysis.....	63
Table 4.11: Example rules generated using template # 2 – meso-level analysis.....	64
Table 4.12: Example rules generated using template # 1 – macro-level analysis	65
Table 5.1: The 4 Vs of given educational data	67
Table 5.2: Data cleansing details for different variables	71
Table 5.3: Attributes modified for consistency.....	73
Table 5.4: Derived attributes.....	73
Table 5.5: Data in long format	73
Table 5.6: Data in wide format	74
Table 6.1: Items in education basket for teacher outcome analysis.....	77
Table 6.2: Resulting rule – Teacher outcome = Good.....	79
Table 6.3: Teachers belonging to rule number 1 – Teacher outcome = Good	80
Table 6.4: Itemsets from parallel coordinates plot – Teacher outcome = Good.....	80

Table 6.5: Interesting rules from scatter plot – Teacher outcome = Bad.....	82
Table 6.6: Interesting rules from matrix plot – Teacher outcome = Bad.....	82
Table 6.7: Teachers belonging to rule number 1 – Teacher outcome = Bad.....	83
Table 6.8: Itemsets from parallel coordinates plot – Teacher outcome = Bad	84
Table 6.9: Summary of rules for teacher outcome analysis.....	85
Table 6.10: Items in education basket for subject outcome analysis.....	86
Table 6.11: Resulting rules – English outcome = Good	88
Table 6.12: Teachers belonging to rule number 2 – English outcome = Good	89
Table 6.13: Itemsets from parallel coordinates plot - English outcome = Good	89
Table 6.14: Resulting rule – English outcome = Bad	91
Table 6.15: Itemsets from parallel coordinates plot - English outcome = Bad.....	91
Table 6.16: Resulting rule – GK outcome = Good	93
Table 6.17: Itemsets from parallel coordinates plot - GK outcome = Good.....	93
Table 6.18: Resulting rules – Religion outcome = Good	95
Table 6.19: Teachers belonging to rule number 3 – Religion outcome = Good.....	97
Table 6.20: Itemsets from parallel coordinates plot – Religion outcome = Good.....	97
Table 6.21: Resulting rules – Maths outcome = Good	100
Table 6.22: Teachers belonging to rule number 1 – Maths outcome = Good	100
Table 6.23: Teachers belonging to rule number 4 – Maths outcome = Good	101
Table 6.24: Itemsets from parallel coordinates plot - Maths outcome = Good	102
Table 6.25: Resulting rule - NL outcome = Good	104
Table 6.26: Itemsets from parallel coordinates plot - NL outcome = Good	104
Table 6.27: Summary of rules for subjects outcome analysis	108
Table 6.28: Interesting rules from scatter plot – Recommended teacher training	112
Table 6.29: Interesting rules from matrix plot – Recommended teacher training	113
Table 6.30: Itemsets from parallel coordinates plot – Recommended teacher training	114
Table 6.31: Summary of rules for teacher training analysis	115
Table 7.1: Items in education basket for school outcome analysis.....	117
Table 7.2: Interesting rules from scatter plot – School outcome = Good	119
Table 7.3: Interesting rules from matrix plot – School outcome = Good	120
Table 7.4: Schools belonging to rule number 7 – School outcome = Good	121
Table 7.5: Itemsets from parallel coordinates plot – School outcome = Good.....	121

Table 7.6: Interesting rules from scatter plot – School outcome = Bad	123
Table 7.7: Interesting rules from matrix plot – School outcome = Bad	124
Table 7.8: Schools belonging to rule number 3 – School outcome = Bad.....	125
Table 7.9: Itemsets from parallel coordinates plot – School outcome = Bad	125
Table 7.10: Summary of rules for school outcome analysis.....	126
Table 7.11: Interesting rules from scatter plot – School size analysis.....	129
Table 7.12: Interesting rule from matrix plot – School size analysis	129
Table 7.13: Schools belonging to rule number 1 – School size = Small	130
Table 7.14: Schools belonging to rule numbers 5 and 6 – School size = Large.....	131
Table 7.15: Itemsets from parallel coordinates plot – School size analysis	131
Table 7.16: Summary of rules for school size analysis	132
Table 8.1: Items in education basket for cluster outcome analysis.....	134
Table 8.2: Resulting rules – Cluster outcome = Good.....	136
Table 8.3: Clusters belonging to rule number 1 – Cluster outcome = Good	137
Table 8.4: Itemsets from parallel coordinates plot – Cluster outcome = Good.....	137
Table 8.5: Resulting rules – Cluster outcome = Bad	139
Table 8.6: Itemsets from parallel coordinates plot – Cluster outcome = Bad.....	140
Table 8.7: Summary of rules for cluster outcome analysis.....	141
Table 9.1: Summary of rules obtained at each level of analysis.....	144

List of Figures

Fig. 2.1: Application of association rules to educational data	27
Fig. 2.2: FARIM algorithm process illustration [55]	32
Fig. 3.1: The data mining process	34
Fig. 3.2: Relationship of stakeholders to different educational levels of analytics	35
Fig. 3.3: Generic model of data integration at macro and meso levels.....	40
Fig. 3.4: Example supermarket rules	41
Fig. 3.5: Example educational items.....	41
Fig. 3.6: Example educational rules.....	42
Fig. 3.7: Confidence values for supermarket rules	43
Fig. 3.8: Template-based educational rule.....	44
Fig. 3.9: An example scatter plot.....	46
Fig. 3.10: An example matrix plot.....	46
Fig. 3.11: An example parallel coordinates plot.....	47
Fig. 4.1: Big picture of education data and stakeholders for the CPD process.....	49
Fig. 4.2: Algorithm for rule discovery on educational data.....	65
Fig. 5.1: The structure of given education data	68
Fig. 5.2: Cluster data extracted in Microsoft Excel Sheet	68
Fig. 5.3: The structure of extracted education data represented as UML diagram.....	69
Fig. 5.4: Dirty data for variable Test Report Issuance %.....	72
Fig. 6.1: Distribution of teachers with different outcomes	77
Fig. 6.2: Scatter plot for Experiment 1 – Teacher outcome = Good.....	79
Fig. 6.3: Matrix plot for Experiment 1 – Teacher outcome = Good.....	79
Fig. 6.4: Scatter plot for Experiment 2 – Teacher outcome = Bad	81
Fig. 6.5: Matrix plot for Experiment 2 – Teacher outcome = Bad	82
Fig. 6.6: Distribution of subjects with different outcomes	85
Fig. 6.7: Scatter plot for Experiment 1 – English outcome = Good	87
Fig. 6.8: Matrix plot for Experiment 1 – English outcome = Good	87
Fig. 6.9: Scatter plot for Experiment 2 – English outcome = Bad.....	90
Fig. 6.10: Matrix plot for Experiment 2 – English outcome = Bad	90
Fig. 6.11: Scatter plot for Experiment 3 – GK outcome = Good.....	92
Fig. 6.12: Matrix plot for Experiment 3 – GK outcome = Good	92

Fig. 6.13: Scatter plot for Experiment 5 – Religion outcome = Good.....	95
Fig. 6.14: Matrix plot for Experiment 5 – Religion outcome = Good	95
Fig. 6.15: Scatter plot for Experiment 7 – Maths outcome = Good	99
Fig. 6.16: Matrix plot for Experiment 7 – Maths outcome = Good.....	99
Fig. 6.17: Scatter plot for Experiment 9 – NL outcome = Good	103
Fig. 6.18: Matrix plot for Experiment 9 – NL outcome = Good	104
Fig. 6.19: Distribution of teachers with different recommended training areas	110
Fig. 6.20: Scatter plot for Experiment 1 – Recommended teacher training	112
Fig. 6.21: Matrix plot for Experiment 1 – Recommended teacher training.....	112
Fig. 7.1: Distribution of schools with different outcomes	117
Fig. 7.2: Scatter plot for Experiment 1 – School outcome = Good	119
Fig. 7.3: Matrix plot for Experiment 1 – School outcome = Good.....	119
Fig. 7.4: Scatter plot for Experiment 2 – School outcome = Bad.....	123
Fig. 7.5: Matrix plot for experiment 2 – School outcome = Bad.....	123
Fig. 7.6: Distribution of schools with different sizes.....	127
Fig. 7.7: Scatter plot for Experiment 1 – School size analysis	128
Fig. 7.8: Matrix plot for Experiment 1 – School size analysis	129
Fig. 8.1: Distribution of clusters with different outcomes	134
Fig. 8.2: Scatter plot for Experiment 1 – Cluster outcome = Good	136
Fig. 8.3: Matrix plot for Experiment 1 – Cluster outcome = Good	136
Fig. 8.4: Scatter plot for Experiment 2 – Cluster outcome = Bad	138
Fig. 8.5: Matrix plot for Experiment 2 – Cluster outcome = Bad.....	139
Fig. A.1: Apriori algorithm [64].....	154

List of Abbreviations

3DM	Data-driven decision making
CFS	Child Friendly School
CPD	Continuous Professional Development
DSD	Directorate of Staff Development
EAAM	Enhanced Apriori Algorithm Model
EDM	Educational data mining
ETL	Extract, Transform, and Load
GK	General Knowledge
HT	Head Teacher
L0M	Level 0-manager
L1M	Level 1-manager
L2M	Level 2-manager
MOOC	Massive Open Online Course
MCRS	MOOC oriented course recommendation system
NL	National language
SOLO	Structure of Observed Learning Outcomes
SS	Social Studies
TTI	Teacher Training Institute

Chapter 1 . Introduction

1.1. Background

Data mining is the process of extracting implicit, previously unknown, and potentially useful information from data [1]. Typically, the term ‘data’ here, refers to ‘big data’, i.e. large amounts of data which can be business or market-related. The goal of data mining is to find patterns, trends, and information from this data and convert it into useful and understandable form. Educational Data Mining (EDM) is a relatively new and emerging discipline and is usually associated with applying a set of data mining techniques to discover how students learn, to predict learning outcomes, and to understand the learning behaviour of individual students or a cohort of students. EDM can be then used to design better and smarter learning technologies, and to better inform the education stakeholders like students, parents, teachers, and administrators [2].

As shown in Table 1.1, educational analytics can be conducted at three distinct levels; macro, meso and micro [3]. The Micro-level analytics is performed for individual students or groups of students within a cohort or a school to identify students’ strengths, and weaknesses, and to predict their success. The Meso-level educational analysis operates at a higher level such as school and groups of schools (e.g., school clusters) to improve school’s performance and business processes for a group of schools. Macro-level analytics is the highest level of education analytics that is performed at a regional or state level by analyzing data from many schools and school districts.

Table 1.1: Different levels of educational analytics

Level of Analysis	Definition	Examples of Analyses
Micro	Analysis of process-level data for individual students or groups of students	Students grades, attendance, library loans and purchases, online activity, social activity analysis etc.
Meso	Analysis at institutional level to improve school’s performance and business processes	Structuring course content and its effectiveness in the learning process, constructing student models, stakeholder perspective analysis
Macro	Enables cross-institutional analysis at region or state level by benchmarking and data integration from meso/micro levels	Evaluating teachers and curricula, organizing institutional resources (human and material), enhancing educational programs

Data-driven decision making (3DM) operates especially at the macro and meso level of analytics [4] where data is gathered from students and schools to enable decision making for the betterment of the entire education landscape.

1.2. Problem Statement

Educational data is typically multi-dimensional, complex and hierarchical as it aggregates data from problems, lessons, activities, curricula, schools, school clusters, regions, and so on. The use of data mining in education emerged later than other fields primarily because of the way educational data is stored. The problem is particularly acute in developing countries. Even now, many schools in developing countries store their data in paper files, and those schools that store their data on systems or online, usually store it in difficult-to-use formats [2]. Education institutions are flooded with data, but this data is very often not used as input into effective educational reforms. This thesis addresses the challenge of how sparse and error-ridden educational data increasingly available in developing countries' educational systems can be used to create meaningful and actionable description at three different levels, as described in Table 1.1.

1.3. Motivation and Goals

To address the problem of obtaining a sense-making structural description of education data, unsupervised learning techniques in data mining can be used. The goal of unsupervised learning is to create structural descriptions that explain how prediction is derived, rather than predicting the outcome for a new instance (termed as supervised learning) [1].

Given an education dataset, with many different environment, process and outcome features of schools, teachers, and clusters (cohort of schools in geographic proximity) from a developing country, a concept description can be obtained. Example questions that can be answered from this kind of a description with respect to various educational entities (students, schools, teachers etc.) are detailed below.

1.3.1. Students performance.

- How does the performance of the students in one school relates to other students in different schools of the same geographical location?
- How does the performance of the students in one location relates to other students in a different location?
- What aspects of school with good ranking and performance can be shared in other schools?

- What is the relationship between groups of teachers with good/average/bad ranking according to students' performances?
- How do the student learning outcomes vary for different schools and locations?
- How do the student learning outcomes vary for different teachers?
- How do the student learning outcomes vary for different subjects?

1.3.2. School characteristics.

- What are the characteristics of schools with different enrolment sizes?
- How does the distance of school from the training and monitoring centre relate to the school and teachers ranking?
- How does the distance of school from the training and monitoring centre relate to the recorded assessment indicators?

1.3.3. Teacher characteristics.

- How do the teacher variables in one school and location relate to teacher variables in different schools and locations?
- Which teachers in a school need training for different subjects?
- Which teachers in a group of schools need training for different subjects?
- How does the training received by teachers relate to their performance and ranking?
- How does teachers' workload relate to their performance?

1.3.4. Mentoring and assessment processes.

- How do assessment indicators vary with different data collecting personnel?
- Are the learning outcomes achieved in all locations? If yes, then to what extent?
- How do the learning outcomes achieved in one location vary as opposed to other locations?

1.4. Expected Nature of Large-Scale Educational Data

Centre for Education Policy and Practice and Centre for Global Education Mentoring, UNESCO, states that large-scale educational data should have the following basic properties [5]:

- The data is standardised to facilitate comparability across students, schools, regions, and in some cases, countries.
- The data is representative of the education structure at some level, be it regional or national level.

For large-scale educational data in developing countries, the nature of the data can be studied using the 4 Vs of Volume, Velocity, Variety, and Veracity [6], as shown in Table 1.2. The 3 Vs of Data Volume, Velocity, and Veracity were proposed in [7] by Gartner, Inc. From the nature of educational data, appropriate data mining algorithm to obtain a structural description can subsequently be determined.

Table 1.2: The expected 4 V's of large-scale education data

Property	Definition	Expected V's in large-scale education data
Volume	Determines how much data is present for each type of the data.	Education data that is collected on large-scale have high volumes of data with data being collected from many schools, groups of schools called clusters, and associated teachers and administrators. This data is often collected over time (e.g., monthly) to enable policy-making at regional or state level.
Velocity	Determines the time with respect to complexity and run-time performance in which the data is analyzed by any algorithm. This is related to how fast the data is changing.	Education data changes over time and is mostly processed in batches of data collected over time. For example, student performance data can change monthly or quarterly depending on frequency of data collection.
Variety	This property describes the form of the data, and the type of attributes that are present in dataset, whether they are numerical or categorical.	Large-scale education data is maintained in files (online or system or paper) for different time periods and contains both numerical and categorical values due to the features of education data having both textual variables like subject names, and numeric variables like marks/scores. The data may also include audio or video snippets of classroom observations.
Veracity	This property describes the reliability, validity, and accuracy of the data and determines the significance of missing values, the reasons for incomplete or incorrect data, and the usefulness of data by checking if it were duplicated or stale.	The reliability of the educational data is dictated by the quality of processes used to collect the data. However, such large-scale data is bound to be dirty due to the variety of data sources, scale, data collection fatigue, performance anxiety, political issues and the involvement of different personnel at various levels of the educational system.

In summary, large scale educational data is expected to have high volume, reasonably high velocity, lower variety, and low veracity.

1.5. Appropriate Data Description Techniques

Two widely used unsupervised learning techniques to obtain structural descriptions are clustering and association learning [8]. The clustering learning refers to grouping similar instances into clusters, while the association learning discovers relationships between different variables in the data [1]. Appropriateness of each technique is discussed next.

1.5.1. Clustering techniques. The application of any unsupervised clustering technique is to find a concept description that can be used to answer large-scale education questions. The clustering algorithms can be broadly divided into these classes [7]:

- **Partitioning-based:** These algorithms divide the data into a number of clusters according to their centroid which is usually the mean, mode, median, or medoid, etc. to represent the centre point of the cluster.
- **Hierarchical-based:** These algorithms categorise data in a hierarchical fashion such that it can be represented by the leaf nodes of a dendrogram (a tree diagram). In this representation, individual data points are present on the leaf nodes of the dendrogram.
- **Density-based:** Here, data objects are divided into arbitrary-shaped clusters based on their region of density.
- **Grid-based:** In these algorithms, the data are divided into grids and clustering is performed on the grid. These algorithms give the advantage of fast processing and high performance due to small size of the grid.
- **Model-based:** These methods try to find a fit between the data and some pre-defined mathematical model or probability distribution. Such methods typically determine the number of clusters automatically based on standard statistics.

To apply any clustering algorithm to the data with the properties as shown in Table 1.2, the algorithm should be apt in handling large datasets, dirty data, high dimensionality data, and should handle categorical and numerical values.

K-means is the most commonly used algorithm for partition-based clustering. K-means algorithm attempts to find a user-specified number of non-overlapping clusters which are represented by their centroid [9]. K-means does not work well with high dimensional data and data with many missing values [10], [11], [12]. In addition, traditional K-means can only handle numerical values as input, however, variants of K-means can be used to handle categorical values [13]. Consequently, K-means will face challenges in handling large-scale educational data.

Hierarchical-based clustering algorithm will also face challenges in application to large-scale educational data because majority of these algorithms require data with

no outliers, and those algorithms that handle noisy data and high dimensionality cannot handle a dataset with both numerical and categorical values. Grid-based, model-based, and density-based algorithms also have issues dealing with non-numeric data.

Finally, when using clustering algorithms for large-educational datasets, the clusters do not provide a higher order data description, and secondary analysis will need to be carried out on the cluster formation to answer the types of questions postulated earlier.

1.5.2. Associative methods. There are different types of relationship mining techniques like association rule mining, sequential pattern mining, correlation mining, and causal data mining [14]. In association rule mining, rules embodying relationships are discovered with respect to the frequency by which they appear in the dataset, by giving some minimum threshold level of their coverage. Alternatively, sequential pattern mining refers to finding the temporal associations between different attributes of the dataset. Finally, the correlation mining and causal data mining refers to finding the linear and causal relationships in data respectively.

Association rule mining encompasses a broad set of analytic techniques such as Apriori algorithm to find patterns in specific objects, which might be the visitors to a website, or products in a store [15]. Market basket analysis is one of the applications of association mining that finds supermarket items that occur together in customer transactions. The market basket analysis has been successfully used by various retail outlets and supermarkets to organize their products to have items that are bought together being placed in close physical proximity, or to design promotions, and thus to increase their profits. The market basket data is usually sparse in nature, for example, a single customer transaction may include about 15 items total Stock Keeping Units (SKU) of 60,000 items available in a typical supermarket.

Sequential pattern mining is applied on problems where sequenced (or time-based) data is compared for similarities or to recover missing information [16]. The correlation mining has applications in finding dependency among large number of signals, images, and video sequences [17]. Finally, the causal mining aims at studying the effect of a particular behaviour and its application to large-scale educational data requires the recording of student learning and behaviour events like taking a quiz before

each class or number of attempted voluntary exercises [18]. As opposed to clustering methods and other relationship mining techniques, associative methods promise to handle the data characteristics of high scale, high velocity, and low variety and veracity. Apriori algorithm is the most common association learning algorithm which has been suitably and widely used for market basket analysis [19], [20], [21], [22]. Market basket analysis has been successful in domains like the retail industry characterized by high volume, high speed, low variety and low-veracity. Low veracity in the industry is typically handled by using Extract Transform and Load (ETL) processes [23] that clean and transform the data before using the Apriori algorithm. Apriori algorithm works by identifying frequent items that co-occur together and producing association rules by using a minimum count of how often these items appear together in a dataset. The details of this algorithm are shown in Appendix A.

1.6. Thesis Statement

Apriori algorithm in conjunction with appropriately designed Extract Transform and Load (ETL) processes can be effectively used to conduct macro, meso-, and micro-level educational analysis for large-scale educational data characterized by high volume, high velocity, low variety and veracity. The specific research objectives are stated below:

1. How to build ETL processes to clean and transform large-scale educational data in the right format for Apriori algorithms to handle the veracity issues?
2. How to model the educational data at different levels that can be used as input to the Apriori algorithm?
3. How useful is the association rules analysis at the micro- to macro-level with respect to the goodness metrics?

1.7. Significance of the Research

Due to ubiquity and affordable prices of mobile phones, internet technologies and storage systems, many developing countries facing educational emergencies are beginning to collect large amounts of educational data at various levels. However, effective usage of this data to inform key educational processes like student learning, teacher training and mentoring, and educational governance lacks severely behind. An effective use of data mining technique will help bring data driven decision-making to these environments.

1.8. Organization of the Thesis

Background and literature review is presented first in Chapter 2. The methodology to prepare educational data for rule mining and the proposed approach and algorithms to apply Apriori algorithm on large-scale educational data is explained in Chapter 3. The framework of available educational data, formation of educational baskets at different analytics levels, and the templates to mine effective association rules are explained in Chapter 4. Chapter 5 discusses the ETL steps used to transform the raw data into educational information in a structured format. Experiments related to the micro-, meso-, and macro-level analyses are presented in Chapters 6, 7, and 8 respectively. Finally, a discussion on this research, the limitation and future directions are given in Chapter 9.

Chapter 2 . Background and Literature Review

This chapter presents a review of relevant literature in prior research in Educational Data Mining (EDM), association mining, and specifically the use of Apriori algorithm in educational data mining.

2.1. Educational Data Mining

A survey of recent literature shows that while a host of data mining techniques have been used for analyzing educational data, the scope of application is mostly limited to the micro or meso-level of analytics [24]. A significant amount of research has been conducted at the meso-level of educational data mining. For example, Yohannes & Halim [25] proposed predictive models to enable the analysis of data from both school stakeholders perspective and students' and their parents perspective. Similarly, Baha et al. [26] used a host of machine learning techniques like Artificial Neural Networks, Support Vector Machines, Decision Trees, and Multinomial Logistic Regression to predict and analyze placement test scores using a large and feature rich dataset from Secondary Education Transition System in Turkey. This work aimed at improving the Secondary Education System by analyzing the structure of placement tests to make more effective and fair placement tests and assessment tools. Another representative work was performed by Alex J. Bowers [4], in which longitudinal data of grades from two districts of United States industrial Midwest state was subjected to hierarchical cluster analysis, and based on this longitudinal data from Kindergarten through Grade-12, students graduation status (e.g., on-time or late), or drop-out before graduation was predicted.

2.2. Apriori Algorithm Applications in EDM

The Apriori algorithm has been widely used to learn interesting trends and patterns in the education data. Literature review indicated that such an analysis was often conducted to satisfy a number of objectives as shown in Fig. 2.1. Applications of Apriori for each objective is briefly described next.

2.2.1. Course recommendation. The purpose of this type of analysis is to find courses that are often taken together or the courses whose results influence each other and to suggest these course grouping to students.

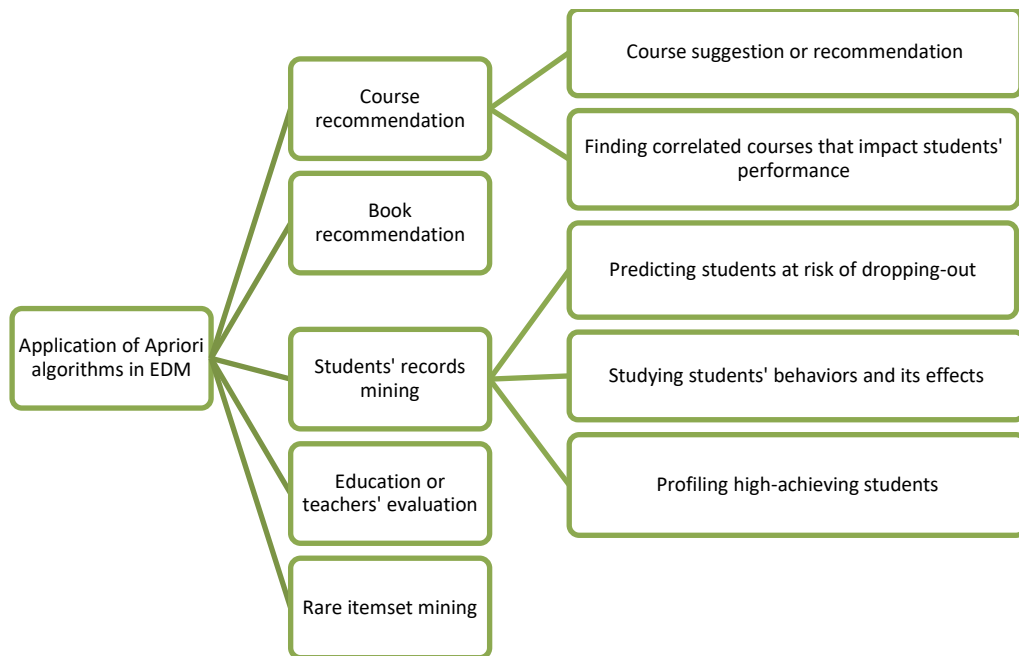


Fig. 2.1: Application of association rules to educational data

In this respect, authors Qiang Yang & Yanhong Hu proposed an improved Apriori algorithm that minimized run-time and scanned the database only once for candidate itemsets [27]. Candidate itemsets are sets of those frequent items between which association need to be identified and that fulfill the minimum coverage threshold. This work also focused on determining course correlations from college students' data. A similar technique of elective courses and research papers suggestion by means of Apriori algorithm was presented in [28]. In this work, a course registration system was implemented. In this system, the students were given elective course suggestions and research papers that were published by the faculty, based on the elective courses taken in the previous semester by the student, the data of elective courses taken in each semester by all students, and the domain of interest of students. Similarly, an Enhanced Apriori algorithm model (EAAM) was implemented in [29] for course suggestion using grade data of students from the previous semester. This model used improved filtration to prune infrequent itemsets and to optimize the computational efficiency of the Apriori algorithm.

Apriori was also used in conjunction with k-Means in many works to perform categorization and association analysis [14], [30], [31]. This is achieved by using k-Means as a preprocessing algorithm to cluster or organize similar records into groups, and then use Apriori algorithm on the obtained clusters to find frequent patterns from

each group. For example, courses can be clustered according to their fields of study or domain areas, and Apriori algorithm can be applied for each field to determine the courses that are taken together. One such work is presented in [32], in which course correlations (i.e., the courses that are mostly taken together by students) were discovered. Advancing this approach, Liu et al. [33] proposed a course recommender scheme for the Agriculture major students in China Open University. The Apriori algorithm was applied in two stages in this research to suggest local and global professional courses that corresponded to the same group from which the student had already selected some courses, or from an entirely different group.

Course correlation schemes can also be adapted to warn students early-on about subjects that are going to be difficult. One such research was performed by Mahatthanachai et al. [34] who used Apriori algorithm and classification to determine course correlations that influenced a high student dropout rate and factors that were responsible for drop-out of students in Chiang Mai Rajabhat University, Thailand. This system took various personal and academic variables into account like gender, occupation of parents, province of residence, previous GPA, previous education field, currently registered courses and results. From the results, it was concluded that the two primary factors for students drop-out were the previous education background and previous results and the courses that had an impact on students dropping out were Computer, English, Mathematics, and Physics courses. A similar micro-analysis on students data was performed by Maniar & Khanna [35] to alert students about courses that were going to be difficult based on the courses taken for the current term. Auto-Adjust Apriori algorithm was employed for this purpose that uses both Apriori algorithm and linear regression. The Apriori algorithm found frequent subjects and formed subject sets from student results, and by linear regression, the difficult courses were predicted for the next term, given the courses students find difficult in the current term.

Course recommendation schemes are not just useful in traditional colleges and universities but are utilized for Massive Open Online Courses (MOOCs) as well. MOOCs have gained popularity nowadays due to their open access via the Internet and unlimited participation prospects. The work by Zhang et al. [36] presents a MOOC-oriented course recommendation system (MCRS) that uses Apriori algorithm on the

Spark framework [37]. This system displayed significant improvement in execution time when compared with the implementation of Apriori algorithm on a Hadoop platform [38].

2.2.2. Studying students' behavior and its effect. This kind of analysis deals with mining general patterns from students' academic data to study overall student performance and the factors causing such behavior. For example, Merceron et al. [39] discussed different interestingness measures of association rules, apart from the traditional goodness measures like support, confidence, and argued that these measures are more suited to educational data. With the help of measures like cosine similarity and lift, teachers can effectively decide on whether to keep a rule or discard it. Cosine similarity uses a similarity between vectors metric to determine relationship between two items [39]. Lift, on the other hand is a measurement of the relative confidence of that pattern [39].

In 2009, Romero et al. [31] mined real web-based educational data with AprioriAll, GSP (Generalized Sequential Pattern algorithm), and PrefixSpan (Prefix-projected Sequential pattern mining) algorithms that work by building subsequences from the data. These algorithms were used to discover personalized recommendation links for students. The sequential mining was performed on the data with and without clustering by using K-means algorithm, and performance of both methods was compared.

The work in [24] by Borkar and Rajeswari employed Apriori algorithm without any variations to a dataset of 60 students, and generated frequent itemsets and rules based on this data. The students' records contained quantitative values that were first translated to nominal subjectively as Good, Average, and Poor. In this work, the correlation between different attributes was also found to provide a big picture of attributes dependability upon each other. Similarly, in 2010, Loraine Charlet & Kumar found frequent itemsets from a class of 28 students who took 74 courses [40]. The Apriori algorithm's detailed account of each step was also provided in this work with the pruning of itemsets before generating final rules. In [41], Parack et al. applied Apriori on student records without any variation to find frequent itemsets and interesting rules with the thresholds of support and confidence.

Similarly, Matetic et al. in [42] used Apriori algorithm to study the effects of course activities like video lectures, quizzes, and self-assessments to predict students' overall performance in the introduction to programming course. These course activities data were generated from log files of a learning platform. Wang et al. [43] used a similar approach in 2016 to study the causal relationship between the behavior of students obtained from the MOOC platform (Massive Open Online Courses) and its effect on score.

2.2.3. Teachers' evaluation. In these studies, the objective was to improve education and teaching quality by analyzing data from student surveys, and questionnaires etc. Using this approach, Deng et al. in [44] used an improved Apriori algorithm to perform mining on teaching evaluation surveys filled by students. This research used teacher's age, degree, designation, teaching attitude to allow decision making and improvement in teaching quality. In 2017, Mao et al. [45] proposed an improved Apriori-gen algorithm to mine student self-evaluation, teachers' evaluation, and environment factors data by using the answers obtained from a questionnaire. These rules were aimed at improving teaching in ideological and political education courses.

2.2.4. Predicting students at risk. This approach identified factors that have a negative impact on student performance. One study was conducted by Ahmed et al. to predict the students' retention, drop-out or graduation status by taking into account both the academic and personal information of students of the Bangladesh University of Engineering and Technology (BUET) [46]. In this research, various socio-economic factors of a developing country were also studied with their effect on students' graduation status. In another work by Guerrero & Ambat [47] Apriori algorithm was used to learn student patterns from Cisco academic records to predict which students will fail the Cisco certification. This research also used regression analysis to determine the most important attributes to predict students' success in the program.

The students at risk of dropping-out were also predicted on a MOOC platform in a study by Srilekshmi et al. [48]. The authors tried to predict students who may drop out at any stage of the MOOC or did not meet the passing criteria to earn certification. This research was based on the students' activity records in HarvardX and MITx courses on the edX platform. The system was refined using different workflows and

obtained associations were also tested for goodness by various measures like curve fitting and root mean square error.

2.2.5. Profiling high-achieving students. High-achieving students can be profiled by mining their personal and academic attributes to serve as a guideline for new students. In the research performed by Mouri et al. in [49], high-achievers were profiled by means of Apriori algorithm using the logs obtained by document viewer Booklooper [50] which is a cloud-service for downloading and reading eBooks. Answers obtained by a questionnaire were also used to analyze daily habits of high-achieving students like the time they woke up, how many hours they studied for, etc. A template-based approach was used to mine rules with defined consequent like “mid-term score=Good”, “final grade=Good”, to mine only data for students with good results. Alternatively, the students who got admitted to world’s top universities with funding were mined in the study performed by Ahmed et al. in [51]. The academic profiles of successful applicants were mined using Predictive Apriori algorithm. The influence of different attributes in the academic profile like undergraduate CGPA, GRE, IELTS, TOEFL and other standardized test scores, research, and job experience were also studied for their predictive accuracy value in this work.

2.2.6. Book recommendation. The result of this analysis were the books that were often lent together from the library. In the research by Teng et al. [52] book recommendations were given from university library circulation database based on A-FAHP, a technique which uses Apriori algorithm as the first step combined with a proposed FAHP (Fuzzy Analytic Hierarchy Process) algorithm as the second step. FAHP uses fuzzy logic to further rank the books recommended by Apriori algorithm according to different criteria of support, confidence, the readers’ school, press release of the book etc. The final books recommended to the user were evaluated by metrics like precision and frequency and compared against collaborative filtering for goodness. Collaborative filtering is based on the idea that if two students A and B borrowed the same book from the circulation library, then A is likely to borrow the same book as B in future, rather than the books rented by other students.

2.2.7. Rare itemset mining. For educational data, it is very common to have both normal and rare (i.e., low-base rate) behaviors. For example, the students passing out of a school are much more than the students dropping out. This imbalanced data

leads to the problem of finding frequent or rare patterns. Both frequent and rare patterns are interesting. The work in [53] by Romero et al. addresses the mining of rare association rules from educational data. The researchers used four different variants of Apriori in this work to mine rare itemsets. These itemsets were used for discovering patterns that do not apply to a large population of the data but were important to rare and crucial learning cases. This kind of association mining is useful to look into imbalanced educational data. In a similar approach, the co-occurrence of courses in which low marks can predict student drop-out were mined by Chen and Chang [54] by means of a fuzzy association rule mining algorithm called Fuzzy Apriori Rare Itemset Mining (FARIM). FARIM was proposed in [55] and works with a new threshold called “Rank” to find rare or infrequent itemsets from students data. FARIM translates the result data of each course into linguistic variables of “Good”, “Medium”, and “Poor” by using some membership functions. Rank and support values are then calculated for candidate itemsets at each level with respect to these functions. The process of FARIM is shown in Fig. 2.2. Then association rules are mined from this data by selecting low rank itemsets to gain information about rare patterns in the learning outcome of students.

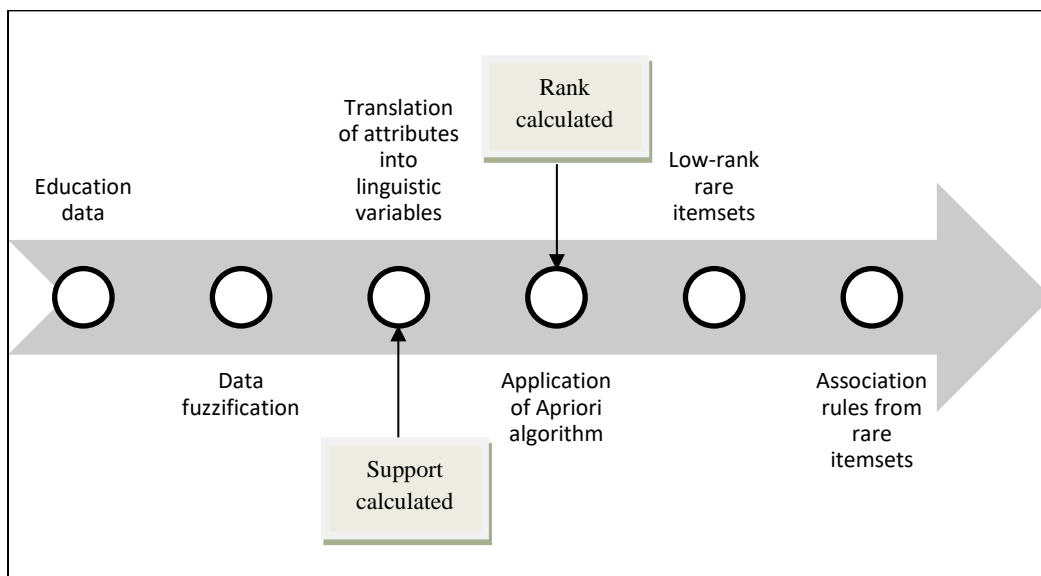


Fig. 2.2: FARIM algorithm process illustration [55]

2.3. Summary and Relevance

A summary of previous work conducted in the area of educational data mining using Apriori algorithm is provided in Table 2.1. The previous research works are

categorized according to their domain areas in Table 2.1 and the itemsets that were supposedly used in these works to mine the educational data are also listed. From the retail point of view, an itemset is the collection of products that are bought in one supermarket transaction. For the research works conducted for the purposes of course recommendation, the itemsets contain the subjects and obtained grades information of past students. The research targeted in the area of students' behavior and effect analysis used students' overall academic records containing homework scores, midterm grades, attendance, and online course activity data. Teachers' background data and data from evaluation surveys and questionnaires were used to conduct research aimed at finding information from teaching evaluation. Students' personal and academic data along with their graduation status were used to predict students who were high-achievers or were at risks of dropping out. Finally, circulation library's data containing various books information were used to recommend books to students and faculty.

Table 2.1: Prior work in EDM using Apriori algorithm

Research Work	Domain	Itemset Description
[27], [28], [29], [32], [33], [34], [35], [36]	Course recommendation or correlation	Subjects registered by students each semester: {Sub ₁ , Sub ₂ , ..., Sub _n } Grades obtained in each subject: {G_Sub ₁ , G_Sub ₂ , ..., G_Sub _n }
[24], [39], [31], [40], [41], [42], [43]	Students' behavior and effect analysis	Students' academic data: For example, {StudID ₁ _GPA, StudID ₁ _MidtermScore, StudID ₁ _HW1Score, StudID ₁ _HW2Score, StudID ₁ _HW3Score, StudID ₁ _Attendance} Students' course activity data (online): {StudID ₁ _VideoLecture1ViewStatus, ... StudID ₁ _VideoLectureXViewStatus, StudID ₁ _Quiz1Score, StudID ₁ _Quiz2Score}
[44], [45]	Education and teaching evaluation	Teachers' characteristic data. For example, designation, qualification, age etc. {TID ₁ _Designation, TID ₁ _Qualification, TID ₁ _Age, TID ₁ _Experience} Data from teachers' surveys, questionnaires for evaluation. For example, {StudID ₁ , Ques ₁ _RespA, Ques ₂ _RespC, Ques ₃ _RespA, Ques ₄ _RespB}
[46], [47], [48], [49], [51]	Analysis to predict students who are high-achievers, or who are at risk of dropping-out or failing	Students' academic data Students' course activity data (online) Students' personal data (e.g. socio-economic factors, age, gender, occupation etc.) {StudID ₁ _Age, StudID ₁ _Gender, StudID ₁ _FatherOccupation, StudID ₁ _FinancialStatus} Students' passing out, dropping-out, or failing status {StudID ₁ _GraduationStatus}
[52]	Book recommendation	Circulation library data {Book ₁ _Name, Book ₁ _PressRelease, Book ₁ _ReaderSchool}
[53], [54], [55]	Rare itemset mining	Any kind of education data that needs to be mined for rare itemsets

Chapter 3 . Proposed Methodology

This chapter outlines the steps of a methodology to mine effective association rules from the large-scale education data. An illustration of this methodology is shown in Fig. 3.1 with the help of a flowchart. Each step in the flow-chart is described next.

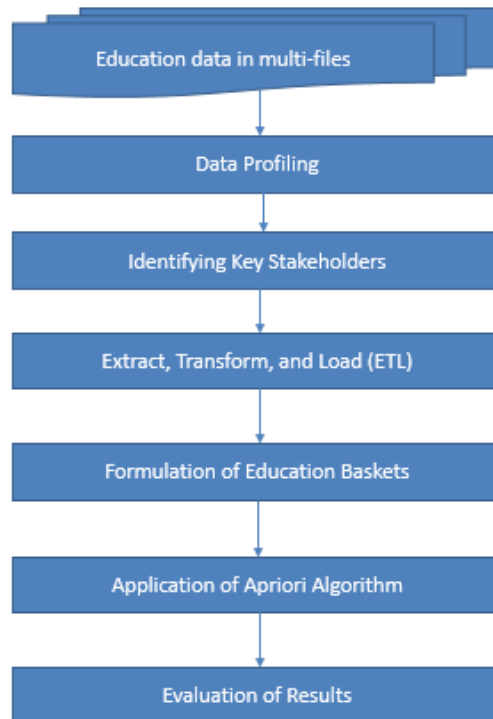


Fig. 3.1: The data mining process

3.1. Data Profiling

Data profiling is the first step in data mining. The purpose of this step is to understand the data and to determine the classes or objects present in the data. For example, it is determined whether the data is temporal (i.e., collected over a period, or not), whether it is to be processed in real time or not, etc. In addition, the type of variables and their values are determined. Finally, the reliability, accuracy, and usefulness of the data is determined. The data profiling is done by studying the 4Vs of data. The details of this step were provided in Table 1.2 in Chapter 1.

3.2. Identifying Key Stakeholders

The second step in the methodology is an identification of key stakeholders. The determination of education questions that can be answered with respect to

different key stakeholders is essential for the rule mining method because the patterns can be mined at different levels, ranging from grade to school to cluster of schools.

The stakeholders in an educational system are the people who take the decisions or are responsible for generating the data to help make informed decisions. As shown in Fig. 3.2, the stakeholders operate at all the levels of educational analytics. For example, at macro-level, the obvious stakeholders are policy makers, politicians, and administrative decision makers. At the meso-level, there could be data collecting personnel and regional heads. At the micro-level, one would expect data creating people within a school like students, parents, teachers, principals etc.

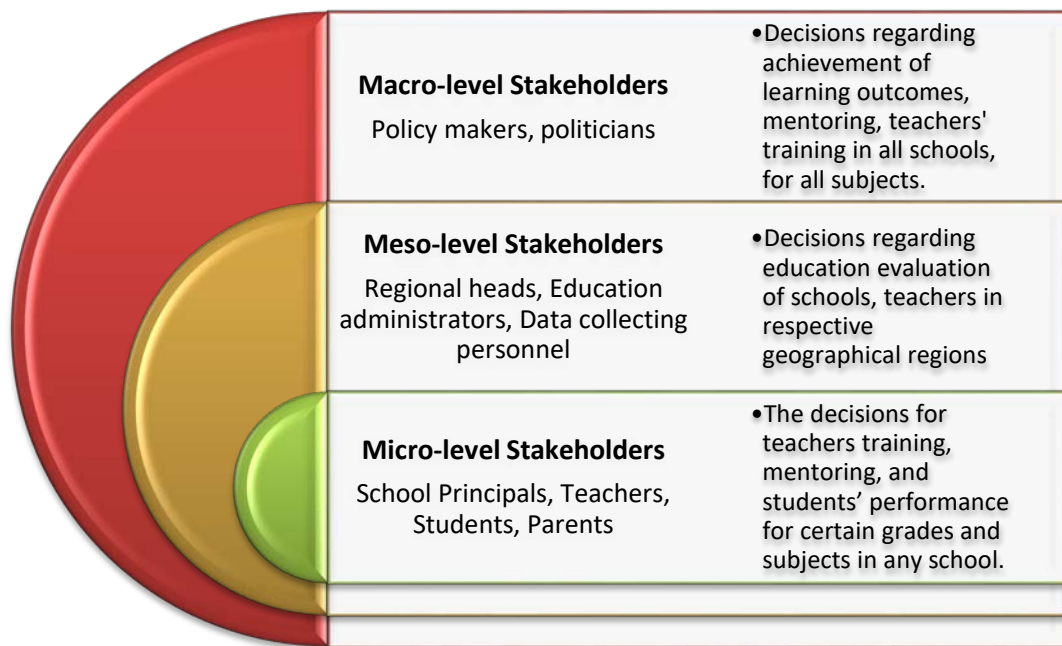


Fig. 3.2: Relationship of stakeholders to different educational levels of analytics

The questions listed in Chapter 1 are from the perspective of these stakeholders operating at different levels of educational analytics. The categorization of the questions with respect to stakeholders at different levels is given below:

3.2.1. Macro-level. The macro-level stakeholders can benefit from the answers to types of questions like:

- How does the performance of students in one location relates to other students in a different location?
- What are the characteristics of schools with different enrolment sizes?

- How do the teacher variables in one location relate to teacher variables in different schools and locations?
- What is the relationship between groups of teachers with good/average/bad ranking?
- How do assessment indicators vary with different data collecting personnel?
- How does teachers' workload relate to their performance?
- How do the student learning outcomes vary for different locations?
- How does the distance of school from the training and monitoring centre relate to the school and teachers ranking?
- How does the distance of school from the training and monitoring centre relate to the recorded assessment indicators?
- Are the learning outcomes achieved in all locations? If yes, then to what extent?
- How do the learning outcomes achieved in one location vary as opposed to other locations?

3.2.2. Meso-level. The meso-level stakeholders can consider issues within their own geographical area of interest with a group of schools. These stakeholders can benefit from the answers to types of questions given below:

- How does the performance of students in one school relates to other students in different schools of the same geographical location?
- What aspects of school with good ranking and performance can be shared in other schools?
- How do the teacher variables in one school relate to teacher variables in different schools?
- Which teachers in a group of schools need training for different subjects?
- How does the training received by teachers relate to their performance and ranking?
- How do the student learning outcomes vary for different schools?
- How do the student learning outcomes vary for different teachers?
- How do the student learning outcomes vary for different subjects?

3.2.3. Micro-level. Finally, the micro-level stakeholders can benefit from the answers to questions like:

- Which students in a school need attention for different subjects?
- Which teachers in a school need training in different areas?

3.3. Extract, Transform, and Load (ETL)

The ETL process consists of all the pre-processing steps that need to be applied on the data to apply the rule learning algorithm. The Extract step of ETL caters to the extraction of unstructured data to some structural form like tables etc., that can be represented in a structured notation such as the Unified Modeling Language (UML). This structural representation will help in the formation of education baskets described later in Section 3.4.

The Transform step in ETL consists of sub-steps which are detailed below:

3.3.1. Data cleansing. This is the first step in the data transformation. The large-scale educational data from a developing country will have low veracity because it is prone to human errors like typographical and measurement errors, and missing data. Because dirty data reduces the reliability of the data, the data needs to be cleansed by replacing missing values with substitutes if possible, or otherwise, removing the attributes or instances that are missing information. The data needs to be checked for consistency and the presence of any outliers for any out-of-range values, misspelled values, etc. These values need to be sought in the data by using bar plots, or histograms, or commands that can generate some sort of summary for each variable.

Missing data pose a major challenge for any data because such data reduces the classification performance of any data mining tool. However, to perform rule-mining on educational data, there is no need to remove the records or variables with missing data, because the data takes the form of education baskets or transactions where a value can be present or absent like a supermarket transaction, and the rules are sought from only the present values of items in a transaction.

3.3.2. Attribute selection. This is the process of removing irrelevant attributes from the data. This process is essential because there are many dependent and irrelevant attributes that can be found in multi-level educational data. For example, variables such as regional head's name may be a single one for each region, so it can be removed when performing analysis for a region. Also, there can be dependent attributes like the ID and Name of teachers, which refer to the same teacher, so these can also be removed since

the goal of this kind of mining is to study the characteristics of different types of data rather than knowing details about the teacher, or school, or subject that is not performing up to the mark.

3.3.3. Attribute discretization. The second step in Transform is the Attribute Discretization which is often called binning. Discretization is the process of putting values into bins with each bin consisting of a specific range of values. To discretize the data, various approaches can be used. For association mining the binning method should be unsupervised because it should not depend on an outcome variable. Some of the unsupervised attribute discretization techniques are listed below (e.g., [56]).

- **Quartiles:** The median is used to divide the ordered attribute into two halves. The median is not included in either half. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.
- **Percentiles:** A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
- **Bi-modal Discretization:** This method combines a mixture of two normal distributions with the same variance but different means.
- **Clustering:** Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups. K-means as explained in Chapter 2 can be used to cluster the variables into various groups or bins for the application of association mining.
- **Equal-interval binning:** Also called equal-width binning. By this approach, the attribute can be divided into k intervals of equal size.
- **Equal-frequency binning:** Also called histogram equalization. This approach divides the data into k groups where each group contains approximately same number of values.
- **Median binning:** This is the usually the next step of equal-interval/frequency binning in which the data in each interval is smoothed by using median as the data label in that interval.

3.3.4. Projections. The final pre-processing step is Projections. This step consists of performing simple transformations on the data for performance improvement. For example, in the educational data some values like the total number of boys and girls in a class may not be useful on their own but taking the ratio of these attributes give information about the boys-to-girls ratio in the classroom. Other projections like, adding noise to the data, selectively removing data, adding ratio of two numeric quantities, taking difference of two date attributes etc., can also be performed.

By performing all the ETL steps, the data will be ready to feed into the rule-mining algorithm for further analysis and evaluation.

3.4. Formulation of Education Baskets

In this section, an approach to form educational baskets at different levels of educational analytics is presented first, followed by the description of key concepts that are used in conjunction with association mining and market basket analysis.

3.4.1. Approach. The education variables are grouped into different baskets according to various objectives at different levels of educational analytics. The objectives at these levels can be to study the outcome variables, or to study the effect of any characteristic of school, teacher etc. on outcome variables, for example. At each level, a data subset or a generic educational basket can be obtained that is representative of the variables that are operational on the particular level of analytics.

According to the definition of different levels of educational analytics in Table 1.1, the macro-level integrates the data from the lower levels of meso and micro to enable educational benchmarking between different regions, and the meso-level integrates the data from the micro-level to compare various schools' performances and results. Finally, the micro-level analysis operates within a school to analyse groups of students' results and academic patterns. Fig. 3.3 shows a generic model of data integration at macro and meso levels that is used to obtain educational baskets at all three levels of analysis.

A formal description of each key concept in market basket analysis is given below:

3.4.2. Items. Items are the objects between which associations need to be identified. For example, for a supermarket, each item i_k is a product in the supermarket.

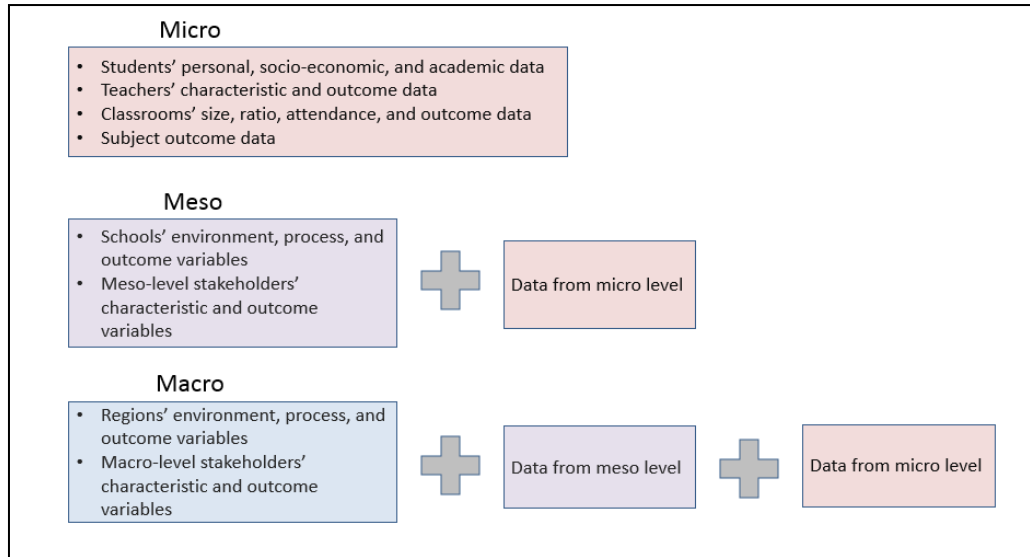


Fig. 3.3: Generic model of data integration at macro and meso levels

3.4.3. Itemset. A group of items i_1, i_2, \dots, i_n that occur together are called an item set.

$$I = \{i_1, i_2, \dots, i_n\} \quad (1)$$

3.4.4. Transactions. Transactions are instances of groups of items co-occurring together. For example,

$$T = \{i_1, i_6, i_{14}\} \quad (2)$$

For a super market, this may consist of the set of products bought together. An example of transactions in a supermarket database is given in Table 3.1.

Table 3.1: Example transactions

TID	Items
1	Milk, sugar
2	Milk, bread, eggs
3	Eggs, milk, butter
4	Juice, eggs, bread
5	Milk, juice, cookies

3.4.5. Rules. Rules are statements of the form:

$$\{i_1, i_2, \dots\} \Rightarrow \{i_k\} \quad (3)$$

That is, if there are items i_1, i_2, \dots , in the item set on the left-hand side (LHS) of the rule, then it is likely that a visitor will be interested in the item on the right-hand side (RHS) of the rule (i.e. i_k). Example rules from the transactions in Table 3.1 are

shown in Fig. 3.4. The first rule in Fig. 3.4 shows that when Milk is bought, one would also expect Eggs to be bought as well.

[1]	Milk \Rightarrow Eggs
[2]	Eggs \Rightarrow Bread

Fig. 3.4: Example supermarket rules

The output of a market basket analysis is generally a set of rules that can be exploited to make business decisions related to marketing or product placement [22], [57].

3.4.6. Educational baskets and rules. In educational data mining, the items are “*variable=value*” pair of all items that are present in an educational basket. The educational basket is created with respect to each level of analysis as explained in Section 3.4.1. Example items in an educational basket are shown in Fig. 3.5.

<i>StudentMarks = Good, TeacherQualification = MASTERS, SchoolLevel = Primary</i>

Fig. 3.5: Example educational items

A set of educational transactions or baskets can be created with the help of different constraints that are specific to an educational question or situation. For example, to study the characteristics of students, teachers, grade etc. that relate to students’ performance in the subject of Mathematics, a basket can be created with the following variables to mine the rules specific to the question:

- Students’ academic profile (Grade, Assessment marks, Homework scores, Quiz scores) in the subject of Mathematics
- Mathematics teachers’ qualification, age, experience, designation
- Environment variables like classroom size, attendance of students etc.

When educational data subset is created with the above variables only, the purpose of formation of educational baskets is achieved. These baskets can contain the example transactions, as shown in Table 3.2. A notable observation in Table 3.2 is that these transactions do not contain identity attributes for students, or teachers, and do not contain all the above listed variables. So, missing values are allowed in educational baskets which simply mean that an item is not present in the transaction.

Finally, rules can be created by the application of Apriori algorithm (details in Section 3.5) from the educational transactions (Table 3.2), as shown in Fig. 3.6.

Table 3.2: Example educational transactions

TID	Educational Items
1	Grade = III, Attendance = Average, AssessmentMarks = Poor, FinalScore = Poor, TeacherAge = Less than 30 years, TeacherQualification = MASTERS, Classroomsize = [20-30] students
2	Grade = III, Attendance = Good, AssessmentMarks = Good, HomeworkScores = Good, FinalScore = Good, TeacherAge = Less than 30 years, TeacherQualification = MASTERS, Classroomsize = [20-30] students
3	Grade = III, Attendance = Good, AssessmentMarks = Poor, QuizScores = Good, HomeworkScores = Good, FinalScore = Average, TeacherAge = Less than 30 years, TeacherQualification = MASTERS, Classroomsize = [20-30] students
4	Grade = IV, Attendance = Average, AssessmentMarks = Good, HomeworkScores = Average, FinalScore = Good, TeacherAge = [30-45) years, TeacherQualification = MATRIC, Classroomsize = [30-40) students
5	Grade = V, AssessmentMarks = Good, QuizScores = Good, FinalScore = Good, TeacherAge = Less than 30 years, TeacherQualification = MASTERS

[1]	$TeacherQualification = MASTERS \Rightarrow TeacherAge = Less\ than\ 30\ years$
[2]	$Grade = III \Rightarrow TeacherAge = Less\ than\ 30\ years$
[3]	$AssessmentMarks = Good \Rightarrow FinalScore = Good$
[4]	$Grade = III, TeacherQualification = MASTERS \Rightarrow TeacherAge = Less\ than\ 30\ years$

Fig. 3.6: Example educational rules

3.5. Application of Apriori Algorithm

The Apriori algorithm works by performing a level-wise search to find frequent itemsets. The algorithm works on the principle that if an itemset is infrequent, then its super sets are infrequent too. The algorithm successively forms longer itemsets from shorter ones with a generate and test strategy against the threshold of minimum support. Minimum support is the frequency of itemsets that constitute a rule. The Apriori Algorithm is applied to educational baskets where, each of the baskets belong to one of the three categories of analyses; macro, micro, or meso.

The Apriori algorithm requires various parameter and appearance inputs. The parameter inputs are the following:

- 1. Minimum support:** specifies the minimum fraction of transactions that contain the itemsets of both antecedents and consequent of the rule. For example, the support values of rules in Fig. 3.4 can be determined as shown in Table 3.3. For example, $Support(Milk) = 4/5$ because Milk occurs in 4 out of 5 transactions. Minimum support is a user specified parameter, with a value in the range of [0,1]. A minimum support of 1 would mean that the item must appear in all transactions. Typically, minimum support values are determined by trial-and-error. A starting minimum support value of 0.1 can be used to see how many rules are generated, and whether there are any interesting rules obtained. If no

rules are generated at this value, the support is decreased to a value where the frequency of itemsets is not too low, and some interesting rules are obtained.

Table 3.3: Example of support measure

TID	Items	Support = Frequency/Total Transactions
1	Milk, sugar	Total transactions =5 Support(Milk) = 4/5 = 0.8 Support(Eggs) = 3/5 = 0.6 Support(Milk, eggs) = 2/5 = 0.4 Support(Eggs, bread) = 2/5 = 0.4
2	Milk, bread, eggs	
3	Eggs, milk, butter	
4	Juice, eggs, bread	
5	Milk, juice, cookies	

- 2. Minimum confidence:** specifies the percentage of transactions that contain the consequent of the rule if the antecedent is present. For example, for the rules in Fig. 3.4, corresponding to the example transactions in Table 3.1, the confidence values are determined as shown in Fig. 3.7. As Table 3.1 shows, in 2 of the four transactions, Milk and Eggs are both present.

$Confidence(Milk \Rightarrow Eggs) = \frac{2}{4} = 0.5$ $Confidence(Eggs \Rightarrow Bread) = \frac{2}{3} = 0.66$

Fig. 3.7: Confidence values for supermarket rules

Again, minimum confidence is determined by trial-and-error methodology, a starting minimum confidence of 0.85 can be used to obtain rules where RHS is present at least 85% of the times LHS is present. The confidence can be increased to obtain high-confidence rules where the relationship of LHS and RHS is true most of the times.

- 3. Maximum itemsets:** refers to the maximum number of itemsets that can be present in the resulting ruleset.
- 4. Maximal time for subset checking:** refers to the maximum time the algorithms spends in subset checking.

3.5.1. Templates for rule generation. The rules generated from the Apriori algorithm can be constrained using the appearance input to the algorithm, where the rule structure can be specified using the position of various itemsets as antecedents or consequents. So, instead of returning all the rules, the algorithm only returns rules that satisfy a user-defined template. This approach is termed as the template-based approach

to rule generation. This is useful when cause and effect relationship between variables is studied.

Continuing the example of educational transactions and rules in Table 3.2 and Fig. 3.6 respectively, if the obtained rules use a template that allows only the variable *FinalScore* to appear as a consequent, then an example rule based on this template is shown in Fig. 3.8.

[1] <i>AssessmentMarks = Good</i> \Rightarrow <i>FinalScore = Good</i>
--

Fig. 3.8: Template-based educational rule

3.6. Evaluation Approach

The application of Apriori algorithm generates a set of rules, among these rules some are useful, while others are not. There are many different methods to evaluate the goodness of rules like objective interestingness measures, subjective interestingness measures, and visualization [58], [59]. The objective measures use measures like support, confidence, and lift to determine the objective goodness of a rule. Alternative objective evaluation techniques like correlation analysis and IS measure also exist, however these measures are more suitable for nominal and symmetric binary variables, and not as well suited for data with numeric and n-ary variables [15]. For this thesis, objective interestingness measures of support, confidence, and lift were used with the aid of visualization. These metrics are defined next.

3.6.1. Objective evaluation metrics. Given below are the evaluation metrics which are used to assess the rules for interestingness.

- **Support:** The support of an item or item set is the fraction of transactions in the data set that contain that item or item set. Intuitively, support measures relative frequency of the items in the dataset. For example, for an itemset that occurs in at least n observations, the support would be:

$$Support = \frac{n}{number\ of\ rows\ in\ the\ transaction} \quad (4)$$

A Support of 1 means that the itemset occurs in all the transactions.

- **Confidence:** The confidence of a rule is the likelihood that in a new transaction the items on the RHS will also be present if the items on the LHS of the rule are

present. Intuitively, confidence measures how good the rule is at prediction. Formally:

$$Confidence(i_m \Rightarrow i_n) = \frac{Support(i_m \cup i_n)}{Support(i_m)} \quad (5)$$

A Confidence of 1 means that both the LHS and RHS of the rule always occur together.

- **Lift:** The lift of a rule is the ratio of the support of the items on the LHS of the rule co-occurring with items on the RHS divided by probability that the LHS and RHS co-occur if the two are independent. Intuitively, lift measures the degree to which LHS and RHS are related to each other.

$$Lift(i_m \Rightarrow i_n) = \frac{Support(i_m \cup i_n)}{Support(i_m) \times Support(i_n)} \quad (6)$$

A Lift of 1 means that i_m and i_n are independent that is no association is present between them. A higher lift value indicates that the co-occurrence of LHS and RHS in a transaction is not random but due to some relationship between them.

3.6.2. Subjective evaluation metrics. The subjective measure uses domain knowledge to discard non-actionable and obvious rules, and this measure can be typically applied by the stakeholders or domain experts. The subjective evaluation for educational data can be conducted by the domain experts to rank the obtained rules based on rules that are obvious, or the rules that advance their knowledge and are interesting, or the rules that contradict their knowledge and should be further explored.

3.6.3. Visualization plots. Visualization plots are generated using the objective evaluation metrics of support, confidence, and lift to highlight interesting rules. The various plots that can be studied using a package like *arulesViz* [60] for different experiments are described below:

- **Scatter plot:** Scatter plot shows the rules by their objective interestingness parameters of support and confidence on the axes, and the color (shading) of the point represents the third parameter lift. For example, in Fig. 3.9, the dark points (red) are the most interesting rules with high lift but low support. These rules have a confidence of 90-100% as denoted by the vertical axis, and a support of 0.01 to 0.05 for the itemsets in both the LHS and RHS of the rule. Some high

support rules (support greater than 0.2) with low lift can also be seen towards the right side of the plot. The interactive version of this plot enables the user to see the rule represented by a specific point.

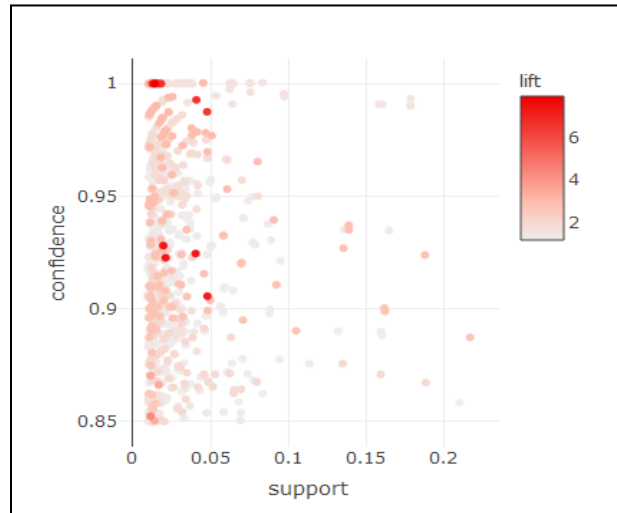


Fig. 3.9: An example scatter plot

- Matrix plot:** The second visualization plot is Matrix plot which arranges the association rules as a matrix with the itemsets (shown by numbers) in the antecedents and consequents on the x and y axes respectively. The color of the box shows support for the itemsets present in both the antecedents and consequent of the rule. In Fig. 3.10, the antecedents 10-12 with consequent 2, antecedents 16, 17 with consequent 4, antecedent 20 with consequent 5, and antecedents 44-48 with consequent 7 formulate high-support rules, for example.

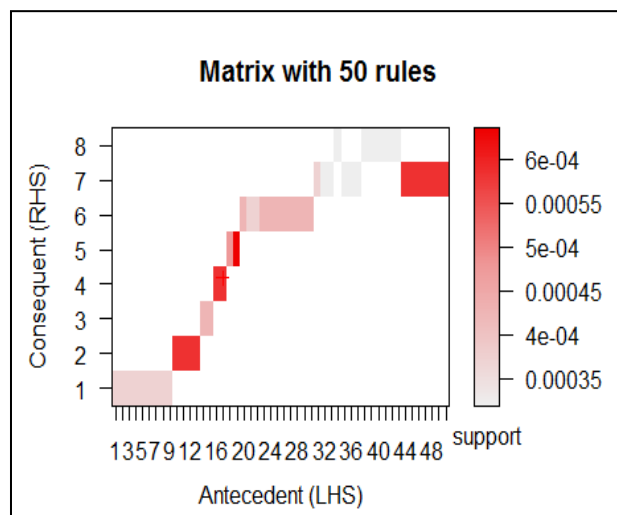


Fig. 3.10: An example matrix plot

- Parallel coordinates plot:** The parallel coordinates plot represents the itemsets on the y axis with their respective positions in the rules on the x axis. The parameter of support for each rule is represented by the darkness of the arrow. For example, Fig. 3.11 shows a parallel coordinates plot which highlights 3 interesting rules with dark red arrows and shows the itemsets that are present at the 1st, 2nd, and 3rd positions as antecedents, and at the 4th position as consequent. The interesting itemsets that can be seen from this plot are given in Table 3.4.

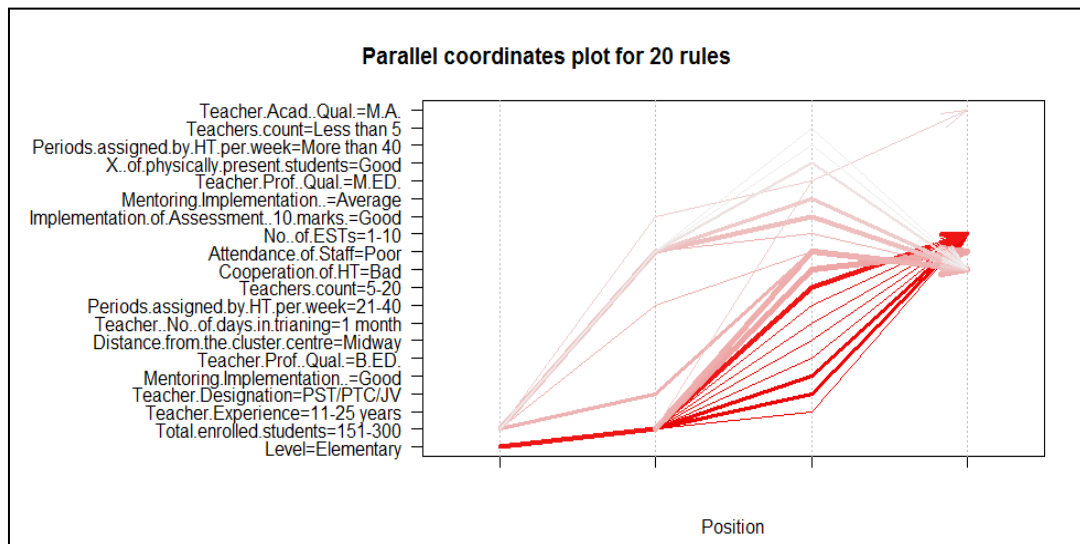


Fig. 3.11: An example parallel coordinates plot

Table 3.4: Interesting itemsets from Fig. 3.11

Position	Interesting itemsets determined by parallel coordinates plot
1	Level = Elementary
2	Total enrolled students = 151-300
	Teacher designation = PST or PTC or JV
3	Mentoring implementation = Good
	Teachers' count = 5-20
4	Number of ESTs = 1-10 (consequent)

3.7. Summary

The steps that are needed to obtain a concept description for the large-scale educational data are summarized in Table 3.5. These steps are not specific to the problem of rule discovery in educational analytics and can be applied to any big data to mine specific association rules that give a solution or description of some pre-defined questions.

Table 3.5: Summary of the rule mining process

S. No.	Steps	Description
1.	Data Profiling	Determine data quality with 4 Vs
2.	Identifying stakeholders	To formulate different scenarios or questions with respect to all stakeholders that can be solved by the rule discovery
3.	ETL	Steps needed to cleanse, pre-process, and transform the data to concoct it as input to the rule mining algorithm
4.	Formulation of Baskets	Creation of data subsets with respect to the given problem or question.
5.	Application of rule mining algorithm	To achieve this step, various thresholds like support and confidence are used that constraint the rules, and then rules are mined for each of the baskets.
6.	Evaluation of Results	Can be objective, subjective, or by visualization (or a combination of these) to determine interestingness of rules: Objective: Uses different metrics like support, confidence, and lift etc. Subjective: Analysis of rules by a domain expert. Visualization: Uses objective evaluation metrics to highlight rules in different plots.

Chapter 4 . Case Study: Applying the Methodology

In this chapter, the proposed methodology is applied to a large-scale educational data set. This study is based on educational data being generated from a Continuous Professional Development (CPD) framework of a developing country. This means that the data-driven decision making needs to help improve the CPD educational process.

4.1. Case Study

The data is obtained from grades 3-5 of primary public schools in one district of a developing country. A district is an administrative division that is managed by the local government of the country. The schools in this district are geographically located in the form of clusters; each school cluster is approximately 6 km in diameter. The number of schools in each cluster can vary from 11-45. The education provided in the government schools of the district follows the CPD framework implemented by the government.

The data created from the CPD process has multiple inputs, assessment, and reports for each cluster, school, subject, teacher, and grade. An illustration of the distribution of the schools and clusters in this education landscape is presented in Fig. 4.1. This figure also shows the scope of various levels of education.

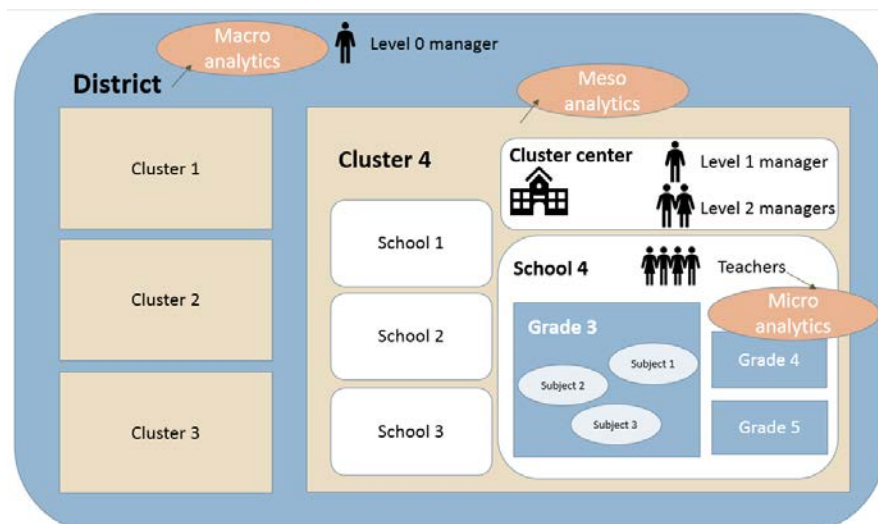


Fig. 4.1: Big picture of education data and stakeholders for the CPD process

Fig. 4.1 shows the educational landscape in a district. Each district has several clusters. Each of which has a cluster centre and 11-45 schools. The cluster centre is the

host for teacher training activities and records maintenance for all the schools. In the CPD process, Level 0-Manager (L0M) works as the district administrator, while the Level 1-Manager (L1M) and Level 2-Managers (L2Ms) work at the cluster centre and are responsible for the educational data maintenance and collection respectively. Data from grades 3, 4, and 5 are recorded for each primary school in the cluster for the subjects of English, Mathematics, Science, Social Studies (SS), General Knowledge (GK), National Language (NL), and Religion. As illustrated in Fig. 4.1, the macro-level analysis is performed across clusters (i.e., for a region or district), the meso-level analysis is performed across schools (i.e., for a school cluster), and finally, the micro-level analysis is performed across teachers, grades, and subjects within a school.

4.2. The CPD Framework

4.2.1. Objective. The objective of CPD framework being studied was to have knowledgeable, well-trained, and motivated teachers and education personnel to impart quality education to students in the government schools of the country.

4.2.2. Organizations and personnel. Under the CPD framework, the education department was responsible for assessment, mentoring, professional development, reporting, coordination, and monitoring of teachers. Each cluster had a cluster centre school that maintained records of all the schools in the cluster. The cluster centre was also the host for round-the-year teacher development activities and was also responsible for reporting the educational data to the district centre and the centralized Teacher Training Institute (TTI) which was the organization responsible for preparation of educational reports at the provincial level.

The management in the CPD framework worked at district and cluster centres. L0M was the district administrator and monitored all the clusters' performance by reviewing their results and observing the clusters in which educational benchmarks were not met. An L1M was responsible for the bookkeeping of the records of all schools in their respective cluster (comprising of 11-45 schools). The schools in each cluster were distributed between 1 to 3 L2Ms. The L2Ms were responsible for the data collection of all the school and teacher indicators from their respective allocated schools. The L2Ms were trained under the CPD framework at the district centres and were trained for performing the effective assessment of the learning system by giving scores to the teachers, head teachers, schools, and clusters. This manual ranking was

then used to generate reports which aided the training, mentoring, and professional development of the teachers. This report was sent to the LOM for timely intervention in clusters which were not achieving good results.

4.2.3. Mentoring areas. Teachers were mentored in various aspects of teaching. Some of these trainings were specific to different subjects, while others were provided to improve the overall education processes in the district. Examples of trainings include:

- Training on use of support material
- Training on activity-based teaching and learning
- Training on interaction with students
- Training on classroom management

4.2.4. Educational levels. The educational process in the CPD can be defined at macro, meso, and micro levels by studying the data, educational stakeholders, their objectives, and decisions that are important at each level.

4.2.4.1. Macro-level. The data at the macro-level of CPD process consist of all the aggregated data from all teachers, schools, and clusters, since such an analysis is performed across all the clusters or regions.

4.2.4.1.1. Stakeholders. The stakeholders of the data at the macro-level are the head of education department and policy makers at the provincial or regional level. In the CPD framework, the LOM is the primary stakeholder for the analysis at the macro-level.

4.2.4.1.2. Objective. The macro-level stakeholders are interested in the achievement of learning outcomes and the completion of mentoring and assessment processes in all the clusters. Their decisions mostly revolve around clusters where the overall aggregated students' marks are low. The LOM enforce increased L2M mentoring visits to these clusters to improve teacher training, and thus get better students' results.

4.2.4.2. Meso-level. The data at the meso-level of the CPD process consist of the educational data from all the schools for each of the clusters. Thus, the meso level

of education in the CPD process can be pictured within each cluster and across all schools in the clusters.

4.2.4.2.1. Stakeholders. The stakeholders of the data at the meso-level of CPD that can benefit from the analysis are the people who are responsible for their own cluster like LIM.

4.2.4.2.2. Objective. The decisions of the LIMs revolve around the education evaluation in each school, mentoring and assessment processes of each school, and proper training of teachers in their respective clusters.

4.2.4.3. Micro-level. The data at the micro-level of the CPD process consist of the educational data in each of the schools containing teachers', grades', and subjects' data. Thus, the analysis of the micro-level data obtained from the CPD process can be conducted across teachers, grades and subjects, for all schools in all the clusters.

4.2.4.3.1. Stakeholders. The stakeholders for this data at the finest level of granularity are the school principals, L2Ms, teachers, students, and parents. However, the primary decision-makers at this level are only the school principal and the L2Ms.

4.2.4.3.2. Objective. The educational stakeholders' decisions at this level revolve around the teachers, for example, which teachers are not performing up to the mark in a school, or which teacher needs training in a certain subject, etc. The decisions related to grade and subjects' results can also be taken by the stakeholders at this level, for example, which training should be given to teachers to improve the grade 4 result or should the activity-based learning introduced to students to augment their interest in the subject of General Knowledge, etc.

4.3. Formation of Education Baskets in the CPD Framework

The education baskets are created at each of the three levels according to the types of variables that operate at each level. Reiterating, the macro-level analysis is performed across clusters, while the meso-level analysis is performed across schools. And finally, the micro-level analysis is performed within the schools across teachers, grades, and subjects. Since the higher-level analysis works by integrating data from lower level analysis, therefore, the aggregated variables are carried upwards to be mined with macro and meso analysis, as shown in Fig. 3.3.

This data aggregation is performed by grouping the variables with respect to each school at meso-level and with respect to each cluster at the macro level. The central tendency of these grouped variables is obtained by finding the mode of the nominal variables, and the median of the ordinal variables. The quantitative variables will be represented by mean and they will be discretized to the levels given in Section 4.4 after aggregation. Table 4.1 describes the techniques which were used to obtain the typical mid-point of variables operating at lower levels of analysis.

Table 4.1: Data aggregation techniques used for different variables

Variable type	Data aggregation technique	Example
Nominal	Mode	Teacher academic qualification = {Grade 10, High School Diploma, Bachelors, Masters }
Ordinal	Median	Level of teacher identified by peers = {1, 2, 3, 4}
Quantitative	Mean	Teacher workload per week = {1-20 hours, 21-40 hours, More than 40 hours }

4.4. Variables in the Educational Data

The units of cluster, school, and grade are the specific data objects of observation under the CPD framework. The variables of educational data under the CPD framework belong to either of the process, environment, people or outcome type. These types of variables are discussed next along with the variables that belong to each type and their transformed values. The transformation details of these variables are given in Chapter 5.

4.3.1. Process variables. The process variables are the indicators of mentoring and assessment completion of schools and clusters. In this case study, these variables operate on the meso- to macro- level of educational analytics. The details of process variables in the educational data are given in Table 4.2.

The process variables which operate at both the higher and lower levels like mentoring completion and assessment completion are not integrated into the higher level due to their counterparts of cluster mentoring completion, cluster assessment completion.

Table 4.2: Process variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Description	Discretized Values
Micro (Grade)	<No process variables are collected for Grade>			
Meso (School)	Direct	SV ₁ : School mentoring completion	This variable shows the achievement of mentoring areas in the school	School.Mentoring.completion = {Good, Average, Bad}
		SV ₂ : School assessment completion	This variable shows the status of assessment conducted by the L2Ms in the school	School.Assessment.implementation = {Good, Bad}
		SV ₃ : Cooperation of Head Teacher (HT) of the School	This variable shows the extent of facilitation provided by HTs to the L2Ms for the assessment and mentoring processes	Cooperation.of.HT = {Good, Average, Bad}
Macro (Cluster)	Aggregated from School	SV ₃ : Average cooperation of Head Teacher (HT)	The mean cooperation of HT across schools of each cluster	Average.Cooperation.of.HT = {Good, Average, Bad}
	Direct	CV ₁ : Cluster mentoring completion	The mentoring achievement in all clusters is shown by this attribute	Cluster.Mentoring.completion = {Good, Bad}
		CV ₂ : Cluster assessment completion	The status of assessment conducted in the cluster	Cluster.Assessment.completion = {Good, Bad}
		CV ₃ : Cluster test report issuance	The status of cluster report that is to be issued to the district centre	Cluster.Test.report.issuance = {Good, Bad}
		CV ₄ : Cluster pre-mentoring status	Cluster learning evaluation before the mentoring process began	Cluster.Pre.mentoring.status = {Good, Bad}

4.3.2. Environment variables. The environment variables are the characteristic variables like enrolment, size, number of teachers etc. of grades 3-5, schools, and clusters. These variables are present at all the three levels of educational analytics. The details of environment variables are given in Table 4.3.

Again, the environment variables which operate at both the higher and lower levels like percentage of present students are not integrated from the lower level of grade into the higher level of school due to its counterpart of percentage of present students in school.

Table 4.3: Environment variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Discretized Values
Micro (Grade)	Direct	GV ₁ : number of students in a grade	Class.size = {1-15, 16-35, More than 35}
		GV ₂ : % of students present students in a grade	Percentage.of.present.students.in.class = {Good, Bad}
		GV ₃ : Boys-to-Girls ratio in a grade	Class.ratio = {All Boys, All Girls, Balanced, More boys, More girls}
Meso (School)	Aggregated from Grade	\overline{GV}_1 : Average of total number of students in a school	Average.class.size = {1-15, 16-35, More than 35}
		\overline{GV}_3 : Average of Boys-to-Girls ratio in a school	Average.class.ratio = {All Boys, All Girls, Balanced, More boys, More girls}
	Direct	SV ₄ : Distance of school from the cluster centre	Distance.from.the.cluster.centre = {Near, Midway, Far}
		SV ₅ : Level of school	Level = {Primary, Elementary, High}
		SV ₆ : Number of teachers in the school	Number.of.teachers = {1-3, 4-10, More than 10}
		SV ₇ : Number of Elementary School Teachers (ESTs) in the school	Number.of.ESTs = {1-5, 6-10, More than 10}
		SV ₈ : Total enrolled students	Total.enrolled.students = {1-150, 151-300, More than 300}
		SV ₉ : % of students present in the school	Percentage.of.present.students.in.school = {Good, Bad}
		SV ₁₀ : Attendance of teachers in School	Attendance.of.Staff = {Good, Bad}
Macro (Cluster)	Aggregated from Grade	Mo.GV ₁ : Mode total number of students*	Mode.class.size = {1-15, 16-35, More than 35}
		Mo.GV ₃ : Mode Boys-to-Girls ratio*	Mode.class.ratio = {All Boys, All Girls, Balanced, More boys, More girls}
	Aggregated from School	\overline{SV}_4 : Average distance of school from the cluster centre	Average.Distance.from.the.cluster.centre = {Near, Midway, Far}
		Mo.SV ₅ : Mode Level of School	Mode.Level = {Primary, Elementary, High}
		\overline{SV}_6 : Average number of teachers per school	Average.Number.of.teachers = {1-3, 4-10, More than 10}
		\overline{SV}_7 : Average number of Elementary School Teachers (ESTs)	Average.Number.of.ESTs = {1-5, 6-10, More than 10}
		\overline{SV}_8 : Average total enrolled students per school	Average.Total.enrolled.students = {1-150, 151-300, More than 300}
		\overline{SV}_9 : Average percentage of present students per school	Average.Percentage.of.present.students.in.school = {Good, Bad}
		\overline{SV}_{10} : Average attendance of teachers per school	Average.Attendance.of.Staff = {Good, Bad}
	Direct	<No environment variables are available for Cluster>	
* Due to the unnormalized distributions of variables Class size and Class ratio when grouped by cluster, their mode was taken after discretization instead of taking Mean.			

4.3.3. People variables. People variables are the information pertaining to teachers, and L2Ms that are operational in the CPD framework. These variables are at all the three levels with stakeholders' personal and professional information. The values of these variables can impact the process, environment, and outcome variables. However, variables with many different values like recommended training to teachers and subject team of teachers are not integrated into the higher-level analysis. The details of these variables are given in Table 4.5.

4.3.4. Outcome variables. The outcome variables are the marks obtained by the students in different subjects in grades 3-5. The aggregation of these marks for different educational units produce respective outcome variables for subjects, teachers, schools, and clusters. The outcome variables are also integrated into higher-level analysis since they can show how the outcome variables at lower levels relate to the outcome variable at a higher level. The details of outcome variables are given in Table 4.4.

Table 4.4: Outcome variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Discretized Values
Micro (Grade)	Direct	SubV _{1,7} : Average marks in subjects English, Mathematics, Science, SS, GK, Religion, and National Language	Avg.marks.per.subject.ENG = {Good, Average, Bad} Avg.marks.per.subject.MATHS = {Good, Average, Bad} Avg.marks.per.subject.SCIENCE = {Good, Average, Bad} Avg.marks.per.subject.SS = {Good, Average, Bad} Avg.marks.per.subject.GK = {Good, Average, Bad} Avg.marks.per.subject.RELIGION = {Good, Average, Bad} Avg.marks.per.subject.NL = {Good, Average, Bad}
		TV ₁₁ : Teacher result	Teacher.result = {Good, Average, Bad}
Meso (School)	Aggregated from Grade	\overline{TV}_{11} : Average Teacher result per school	Average.Teacher.result = {Good, Average, Bad}
	Direct	SV ₁₁ : School result	School.result = {Good, Average, Bad}
Macro (Cluster)	Aggregated from Grade	\overline{TV}_{11} : Average Teacher result per cluster	Average.Teacher.result = {Good, Average, Bad}
	Aggregated from School	\overline{SV}_{11} : Average School result per cluster	Average.School.result = {Good, Average, Bad}
	Direct	CV ₅ : Cluster result	Cluster.result = {Good, Average, Bad}

Table 4.5-A: People variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Discretized Values
Micro (Grade)	Direct	TV ₁ : Teacher designation	Teacher.designation = {DYHM, ESE, EST, HM, PST, SESE, SSE}
		TV ₂ : Teacher academic qualification (degrees)	Teacher.academic.qualification = {Grade 10, High School Diploma, Bachelors, Masters}
		TV ₃ : Teacher professional qualification (degrees)	Teacher.professional.qualification = {PTC or JV or CT, B.Ed., M.Ed.}
		TV ₄ : Teacher workload per week	Teacher.workload.per.week = {1-20 hours, 21-40 hours, More than 40 hours}
		TV ₅ : Teacher age	Teacher.age = {Upto 30 years, 31-50 years, More than 50 years}
		TV ₆ : Teacher experience	Teacher.experience = {Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
		TV ₇ : Teacher training duration	Teacher.training.duration = {Upto 2 weeks, 1 month, More than a month}
		TV ₈ : Level of teacher identified by peers (1 being the best)	Level.of.teacher.identified.by.peers = {1, 2, 3, 4}
		TV ₉ : Training recommended for teacher (list)	There are 31 areas in which the teachers are recommended trainings. Some of them are: Recommended.teacher.training = {English, Maths, Social Studies, Lesson planning, activity-based teaching and learning, classroom management, multi-grade teaching, child friendly school}
		TV ₁₀ : Subject team of teacher	There are 14 subject teams in total. Some of the teams are listed below: Subject.team.of.teacher = {All, English, English + Maths + Science, NL, NL + Religion + SS}
DYHM: Deputy Head Master HM: Head Master ESE: Elementary School Educator EST: Elementary School Teacher PTC or JV or CT: Primary Teacher Certificate/Junior Vernacular/Certificate of teaching		SSE: Secondary School Educator PST: Primary School Teacher B.Ed.: Bachelors in Education M.Ed.: Masters in Education SESE: Senior Elementary School Educator SST: Secondary School Teacher	

Table 4.5-B: People variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Discretized Values
Meso (School)	Aggregated from Grade	Mo.TV ₁ : Mode Teacher designation	Mode.Teacher.designation = {DYHM, ESE, EST, HM, PST, SESE, SSE}
		Mo.TV ₂ : Mode Teacher academic qualification (degrees)	Mode.Teacher.academic.qualification = {Grade 10, High School Diploma, Bachelors, Masters}
		Mo.TV ₃ : Mode Teacher professional qualification (degrees)	Mode.Teacher.professional.qualification = {PTC or JV or CT, B.Ed., M.Ed.}
		\overline{TV}_4 : Average Teacher workload per week	Average.Teacher.workload.per.week = {1-20 hours, 21-40 hours, More than 40 hours}
		\overline{TV}_5 : Average Teacher age	Average.Teacher.age = {Upto 30 years, 31-50 years, More than 50 years}
		\overline{TV}_6 : Average Teacher experience	Average.Teacher.experience = {Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
		\overline{TV}_7 : Average Teacher training duration	Average.Teacher.training.duration = {Upto 2 weeks, 1 month, More than a month}
		\overline{TV}_8 : Median Level of teacher identified by peers (1 being the best)	Median.teacher.peer.ranking = {1, 2, 3, 4}
	Direct	L2V ₁ : L2M designation	L2M.designation = {ESE, EST, PST, SESE, SSE, SST}
		L2V ₂ : L2M academic qualification (degrees)	L2M.academic.qualification = {Bachelors, Masters}
		L2V ₃ : L2M professional qualification (degrees)	L2M.professional.qualification = {B.Ed., M.Ed., CT}
		L2V ₄ : L2M age (years)	L2M.age = {25-35, 36-45, 46-55}
		L2V ₅ : L2M experience	L2M.experience = {Less than 5 years, 5-10 years}
		L2V ₆ : L2M training duration	L2M.training.duration = {Upto 2 weeks, 1 month, More than a month}
		L2V ₇ : L2M attendance	L2M.attendance = {Good, Bad}

Table 4.5-C: People variables for the CPD process in the educational data

Level of Analytics	Kind of Measure	Variable	Discretized Values
Macro (Cluster)	Aggregated from Grade	Mo.TV ₁ : Mode Teacher designation	Mode.Teacher.designation = {DYHM, ESE, EST, HM, PST, SESE, SSE}
		Mo.TV ₂ : Mode Teacher academic qualification (degrees)	Mode.Teacher.academic.qualification = {Grade 10, High School Diploma, Bachelors, Masters}
		Mo.TV ₃ : Mode Teacher professional qualification (degrees)	Mode.Teacher.professional.qualification = {PTC or JV or CT, B.Ed., M.Ed.}
		\overline{TV}_4 : Average Teacher workload per week	Average.Teacher.workload.per.week = {1-20 hours, 21-40 hours, More than 40 hours}
		\overline{TV}_5 : Average Teacher age	Average.Teacher.age = {Upto 30 years, 31-50 years, More than 50 years}
		\overline{TV}_6 : Average Teacher experience	Average.Teacher.experience = {Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
		\overline{TV}_7 : Average Teacher training duration	Average.Teacher.training.duration = {Upto 2 weeks, 1 month, More than a month}
		\overline{TV}_8 : Median Level of teacher identified by peers (1 being the best)	Median.teacher.peer.ranking = {1, 2, 3, 4}
	Aggregated from School	Mo.L2V ₁ : Mode L2M designation	Mode.L2M.designation = {ESE, EST, PST, SESE, SSE, SST}
		Mo.L2V ₂ : Mode L2M academic qualification (degrees)	Mode.L2M.academic.qualification = {Bachelors, Masters}
		Mo.L2V ₃ : Mode L2M professional qualification (degrees)	Mode.L2M.professional.qualification = {B.Ed., M.Ed., CT}
		$\overline{L2V}_4$: Average L2M age (years)	Average.L2M.age = {25-35, 36-45, 46-55}
		$\overline{L2V}_5$: Average L2M experience	Average.L2M.experience = {Less than 5 years, 5-10 years}
		$\overline{L2V}_6$: Average L2M training duration	Average.L2M.training.duration = {Upto 2 weeks, 1 month, More than a month}
		$\overline{L2V}_7$: Average L2M attendance	Average.L2M.attendance = {Good, Bad}
Direct	<L1M operates at the cluster level. But the L1M variables are not used because of their role in the CPD framework according to which they are only responsible for bookkeeping and thus do not impact or relate to any other variables>		

The variables of the CPD framework that are present in the given high-dimension educational dataset at different levels of analytics are shown in Table 4.6.

Table 4.6: Variables at different levels of educational analytics

Level of Analytics	Type of Variables			
	People	Process	Environment	Outcome
Macro		CV ₁ - CV ₄		CV ₅
Meso	L2V ₁ - L2V ₇	SV ₁ - SV ₃	SV ₄ - SV ₁₀	SV ₁₁
Micro	TV ₁ - TV ₁₀		GV ₁ -GV ₃	TV ₁₁ , SubV ₁ - SubV ₇

↑ : variables are aggregated from the micro to macro level
 ↑ : variables are aggregated from the meso to macro level
 ↑ : variables are aggregated from the micro to meso level

So, the macro-level basket will contain the cluster variables alongwith the aggregated school and grade variables. The meso-level basket will contain the school variables alongwith aggregated grade variables. And finally, the micro-level basket will contain only grade variables. The itemsets present in these baskets can be updated with respect to different educational rule mining objectives and constraints. The consequent (RHS) of the rules are mined with outcome variables distinctive to each level of analysis.

For each of the data basket, a set of transactions is obtained on which Apriori algorithm is applied. Each variable has a set of values, and transactions are *variable = value* pairs of all variables as explained in Chapter 3.

Table 4.7 shows a sample micro-level educational basket with 5 transactions from the educational data that is generated from the primary schools in the CPD framework.

Table 4.7: Sample micro-level educational basket created from educational data generated by CPD framework

Transaction ID	Teacher designation	Teacher workload per week	Teacher academic qualification	Teacher professional qualification	Teacher age	Teacher experience	Teacher training duration	Recommended teacher training	Subject team of teacher	Level of teacher (peer ranking)	Class size	% of present students in class	Class ratio	Teacher result
1	ESE	Upto 20	M.A.	<NA>	30-49 years	5-10 years	1 month	Training on lesson planning	Maths + Science	2	1-15	Good	Balanced	Bad
2	PST OR PTC OR JV	21-40	F.A.	CT	50 and Up years	26-35 years	More than a month	<NA>	English	1	16-35	Good	All girls	Good
3	PST OR PTC OR JV	More than 40	F.A.	PTC/JV	30-49 years	11-25 years	1 month	<NA>	<NA>	<NA>	16-35	Good	All boys	<NA>
4	DYHM	More than 40	B.A.	<NA>	Less than 30 years	Less than 5 years	<NA>	Training in subject of English	Maths	<NA>	<NA>	Bad	More boys	Average
5	PST OR PTC OR JV	More than 40	M.A.	B.Ed.	30-49 years	11-25 years	1 month	Training on activity-based teaching and learning	Maths + Science	4	More than 35	<NA>	All boys	Bad

4.5. Formulation of Rule Templates in the CPD Framework

The rule templates are created for each level of analysis to mine rules that represent the education objectives and questions in Section 4.2.4. The rules formed from these templates have specific itemsets in the antecedent or consequent of the rule depending on the educational question.

4.5.1. Rule templates at the micro-level. The rule templates used for micro-level analytics are given below.

4.5.1.1. Template # 1 – Teacher and subjects outcome analysis.

4.5.1.1.1. Motivation. The outcome variables for the micro analysis are teacher result and average marks obtained by students in all the subjects. Among all variables, it is interesting to find out the relationship between environment and people variables whose co-occurrence together can impact the subject or teacher outcome to be “Good” or “Bad.”

4.5.1.1.2. Template. The rules mined from this template have the outcome variables in the consequent (RHS) of the rule.

$$\{Environment\ variables\}, \{People\ variables\} \Rightarrow \{Teacher\ outcome = \text{Good/Bad}\} \text{ or } \{Subject\ outcome = \text{Good/Bad}\} \quad (7)$$

4.5.1.1.3. Example. Some example rules or associations that are expected to be seen using this rule template are shown in Table 4.8.

Table 4.8: Example rules generated using template # 1 – micro-level analysis

Teacher.academic.qualification = Masters, Teacher.workload.per.week = 21-40 hours, Class.size = 21-40 ⇒ Teacher.result = Good
Class.size = 1-20, Teacher.designation = PST ⇒ Avg.marks.per.subject.NL = Bad
Teacher.professional.qualification = M.Ed., Subject.team.of.teacher = English + Maths + Science ⇒ Avg.marks.per.subject.ENG = Good

4.5.1.2. Template # 2 – Recommended teacher training analysis.

4.5.1.2.1. Motivation. Different trainings are proposed for various teachers under the CPD framework. Some trainings are for specific subjects like English, Science etc., and others are for overall professional development of teachers. This template explores the impact of these recommended trainings in combination with other teachers’ characteristics on the performance of teachers.

4.5.1.2.2. *Template.* The rules mined using this rule template have the recommended teacher training area in the consequent (LHS) of the rule and outcome variable for teacher at the RHS.

$$\{Recommended\ teacher\ training\ area\}, \{Environment\ variables\}, \quad (8)$$

$$\{People\ variables\} \Rightarrow \{Teacher\ outcome = Good/Bad\}$$

4.5.1.2.3. *Example.* Example associations that can be seen using the template for recommended teacher training analysis are shown in Table 4.9.

Table 4.9: Example rules generated using template # 2 – micro-level analysis

Teacher.age = 50 and up, Recommended.teacher.training = Training on activity based teaching and learning, Teacher.professional.qualification = PTC/JV \Rightarrow Teacher.result = Bad
Recommended.teacher.training = Training on Child Friendly School (CFS), Teacher.experience = Less than 5 years, Teacher.workload.per.week = More than 40 hours \Rightarrow Teacher.result = Good
Subject.team.of.teacher = All, Level.of.teacher.identified.by.peers = 2, Recommended.teacher.training = Training in subject of English \Rightarrow Teacher.result = Good

4.5.2. Rule templates at the meso-level. The meso level analysis is performed across schools and has data aggregated from the micro-level. At this level, the major unit of analysis are the schools, so it is useful to know how different variables of the schools, teachers, and L2Ms affect the school outcomes.

4.5.2.1. Template # 1 – School outcome analysis.

4.5.2.1.1. *Motivation.* The outcome variable for the meso analysis is the school result. The relationship between the environment, people, and process variables whose co-occurrence together can impact the school outcome to be Good or Bad can be examined using this rule template.

4.5.2.1.2. *Template.* The rules mined from this template have the school outcome in the consequent of the rule.

$$\{Environment\ variables\}, \{People\ variables\}, \{Process\ variables\} \Rightarrow \quad (9)$$

$$\{School\ outcome = Good/Bad\}$$

4.5.2.1.3. *Example.* Some example rules or associations that are expected to be seen using this rule template are shown in Table 4.10.

Table 4.10: Example rules generated using template # 1 – meso-level analysis

Total.enrolled.students = More than 300, School.Assessment.implementation = Poor \Rightarrow School.result = Bad
Number.of.teachers = 4-10, L2M.academic.qualification = Bachelors, Cooperation.of.HT = Good \Rightarrow School.result = Good
Percentage.of.present.students.in.school = Good., L2M.designation=ESTG, L2M.age=45-55 \Rightarrow School.result = Good

4.5.2.2. *Template # 2 – School size analysis.*

4.5.2.2.1. *Motivation.* The schools under the CPD framework have different number of enrolled students and are small, medium, or large-sized. For example, there are schools who have 1-150 students (Small), or 151-300 enrolled students (Medium), or more than 300 students (Large). Using this rule template, it can be determined if the school size along with other process, people and environment characteristics are related to the school outcome.

4.5.2.2.2. *Template.* The rules mined from this template have the school size in the antecedent (LHS) of the rule with other variables, and the school outcome in the consequent (RHS) of the rule.

$$\{School\ size\}, \{Environment\ variables\}, \{People\ variables\}, \quad (10)$$

$$\{Process\ variables\} \Rightarrow \{School\ outcome = Good/Bad\}$$

4.5.2.2.3. *Example.* Example associations that are expected to be seen using this rule template are shown in Table 4.11.

Table 4.11: Example rules generated using template # 2 – meso-level analysis

Number.of.teachers = More than 10, Total.enrolled.students = More than 300 \Rightarrow School.result = Good
Total.enrolled.students = 151-300, Level = Elementary, L2M.academic.qualification = Masters \Rightarrow School.result = Good
Distance.from.the.cluster.centre = Midway, Total.enrolled.students = 1-150, L2M.designation = PST \Rightarrow School.result = Bad

4.5.3. Rule templates at the macro-level. The macro-level analysis has the major unit of cluster, and data integrated from both meso and micro levels. The following rule template is used for this analysis.

4.5.3.1. *Template # 1 – Cluster outcome analysis.*

4.5.3.1.1. *Motivation.* At this level, there are no cluster characteristic (cluster environment) variables available, so this rule template is used to perform outcome analysis on cluster level macro data to study the process variables of clusters when their respective outcome is Good or Bad.

4.5.3.1.2. *Template.* The rules mined from this template have the cluster outcome in the consequent of the rule.

$$\{Environment\ variables\}, \{People\ variables\}, \{Process\ variables\} \Rightarrow \quad (11)$$

$$\{Cluster\ outcome = Good/Bad\}$$

4.5.3.1.3. *Example.* Some example rules or associations that are expected to be seen using this rule template are shown in Table 4.12.

Table 4.12: Example rules generated using template # 1 – macro-level analysis

Pre.mentoring.status = Good, L2M.academic.qualification = Masters \Rightarrow Cluster.result = Good
C.Assessment.completion = Bad, C.Test.report.issuance = Good, L2M.designation = SRHM \Rightarrow Cluster.result = Bad

4.6. Algorithm for Rule Discovery on Educational Data

The algorithm to implement this approach of rule discovery in educational analysis is given in Fig. 4.2.

Algorithm: Association rule mining approach to mine data at different levels

Input: $E := \{E_{mac}, E_{mes}, E_{mic}\}$ an education dataset with macro, meso, and micro subsets, $constraint(E) := \{value_1, value_2, \dots, value_n\}$ the values of the constrictive itemsets that should be present in the LHS or RHS of the rule

Output: $RS := \{RS_1, RS_2, \dots, RS_n\}$ Resulting rule set for each constraint

- 1: **foreach** $E_i \in E$ **do**
 - 2: Convert E_i to transactions T
 - 3: **foreach** $value_i \in constraint(E)$ **do**
 - 4: Apply rule template by putting $value_i$ in the LHS or RHS of the rule structure
 - 5: Set $minlen = 2, maxlen = 4, minconf = 0.85$
 - 6: Establish $minsupp$, starting with a value of 0.1, and decreasing it until some rules are obtained
 - 7: Mine association rules RS_i from T with $value_i$ in the antecedent or consequent
 - 8: Remove redundant rules from RS_i
 - 9: $RS_i = \text{sort by lift}(RS_i)$
 - 10: Analyse RS
-

Fig. 4.2: Algorithm for rule discovery on educational data

For given educational baskets at macro (E_{mac}), meso (E_{mes}), and micro (E_{mic}) levels, the resulting rule sets can be obtained for each rule template by converting the educational subset into transactions T . The rule structure is specified by giving the value of $constraint(E)$ in the appearance input of the Apriori algorithm. The $minsupp$ (minimum

support) and other parameter inputs *minlen* (minimum number of itemsets in the rules), *maxlen* (maximum number of itemsets in the rules), *minconf* (minimum confidence) of the Apriori algorithm are established. Association rules are mined and resulting rule sets *RS* are obtained for each value of *constraint(E)*. Redundant rules are then removed, and rules are sorted by lift. This process is repeated for all educational baskets. Finally, the resulting rule sets are analysed with visualization plots.

4.7. Summary

In this chapter, the case study and framework of education in primary schools of developing countries is presented. The hierarchy of schools, clusters, and district is observed to characterise different levels of analytics. The CPD framework is studied by different levels, stakeholders, objectives, and variables. The different kinds of variables can be either of the process, outcome, people, or environment type in this framework.

The formulation of education baskets at different levels of analysis is discussed by the aggregation of variables from different levels. The rule templates that are to be used to mine rules for various stakeholders' objectives are discussed. Finally, the template-based approach of problem solving, and association rules mining is presented with the algorithm that is used to mine rules for each rule template.

Chapter 5 . ETL

This chapter describes the ETL steps to extract the data from multiple education files, transforming it to the point where it is suitable to be used as an input to the rule mining algorithm, and loading steps to apply the rule mining algorithm.

Before the ETL process, the given educational data was profiled and studied using the 4 Vs which are detailed below in Table 5.1.

Table 5.1: The 4 Vs of given educational data

Property	V in given education data
Volume	The data had been collected from 59 clusters, 1391 primary schools, and 2613 teachers for the subjects of English, Mathematics, Science, SS, GK, Religion, and NL in Grades 3-5.
Velocity	The educational data was collected for every month from all the 59 clusters.
Variety	The education data were maintained in Excel files and contained 23 numeric, 10 qualitative, and 4 date attributes.
Veracity	The dataset is reliable and valid because it is already being used to derive reports for certain assessment and learning measures. This dataset has fine granularity since all indicative variables are recorded by sub-division into various fields. However, it is also missing values for variables, which makes it less meaningful. The details of missing values are given in Table 5.2.

5.1. Data Extraction

The education data used for the rule discovery in this thesis was extracted from the original educational data which was in Microsoft Excel files using the Visual Basic for Applications (VBA).

The data stored in the original data files had a different structure and the data was maintained in separate Excel files for each cluster (the area in which schools are geographically co-located). These files had an unstructured format with variables of different educational units of analysis like cluster, school, grade etc. spread over different Excel sheets. For example, as Fig. 5.1 shows, the environment variables were stored as Input data in the preliminary sheets, followed by the assessment indicators, and finally the reports generated from this data were stored as outcome variables in the final sheets. The sheets were sorted with respect to the cluster, school, grade, teacher, and head teacher indicators. This data needed to be extracted in some structured format to aid the next step of ETL which is the transformation of data in an appropriate format. For this thesis, various VBA scripts were used to store each major education unit like cluster, school, teacher, L2M as tables. These tables had rows as instances in the education data, and columns as the attribute values. These tables were stored in a relational database system (RDMS).

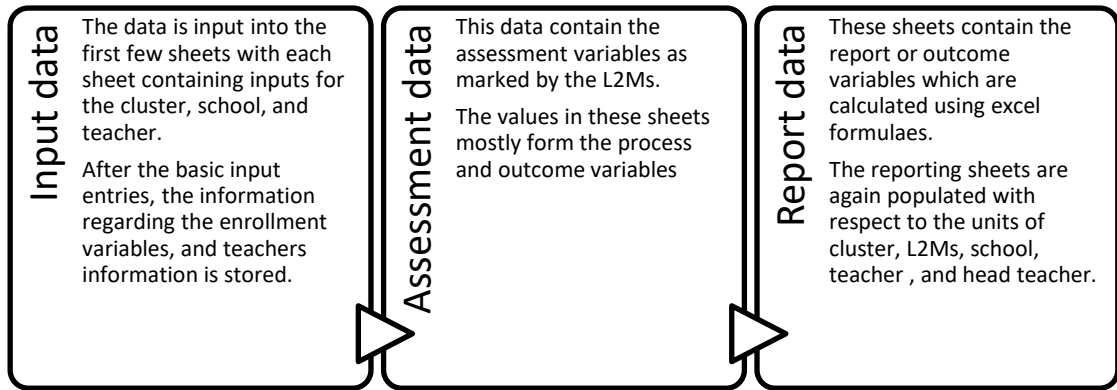


Fig. 5.1: The structure of given education data

Relational data [14] consists of tables or relations containing rows and columns. Each table has a unique identifier called the primary key and relations are linked to other relations by using foreign key. Each row in a relational data represent a record for an object. The UML diagram depicting the structure of the extracted education data is shown in Fig. 5.3.

The VBA scripts used the Microsoft Scripting Runtime to use the File System Objects which enables the access to computer's file system to perform operations like opening files, reading files, iterating through folders, etc. The data values respective to the unit of analysis were copied, and the data after extraction was obtained with separate sheets for Cluster, L1 manager, School, Teacher, L2 manager, Grade, and Subject. In each sheet, a row referred to details of all types of variables (people, environment, process) of that unit only. Fig. 5.2 shows structured extracted data for cluster information in Microsoft Excel.

CTSC ID	Cluster Name	District Name	Mentoring Completion	Assessment Completion	Test Report Issuance	Pre mentoring status	DTE Attendance	Avg. marks per subject for cluster
56	GGHS*****HIB		96.875	100	100	36.68797348	#N/A	6.578808446
57	GHS*****HIB		77.08333333	94.11764706	95.83333333	18.11371528	#N/A	5.387656702
58	#N/A		96.98275862	#REF!	100	24.50182763	#N/A	#REF!
59	GHS*****HIB		48.07692308	50	50	30.63782051	#N/A	4.970384995
60	GHS*****URA		5693.103448	48.27586207	48.27586207	34.60758377	91.66666667	2.6875829
61	GHS*****URA		13780	100	0	18.88888889	96.15384615	6.14638327
62	GHS*****URA		5710.714286	93.69747899	100	24.1622575	89.13043478	7.629819563
63	GHS*****URA		4578.26087	94.11764706	52.17391304	25.51440329	91.66666667	5.568801522
64	GHS*****URA		5200	99.46524064	100	23.48484848	86.95652174	4.94333504
65	GHS*****URA		6586.666667	98.62745098	100	36.56378601	86.95652174	5.48176114
66	GHS*****URA		11640.90909	98.39572193	#N/A	#N/A	#N/A	4.624365482
67	GHS*****URA		4727.272727	59.09090909	59.09090909	29.1005291	100	3.196876952
68	GHS*****URA		#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#N/A	#DIV/0!
69	GHS*****URA		0	49.32126697	50	#DIV/0!	92.30769231	1.338915319
70	GGHS*****URA		0	95.24886878	100	#DIV/0!	86.95652174	4.103736941
71	GES*****URA		#DIV/0!	#DIV/0!	#N/A	#N/A	#N/A	#DIV/0!
72	GHS*****URA		6128.571429	100	0	29.4640853	94.44444444	4.361167158
73	GHS*****URA		0	75.40106952	90.90909091	#DIV/0!	91.66666667	4.64063745
74	GGHS*****URA		3954.166667	50	50	23.7654321	95.65217391	2.424904701
75	GGHS*****URA		0	97.32620321	#N/A	#N/A	#N/A	5.426319937
76	GHS*****URA		5850	50	50	50.83095916	95.65217391	3.748782536
77	GHS*****URA		12675	91.66666667	#N/A	#N/A	#N/A	4.569117871
78	GHS*****URA		7057.142857	87.11484594	100	34.4486532	90.38461538	4.764128063

Fig. 5.2: Cluster data extracted in Microsoft Excel Sheet

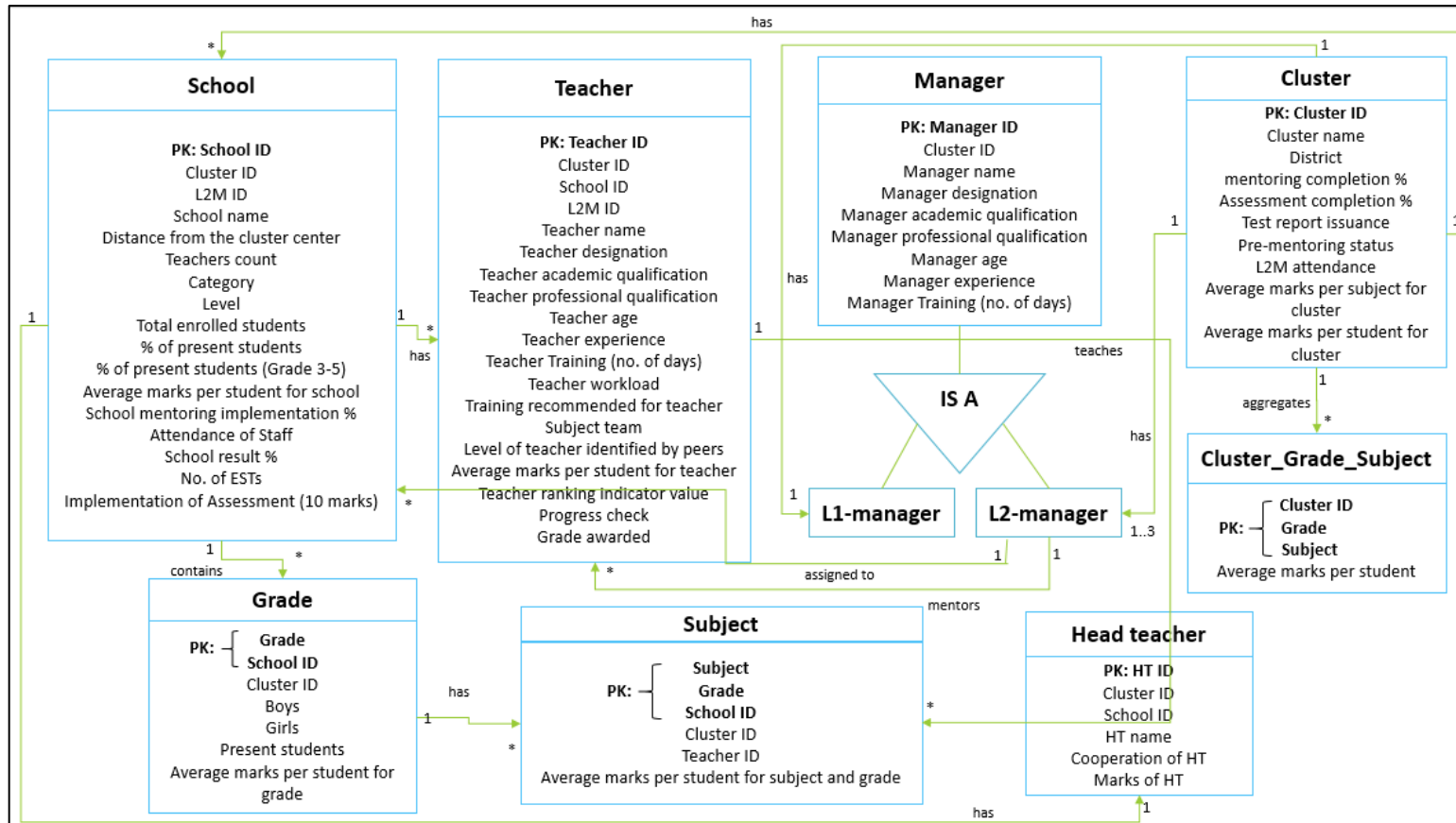


Fig. 5.3: The structure of extracted education data represented as UML diagram

5.2. Data Transformation

Big data needs to be cleansed, normalized, and checked for consistency before applying any data mining algorithm. This is because real-world data contains noise, outliers, data in different formats etc. Therefore, different processes are used to transform data into a format that is suitable to the algorithm that is to be applied. In the context of this research, the Apriori algorithm works only with nominal that is qualitative values of data, so quantitative variables needed to be transformed into nominal ones by various data discretization methods.

The extracted education data with different educational units was then transformed using various data transformation techniques. These data transformations were performed in RStudio using R version 3.4.1 [61]. A selective series of data transformation steps for class “Teacher” using R are given in Appendix B. To perform the transformation steps, each educational class was loaded into R using the *xlsx* [62] package’s function *read.xlsx*.

The various steps with which the given data was transformed to prepare it for rule mining by Apriori algorithm are listed below:

5.2.1. Data cleansing. The attributes in each of the classes had several invalid and out-of-range values. These values needed to be dropped and changed to missing values to improve the quality of resulting association rules. The examples of attribute values where data cleansing was required were:

- Attributes which represented a percentage and had values above 100. For example, Fig. 5.4 shows dirty data where the percentage value was more than 100 for one instance.
- Missing values coded as 0.
- Invalid values in qualitative variables like teacher qualification and designation.
- Invalid age and experience values for teachers and L2Ms like an age of over 100 years and a negative value of experience.

Table 5.2 lists the percentage of values that were changed to missing for each of the cleansed variables and the total percentage of missing values after data cleansing. Most of the invalid values were cleansed for teacher age and experience which had resulted due to the wrong or missing record entry of date of birth and date of joining of

teachers. The teacher result had about 29% missing values which were coded as 0 and needed to be dropped.

Table 5.2: Data cleansing details for different variables

Educational Unit	Variable	Percentage of values that were dropped	Percentage of total missing values
Cluster	Cluster mentoring completion	3.38%	6.78%
	Cluster assessment completion	1.69%	8.47%
	Pre-mentoring status	11.8%	15.2%
	Test report issuance	1.69%	11.86%
	Cluster result	1.69%	6.78%
School	Distance of school from the cluster centre	-	1.29%
	Number of teachers	1.8%	1.8%
	Level	3.89%	4.02%
	Total enrolled students	2.92%	3.05%
	Percentage of present students	27.7%	30.8%
	School mentoring completion	2.53%	8.7%
	School assessment completion	-	2.2%
	Attendance of teachers	-	6.2%
	Cooperation of HT	-	7.15%
	School result	0.77%	10.5%
L2M	L2M designation	7.2%	8.1%
	L2M academic qualification	-	3.6%
	L2M professional qualification	-	6.3%
	L2M age	8.1%	9%
	L2M experience	9.9%	11.7%
	L2M attendance	0.9%	0.9%
	L2M training duration	2.7%	33.3%
Teacher	Teacher designation	0.92%	3.15%
	Teacher workload per week	0.14%	24.4%
	Teacher academic qualification	0.1%	16%
	Teacher professional qualification	0.74%	21.9%
	Teacher age	30.2%	34.3%
	Teacher experience	30.8%	34.8%
	Teacher training duration	0.88%	46.4%
	Training recommended for teacher	0.03%	44%
	Subject team of teacher	2.83%	49.8%
	Level of teacher identified by peers	1.31%	58.9%
	Teacher result	28.6%	30.5%
Grade	Average Boys-to-Girls ratio	0.63%	4.3%
	Average total number of students	4.3%	4.32%
	Average percentage of present students	29.4%	33.7%
Subject	Average marks in subjects English, Mathematics, Science, SS, GK, Religion, and National Language	2.4%	4.9%

The cluster pre-mentoring status which is the cluster education status before the mentoring process began also had many out-of-range values. In addition, the percentage

of present students for schools were obtained incorrectly 27.7%, and for classrooms 29% of the time.

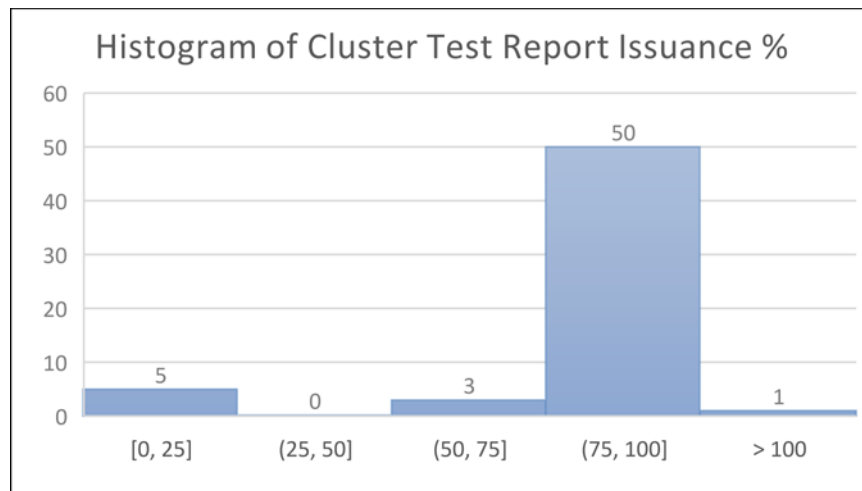


Fig. 5.4: Dirty data for variable Test Report Issuance %

The data instances or attributes containing missing values were not removed because the educational data will take the form of transactions where an item can be present or absent, so missing values will represent the items that are not present in a transaction. However, a large number of missing values for most of the teacher attributes, process variables like cluster pre-mentoring status and test report issuance, and the attendance of students in each grade and schools indicate that the L2Ms responsible for recording the educational data were not keen on recording these variables.

5.2.2. Attributes consistency. The features related to the designation, academic qualification, and professional qualification of L2Ms and teachers were modified for consistency. For example, degrees MS.ED. and M.ED. were combined into a single value M.Ed. In addition, attributes that had misspelled values, or same values in both upper and lowercase were modified. The variables that were modified to maintain consistency are given in Table 5.3. A few values that were misspelled were also modified for teachers' training duration, recommended training and subject team.

5.2.3. Data derivation. In this pre-processing step, the derived attributes or new attributes were created from existing attributes in the data. The attributes derived from the given educational data are listed in Table 5.4 below:

Table 5.3: Attributes modified for consistency

Educational Unit	Variable	Percentage of values that were modified
L2M	L2M designation	90.9%
	L2M academic qualification	96.4%
	L2M professional qualification	93.7%
Teacher	Teacher designation	76%
	Teacher academic qualification	83.9%
	Teacher professional qualification	78%
	Teacher training duration	0.07%
	Training recommended for teacher	1.6%
	Subject team of teacher	0.2%

Table 5.4: Derived attributes

Attribute in data	Derived attribute
Date of birth of teacher/L2M	Teacher/L2M age (in years)
Date of joining as teacher/L2M	Teacher/L2M experience (in years)
Number of Primary School Teachers (PSTs)	Number of teachers
Number of Elementary School Teachers (ESTs)	
Number of boys and girls in Grades 3, 4, and 5	Average total number of students
	Average boys-to-girls ratio
Number of present students in Grades 3, 4, and 5	Average percentage of present students
Number of days spent in individual trainings by teacher/L2M	Teacher/L2M total training duration
Marks in subjects English, Mathematics, Science, SS, GK, Religion, and National Language in Grades 3, 4, and 5	Average marks in subjects English, Mathematics, Science, SS, GK, Religion, and National Language

5.2.4. Reshaping data. The subject-wise average marks per student for each of the Grades 3, 4, and 5 were recorded initially in the long format, as shown in Table 5.5. This data was converted to wide format, as shown in Table 5.6, so that each row represented one school in each cluster. The wide format data was used to integrate the marks data with Teachers' data to perform micro analysis.

Table 5.5: Data in long format

Subject	Grade	School ID	Cluster ID	Average marks per student
English	3	35---1--	35**00--	4.43
Maths	3	35---1--	35**00--	3.43
English	4	35---1--	35**00--	4.95
Maths	4	35---1--	35**00--	4.32
Science	4	35---1--	35**00--	4.14
English	5	35---1--	35**00--	4.74
Maths	5	35---1--	35**00--	5.00
Science	5	35---1--	35**00--	4.34

Table 5.6: Data in wide format

School ID	Cluster ID	3_English	3_Maths	4_English	4_Maths	4_Science	5_English	5_Maths	5_Science
35---1--	35**00--	4.43	3.43	4.95	4.32	4.14	4.74	5.00	4.34

5.2.5. Attribute discretization. The attribute discretization refers to the binning of attribute values that is placing each value in its respective bin according to the range of bin. The details of this step are provided in Chapter 3. This step was performed on all numeric attributes since Apriori algorithm accepts only categorical or qualitative data. The numeric discretized values were also assigned labels, such as “Bad”, “Average”, or “Good” for better readability.

The binning method used for the numeric attributes was Manual Discretization [14], which lets the user decide the range of each bin. This method is especially used for education data for variables such as marks, GPA etc., because it is not suitable to bin such variables with respect to equal-interval or equal-frequency methods.

The average marks per student for all the educational classes in the range of 0-10 were discretized using the given criteria:

$$\text{Average marks per student} = \begin{cases} \text{Bad,} & \text{if value} \leq 4 \\ \text{Average,} & \text{if value} > 4 \text{ and } \leq 6.5 \\ \text{Good,} & \text{if value} > 6.5 \end{cases} \quad (12)$$

Similarly, cut-off points to discretize other numeric attributes were also manually decided using the domain knowledge of education data and by analysing the natural distribution of values as observed by the histograms of all values. For example, the distance of schools from the cluster centre was discretized as:

$$\text{Distance} = \begin{cases} \text{Near,} & \text{if value} \leq 5 \text{ km} \\ \text{Midway,} & \text{if value} > 5 \text{ and } \leq 10 \text{ km} \\ \text{Far,} & \text{if value} > 10 \text{ km} \end{cases} \quad (13)$$

5.2.6. Attribute selection. This step was performed to remove the dependent and irrelevant attributes from the data. The identity attributes in each table were not useful for association mining because the motivation of this thesis was to mine rules that shows the characteristics of different classes, rather than rules which determine the identity of any class performing good or bad. Therefore, all identity attributes

comprising of primary keys, foreign keys, and name attributes were removed from the data after the formulation of education baskets. Since, these attributes were required to join or make a subset of different educational classes, the identity attributes were removed just before applying the rule mining algorithm.

Along with the identity attributes, the variables that were dependent on some other attributes were also removed. For example, Teacher grade which was entirely dependent on Teacher result was removed.

5.3. Data Load

This step is performed after the pre-processing to convert the data into a suitable format to be given as input to the rule mining algorithm. For example, converting the data to *.csv* or *.arff* formats if the analysis were to be performed in *Weka* [63]. However, in this thesis the association rule mining was also performed in RStudio like the data transformation, so there was no need to change the data format or load it in some other software.

Data was available after all the ETL steps for 10 months. However, for the experiments conducted in the following experiments one month's data was used for analysis because time-series or sequential analysis by rule mining algorithm was not within the scope of this thesis.

5.4. Summary

The ETL steps that were performed to transform the data into a form that was suitable for the application of Apriori algorithm were discussed in this chapter. The data was first extracted into a structured format in which each educational unit of analysis can be represented by a class in the UML diagram. Then, it was transformed by various processes. First, the data was cleansed to remove any out-of-range values. After the data cleansing, various attributes were modified for consistency. Data derivation was performed to create new variables that added more information to the analysis and subject marks variable was reshaped to obtain students' average marks across each subject. Then, different data bins were created for each of the numeric variables using manual discretization. And finally, the attributes selection was performed to remove the dependent and identity attributes before the application of rule mining algorithm.

Chapter 6 . Micro-level Educational Analytics

This chapter provides the details of the micro analysis performed with the educational data. For micro analysis, data from three entities were used, namely; Teacher, Grade, and Subject. These three entities correspond to the finest level details in the data by means of which analysis within a school can be performed. The students' academic and personal details were not available, however if available that data could also operate on the micro-level of analysis. This data was cleansed and transformed using the techniques detailed in Chapter 5. The characteristic and environmental variables of all three classes (Teacher, Grade, and Subject) were merged, and Apriori algorithm was run on the merged dataset to obtain rules across these classes. The rules were analyzed for interestingness on the basis of objective interestingness measures by the means of visualization plots, and by this analysis different variables that impact the outcome features at micro-level were determined.

6.1. Teacher Outcome Analysis

The teacher outcome is the Teacher result variable in the CPD framework. This variable is based on the average marks obtained by students in grades 3-5 in all subjects that were being taught by a teacher. So, with this variable the students results are used as a proxy for teachers overall result or performance. The teacher result had many missing values which were initially coded as 0 and some outliers that were above 100. These values were dropped and changed to missing, and the values in the range of 1-100 were translated to the following levels using the domain knowledge of grading system in the developing country.

$$Teacher\ result = \begin{cases} Bad, & \text{if value} \leq 40 \\ Average, & \text{if value} > 40 \text{ and } \leq 65 \\ Good, & \text{if value} > 65 \end{cases} \quad (14)$$

The distribution of teachers' outcome in the educational data is given below in Fig. 6.1.

6.1.1. Educational basket for teacher outcome analysis. The education basket used to study teachers' outcome relationship to other variables consist of the items shown in Table 6.1. This basket has a total of 2,613 transactions with each transaction corresponding to one teacher's details.

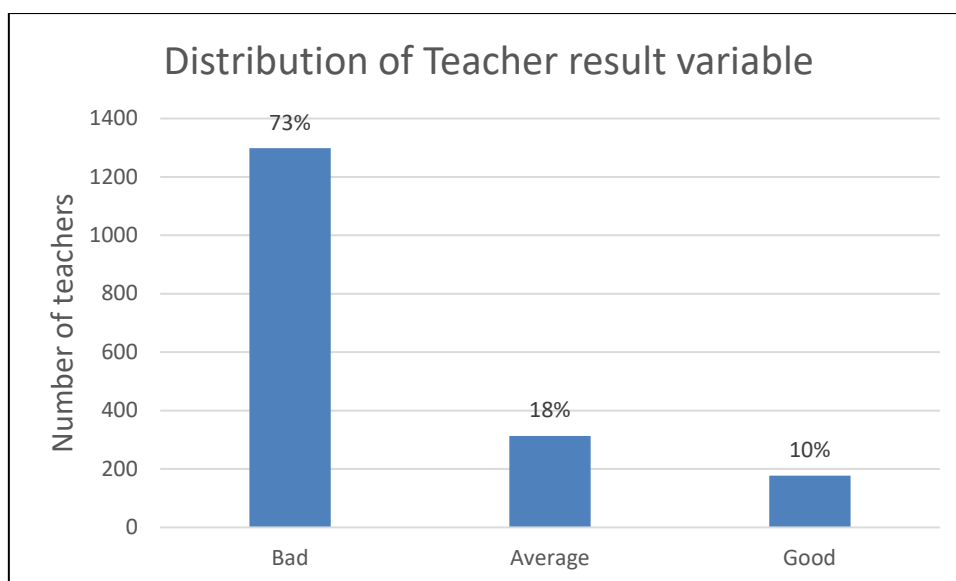


Fig. 6.1: Distribution of teachers with different outcomes

Table 6.1: Items in education basket for teacher outcome analysis

Items	Potential Values
Teacher designation	{DYHM, ESE, EST, HM, PST, SESE, SSE}
Teacher workload per week	{1-20 hours, 21-40 hours, More than 40 hours}
Teacher academic qualification	{Grade 10, High School Diploma, Bachelors, Masters}
Teacher professional qualification	{PTC/JV/CT, B.Ed., M.Ed., Other}
Teacher age	{Upto 30 years, 31-50 years, More than 50 years}
Teacher experience	{Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
Teacher training duration	{Upto 2 weeks, 1 month, More than a month}
Recommended teacher training	There are 34 training areas. Some of them are: {English, Maths, Social Studies, lesson planning, activity-based teaching and learning, classroom management, multi-grade teaching, child friendly school, ...}
Subject team of teacher	There are 14 subject teams for teachers. Some of them are: {All, English, English + Maths + Science, ...}
Level of teacher identified by peers	{1, 2, 3, 4}; 1 being the best
Class size	{1-15, 16-35, More than 35}
Percentage of present students in class	{Bad, Good}
Class ratio	{All Boys, All Girls, Balanced, More boys, More girls}
Teacher result	{Bad, Average, Good}

6.1.2. Experiment 1 – Teacher outcome = Good.

6.1.2.1. Objective. The objective of this experiment is to study the teacher and classroom characteristics that led to a teacher’s outcome being Good.

6.1.2.2. Constraints. For this and all the subsequent experiments, the rules were mined using the *arules* [64] package in R. The following constraints were used to mine the rules for this experiment:

- Support=0.0015, a very low support value used eventually by decreasing the starting minimum support value of 0.1 to find interesting rules.
- Confidence=85%, value used to obtain high-confidence rules which are constrained to be true most of the time.
- Absolute minimum support count = 3 rows out of 2,613 transactions. Since, only 3 rows are covered with a minimum support of 0.0015 (calculated from Equation 4).
- Rule template: {LHS: All items from the basket in Table 6.1 except for Teacher result, RHS: Teacher result = Good}

A template-based approach was used to mine the rules with the RHS of the resulting rules pre-specified to obtain the rules with itemsets that relate to the teacher outcome to be Good.

6.1.2.3. Results. Only 1 rule was obtained in this experiment when Apriori algorithm was run with the above listed constraints. This rule has a very low support of 0.0019 but high confidence and lift values of 100% and 14.76 respectively. The high lift and low support means that the antecedents and consequent of this rule co-occur only a few times but are strongly related. The scatter plot of this rule is given in Fig. 6.2, matrix plot in Fig. 6.3, and the itemsets from the parallel coordinates plot in Table 6.4.

The scatter (Fig. 6.2) and matrix (Fig. 6.3) plots represent the resulting rule by a point and bar respectively with a support of 0.0019. The resulting rule is given in Table 6.2.

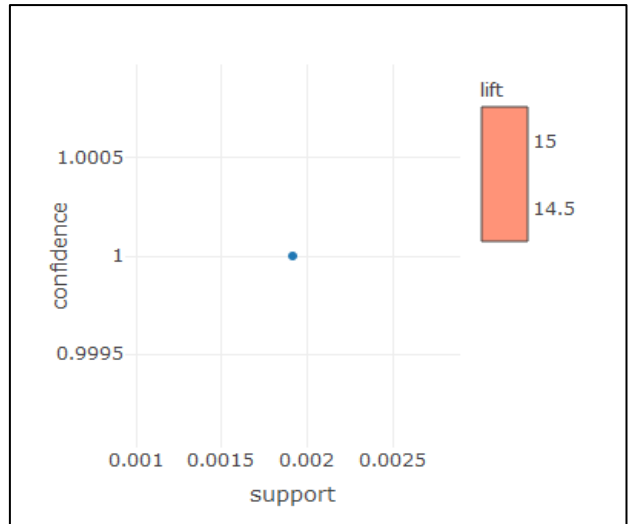


Fig. 6.2: Scatter plot for Experiment 1 – Teacher outcome = Good

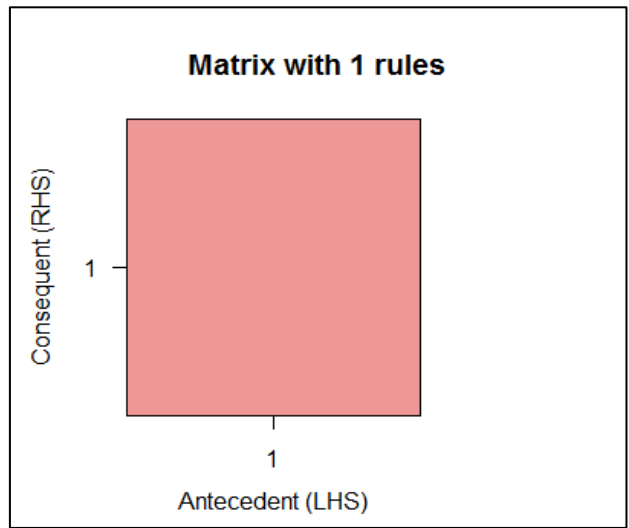


Fig. 6.3: Matrix plot for Experiment 1 – Teacher outcome = Good

Table 6.2: Resulting rule – Teacher outcome = Good

Rule	Support	Confidence	Lift
[1] {Teacher.experience=Upto 5 years, Teacher.training.duration=1 month, Level.of.teacher.identified.by.peers=1} ⇒ {Teacher.result=Good}	0.0019	100%	14.76

6.1.2.4. Discussion. The obtained rule in Table 6.2 is applicable on a group of 5 teachers for whom good result was observed when they had up to 5 years of experience, had acquired training for a month’s duration, and were ranked as the best (1) by their peers. An identification of this group of teachers is interesting for LOM who may want to know from which schools and clusters these teachers belong, and what are the other characteristics of these teachers. This set of teachers belonged to three

different geographical clusters and 3 out of 5 teachers were mentored by the same L2M. As shown in Table 6.3, the designation of these teachers was ESE (Elementary School Educator) which is a senior designation and these teachers had good academic and professional qualifications and most of them had a workload of more than 40 hours per week.

Table 6.3: Teachers belonging to rule number 1 – Teacher outcome = Good

S. No.	Teacher designation	Teacher workload per week	Teacher academic qualification	Teacher professional qualification	Teacher age	Subject team of teacher
1	ESE	More than 40 hours	Bachelors	M.Ed.	31-50 years	<NA>
2	ESE	21-40 hours	Masters	B.Ed.	31-50 years	<NA>
3	ESE	More than 40 hours	<NA>	B.Ed.	Upto 30 years	All
4	ESE	More than 40 hours	Masters	B.Ed.	Upto 30 years	All
5	ESE	More than 40 hours	Bachelors	B.Ed.	31-50 years	All

Table 6.4: Itemsets from parallel coordinates plot – Teacher outcome = Good

Position	Itemsets	Comments
1	Teacher experience = Upto 5 years	Average teacher experience is 22 years
2	Teacher training duration = 1 month	Average teacher training duration is 28 days
3	Level of teacher identified by peers = 1	Peer ranking of teacher where 1 is the best
4	Teacher result = Good	Teacher result being Good among {Good, Average, Bad}

6.1.3. Experiment 2 – Teacher outcome = Bad.

6.1.3.1. Objective. The objective of this experiment was to determine the teacher and classroom attributes that were correlated with teacher's outcome being Bad.

6.1.3.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.007, value established using trial-and-error with a starting minimum support of 0.1 and decreasing it until some interesting rules were obtained.
- Confidence=85%, value used in all experiments to obtain high-confidence rules which are constrained to be true most of the time.
- Absolute minimum support count = 18 rows out of 2,613 transactions (from Equation 4).

- Rule template: {LHS: All items from the basket in Table 6.1 except for Teacher result, RHS: Teacher result = Bad }

The RHS of the resulting rules is pre-specified to obtain the rules with itemsets that relate to the teacher outcome to be Bad.

6.1.3.3. Results. A total set of 18 rules was generated from the Apriori algorithm using the above listed constraints. The resulting rules are interesting due to their high lift and confidence values. The scatter plot is shown in Fig. 6.4, matrix plot in Fig. 6.5, and the interesting itemsets from the parallel coordinates plot in Table 6.8. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot (Fig. 6.4) for the rule set generated for bad teacher outcome shows two high-lift rules at the top with high confidence but low support. These interesting rules are given in Table 6.5.

The matrix plot shows high support rules when bad teacher result is observed. These rules are at antecedents 12, 16, and 7. The rules that are deemed interesting from the matrix plot are given in Table 6.6.

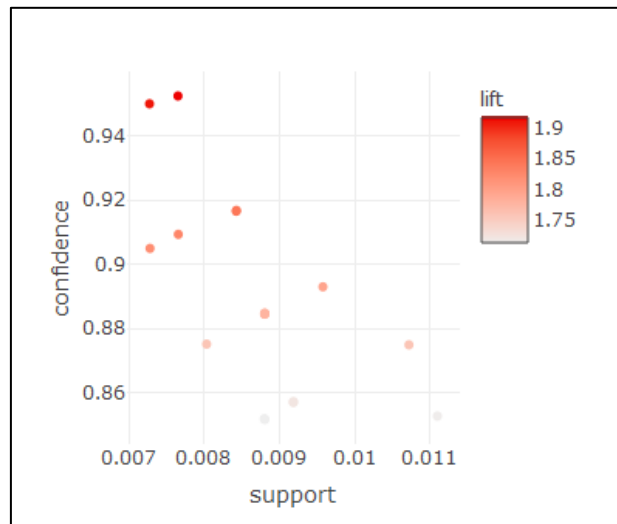


Fig. 6.4: Scatter plot for Experiment 2 – Teacher outcome = Bad

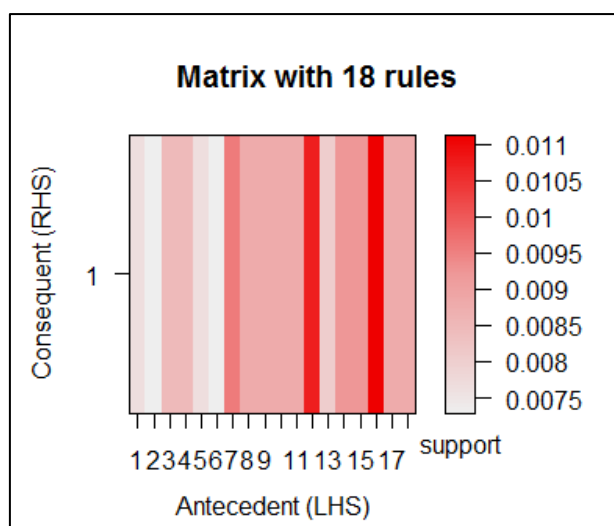


Fig. 6.5: Matrix plot for Experiment 2 – Teacher outcome = Bad

Table 6.5: Interesting rules from scatter plot – Teacher outcome = Bad

Rule	Support	Confidence	Lift
[1] {Teacher.workload.per.week=21-40 hours, Teacher.professional.qualification=PTC/JV/CT, Subject.team.of.teacher=NL + Religion + SS} ⇒ {Teacher.result=Bad}	0.0076	95.2%	1.91
[2] {Teacher.designation=PST, Recommended.teacher.training=Training on use of support material} ⇒ {Teacher.result=Bad}	0.0072	95%	1.91

Table 6.6: Interesting rules from matrix plot – Teacher outcome = Bad

Rule	Support	Confidence	Lift
[7] {Teacher.workload.per.week=21-40 hours, Teacher.training.duration=1 month, Level.of.teacher.identified.by.peers=4} ⇒ {Teacher.result=Bad}	0.009	89.3%	1.79
[12] {Teacher.experience=16-30 years, Level.of.teacher.identified.by.peers=4, Class.ratio=All Boys} ⇒ {Teacher.result=Bad}	0.01	87.5%	1.76
[16] {Subject.team.of.teacher=English + SS, Class.ratio=All Boys} ⇒ {Teacher.result=Bad}	0.01	85.3%	1.71

6.1.3.4. Discussion. The rules obtained from scatter and matrix plots represented various overlapping groups of teachers for whom bad outcome was related to certain classroom and teacher characteristics.

The first rule (Table 6.5) identified a set of 21 teachers for whom a workload of 21-40 hours per week, low professional qualification of PTC/JV/CT, and being an expert in the subjects of NL, Religion and SS were related to their bad results. The examination of other characteristics of this cluster of teachers showed that all these

teachers had a low designation of PST (Primary School Teacher) and were relatively older that is they were aged between 31 to more than 50 years. A violation of fidelity was also observed for this group of teachers since most of them had good peer ranking values of 1 or 2 but obtained bad results. This could mean that the L2Ms might be making up data for these teachers. Table 6.7 shows some attributes of the top 8 teachers belonging to the group formed by this rule.

Table 6.7: Teachers belonging to rule number 1 – Teacher outcome = Bad

S. No.	Teacher designation	Teacher academic qualification	Teacher age	Recommended teacher training	Level of teachers (peer ranking)
1	PST	Bachelors	<NA>	Child Friendly School (CFS)	1
2	PST	Grade 10	<NA>	Whole School Development Plan	1
3	PST	High School Diploma	31-50 years	English	1
4	PST	Grade 10	31-50 years	Maths	3
5	PST	Masters	31-50 years	English	1
6	PST	Grade 10	More than 50 years	Maths	4
7	PST	High School Diploma	More than 50 years	English	4
8	PST	Grade 10	More than 50 years	English	2

As for rule number 2 (Table 6.5), the 20 teachers who got bad results were associated to the low designation of PST and the recommended training on use of support material.

Rule number 7 in Table 6.6 associated the workload of 21-40 hours per week, training duration of 1 month and a bad peer ranking of 4, to 28 teachers who got bad results.

According to rule number 12 (Table 6.6) which is a high-support rule, the 32 teachers who were at their senior-career level and had a considerably good teaching experience of 16-30 years and bad peer ranking of 4 were linked to bad results. These teachers taught in all-boys classrooms.

Finally, rule number 16 (Table 6.6), again a high-support rule with a support value of 0.01, refers to a group of 34 teachers for whom bad results were associated to the teachers being experts in the subjects of English and SS and all-boys classrooms.

The itemsets from the parallel coordinates plot (Table 6.8) also identified a class size of 16-35 students as a factor that is associated to bad teacher outcome. This characteristic is comprehensible and can be controlled to achieve good teacher result and improved learning for students.

Table 6.8: Itemsets from parallel coordinates plot – Teacher outcome = Bad

Position	Itemsets	Comments
1	Class size = 16-35	Average class size is 21 students
	Class ratio = All boys	Ratio: { All boys, All girls, More boys, More girls, Balanced }
2	Subject team of teacher = English + SS	Teacher is an expert in the subjects of English and Social Studies
	Teacher experience = 16-30 years	Average teacher experience is 22 years
	Level of teacher identified by peers = 4	Peer ranking of teacher where 4 is the worst
4	Teacher result = Bad	Teacher result being Bad among { Good, Average, Bad }

6.1.4. What distinguished good and bad teachers. A comparison of experiments performed for teacher outcome analysis show that the same training duration of 1 month was observed for both kinds of teachers, however, the teachers who got good results held a designation of ESE (Elementary School Educator) and an experience of up to 5 years with a good peer ranking value of 1. Alternatively, the teachers who got bad results had a low designation of PST (Primary School Teacher), were more experienced (16-30 years), and had obtained a bad peer ranking of 4 on the scale of 1-4. In addition, teachers with bad outcomes were observed to have low professional qualification of PTC/JV/CT, they were deemed experts in the subjects of NL, SS, Religion and English, and they taught in boys schools.

6.1.5. Conclusion. The experiments conducted for teacher outcome analysis gave only 1 rule for itemsets that were related to the good teacher outcomes with a very low support of 0.0019. For teachers with bad results, a mean support of 0.008738 (standard deviation of 0.00099) was observed which was relatively higher and thus these rules were applicable to a larger group of teachers. For both kinds of teachers, the mean confidence was high. The mean lift of 1.79 for bad results across 2,613 teachers indicate good dependability between the antecedents and consequent of the obtained rules. Due to the very low support, the obtained rules do not represent the complete data set and are representative of only the group of teachers they are applicable to. Table 6.9 shows the mean quality measures for experiments performed for teacher outcome analysis. Steps for rule generation and visualization of experiments performed for Teacher outcome analysis are given in Appendix C.

Table 6.9: Summary of rules for teacher outcome analysis

Teacher outcome	Good				Bad			
No. of rules	1				18			
Quality measure	Min	Max	Mean	StDev	Min	Max	Mean	StDev
Support	0.0019	0.0019	0.0019	-	0.0073	0.011	0.0087	0.00099
Lift	14.76	14.76	14.76	-	1.71	1.92	1.79	0.061
Confidence	1.00	1.00	1.00	-	0.852	0.95	0.89	0.03

6.2. Subjects Outcome Analysis

The subjects outcome analysis is based on the variables of average marks obtained by students in the subjects of English, Science, Social Studies, General Knowledge, National Language, and Religion. The subject results are expressed by students' results in this analysis. The subjects outcome ranged from 1-10, but also contained missing values coded as 0 and some outliers that were above 10. These values were cleansed and the subjects outcome was discretized as follows:

$$\text{Average marks per subject} = \begin{cases} \text{Bad,} & \text{if value} \leq 4 \\ \text{Average,} & \text{if value} > 4 \text{ and } \leq 6.5 \\ \text{Good,} & \text{if value} > 6.5 \end{cases} \quad (15)$$

The distribution of outcomes of all subjects in the educational data are given below in Fig. 6.6.

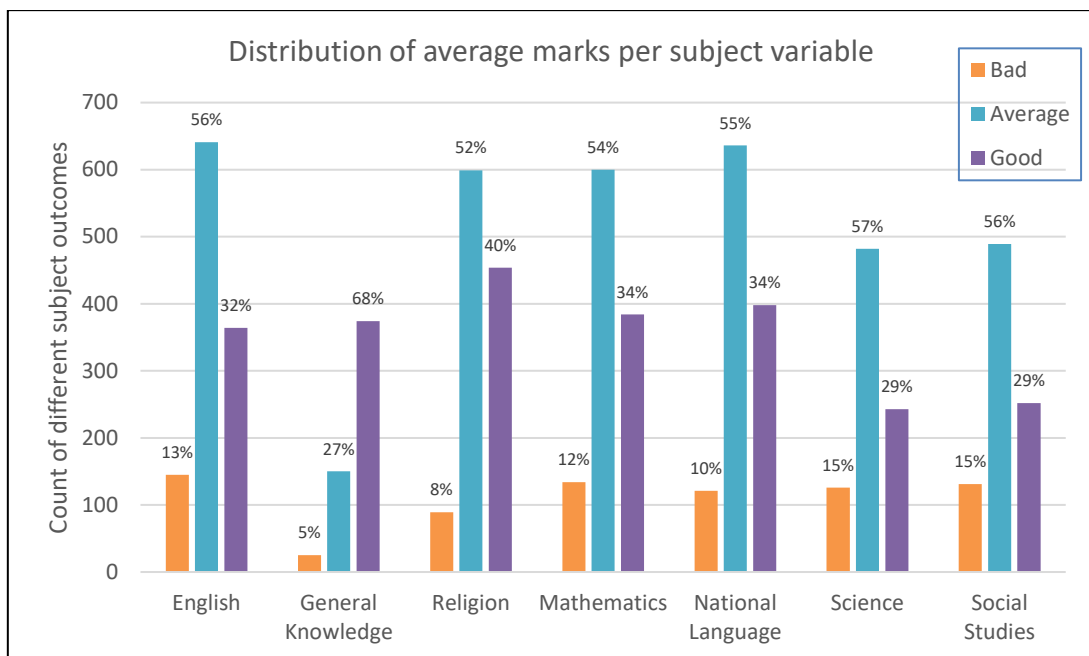


Fig. 6.6: Distribution of subjects with different outcomes

6.2.1. Educational basket for subjects outcome analysis. The education basket used to study all subjects’ outcomes relationship to other variables is similar to the basket shown in Table 6.1, except for the outcome variable which is replaced by different subjects’ outcome. This basket has a total of 2,613 transactions and is given in Table 6.10.

Table 6.10: Items in education basket for subject outcome analysis

Items	Potential Values
Teacher designation	{DYHM, ESE, EST, HM, PST, SESE, SSE}
Teacher workload per week	{1-20 hours, 21-40 hours, More than 40 hours}
Teacher academic qualification	{Grade 10, High School Diploma, Bachelors, Masters}
Teacher professional qualification	{PTC/JV/CT, B.Ed., M.Ed., Other}
Teacher age	{Upto 30 years, 31-50 years, More than 50 years}
Teacher experience	{Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
Teacher training duration	{Upto 2 weeks, 1 month, More than a month}
Recommended teacher training	There are 34 training areas. Some of them are: {English, Maths, Social Studies, lesson planning, activity-based teaching and learning, classroom management, multi-grade teaching, child friendly school, ...}
Subject team of teacher	There are 14 subject teams for teachers. Some of them are: {All, English, English + Maths + Science,...}
Level of teacher identified by peers	{1, 2, 3, 4}; 1 being the best
Class size	{1-15, 16-35, More than 35}
Percentage of present students in class	{Bad, Good}
Class ratio	{All Boys, All Girls, Balanced, More boys, More girls}
Average marks per subject English or GK or Religion or Maths or NL or Science or SS	{Bad, Average, Good}

6.2.2. Experiment 1 – English outcome = Good.

6.2.2.1. Objective. The objective of this experiment was to study the teacher and classroom characteristics that are related to student learning outcomes that are “Good” in the subject of English.

6.2.2.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, using trial-and-error with a starting minimum support of 0.1.
- Confidence=85%
- Absolute minimum support count = 5 rows out of 2,613 transactions (Equation 4).
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per subject English, RHS: Average marks per subject English = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the subject English outcome to be Good.

6.2.2.3. Results. A set of 3 rules was generated from the Apriori algorithm using the above listed constraints. The minimum support value of 0.002 was low with a coverage of only 5 teachers, but the obtained rules were interesting based on their very high lift and confidence values. The scatter plot is shown in Fig. 6.7, matrix plot in Fig. 6.8, and the itemsets from the parallel coordinates plot in Table 6.13.

The scatter plot in Fig. 6.7 shows only one blue point, since all the obtained rules have the same values of support, confidence, and lift. Similarly, the matrix plot in Fig. 6.8 shows the three rules with equally dark bars due to the same support values. The resulting rules are given in Table 6.11.

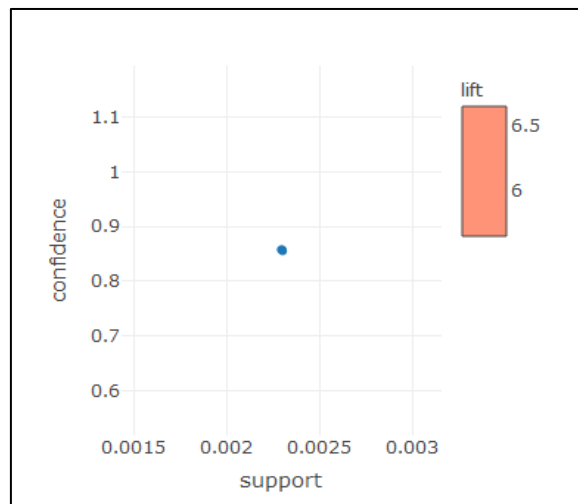


Fig. 6.7: Scatter plot for Experiment 1 – English outcome = Good

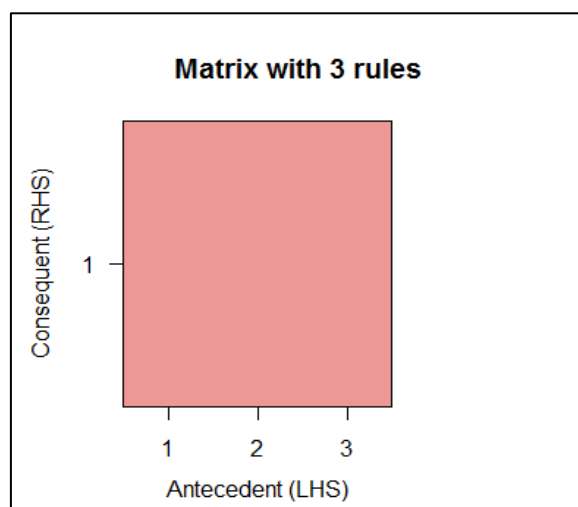


Fig. 6.8: Matrix plot for Experiment 1 – English outcome = Good

Table 6.11: Resulting rules – English outcome = Good

Rule	Support	Confidence	Lift
[1] {Teacher.designation=ESE, Recommended.teacher.training=Training on Multi-grade teaching, Percentage.of.present.students.in.class=Good} ⇒ {Avg.marks.per.subject.ENG=Good}	0.0023	85.7%	6.15
[2] {Teacher.professional.qualification=PTC/JV/CT, Teacher.training.duration=Upto 2 weeks, Recommended.teacher.training=Training on lesson planning} ⇒ {Avg.marks.per.subject.ENG=Good}	0.0023	85.7%	6.15
[3] {Teacher.training.duration=Upto 2 weeks, Recommended.teacher.training=Training on lesson planning, Percentage.of.present.students.in.class=Good} ⇒ {Avg.marks.per.subject.ENG=Good}	0.0023	85.7%	6.15

6.2.2.4. Discussion. Since the above obtained rules have a very low support covering only 5 to 6 teachers, these rules can be seen as three clusters which have some specific teacher and classroom characteristics that led to good outcome in the subject of English. The first rule is a cluster of 6 teachers who held a designation of ESE and were recommended training on multi-grade teaching. Multi-grade teaching is a common situation in developing countries and refers to the teaching of different grades at the same time in the same classroom. A good attendance of students was also observed for such teachers who were associated to good English outcome.

The second and third rule identified the same cluster of 6 teachers who had been trained for upto 2 weeks only, had a professional qualification of PTC/JV/CT, and were recommended training on lesson planning. A good attendance of students was also observed in classrooms that were taught by these teachers. Upon identifying the teachers in this cluster, it was observed that these teachers had varied academic qualification ranging from Grade 10 (lowest qualification) to Masters (highest qualification). However, most of these teachers taught in all-girls schools and belonged to the subject team of English, as shown in Table 6.12.

Table 6.12: Teachers belonging to rule number 2 – English outcome = Good

S. No.	Teacher workload per week	Teacher academic qualification	Subject team of teacher	Level of teacher identified by peers	Class size	Class ratio
1	More than 40 hours	Masters	English + Maths + Science	3	1-15	All girls
2	More than 40 hours	High School Diploma	English	2	1-15	All girls
3	21-40 hours	High School Diploma	Science	2	More than 35	More girls
4	More than 40 hours	Bachelors	English + SS	<NA>	16-35	All girls
5	More than 40 hours	High School Diploma	English	<NA>	16-35	All girls
6	More than 40 hours	Grade 10	SS	<NA>	16-35	All boys

Table 6.13: Itemsets from parallel coordinates plot - English outcome = Good

Position	Itemsets	Comments
1	Teacher designation = ESE	Teacher designation of Elementary School Educator (mid-career level)
	% of present students in class = Good	The % of present students can be Good or Bad
	Teacher professional qualification = PTC/JV/CT	Low professional qualification
2	Recommended teacher training = Training on Multi-grade teaching	A training area recommended to teacher out of 34 training areas
	Teacher training duration = Upto 2 weeks	Average teacher training duration is 28 days
3	Recommended teacher training = Training on lesson planning	A training area recommended to teacher out of 34 training areas
4	Average marks per subject English = Good	English outcome being Good among {Bad, Average, Good}

6.2.3. Experiment 2 – English outcome = Bad.

6.2.3.1. Objective. The objective of this experiment was to study the teacher and classroom characteristics that are related to student learning outcomes that are “Bad” in the subject of English.

6.2.3.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value established using trial-and-error
- Confidence=85%, value used to achieve high-confidence rules in all experiments
- Absolute minimum support count = 5 rows out of 2,613 (Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per subject English, RHS: Average marks per subject English = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the English outcome to be Bad.

6.2.3.3. Results. Only 1 rule was generated from the Apriori algorithm using the above listed constraints. This rule had a high lift of 15.4 and a good confidence of 85.7%, but a low support of 0.002. These parameters indicate that the antecedents and consequent are infrequent but strongly dependent. The scatter plot is shown in Fig. 6.9 and the matrix plot in Fig. 6.10, and the itemsets from the parallel coordinates plot in Table 6.15.

The scatter plot in Fig. 6.9 and matrix plot in Fig. 6.10 show the resulting rule by a blue point and a bar respectively with a support of more than 0.002. This rule is listed in Table 6.14.

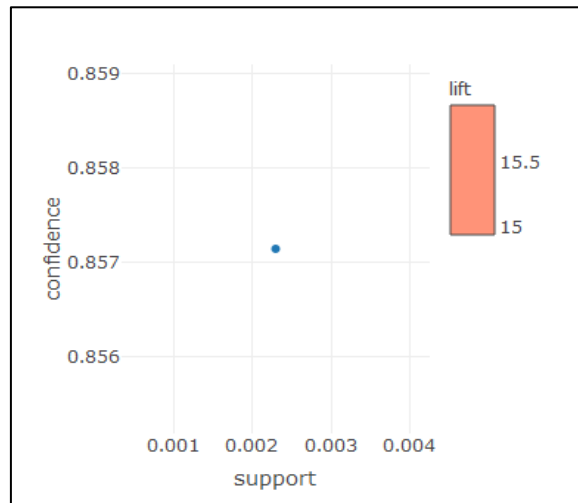


Fig. 6.9: Scatter plot for Experiment 2 – English outcome = Bad

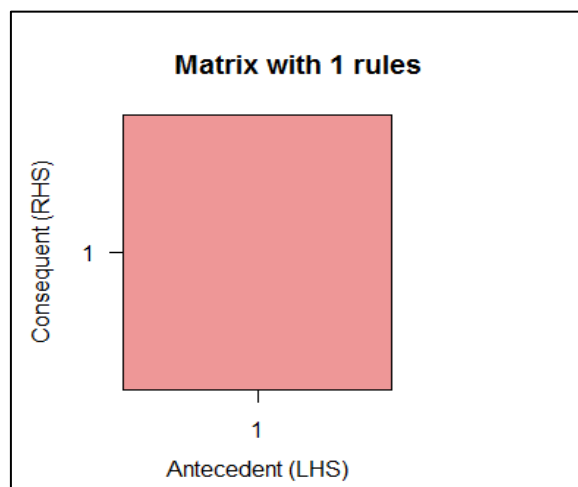


Fig. 6.10: Matrix plot for Experiment 2 – English outcome = Bad

Table 6.14: Resulting rule – English outcome = Bad

Rule	Support	Confidence	Lift
[1] {Subject.team.of.teacher=NL + Religion + SS, Class.size=1-15, Percentage.of.present.students.in.class=Bad} ⇒ {Avg.marks.per.subject.ENG=Bad}	0.0023	85.7%	15.45

6.2.3.4. Discussion. The obtained rule gives only one teacher cluster in which the small class size with a total of 1-15 students, bad students’ attendance and teacher being an expert in the subjects of National Language, Religion, and Social Studies were associated with the bad students’ marks in the subject of English.

Table 6.15: Itemsets from parallel coordinates plot - English outcome = Bad

Position	Itemsets	Comments
1	Subject team of teacher = NL + Religion + SS	Teacher is an expert in the subjects of National Language, Religion, and Social Studies
2	Class size = 1-15	Average class size is 21 students
3	% of present students in class = Bad	The % of present students can be Good or Bad
4	Average marks per subject English = Bad	English outcome being Bad among {Bad, Average, Good}

6.2.4. Comparison of Good vs. Bad English outcome. The teachers who were related to good or bad student learning in the subject of English were mainly distinguished by the percentage of present students in the class. Good pupil’s attendance was observed in classes that achieved good outcomes in the subject of English and vice-versa.

6.2.5. Experiment 3 – General Knowledge (GK) outcome = Good.

6.2.5.1. Objective. The objective of this experiment was to study the teacher and classroom characteristics that are related to student learning outcomes that are “Good” in the subject of GK.

6.2.5.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, using trial-and-error with a starting minimum support of 0.1
- Confidence=85%
- Absolute minimum support count = 5 rows out of 2,613 (Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per subject GK, RHS: Average marks per subject GK = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the GK outcome to be Good.

6.2.5.3. Results. One rule was generated from the Apriori algorithm using the above listed constraints. This rule has a lift of 5.98 and a confidence of 85.7% which is fairly high and suggest that the rule is a strong rule, even though the support is low and covers only 5 transactions out of the 2,613. The antecedents and consequent in the rule are infrequent (due to low support), but it is true most of the time. The rule is shown in Table 6.16.

The scatter plot is shown in Fig. 6.11, matrix plot in Fig. 6.12, and the itemsets with their respective positions in the parallel coordinates plot are shown in Table 6.17.

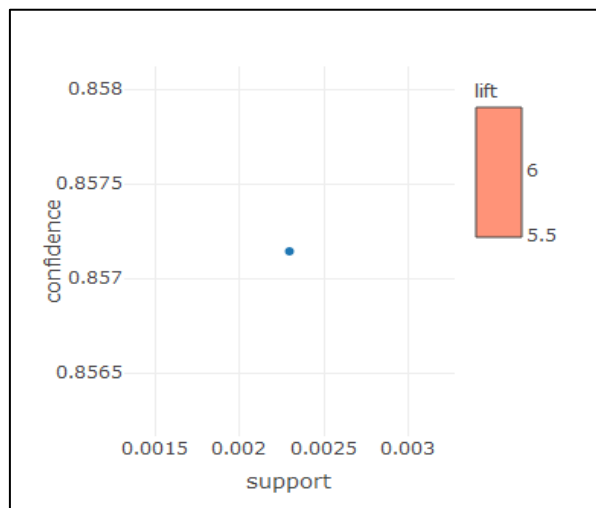


Fig. 6.11: Scatter plot for Experiment 3 – GK outcome = Good

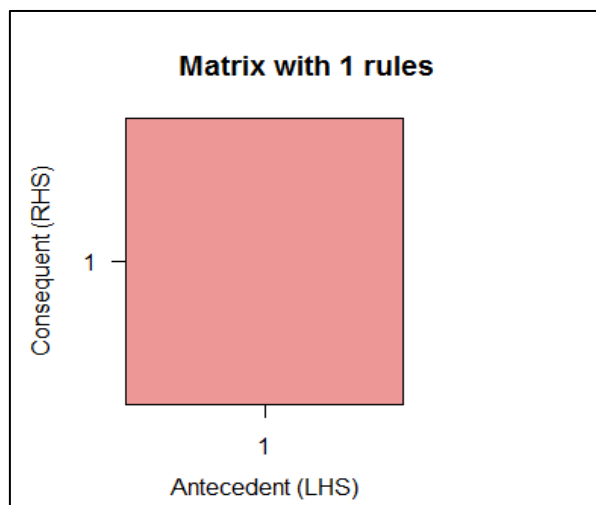


Fig. 6.12: Matrix plot for Experiment 3 – GK outcome = Good

Table 6.16: Resulting rule – GK outcome = Good

Rule	Support	Confidence	Lift
[1] {Recommended.teacher.training=Training on activity-based teaching and learning, Subject.team.of.teacher=All, Class.size=1-15} ⇒ {Avg.marks.per.subject.GK=Good}	0.0023	85.7%	5.98

6.2.5.4. Discussion. The obtained rule gives one teacher cluster and suggests that the good outcome in GK is related to the recommended training in the area of activity-based teaching and learning and teacher being an expert in teaching all Primary-level subjects. In addition, student learning in the subject of GK is seen to improve when the class size is small and consist of 1-15 students only.

Table 6.17: Itemsets from parallel coordinates plot - GK outcome = Good

Position	Itemsets	Comments
1	Recommended teacher training = Training on activity based teaching and learning	Teacher is an expert in the subjects of National Language, Religion, and Social Studies
2	Subject team of teacher = All	Teacher has expertise in teaching all subjects
3	Class size = 1-15	Average class size is 21 students
4	Average marks per subject GK = Good	GK outcome being Good among {Bad, Average, Good}

6.2.6. Experiment 4 – General Knowledge (GK) outcome = Bad.

6.2.6.1. Objective. The objective of this experiment was to determine the teacher and classroom variables that relate to bad results in GK.

6.2.6.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by decreasing the starting minimum support value of 0.1 in order to find rules.
- Confidence=85%, high-confidence rules that are not constrained to be true all the time.
- Absolute minimum support count = 5 rows out of 2,613 (Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per subject GK, RHS: Average marks per subject GK = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the GK outcome to be Bad.

6.2.6.3. Results. No rules were obtained in this experiment, even with a very low support of 0.002 which covered only 5 transactions.

6.2.7. Experiment 5 – Religion outcome = Good.

6.2.7.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that are related to good outcome in the subject of Religion.

6.2.7.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value established using trail-and-error with a starting minimum support of 0.1.
- Confidence=85%, to constrain the resulting rules to have high-confidence, that is they are true most of the time.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Religion, RHS: Average marks per Religion = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the good outcome in Religion.

6.2.7.3. Results. A set of 6 rules was generated from the Apriori algorithm using the above listed constraints. These rules have low minimum support, but high lift and confidence values indicating strong dependence between the antecedents and consequent. The scatter plot is shown in Fig. 6.13, matrix plot in Fig. 6.14, and itemsets from the parallel coordinates plot in Table 6.20. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot in Fig. 6.13 points out the most interesting rule with the highest lift at the top of the plot with a confidence of 100%. Another rule with a support of 0.003 and confidence of 88% is also highlighted by this plot. The matrix plot in Fig. 6.14 highlights one rule with the highest support with a red bar. The resulting rules are given in Table 6.18.

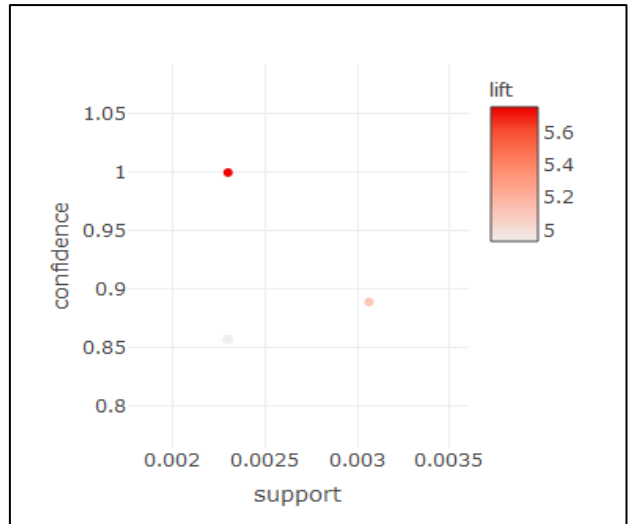


Fig. 6.13: Scatter plot for Experiment 5 – Religion outcome = Good

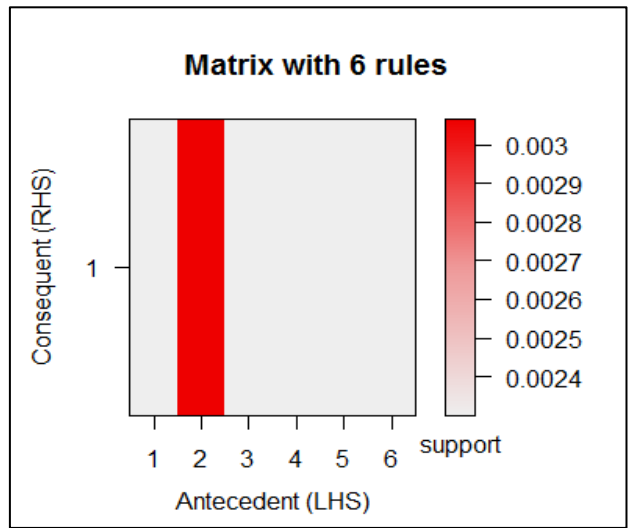


Fig. 6.14: Matrix plot for Experiment 5 – Religion outcome = Good

Table 6.18-A: Resulting rules – Religion outcome = Good

Rule	Support	Confidence	Lift
[1] {Teacher.training.duration=Upto 2 weeks, Recommended.teacher.training=Training in subject of English, Subject.team.of.teacher=English + Maths + Science} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.0023	100%	5.75
[2] {Teacher.academic.qualification=High School Diploma, Subject.team.of.teacher=English, Class.ratio=All Girls} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.003	88.8%	5.12
[3] {Recommended.teacher.training=Training on Multi-grade teaching, Subject.team.of.teacher=Maths} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.0023	85.7%	4.93

Table 6.18-B: Resulting rules – Religion outcome = Good

Rule	Support	Confidence	Lift
[4] {Teacher.professional.qualification=PTC/JV/CT, Teacher.training.duration=Upto 2 weeks, Recommended.teacher.training=Training on lesson planning} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.0023	85.7%	4.93
[5] {Teacher.designation=PST, Teacher.training.duration=Upto 2 weeks, Recommended.teacher.training=Training on lesson planning} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.0023	85.7%	4.93
[6] {Teacher.age=31-50 years, Subject.team.of.teacher=Maths, Class.ratio=Balanced} ⇒ {Avg.marks.per.subject.RELIGION=Good}	0.0023	85.7%	4.93

6.2.7.4. Discussion. The obtained rules give 6 clusters of teachers who are associated to good outcomes in the subject of religion. The first cluster with the highest confidence of 100% contains 6 teachers from the same cluster and mentored by the same L2M. All these teachers worked at different schools and were recommended training in the subject of English when they were deemed experts in the subjects of English, Maths, and Science, and were trained for upto 2 weeks.

The second cluster given by rule number 2 having the highest support of 0.003 and a confidence of 88.8%, comprised of 9 teachers with an academic qualification of High School Diploma that is 12 years of education. These teachers were experts in the subject of English and taught in all-girls classrooms.

The third cluster of teachers who were related to good outcome in the subject of Religion comprised of 7 teachers who were recommended training on multi-grade teaching and were subject experts of Mathematics. Upon further analysis of teachers in this cluster, it was observed that these teachers belonged to the same geographical cluster and were mentored and assessed by the same L2M. Furthermore, these teachers had a workload of more than 40 hours per week and a peer ranking of 2. However, the academic qualification of these teachers displayed quite a contrast where some teachers were highly qualified with an academic qualification of Masters and others studied only till Grade 10 and had the lowest qualification. Some of the teacher characteristics of teachers in this cluster are shown in Table 6.19.

Table 6.19: Teachers belonging to rule number 3 – Religion outcome = Good

S. No.	Teacher designation	Teacher workload per week	Teacher academic qualification	Teacher age	Level of teacher identified by peers	Class size	Class ratio
1	ESE	More than 40 hours	Masters	31-50 years	2	16-35	Balanced
2	ESE	More than 40 hours	Masters	31-50 years	2	16-35	Balanced
3	PST	21-40 hours	Grade 10	31-50 years	2	1-15	More girls
4	PST	More than 40 hours	<NA>	More than 50 years	2	1-15	More girls
5	PST	More than 40 hours	Grade 10	More than 50 years	2	1-15	All boys
6	PST	More than 40 hours	Grade 10	More than 50 years	2	1-15	All boys
7	ESE	More than 40 hours	Masters	31-50 years	2	1-15	More boys

Rule number 4 and 5 represent the same cluster of 6 teachers who had a designation of PST and a professional qualification of PTC/JV/CT. These teachers were trained for up to 2 weeks and were recommended training on lesson planning.

Finally, the cluster of 7 teachers represented by rule number 6 were aged between 31-50 years. These teachers were subject experts of Mathematics and taught in co-education schools with balanced number of boys and girls in the classrooms.

Table 6.20: Itemsets from parallel coordinates plot – Religion outcome = Good

Position	Itemsets	Comments
1	Teacher academic qualification = High school diploma	Low academic qualification
2	Subject team of teacher = English	Teacher is an expert in the subject of English
3	Class ratio = All girls	Ratio: {All boys, All girls, More boys, More girls, Balanced}
4	Average marks per subject Religion = Good	Religion outcome being Good among {Bad, Average, Good}

6.2.8. Experiment 6 – Religion outcome = Bad.

6.2.8.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that are related to bad outcome in the subject of Religion.

6.2.8.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by decreasing the starting minimum support of 0.1 to find rules.

- Confidence=85%, to constrain the resulting rules to have high-confidence.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Religion, RHS: Average marks per Religion = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the bad outcome in Religion.

6.2.8.3. Results. No rules were found in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.9. Experiment 7 – Mathematics outcome = Good.

6.2.9.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to good outcome in the subject of Mathematics.

6.2.9.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by trial-and-error with a starting minimum support of 0.1.
- Confidence=85%, to constrain the resulting rules to have high-confidence and be true most of the time.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Mathematics, RHS: Average marks per Mathematics = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the good outcome in Mathematics.

6.2.9.3. Results. A set of 4 rules was generated in this experiment when Apriori algorithm was run with the above listed constraints. The rules had a low support of 0.002, but their good confidence and lift values deem them as interesting and suggest that though the LHS and RHS of the rules are infrequent, but whenever they occur together, they are almost always true. The scatter plot for this experiment is shown in

Fig. 6.15, matrix plot in Fig. 6.16, and the itemsets from the parallel coordinates plot are given in Table 6.24.

Both the scatter and matrix plots highlight the first two rules as important with dark points and bars respectively. However, since there is not much difference in the parameter values of all rules, all of them are analysed and are given in Table 6.21.

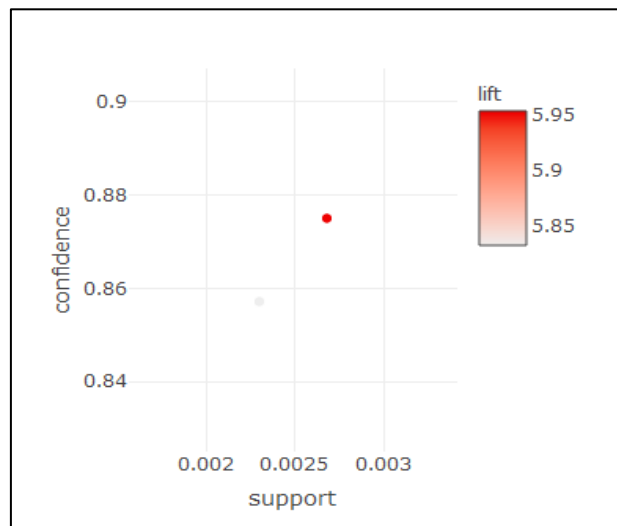


Fig. 6.15: Scatter plot for Experiment 7 – Maths outcome = Good

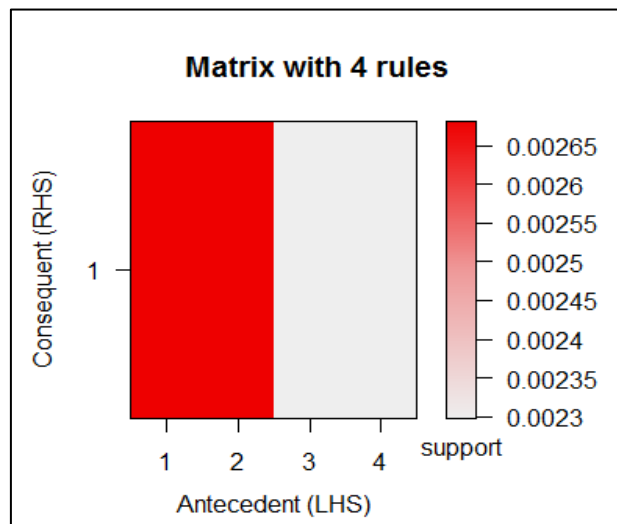


Fig. 6.16: Matrix plot for Experiment 7 – Maths outcome = Good

Table 6.21: Resulting rules – Maths outcome = Good

Rule	Support	Confidence	Lift
[1] {Teacher.experience=Upto 5 years, Teacher.training.duration=1 month, Class.size=1-15} ⇒ {Avg.marks.per.subject.MATHS=Good}	0.0027	87.5%	5.95
[2] {Teacher.experience=16-30 years, Recommended.teacher.training=Training in subject of English, Subject.team.of.teacher=English} ⇒ {Avg.marks.per.subject.MATHS=Good}	0.0027	87.5%	5.95
[3] {Teacher.experience=Upto 5 years, Class.size=1-15, Class.ratio=All Girls} ⇒ {Avg.marks.per.subject.MATHS=Good}	0.0023	85.7%	5.83
[4] {Teacher.experience=More than 30 years, Subject.team.of.teacher=English + SS, Class.ratio=More girls} ⇒ {Avg.marks.per.subject.MATHS=Good}	0.0023	85.7%	5.83

6.2.9.4. Discussion. The resulting rules for teachers and classroom attributes that were correlated to good learning outcome in the subject of Mathematics gave 4 clusters of teachers. The first cluster identified 8 teachers who were less experienced (up to 5 years) but trained for up to a month’s duration. These teachers taught in small sized classrooms with 1-15 students in each class. Upon further examining these teachers, it was found that they were more qualified (Bachelors or Masters), held a designation of ESE (Elementary School Educator) mostly, had a workload of more than 40 hours a week, and mostly taught in all-girls classrooms or classrooms where more girls were present than boys, as shown in Table 6.22.

Table 6.22: Teachers belonging to rule number 1 – Maths outcome = Good

S. No.	Teacher designation	Teacher workload per week	Teacher academic qualification	Teacher professional qualification	Teacher age	Level of teacher identified by peers	Class ratio
1	ESE	More than 40 hours	Masters	B.Ed.	Upto 30 years	4	All girls
2	ESE	More than 40 hours	Masters	PTC/JV/CT	Upto 30 years	<NA>	All girls
3	PST	More than 40 hours	Bachelors	B.Ed.	31-50 years	3	Balanced
4	ESE	More than 40 hours	Bachelors	B.Ed.	Upto 30 years	4	All girls
5	ESE	More than 40 hours	Bachelors	B.Ed.	Upto 30 years	<NA>	All girls
6	ESE	More than 40 hours	<NA>	B.Ed.	Upto 30 years	1	All girls
7	ESE	More than 40 hours	Masters	B.Ed.	Upto 30 years	1	All boys
8	ESE	More than 40 hours	Bachelors	B.Ed.	31-50 years	1	More girls

The second cluster of 8 teachers linked to good Mathematics outcome were more experienced than teachers belonging to the cluster of rule number 1 and had an experience of 16-30 years. These teachers were recommended training in the subject of English, while they were also deemed as subject experts in English.

In the third cluster of 7 teachers obtained from rule number 3, again the teachers were less experienced (up to 5 years) and taught in small-sized classrooms comprising of 1-15 students. However, these teachers also taught in all-girls classrooms which is related to the good student learning in Mathematics. Other teacher characteristics of these teachers were found to be similar to teachers belonging to the cluster formed by rule number 1, that is they were more qualified, held a designation of ESE, and were relatively younger (aged upto 30 years).

The fourth cluster comprised of 7 teachers who were very experienced (more than 30 years), were deemed subject experts in English and SS, and taught in co-education schools with considerably more girls in the classrooms. From further identification of these teachers, it was found that they were relatively older (aged more than 50 years), held a designation of PST and acquired mostly low qualification (Grade 10), as shown in Table 6.23.

Table 6.23: Teachers belonging to rule number 4 – Maths outcome = Good

S. No.	Teacher designation	Teacher workload per week	Teacher academic qualification	Teacher professional qualification	Teacher age	Level of teacher identified by peers	Class size
1	PST	<NA>	Grade 10	PTC/JV/CT	More than 50 years	2	1-15
2	PST	21-40 hours	Grade 10	PTC/JV/CT	More than 50 years	2	1-15
3	PST	More than 40 hours	Grade 10	PTC/JV/CT	More than 50 years	2	1-15
4	PST	More than 40 hours	Grade 10	PTC/JV/CT	More than 50 years	4	16-35
5	PST	More than 40 hours	High School Diploma	PTC/JV/CT	More than 50 years	1	16-35
6	PST	More than 40 hours	Grade 10	PTC/JV/CT	More than 50 years	4	16-35
7	PST	21-40 hours	Masters	B.Ed.	More than 50 years	2	More than 35

Table 6.24: Itemsets from parallel coordinates plot - Maths outcome = Good

Position	Itemsets	Comments
1	Teacher experience = More than 30 years	Average teacher experience is 22 years
2	Class size = 1-15	Average class size is 21 students
	Subject team of teacher = English + SS	Teacher is an expert in the subjects of English and Social Studies
3	Class ratio = All girls, More girls	Ratio: {All boys, All girls, More boys, More girls, Balanced}
4	Average marks per subject Maths = Good	Maths outcome being Good among {Bad, Average, Good}

6.2.10. Experiment 8 – Mathematics outcome = Bad.

6.2.10.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that are related to bad outcome in the subject of Mathematics.

6.2.10.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Mathematics, RHS: Average marks per Mathematics = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the bad outcome in Mathematics.

6.2.10.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.11. Experiment 9 – National language (NL) outcome = Good.

6.2.11.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to good outcome in the subject of NL.

6.2.11.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.003, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.

- Confidence=85%, to constrain resulting rules to have high confidence.
- Absolute minimum support count = 7 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per NL, RHS: Average marks per NL = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the good outcome in National language.

6.2.11.3. Results. One rule was generated in this experiment when Apriori algorithm was run with the above listed constraints. This rule has a lift of 5.83, a confidence of 88.8%, and a support of 0.003. These parameters indicate that the antecedents in the rule do not occur together very often, but when they do, the outcome for the subject of National language is Good. The scatter plot for this experiment is shown in Fig. 6.17, matrix plot in Fig. 6.18, and the itemsets from the parallel coordinates plot in Table 6.26.

The scatter plot in Fig. 6.17 and matrix plot in Fig. 6.18 show the resulting rule by a blue point and a bar respectively with a support of more than 0.003. This rule is listed in Table 6.25.

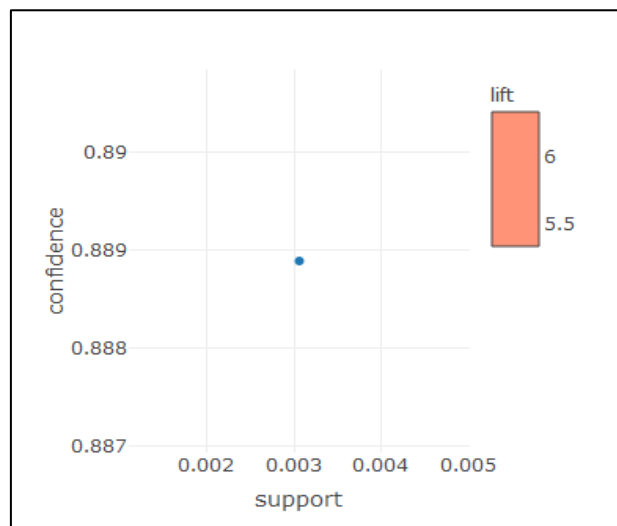


Fig. 6.17: Scatter plot for Experiment 9 – NL outcome = Good

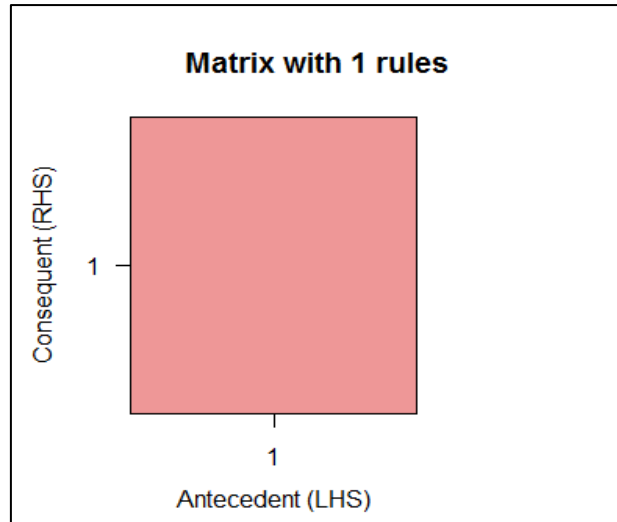


Fig. 6.18: Matrix plot for Experiment 9 – NL outcome = Good

Table 6.25: Resulting rule - NL outcome = Good

Rule	Support	Confidence	Lift
[1] {Teacher.age=31-50 years, Subject.team.of.teacher=English, Class.ratio=More girls} ⇒ {Avg.marks.per.subject.NL=Good}	0.003	88.8%	5.83

6.2.11.4. Discussion. The resulting rule represents one cluster with 9 teachers who were aged between 31-50 years and were considered subject experts in English were associated to good outcome in the subject of NL. Furthermore, these teachers taught in classrooms where there were more girls than boys.

Table 6.26: Itemsets from parallel coordinates plot - NL outcome = Good

Position	Itemsets	Comments
1	Teacher age = 31-50 years	Average teacher age is 46 years
2	Subject team of teacher = English	Teacher is an expert in the subject of English
3	Class ratio = More girls	Ratio: {All boys, All girls, More boys, More girls, Balanced}
4	Average marks per subject National language = Good	National language outcome being Good among {Bad, Average, Good}

6.2.12. Experiment 10 – National language (NL) outcome = Bad.

6.2.12.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to bad outcome in the subject of NL.

6.2.12.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per NL, RHS: Average marks per NL = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the bad outcome in National language.

6.2.12.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.13. Experiment 11 – Science outcome = Good.

6.2.13.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to good outcome in the subject of Science.

6.2.13.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Science, RHS: Average marks per Science = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the good outcome in Science.

6.2.13.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.14. Experiment 12 – Science outcome = Bad.

6.2.14.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to bad outcome in the subject of Science.

6.2.14.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per Science, RHS: Average marks per Science = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the bad outcome in Science.

6.2.14.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.15. Experiment 13 – Social Studies (SS) outcome = Good.

6.2.15.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to good outcome in the subject of SS.

6.2.15.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per SS, RHS: Average marks per SS = Good}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the good outcome in SS.

6.2.15.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.16. Experiment 14 – Social Studies (SS) outcome = Bad.

6.2.16.1. Objective. The objective of this experiment was to determine the teacher and grade characteristics that were related to bad outcome in the subject of SS.

6.2.16.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.002, value obtained by starting with a minimum support of 0.1 and decreasing it to find some rules.
- Confidence=85%, to find high-confidence rules.
- Absolute minimum support count = 5 rows out of 2,613 (from Equation 4)
- Rule template: {LHS: All items from the basket in Table 6.10 except for Average marks per SS, RHS: Average marks per SS = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the bad outcome in SS.

6.2.16.3. Results. No rules were generated in this experiment when Apriori algorithm was run with the above listed constraints.

6.2.17. Conclusion. The experiments conducted for subject outcome analysis did not give any rules for most of the experiments. Table 6.27 shows a summary of experiments performed for subject outcome analysis. The obtained rules have high lift and confidence values showing strong dependence between the antecedents and consequent. However, the very low support values (0.002-0.003) indicate that the rules are applicable to only a certain group of transactions and are not representative of the entire dataset.

Table 6.27: Summary of rules for subjects outcome analysis

Experiment	Subject	Outcome	Number of rules generated	Support				Lift				Confidence			
				Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev
1	English	Good	3	0.002296	0.002296	0.002296	0	6.153	6.153	6.153	0	0.857	0.857	0.8571	0
2		Bad	1	0.002296	0.002296	0.002296	-	15.45	15.45	15.45	-	0.857	0.857	0.8571	-
3	GK	Good	1	0.002296	0.002296	0.002296	-	5.989	5.989	5.989	-	0.857	0.857	0.8571	-
4		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-
5	Religion	Good	6	0.002296	0.0031	0.002424	0.00028	4.93	5.76	5.10	0.3	0.857	1.00	0.8862	0.052
6		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-
7	Maths	Good	4	0.002296	0.00268	0.002488	0.000019	5.83	5.95	5.893	0.061	0.857	0.875	0.8661	0.0089
8		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-
9	NL	Good	1	0.003062	0.003062	0.003062	-	5.836	5.836	5.836	-	0.889	0.889	0.889	-
10		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-
11	Science	Good	0	-	-	-	-	-	-	-	-	-	-	-	-
12		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-
13	SS	Good	0	-	-	-	-	-	-	-	-	-	-	-	-
14		Bad	0	-	-	-	-	-	-	-	-	-	-	-	-

6.3. Teacher Training Analysis

The recommended teacher training analysis was performed by taking into account different values of the 'Training recommended for teacher' variable in the CPD framework. The values for the recommended trainings were assigned by L2M to each of the teachers. The relationship of these recommended trainings with the teacher outcome will be studied in the following experiments.

The bar plot of the distribution of recommended teacher training areas in the micro-education basket in Fig. 6.19 indicates that most of the teachers require training in the following areas:

- Training in subject of English
- Training in subject of Maths
- Training in subject of Science
- Training on lesson planning
- Training on activity-based teaching and learning
- Training on multi-grade teaching

6.3.1. Educational basket for teacher training analysis. The education basket used to study teacher outcome with respect to different recommended training areas is the same basket that was also used for teacher outcome analysis. The items in this basket are given in Table 6.1. This basket has a total of 2,613 transactions.

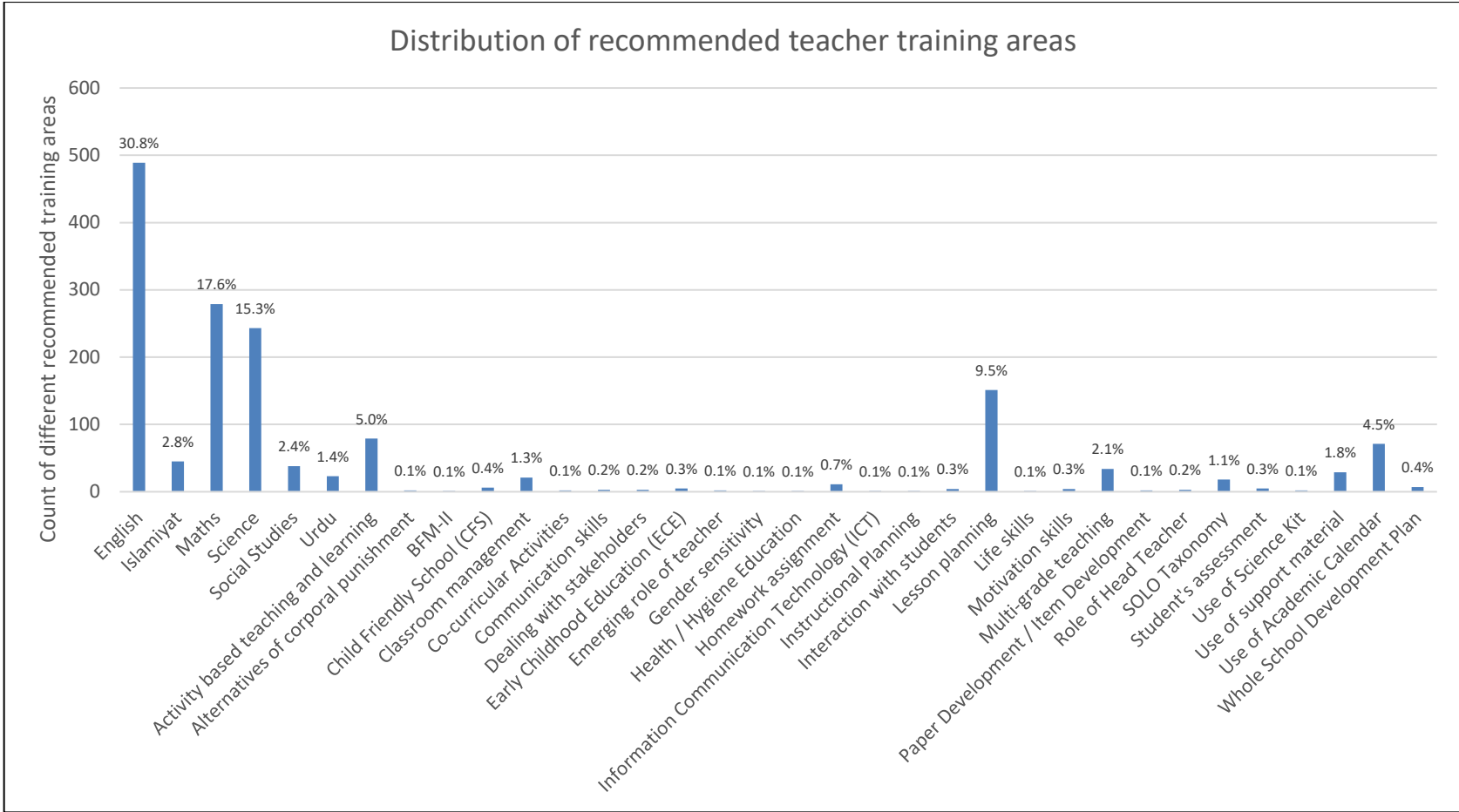


Fig. 6.19: Distribution of teachers with different recommended training areas

6.3.2. Experiment 1 – Recommended teacher training.

6.3.2.1. Objective. The objective of this experiment is to study the relationship of recommended training area alongwith other teacher and classroom characteristics with the teacher outcomes.

6.3.2.2. Constraints. For this experiment, the following constraints were used to mine the rules:

- Support=0.0035, value obtained by decreasing the minimum support from 0.1 until some interesting rules were obtained.
- Confidence=85%, value used in all experiments to obtain high-confidence rules.
- Absolute minimum support count = 9 rows out of 2,613 (Equation 4)
- Rule template: {LHS: Recommended teacher training AND all items from the basket in Table 6.1 except for Teacher result, RHS: Teacher result = “Bad” or “Average” or “Good”}.

The LHS and RHS of the resulting rules were pre-specified to obtain the rules with Recommended teacher training and Teacher result respectively.

6.3.2.3. Results. A set of 21 rules were obtained when Apriori algorithm was run with the above listed constraints. The minimum support value is low indicating that the antecedents do not occur frequently. The highest lift value observed in these rules is 2 with a confidence of 100% showing strong association between the LHS and the RHS. The scatter plot for this experiment is shown in Fig. 6.20, matrix plot in Fig. 6.21, and the itemsets from the parallel coordinates plot in Table 6.30.

The scatter plot in Fig. 6.20 shows 4 interesting rules with the highest lift of 2 at the top of the plot. These rules have support values ranging from 0.0038 to 0.0057, and a confidence of 100%. These rules are listed in Table 6.28.

The matrix plot in Fig. 6.21 shows 5 high-support rules that are formed by the combination of antecedents 5, 10, 15, 19, and 21 with the consequent of 1. These rules are given in Table 6.29.

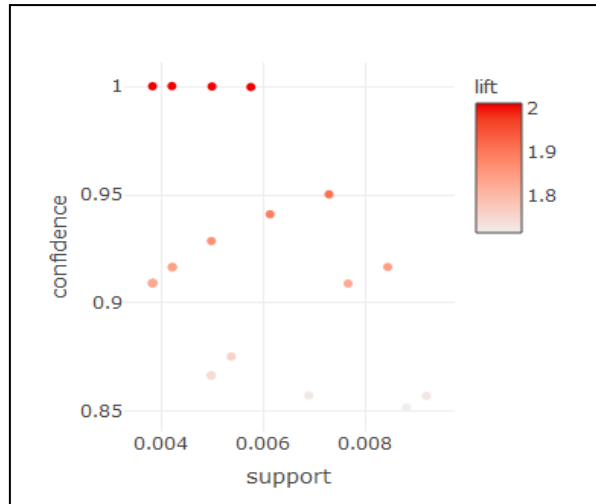


Fig. 6.20: Scatter plot for Experiment 1 – Recommended teacher training

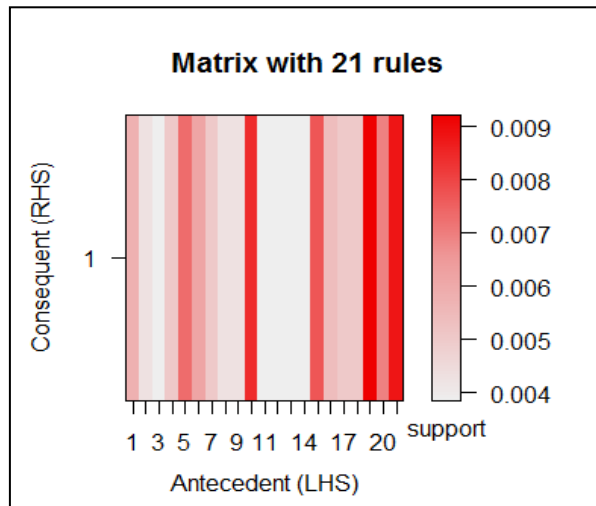


Fig. 6.21: Matrix plot for Experiment 1 – Recommended teacher training

Table 6.28: Interesting rules from scatter plot – Recommended teacher training

Rule	Support	Confidence	Lift
[1] {Teacher.professional.qualification=PTC/JV/CT, Recommended.teacher.training=Training on use of support material} ⇒ {Teacher.result=Bad}	0.0057	100%	2
[2] {Teacher.designation=PST, Recommended.teacher.training=Training on use of support material, Class.ratio=All Boys} ⇒ {Teacher.result=Bad}	0.0042	100%	2
[3] {Recommended.teacher.training=Training on use of support material, Class.size=16-35, Percentage.of.present.students.in.class=Good} ⇒ {Teacher.result=Bad}	0.0038	100%	2
[4] {Teacher.designation=PST, Recommended.teacher.training=Training on use of support material, Percentage.of.present.students.in.class=Good} ⇒ {Teacher.result=Bad}	0.0049	100%	2

Table 6.29: Interesting rules from matrix plot – Recommended teacher training

Rule	Support	Confidence	Lift
[5] {Teacher.designation=PST, Recommended.teacher.training=Training on use of support material} ⇒ {Teacher.result=Bad}	0.0073	95%	1.91
[10] {Recommended.teacher.training=Training on lesson planning, Class.size=More than 35, Class.ratio=All Girls} ⇒ {Teacher.result=Bad}	0.0084	91.6%	1.84
[15] {Teacher.workload.per.week=More than 40 hours, Recommended.teacher.training=Training on lesson planning, Class.size=More than 35} ⇒ {Teacher.result=Bad}	0.0076	90.9%	1.83
[19] {Recommended.teacher.training=Training on use of support material} ⇒ {Teacher.result=Bad}	0.0092	85.7%	1.72
[21] {Teacher.training.duration=1 month, Recommended.teacher.training=Training on lesson planning, Class.ratio=All Girls} ⇒ {Teacher.result=Bad}	0.0088	85.2%	1.71

6.3.2.4. Discussion. The associations obtained in this experiment are only the training areas that are recommended to teachers with bad outcomes. No rules were obtained for good and average teacher outcomes with recommended teacher training as one of the antecedents. The interesting rules listed in Table 6.28 and Table 6.29 shows 9 clusters of teachers with varying support, confidence, and lift values.

Rule numbers 1-4 in Table 6.28 identified overlapping clusters of teachers who achieved bad outcomes and were recommended trainings on use of support material. The teachers also had a professional qualification of PTC/JV/CT, according to rule number 1. By examining other teacher characteristics for this teacher cluster comprising of 15 teachers, it was found that these teachers had low academic qualification of Grade 10 or High School Diploma.

From rule number 2 (Table 6.28), 11 teachers were identified who were recommended training on use of support material and were associated with bad outcomes. These teachers held a designation of PST and taught in all-boys classrooms.

The 10 teachers who were recommended training on use of support material and taught in medium-sized classrooms containing 16-35 students with good students' attendance were associated to bad outcomes, according to rule number 3 (Table 6.28).

Rule number 4 (Table 6.28) identified 13 teachers who were recommended training on use of support material, held a designation of PST, had a good attendance of students, and were related to bad outcomes.

Rules 5, 10, 15, 19, and 21 in Table 6.29 lists high-support rules. Rule number 5 forming a cluster of 20 teachers suggests that low teacher designation of PST alongwith training on use of support material are related to bad teacher outcome.

The 24 teachers who were recommended a training on lesson planning and taught in large-sized, all-girls classrooms with more than 35 students were associated with bad results, according to rule number 10 (Table 6.29). Rule number 15 (Table 6.29) identified 22 teachers who had a workload of more than 40 hours per week. These teachers taught in large-sized classrooms (more than 35 students), were recommended a training on lesson planning and were related to bad results. Upon examining other teacher characteristics of the group of teachers formed by rule numbers 10 and 15, it was found that many of these teachers had high academic qualifications (Masters or Bachelors).

According to rule number 19 (Table 6.29), the 28 teachers who were recommended training on use of support material achieved bad outcomes 85.7% of the time.

Finally, the 27 teachers who were recommended a training on lesson planning, taught in all-girls classrooms, and had attended trainings for upto a month's duration were linked to bad teacher outcomes, as per rule number 21 (Table 6.29).

Table 6.30: Itemsets from parallel coordinates plot – Recommended teacher training

Position	Itemsets	Comments
1	Teacher training duration = 1 month	Average teacher training duration is 28 days
	Recommended teacher training = Training in subject of Maths	A training area recommended to teacher out of 34 training areas
2	Recommended teacher training = Training on lesson planning	A training area recommended to teacher out of 34 training areas
	Level of teacher identified by peers = 3	Peer ranking of teacher where 4 is the worst
	Class size = More than 35	Average class size is 21 students
3	Class ratio = All girls	Ratio: {All boys, All girls, More boys, More girls, Balanced}
	Recommended teacher training = Training on use of support material	A training area recommended to teacher out of 34 training areas
	Class size = 16-35	Average class size is 21 students
	Teacher workload per week = More than 40 hours	Average teacher workload is 38 hours
4	Teacher result = Bad	Teacher result being Bad among {Good, Average, Bad}

6.3.3. Conclusion. The experiment conducted for teacher training analysis did not give any rules for good and average teacher outcomes. Table 6.31 shows a summary of quality measures obtained in teacher training analysis. The obtained rules had low support values indicating their low coverage and representation of the educational dataset. The lift values are high and indicate dependence between the associations of recommended teacher training area to bad teacher outcome.

Table 6.31: Summary of rules for teacher training analysis

Teacher training analysis				
No. of rules	21			
Quality measure	Min	Max	Mean	StDev
Support	0.0038	0.0092	0.0056	0.0017
Lift	1.7	2.0	1.849	0.096
Confidence	0.85	1.0	0.919	0.047

6.4. Summary

Analysis on the micro-level educational data was performed in this chapter with respect to the rule templates discussed in Chapter 4. The educational units operational at this level were the teacher and subjects. For each analysis, the respective distribution of the variable of interest was observed with a description of the education baskets that were to be used in the following experiments. Association rules were analyzed based on the scatter, matrix, and parallel coordinates plots.

The teacher outcome analysis was performed to study the teacher characteristics that distinguish the good teachers from the bad ones. Similarly, the subjects outcome analysis for the subjects of English, GK, Religion, NL, Maths, Science, and SS were conducted to determine the classroom and teacher characteristic variables that pertain to good and bad subject results. Finally, an attempt was made to study the recommended trainings to teachers with respect to their obtained results.

Chapter 7 . Meso-level Educational Analytics

In this chapter, details of the meso analysis performed at the education data are provided. The meso analysis of the educational data is performed at the School level. In other words, data across different schools are mined. The school data contains school environment, process, and outcome attributes. The micro-level data was also integrated at this level to mine rules at school level with grade and teacher variables as well. This data was cleansed and transformed using the techniques detailed in Chapter 5. The characteristic and environment variables of School and L2M were merged for this analysis and Apriori algorithm was run on the merged dataset to obtain rules across all schools. Finally, the rules were analysed based on objective interestingness measures, and interesting associations of school outcome with other variables were studied.

7.1. School Outcome Analysis

The school outcome is the School result variable in the CPD framework. This variable is based on the average marks obtained by students in all the subjects of grades 3 - 5. So, the overall school result is expressed by means of students' performance across all grades and subjects. The school outcome had noisy data with values above 100, and missing values which were incorrectly coded as 0. These values were changed to missing values and the values in the range of 1-100 were translated to the following levels using the domain knowledge of grading system in the developing country.

$$School\ result = \begin{cases} Bad, & \text{if value} \leq 40 \\ Average, & \text{if value} > 40 \text{ and } \leq 65 \\ Good, & \text{if value} > 65 \end{cases} \quad (16)$$

The distribution of school outcome in the educational data is given below in Fig. 7.1.

7.1.1. Educational basket for school outcome analysis. The meso-level education basket formulated to study the relationship of schools' outcomes with different variables consists of the items given in Table 7.1. This basket has a total of 1,391 transactions and each transaction corresponds to one school.

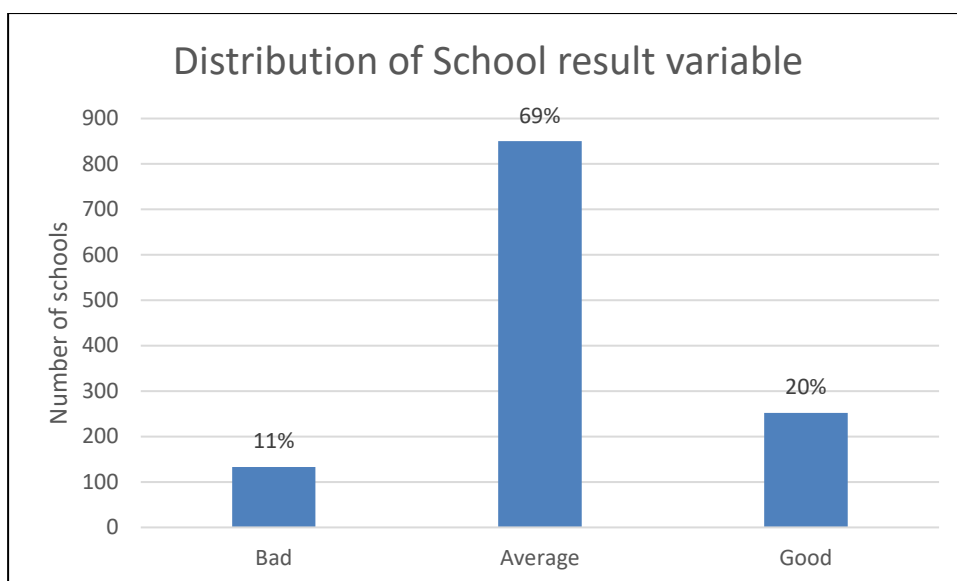


Fig. 7.1: Distribution of schools with different outcomes

Table 7.1: Items in education basket for school outcome analysis

Items	Potential Values
Distance of school from the cluster centre	{Near, Midway, Far}
Level	{Primary, Elementary, High}
Number of teachers	{1-3, 4-10, More than 10}
Number of ESTs	{1-5, 6-10, More than 10}
Total enrolled students	{1-150, 151-300, More than 300}
Percentage of present students in school	{Good, Bad}
Attendance of teaching staff	{Good, Bad}
School mentoring implementation	{Good, Average, Bad}
School assessment implementation	{Good, Bad}
Cooperation of HT	{Good, Average, Bad}
L2M designation	{EST, PST, SESE, SSE, SST}
L2M academic qualification	{Bachelors, Masters}
L2M professional qualification	{B.Ed., M.Ed.}
L2M age	{25-35, 36-45, 46-55}
L2M experience	{Less than 5 years, 5-10 years}
L2M training duration	{Upto 2 weeks, 1 month, More than a month}
L2M attendance	{Good, Bad}
Mode teacher designation	{DYHM, ESE, EST, HM, PST, SESE, SSE}
Average teacher workload per week	{1-20 hours, 21-40 hours, More than 40 hours}
Mode teacher academic qualification	{Grade 10, High School Diploma, Bachelors, Masters}
Mode teacher professional qualification	{PTC/JV/CT, B.Ed., M.Ed., Other}
Average teacher age	{Upto 30 years, 31-50 years, More than 50 years}
Average teacher experience	{Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
Average teacher training duration	{Upto 2 weeks, 1 month, More than a month}
Median level of teacher identified by peers	{1, 2, 3, 4}; 1 being the best
Average teacher result	{Good, Average, Bad}
Average class ratio	{All Boys, All Girls, Balanced, More boys, More girls}
Average class size	{1-15, 16-35, More than 35}
School result	{Good, Average, Bad}

7.1.2. Experiment 1 – School outcome = Good.

7.1.2.1. Objective. The objective of this experiment is to study the school, L2M, teacher, and grade characteristics that led to a school's outcome being Good.

7.1.2.2. Constraints. For this experiment, the following constraints were used to mine the association rules:

- Support=0.005, value established using trial-and-error with a starting support of 0.1.
- Confidence=85%, value used to achieve high-confidence rules that are true most of the time.
- Absolute minimum support count = 6 rows out of 1,391 (from Equation 4)
- Rule template: {LHS: All items present in Table 7.1 except School result, RHS: School result = Good}

A template-based approach was used to mine the rules with the RHS of the resulting rules pre-specified to obtain the rules with itemsets that were related to the school outcome being 'Good'.

7.1.2.3. Results. A total set of 10 rules was generated from the Apriori algorithm using the above listed constraints. The obtained rules had good lift values ranging from 5.52 – 4.91 indicating strong associations between the antecedents and consequent. The scatter plot is shown in Fig. 7.2, matrix plot in Fig. 7.3, and the itemsets from the parallel coordinates plot are shown in Table 7.5. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot in Fig. 7.2 highlights 6 high-lift overlapping points at the top with a confidence of 100% and support values ranging from 0.005 – 0.0072. These rules are given in Table 7.2.

The matrix plot highlights the high support rules formed by antecedents 1, 4, 6, 7, 8, 9, and 10. The rules formed by the combination of these antecedents and consequent of good school result are listed in Table 7.3.

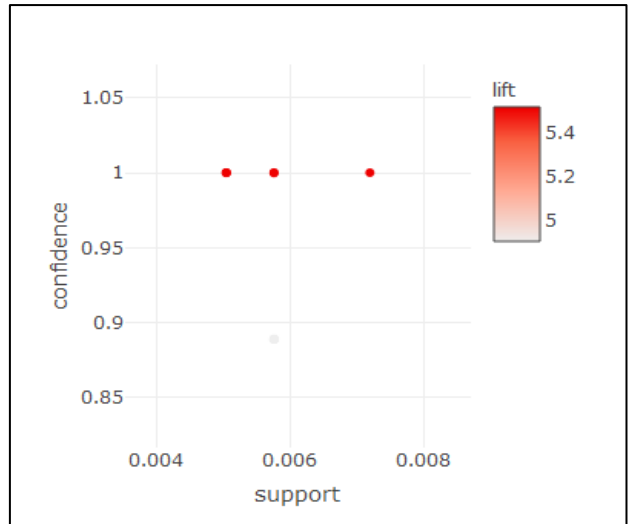


Fig. 7.2: Scatter plot for Experiment 1 – School outcome = Good

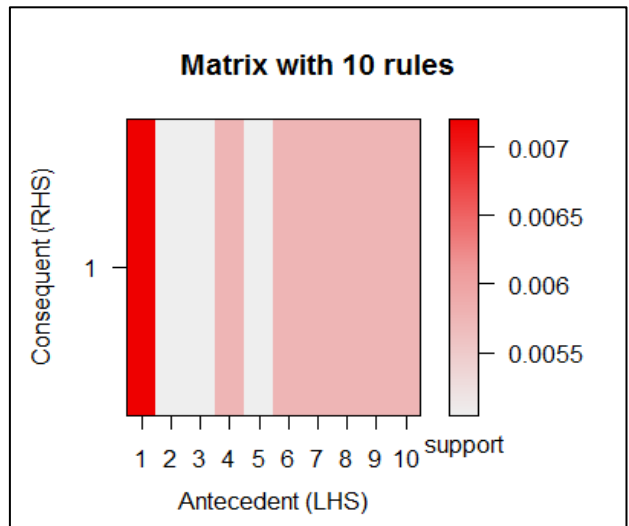


Fig. 7.3: Matrix plot for Experiment 1 – School outcome = Good

Table 7.2-A: Interesting rules from scatter plot – School outcome = Good

Rule	Support	Confidence	Lift
[1] {L2M.training.duration=More than a month, Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0072	100%	5.52
[2] {Mentoring.Implementation..=Good, L2M.academic.qualification=Bachelors, Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.005	100%	5.52
[3] {Mentoring.Implementation..=Good, L2M.professional.qualification=B.Ed., Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.005	100%	5.52
[4] {L2M.training.duration=More than a month, Median.teacher.peer.ranking=2.5, Mode.teacher.academic.qualification=High School Diploma} ⇒ {School.result=Good}	0.0057	100%	5.52

Table 7.2-B: Interesting rules from scatter plot – School outcome = Good

Rule	Support	Confidence	Lift
[5] {Mentoring.Implementation..=Good, L2M.academic.qualification=Bachelors, Average.teacher.workload=21-40 hours} ⇒ {School.result=Good}	0.005	100%	5.52
[6] {L2M.academic.qualification=Bachelors, L2M.training.duration=More than a month, Average.teacher.workload=21-40 hours} ⇒ {School.result=Good}	0.0057	100%	5.52

Table 7.3: Interesting rules from matrix plot – School outcome = Good

Rule	Support	Confidence	Lift
[7] {Distance.from.the.cluster.centre=Near, L2M.academic.qualification=Bachelors, Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0057	88.8%	4.91
[8] {Distance.from.the.cluster.centre=Near, L2M.professional.qualification=B.Ed., Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0057	88.8%	4.91
[9] {Mentoring.Implementation..=Good, L2M.designation=EST, L2M.academic.qualification=Bachelors} ⇒ {School.result=Good}	0.0057	88.8%	4.91
[10] {Mentoring.Implementation..=Good, L2M.academic.qualification=Bachelors, L2M.age=46-55} ⇒ {School.result=Good}	0.0057	88.8%	4.91

7.1.2.4. Discussion. The obtained rules for this experiment identified overlapping subsets of 11 schools from rule numbers 1, 2, 3, 5, and 6 (Table 7.2) and rule numbers 7, 8, 9, and 10 (Table 7.3). According to these rules, schools that were near to the cluster centre and had good status of mentoring completion were linked to good school results. The L2Ms in these schools had a designation of EST (Elementary School Teacher), an academic qualification of Bachelors, and a professional qualification of B.Ed. In addition, longer L2M training duration of more than a month, moderate teacher workload of 21-40 hours per week, and a low aggregated teacher peer ranking value of 4 were related to good school result, according to these rules. The further exploration of these rules is useful for LOM who can point out the clusters and schools where this rule is applicable, since these rules point towards a violation of fidelity because the schools attained good results despite the low teacher ranking. This observation could mean that the L2Ms might be making up teacher peer ranking values in one particular geographical cluster, since most of these rules were obtained from schools that were mentored by the same L2M and belonged to the same cluster. Some details of the schools where these rules were applicable are given in Table 7.4.

Table 7.4: Schools belonging to rule number 7 – School outcome = Good

S. No.	Distance from the cluster centre	Level	Total enrolled students	Percentage of present students	Number of teachers	L2M academic qualification	L2M experience
1	Near	<NA>	151-300	Good	4-10	Bachelors	5-10 years
2	Near	Primary	1-150	Good	4-10	Bachelors	5-10 years
3	Near	High	More than 300	Good	More than 10	Bachelors	5-10 years
4	Near	Primary	151-300	Good	4-10	Bachelors	5-10 years
5	Near	High	More than 300	Good	More than 10	Bachelors	5-10 years
6	Near	Primary	151-300	Good	4-10	Bachelors	5-10 years
7	Near	Primary	151-300	Bad	4-10	Bachelors	5-10 years
8	Near	Primary	151-300	Good	4-10	Bachelors	5-10 years
9	Near	Elementary	More than 300	Bad	More than 10	Masters	5-10 years
10	Near	Elementary	151-300	Bad	More than 10	Masters	5-10 years
11	Near	Primary	1-150	Bad	4-10	Bachelors	5-10 years

Rule number 4 (Table 7.2) identified a group of 8 schools from different geographical clusters in which the longer L2M training duration of more than a month was associated to good school result. In addition, most of the teachers working in these schools had a low academic qualification of High School Diploma that is 12 years of education and they acquired an average of 2.5 peer ranking on a scale of 1-4 (1 being the best).

Table 7.5: Itemsets from parallel coordinates plot – School outcome = Good

Position	Itemsets	Comments
1	Distance from the cluster centre = Near	“Near” is 0-5 km distance of school from the cluster centre. Average distance of schools from the cluster centre is 6.4 km.
2	L2M academic qualification = Bachelors	Relatively lower academic qualification. All levels = {Bachelors, Masters}
	Average level of teacher identified by peers = 4	Peer ranking of teacher where 4 is the worst
	School mentoring completion = Good	Levels = {Bad, Good}
3	L2M age = 46-55 years	Average L2M age is 42 years
	L2M professional qualification = B.Ed.	Relatively lower professional qualification. All levels = {B.Ed., M.Ed.}
	L2M designation = EST	L2M designation of Elementary School Teacher (mid-career level)
4	School result = Good	School result being Good among {Good, Average, Bad}

7.1.3. Experiment 2 – School outcome = Bad.

7.1.3.1. Objective. The objective of this experiment was to study the school, L2M, teacher, and grade characteristics that were related to bad school outcome.

7.1.3.2. Constraints. For this experiment, the following constraints were used to mine the association rules:

- Support=0.004, value established using trial-and-error with a starting support of 0.1.
- Confidence=85%, value used in all experiments to achieve high-confidence rules that are true most of the time.
- Absolute minimum support count = 5 rows out of 1,391 (from Equation 4)
- Rule template: {LHS: All items present in Table 7.1 except School result, RHS: School result = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the school outcome to be Bad.

7.1.3.3. Results. A total set of 14 rules was generated from the Apriori algorithm using the above listed constraints. A support value of 0.004 was used in this experiment which covered at least 5 rows of the total 1,391 transactions. The obtained rules have high lift (10.46 – 8.96) indicating strong dependence between antecedents and consequent. The scatter plot is shown in Fig. 7.4, matrix plot in Fig. 7.5, and itemsets from the parallel coordinates plot in Table 7.9. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot in Fig. 7.4 for the rule set generated for bad school outcome shows three overlapping high-lift rules represented by dark red points at the top with a support of more than 0.004 – 0.005 and a confidence of 100%. These rules are given in Table 7.6.

The matrix plot in Fig. 7.5 shows the antecedents 4 and 5 forming the highest support rules followed by antecedents 1, 6, 7, 8, 9, and 10. The rules formed by the combination of these antecedents with bad school outcome are presented in Table 7.7.

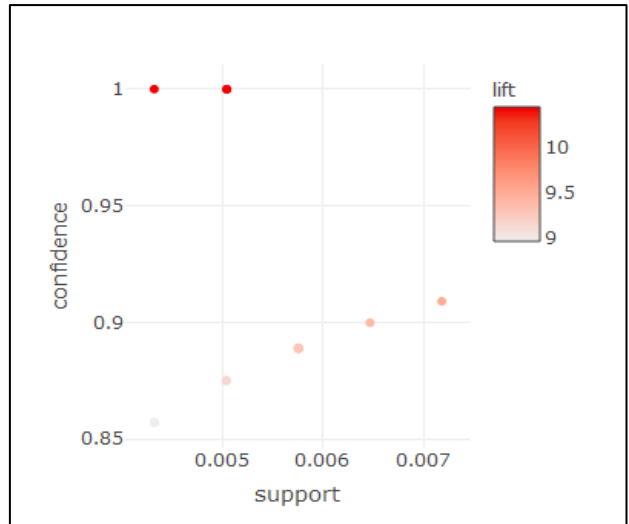


Fig. 7.4: Scatter plot for Experiment 2 – School outcome = Bad

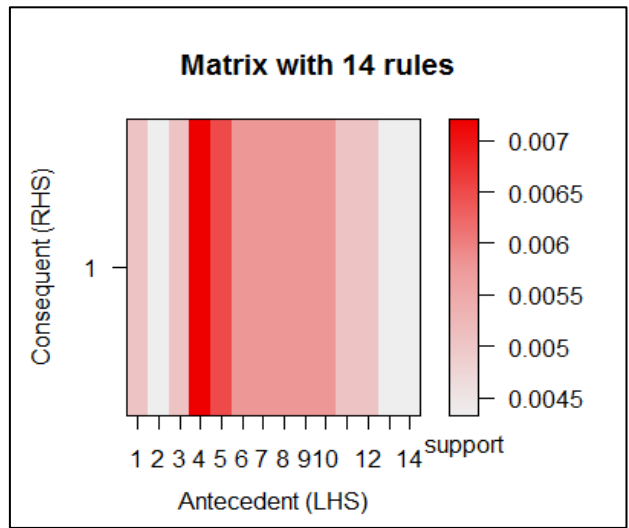


Fig. 7.5: Matrix plot for experiment 2 – School outcome = Bad

Table 7.6: Interesting rules from scatter plot – School outcome = Bad

Rule	Support	Confidence	Lift
[1] {Percentage.of.physically.present.students=Bad, Assessment.implementation=Good, Cooperation.of.HT=Bad} ⇒ {School.result=Bad}	0.005	100%	10.46
[2] {Mentoring.Implementation.=Bad, Cooperation.of.HT=Bad, L2M.experience=5-10 years} ⇒ {School.result=Bad}	0.0043	100%	10.46
[3] {L2M.experience=Less than 5 years, Average.teacher.result=Bad, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.005	100%	10.46

Table 7.7: Interesting rules from matrix plot – School outcome = Bad

Rule	Support	Confidence	Lift
[4] {L2M.age=25-35, L2M.training.duration=1 month, Average.teacher.result=Bad} ⇒ {School.result=Bad}	0.0072	90.9%	9.51
[5] {Cooperation.of.HT=Good, L2M.experience=Less than 5 years, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0065	90%	9.41
[6] {L2M.age=25-35, L2M.training.duration=1 month, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0057	88.8%	9.29
[7] {L2M.experience=Less than 5 years, L2M.training.duration=1 month, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0057	88.8%	9.29
[8] {L2M.professional.qualification=B.Ed., L2M.training.duration=1 month, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0057	88.8%	9.29
[9] {Percentage.of.physically.present.students=Bad, L2M.experience=Less than 5 years, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0057	88.8%	9.29
[10] {Attendance.of.Staff=Good, L2M.experience=Less than 5 years, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.0057	88.8%	9.29

7.1.3.4. Discussion. The interesting rules obtained from the scatter and matrix plots are formed from overlapping groups of schools from different geographical clusters. Rule number 1 (Table 7.6) identified a group of 7 schools with bad outcomes. According to this rule, the attendance of students being low, bad cooperation of headteacher, and good assessment completion indicator was linked with bad outcomes.

Rule number 2 (Table 7.6) identified a group of 6 schools in which the bad status of mentoring completion, bad cooperation of headteachers, and L2Ms having an experience of 5-10 years were associated to bad school results.

Rule number 3 (Table 7.6) and the high-support rules 4, 5, 6, 7, 8, 9, and 10 (Table 7.7) picked out schools that mostly belonged to the same cluster and were mentored by the same L2M who had an age of 25-35 years, an experience of less than 5 years, had a professional qualification of B.Ed., and was trained for upto a month's duration. According to these rules, bad teacher result, an aggregated teachers' peer ranking value of 2, bad attendance of students, good cooperation of headteacher, and good attendance of teaching staff were related to bad school results. Upon further examination of these rules, it was found that many of these schools were Primary-level

schools with an average teacher workload of more than 40 hours per week. The schools for the grades 1 to 5 are primary-level schools in the developing country. Some of the details of these schools are given in Table 7.8.

Table 7.8: Schools belonging to rule number 3 – School outcome = Bad

S. No.	Distance from the cluster centre	Level	Total enrolled students	Percentage of present students	Attendance of Staff	Average teacher workload per week
1	Near	High	151-300	Bad	Good	More than 40 hours
2	Near	<NA>	1-150	Bad	Good	More than 40 hours
3	Near	Primary	151-300	Bad	Bad	More than 40 hours
4	Midway	Primary	1-150	Bad	Good	More than 40 hours
5	Midway	Primary	1-150	Bad	Good	More than 40 hours
6	Near	Primary	1-150	Bad	Good	More than 40 hours
7	Near	Primary	1-150	Bad	Good	More than 40 hours

Finally, from the itemsets obtained from the parallel coordinates plot in Table 7.9, it can be seen that the schools with bad outcomes were also Primary-level schools and had other features already picked out by scatter and matrix plots.

Table 7.9: Itemsets from parallel coordinates plot – School outcome = Bad

Position	Itemsets	Comments
1	L2M training duration = 1 month	Average L2M training duration is 22 days.
	Average level of teacher identified by peers = 2	Peer ranking of teacher where 4 is the worst.
2	L2M experience = Less than 5 years	Average L2M experience is 6 years.
	L2M age = 25-35 years	Average L2M age is 42 years.
3	Total enrolled students = 1-150	Small-sized schools. An average of 159 students are enrolled in schools.
	Level = Primary	Levels = {Primary, Elementary, High}
	L2M professional qualification = B.Ed.	Relatively lower professional qualification. All levels = {B.Ed., M.Ed.}
	Percentage of present students = Bad	Levels = {Good, Bad}
	Average teacher result = Bad	Aggregated teacher result being Bad among {Good, Average, Bad}
	Cooperation of HT = Good	Levels = {Good, Average, Bad}
4	Attendance of Staff = Good	Levels = {Good, Bad}
	School result = Bad	School result being Bad among {Good, Average, Bad}

7.1.4. What distinguished Good from Bad schools? The comparison of different school outcomes from experiments 1 and 2 show that the schools with good and bad results were distinguished by their process variables like status of mentoring completion and L2M variables. The schools with good outcomes showed progress in the achievement of various mentoring areas and vice-versa. The schools with bad outcomes were assigned L2Ms that were trained for a shorter duration (up to 1 month)

and were younger (aged 25-35). While, the L2Ms in good schools were relatively older (aged 46-55 years), held a designation of EST (Elementary School Teacher), and had received trainings for more than a month. In addition, a bad attendance of students, bad cooperation of headteacher, and bad teacher results were also observed in the schools with bad outcomes. The teachers in good schools had an average teacher peer ranking value of 4, as opposed to bad schools which had an average teacher peer ranking value of 2, which is an anomaly that was observed in these experiments. Finally, the school characteristics that was observed for schools achieving good results was that these schools were near (upto 5 km distance) to the cluster centre. And, the schools with bad outcomes were mostly Primary level schools for Grades 1 to 5 having an average class size of 1-15 students.

7.1.5. Conclusion. The experiment conducted for school outcome analysis with good outcome gave rules with relatively better support values. These rules had good lift values indicating strong dependence between the antecedents and consequent of good school result. The obtained rules for bad school outcome had low support but very high lift values. However, by the exploration of these rules, it was found that these rules were applicable to schools that were mostly mentored by the same L2M and thus, they were representative of good or bad school result in only particular geographical clusters. Table 7.10 shows a summary of these experiments:

Table 7.10: Summary of rules for school outcome analysis

School outcome	Good				Bad			
No. of rules	10				14			
Quality measure	Min	Max	Mean	StDev	Min	Max	Mean	StDev
Support	0.005	0.0072	0.0057	0.0006	0.0043	0.0072	0.0054	0.0008
Lift	4.91	10.0	5.27	0.3	8.96	10.46	9.5	0.52
Confidence	0.889	1.0	0.95	0.054	0.86	1.0	0.91	0.049

7.2. School Size Analysis

The school size analysis was performed by studying the total number of enrolled students in a school variable. This variable had missing values incorrectly coded as 0, and the remaining values in the range of 1-930 corresponding to number of students in the school were translated to the following levels after discretization:

$$School\ Size = \begin{cases} Small, & \text{if value } \geq 1 \text{ and } \leq 150 \\ Medium, & \text{if value } \geq 151 \text{ and } \leq 300 \\ Large, & \text{if value } > 300 \end{cases} \quad (17)$$

The bar plot of the school size in Fig. 7.6 indicates that 60% of the schools are small-sized with a total of 1-150 students, while 29% of the schools are medium-sized with 151-300 enrolled students, and 11% schools are large-sized with more than 300 enrolled students.

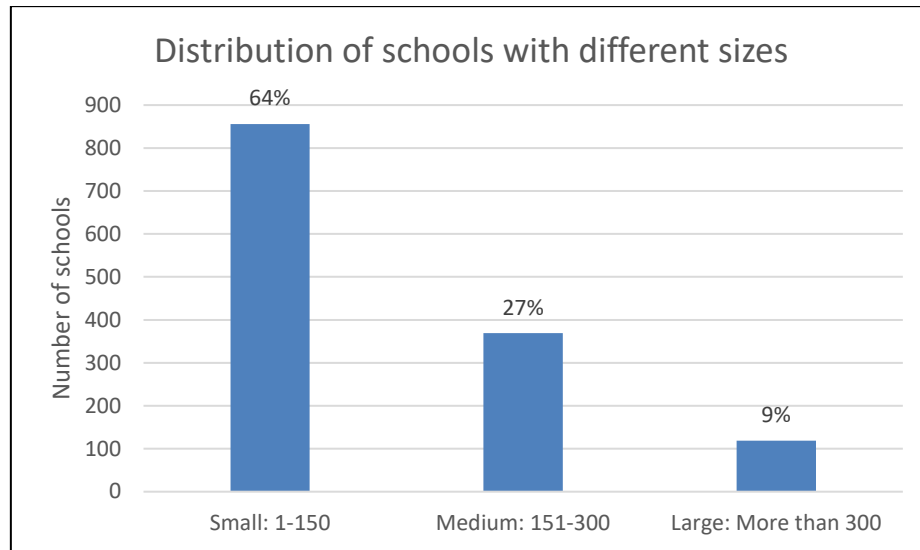


Fig. 7.6: Distribution of schools with different sizes

7.2.1. Educational basket for school size analysis. The items that were used to formulate the education basket to study school characteristics with respect to different sizes are the same as the ones given in Table 7.1 for school outcome analysis with 1,391 transactions.

7.2.2. Experiment 1 – School size.

7.2.2.1. Objective. The objective of this experiment was to study the relationship of school size variable alongwith other school and teacher characteristics on the school outcome.

7.2.2.2. Constraints. For this experiment, the following constraints were used to mine the association rules:

- Support = 0.003, value established using trial-and-error with a starting support of 0.1.

- Confidence = 85%, value used to achieve high-confidence rules that are true most of the time.
- Absolute minimum support = 4 rows out of 1,391 (from Equation 4)
- Rule template: {LHS: School size AND all items from the basket in Table 7.1 except for School result, RHS: School result = “Bad” or “Average” or “Good”}

The items in the LHS and RHS were pre-specified to include School size from the cluster centre and School result respectively.

7.2.2.3. Results. A set of 299 rules was generated when Apriori algorithm was run with the above listed constraints. Some high lift rules were obtained in this experiment showing strong association between the antecedents and consequent. The scatter plot for this experiment is shown in Fig. 7.7, matrix plot in Fig. 7.8, and itemsets from the parallel coordinates plot are given in Table 7.15. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot in Fig. 7.7 highlights 6 interesting rules by dark red points with confidence ranging from 85 – 100% and support ranging from 0.003-0.004. These rules are given in Table 7.11.

The matrix plot highlights only one high-support rule with respect to consequent 293 by the dark red line. This rule is shown in Table 7.12.

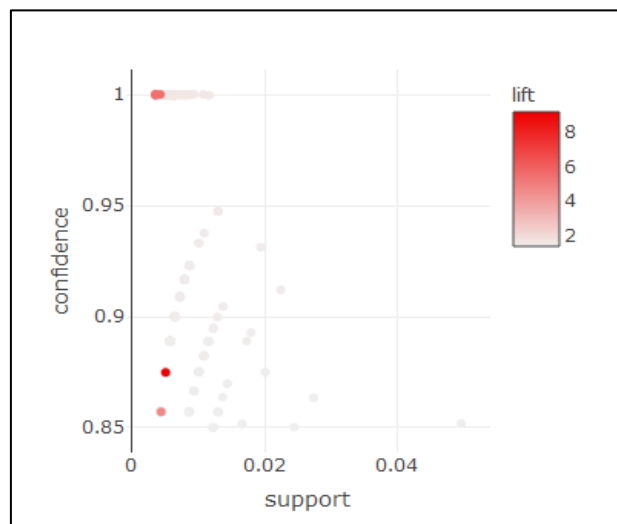


Fig. 7.7: Scatter plot for Experiment 1 – School size analysis

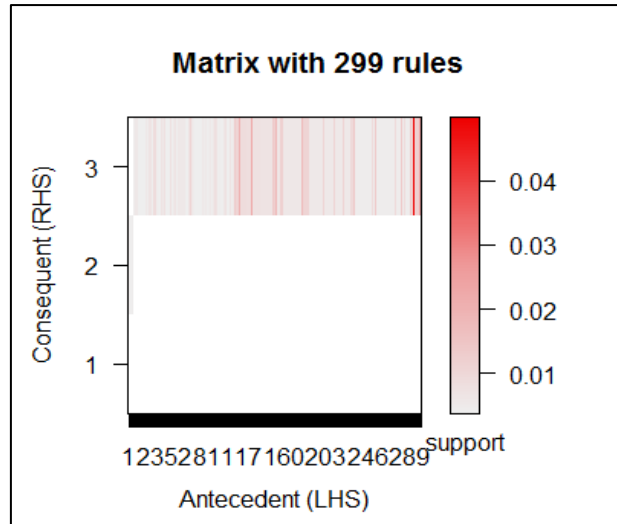


Fig. 7.8: Matrix plot for Experiment 1 – School size analysis

Table 7.11: Interesting rules from scatter plot – School size analysis

Rule	Support	Confidence	Lift
[1] {Total.enrolled.students=1-150, L2M.experience=Less than 5 years, Median.teacher.peer.ranking=2} ⇒ {School.result=Bad}	0.005	87.5%	9.15
[2] {Total.enrolled.students=151-300, L2M.academic.qualification=Bachelors, Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0036	100%	5.52
[3] {Total.enrolled.students=151-300, L2M.training.duration=More than a month, Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0043	100%	5.52
[4] {Total.enrolled.students=151-300, L2M.professional.qualification=B.Ed., Median.teacher.peer.ranking=4} ⇒ {School.result=Good}	0.0036	100%	5.52
[5] {Total.enrolled.students=More than 300, Average.teacher.result=Average, Median.teacher.peer.ranking=2.5} ⇒ {School.result=Good}	0.0036	100%	5.52
[6] {Total.enrolled.students=More than 300, Percentage.of.physically.present.students=Good, Average.teacher.result=Average} ⇒ {School.result=Good}	0.0043	85.7%	4.73

Table 7.12: Interesting rule from matrix plot – School size analysis

Rule	Support	Confidence	Lift
[293] {Total.enrolled.students=151-300, L2M.professional.qualification=M.Ed., Average.teacher.result=Bad} ⇒ {School.result=Average}	0.05	85.2%	1.39

7.2.2.4. Discussion. The purpose of this experiment was to identify if there was a difference in the mentoring and assessment processes, assignment of L2Ms at schools which are different sized, and thus producing good or bad overall results. Rule number

1 (Table 7.11) is the highest lift rule which is applicable on 8 schools. This rule suggests that small-sized schools with less-experienced L2Ms (less than 5 years), and a good aggregated teachers' peer ranking value of 2 are associated to bad school results. The further exploration of this rule led to the observation that most of these schools were co-education Primary level schools with bad students' attendance and teachers having an average workload of more than 4 hours per week, as shown in Table 7.13.

Table 7.13: Schools belonging to rule number 1 – School size = Small

S. No.	Distance from the cluster centre	Level	Percentage of present students	Mentoring completion	Average teacher result	Average teacher workload per week	Average class ratio
1	Midway	Primary	<NA>	Bad	Average	<NA>	More girls
2	Near	<NA>	Bad	Average	Bad	More than 40 hours	More girls
3	Near	Primary	Bad	Bad	<NA>	More than 40 hours	More boys
4	Near	Primary	Bad	Bad	Good	More than 40 hours	More girls
5	Midway	Primary	Bad	Bad	Bad	More than 40 hours	More boys
6	Midway	Primary	Bad	Average	Bad	More than 40 hours	More girls
7	Near	Primary	Bad	Bad	Bad	More than 40 hours	Balanced
8	Near	Primary	Bad	Bad	Bad	More than 40 hours	More boys

Rule numbers 2, 3, and 4 (Table 7.11) are applicable on a set of 6 schools in which good results were linked to the schools being medium-sized with a total of 151-300 enrolled students and being mentored by L2Ms trained for longer durations of more than a month and having an academic qualification of Bachelors, a professional qualification of B.Ed., and attaining an aggregated teacher peer ranking value of 4.

The rules 5 and 6 (Table 7.11) are applicable on a group of 7 schools achieving good results. According to these rules, the large-sized schools with good attendance of students, aggregated teacher results of Average, and aggregated teacher peer ranking of 2.5 (most of the teachers ranked 2 or 3) are related to school outcome being Good. Upon further exploration of this rule, it was found that most of these schools were Elementary or High level, all-girls or all-boys schools. Table 7.14 shows some details of these large-sized schools with more than 300 enrolled students.

Table 7.14: Schools belonging to rule numbers 5 and 6 – School size = Large

S. No.	Distance from the cluster centre	Level	Number of teachers	Mode teacher academic qualification	Average teacher experience	Average teacher workload per week	Average class ratio
1	Near	Primary	More than 10	High School Diploma	16-30 years	More than 40 hours	All girls
2	Near	Elementary	More than 10	Grade 10	16-30 years	21-40 hours	All boys
3	Near	High	4-10	High School Diploma	More than 30 years	21-40 hours	All boys
4	Midway	High	More than 10	High School Diploma	16-30 years	21-40 hours	All girls
5	Midway	Elementary	More than 10	Masters	16-30 years	21-40 hours	All boys
6	Midway	Primary	4-10	Grade 10	16-30 years	More than 40 hours	All boys
7	Midway	Elementary	4-10	Grade 10	16-30 years	21-40 hours	More girls

Finally, rule number 293 (Table 7.12), the highest support rule refers to a group of 81 medium-sized schools in which the professional qualification of L2Ms being M.Ed. (high qualification), and aggregated teacher result being Bad are associated to school outcomes being Average.

The itemsets obtained in Table 7.15 are taken from the top 15 high-lift rules. These itemsets suggest that small and medium-sized schools, and schools that are far from the cluster centre, and L2Ms having low professional qualification of CT (Certificate of teaching) and high designation of ESE (Elementary School Educator) relate to school results being Average.

Table 7.15: Itemsets from parallel coordinates plot – School size analysis

Position	Itemsets	Comments
1	Total enrolled students = 1-150	Small-sized schools. An average of 159 students are enrolled in schools.
2	Total enrolled students = 151-300	Medium-sized schools. An average of 159 students are enrolled in schools.
	Distance from the cluster centre = Far	“Far” is more than 10 km distance of school from the cluster centre. Average distance of schools from the cluster centre is 6.4 km.
3	L2M designation = ESE	High designation of Elementary School Educator.
	L2M professional qualification = CT	Low professional qualification of Certificate of teaching.
4	School result = Average	School result being Average among {Good, Average, Bad}

7.2.3. Comparison of different school sizes. A comparison of rules obtained from the above experiment shows that medium and large-sized schools tend to achieve good results, while small-sized schools achieved bad results. The rules also highlighted some features of different sized schools like small sized schools were assigned less experienced L2Ms with an experience of less than 5 years. On the other hand, medium-

sized schools were assigned well-trained L2Ms with longer training durations of more than a month and they had a professional qualification of B.Ed. Finally, a good attendance of students was observed in large-sized schools and these schools mostly had a combined teacher result of average.

7.2.4. Conclusion. The experiment conducted for school size analysis gave rules for different school outcomes. Table 7.16 shows a summary of quality measures obtained in school size analysis. The obtained rules had low support and good lift and confidence values indicating strong association between the antecedents and consequent of the rules. The lift values have a high standard deviation since only 6 rules had high lift deeming them as interesting, while the remaining rules had low lift values.

Table 7.16: Summary of rules for school size analysis

School size analysis				
No. of rules	299			
Quality measure	Min	Max	Mean	StDev
Support	0.0036	0.049	0.0066	0.0043
Lift	1.39	9.15	1.6	0.67
Confidence	0.85	1.0	0.92	0.061

7.3. Summary

In this chapter, meso-level analysis was performed on education baskets with respect to the rule templates discussed in Chapter 4. The educational units that were of interest at meso-level were the school and L2M. The micro-level data of teachers was also integrated at this level and outcome and size analysis of schools were performed. The rules obtained from the conducted experiments were examined for interestingness by means of scatter, matrix, and parallel coordinates plots.

The school outcome analysis was performed to determine which itemsets particularly distinguished good schools from bad ones. Similarly, with the help of school size analysis, characteristics of teachers and schools that can be observed in schools with different sizes were determined.

Chapter 8 . Macro-level Educational Analytics

In this chapter, details of the macro analysis performed at the education data are provided. The macro analysis of the educational data was performed at cluster level which means that data across different clusters was mined. The cluster data had no cluster-characteristic attributes, but only process and outcome attributes. The micro and meso levels data was also integrated at this level to mine rules containing school, L2M, teacher and grade quality variables, alongwith cluster variables. This data was cleansed and transformed using the techniques detailed in Chapter 5, and the Apriori algorithm was run on this dataset to obtain rules that were further analysed for interestingness by means of various visualization plots.

8.1. Cluster Outcome Analysis

The cluster outcome is the Cluster result variable in the CPD framework. This variable is based on the average marks obtained by students in different subjects in grades 3-5 in each cluster. The students' marks were used to express cluster's overall performance. The cluster outcome had values in the range of 0-10, which were cleansed by removing the values equal to 0 and discretizing the remaining values to the following levels:

$$Cluster\ result = \begin{cases} Bad, & \text{if value} \leq 4 \\ Average, & \text{if value} > 4 \text{ and } \leq 6.5 \\ Good, & \text{if value} > 6.5 \end{cases} \quad (18)$$

The distribution of clusters with respect to different cluster outcomes is shown in Fig. 8.1. Most of the clusters (57%) had Average or Good (34%) result, while only 9% of the clusters had achieved Bad results.

8.1.1. Educational basket for cluster outcome analysis. The education basket used to perform experiments for cluster outcome analysis consisted of the variables shown in Table 8.1. This basket has a total of 59 transactions with each transaction corresponding to record of one cluster.

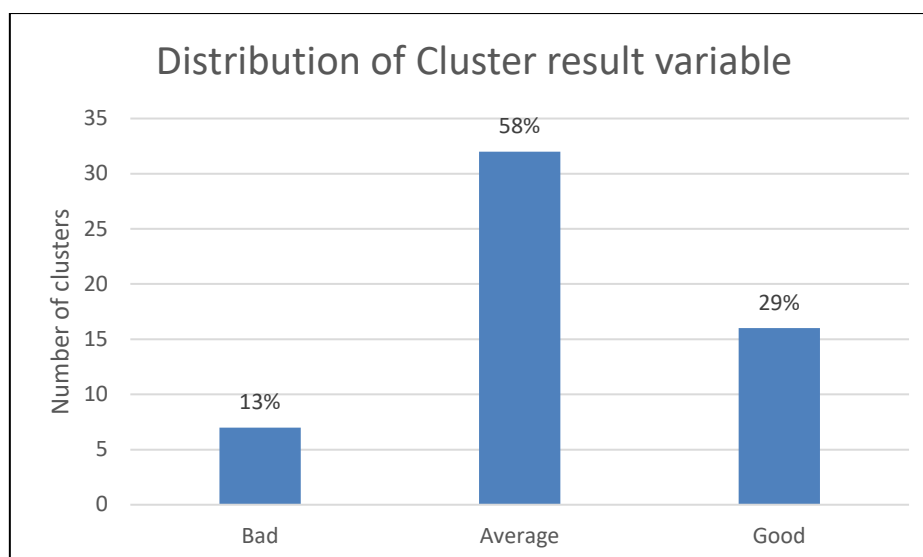


Fig. 8.1: Distribution of clusters with different outcomes

Table 8.1: Items in education basket for cluster outcome analysis

Items	Potential Values
Cluster mentoring completion	{Bad, Good}
Cluster assessment completion	{Bad, Good}
Cluster test report issuance	{Bad, Good}
Cluster pre-mentoring status	{Bad, Good}
Average distance of school from the cluster centre	{Near, Midway, Far}
Mode level	{Primary, Elementary, High}
Average number of teachers	{1-3, 4-10, More than 10}
Average total enrolled students	{1-150, 151-300, More than 300}
Average percentage of present students	{Good, Bad}
Average attendance of teaching staff	{Good, Bad}
Average cooperation of HT	{Good, Average, Bad}
Average school result	{Good, Average, Bad}
Mode L2M designation	{EST, PST, SESE, SSE, SST}
Mode L2M academic qualification	{Bachelors, Masters}
Mode L2M professional qualification	{B.Ed., M.Ed.}
Average L2M age	{25-35, 36-45, 46-55}
Average L2M experience	{Less than 5 years, 5-10 years}
Average L2M training duration	{Upto 2 weeks, 1 month, More than a month}
Average L2M attendance	{Good, Bad}
Mode teacher designation	{DYHM, ESE, EST, HM, PST, SESE, SSE}
Average teacher workload per week	{1-20 hours, 21-40 hours, More than 40 hours}
Mode teacher academic qualification	{Grade 10, High School Diploma, Bachelors, Masters}
Mode teacher professional qualification	{PTC/JV/CT, B.Ed., M.Ed., Other}
Average teacher age	{Upto 30 years, 31-50 years, More than 50 years}
Average teacher experience	{Upto 5 years, 6-15 years, 16-30 years, More than 30 years}
Average teacher training duration	{Upto 2 weeks, 1 month, More than a month}
Median level of teacher identified by peers	{1, 2, 3, 4}; 1 being the best
Average teacher result	{Good, Average, Bad}
Mode class ratio	{All Boys, All Girls, Balanced, More boys, More girls}
Mode class size	{1-15, 16-35, More than 35}
Cluster result	{Good, Average, Bad}

8.1.2. Experiment 1 – Cluster outcome = Good.

8.1.2.1. Objective. The objective of this experiment was to study the cluster, school, L2M, and teacher characteristics that were related to good cluster outcomes.

8.1.2.2. Constraints. For this experiment, the following constraints were used to mine the association rules:

- Support=0.09, value established using trial-and-error with a starting support of 0.1.
- Confidence=85%, value used to achieve high-confidence rules that are true most of the time.
- Absolute minimum support count = 5 rows out of 59 (Equation 4)
- Rule template: {LHS: All items present in Table 8.1 except Cluster result, RHS: Cluster result = Good}

A template-based approach was used to mine the rules with the RHS of the resulting rules pre-specified to obtain the rules with itemsets that relate to the cluster outcome to be Good.

8.1.2.3. Results. A set of 2 rules was generated from the Apriori algorithm using the above listed constraints. The obtained rules have good support values, high confidence, and high lift deeming these rules as interesting. The lift values range from 3.7 – 3.2 implying strong association between the antecedents and consequent in the obtained rules. The scatter plot is shown in Fig. 8.2, matrix plot in Fig. 8.3, and the itemsets from the parallel coordinates plot in Table 8.4. These plots were generated on the resulting rule set to find the most interesting rules.

The scatter plot in Fig. 8.2 of the resulting rules extracted from macro-level education data shows the most interesting rule with a confidence of 100% and a lift of 3.7 at the top of the plot. This rule has a support of 0.12. The matrix plot in Fig. 8.3 highlights the same rule as the scatter plot (Fig. 8.2). The resulting rules are given in Table 8.2.

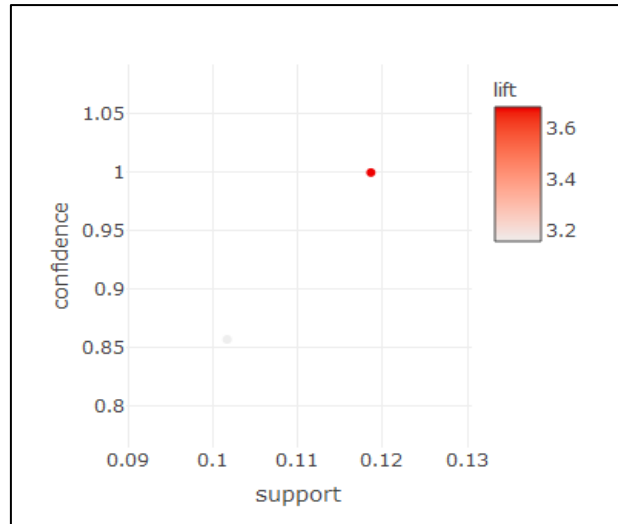


Fig. 8.2: Scatter plot for Experiment 1 – Cluster outcome = Good

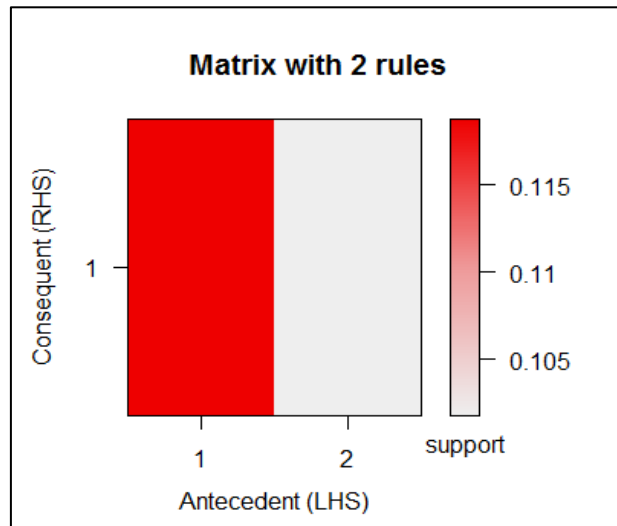


Fig. 8.3: Matrix plot for Experiment 1 – Cluster outcome = Good

Table 8.2: Resulting rules – Cluster outcome = Good

Rule	Support	Confidence	Lift
[1] {Average.schools.result=Good} ⇒ {Cluster.result=Good}	0.12	100%	3.69
[2] {Average.percentage.of.present.students.in.schools=Good, Average.attendance.of.staff.in.schools=Good, Average.number.of.ESTs.in.schools=6-10} ⇒ {Cluster.result=Good}	0.1	85.7%	3.16

8.1.2.4. Discussion. Rule numbers 1 (Table 8.2) identified a group of 7 clusters in which good result was found to be associated with the aggregated results of schools being good. Certain attributes of this group of clusters are listed in Table 8.3, which indicates that some of these clusters had bad status of process variables. This

information is useful for the LOM who may want to investigate as to why these clusters have bad assessment completion indicator when they are achieving good results.

Table 8.3: Clusters belonging to rule number 1 – Cluster outcome = Good

S. No.	Mentoring completion	Assessment completion	Test report issuance	Pre-mentoring status	Average L2M attendance	Average L2M experience
1	Good	Good	Bad	Bad	Good	5-10 years
2	Good	Good	Good	Good	Good	5-10 years
3	Bad	Bad	Good	Bad	Good	5-10 years
4	Good	Bad	Bad	<NA>	Good	5-10 years
5	Bad	Good	<NA>	Bad	Good	<NA>
6	Good	Good	Good	Good	Good	5-10 years
7	Good	Good	Good	Bad	Good	5-10 years

The rule number 2 (Table 8.2) represents 7 cluster groups in which the good attendance of students in schools, good attendance of teaching staff in schools, and the presence of 6-10 Elementary School Teachers in the schools led to a good overall cluster result.

Table 8.4: Itemsets from parallel coordinates plot – Cluster outcome = Good

Position	Itemsets	Comments
1	Average percentage of present students in schools = Good	Levels = {Bad, Good}
2	Average attendance of staff in schools = Good	Levels = {Bad, Good}
3	Average number of ESTs in schools = 6-10	An average of 6 ESTs worked in schools
4	Cluster result = Good	Cluster result being Good among {Good, Average, Bad}

8.1.3. Experiment 2 – Cluster outcome = Bad.

8.1.3.1. Objective. The objective of this experiment was to study the cluster, school, L2M, and teacher characteristics that were related to bad cluster outcomes.

8.1.3.2. Constraints. For this experiment, the following constraints were used to mine the association rules:

- Support=0.06, value established using trial-and-error with a starting support of 0.1.
- Confidence=85%, value used to achieve high-confidence rules that are true most of the time.
- Absolute minimum support count = 3 rows out of 59 (Equation 4)

- Rule template: {LHS: All items present in Table 8.1 except Cluster result, RHS: Cluster result = Bad}

The RHS of the resulting rules was pre-specified to obtain the rules with itemsets that relate to the cluster outcome to be Bad.

8.1.3.3. Results. A set of 5 rules was generated from the Apriori algorithm using the above listed constraints. The minimum support was relatively lower than experiment 1 because of the distribution of clusters with different outcomes in Fig. 8.1 according to which only 13% (7) clusters had bad results. The resulting rules have the same lift value of 8.43 and 100% confidence for all rules showing strong relationship among the antecedents and consequent. The scatter plot is shown in Fig. 8.4, matrix plot in Fig. 8.5, and interesting itemsets from the parallel coordinates plot are given in Table 8.6. These plots were generated on the resulting rule set to find the most interesting rules.

Fig. 8.4 shows a scatter plot for the rule set generated for bad cluster outcome. Since all rules have the same support, lift and confidence values, they are represented as blue points on the plot with at a support of 0.068. The matrix plot in Fig. 8.5 shows equally dark bars since all the rules have the same support. All the resulting rules are given in Table 8.5.

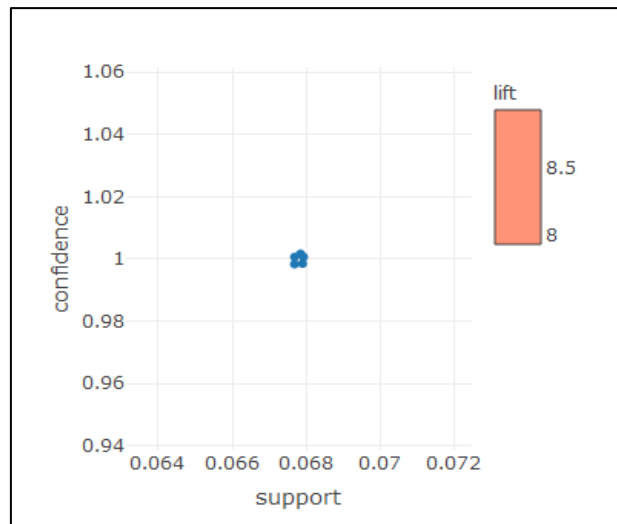


Fig. 8.4: Scatter plot for Experiment 2 – Cluster outcome = Bad

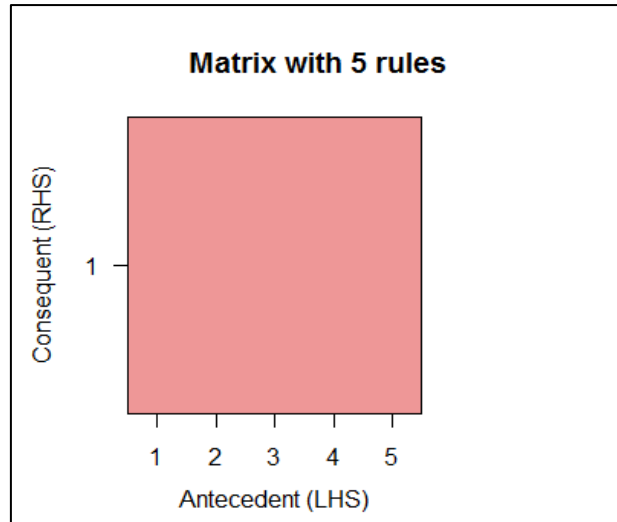


Fig. 8.5: Matrix plot for Experiment 2 – Cluster outcome = Bad

Table 8.5: Resulting rules – Cluster outcome = Bad

Rule	Support	Confidence	Lift
[1] {Average.L2M.experience=Less than 5 years, Median.teacher.peer.ranking=2} ⇒ {Cluster.result=Bad}	0.068	100%	8.43
[2] {Average.L2M.age=25-35, Mode.L2M.professional.qualification=B.Ed., Average.teacher.age=31-50 years} ⇒ {Cluster.result=Bad}	0.068	100%	8.43
[3] {Average.L2M.age=25-35, Median.teacher.peer.ranking=2, Mode.class.size.in.cluster=1-15} ⇒ {Cluster.result=Bad}	0.068	100%	8.43
[4] {Average.L2M.age=25-35, Average.number.of.teachers.in.schools=4-10, Median.teacher.peer.ranking=2} ⇒ {Cluster.result=Bad}	0.068	100%	8.43
[5] {C.Test.report.issuance=Good, C.Pre.mentoring.status=Bad, Average.L2M.age=25-35} ⇒ {Cluster.result=Bad}	0.068	100%	8.43

8.1.3.4. Discussion. Rule numbers 1-5 in Table 8.5 brought out the same group of 4 clusters with bad results, hence the same lift and support. According to these rules, the clusters with bad results had less experienced L2Ms with an experience of less than 5 years, an age of 25-35 years, and a professional qualification of B.Ed. In addition, an average teacher peer ranking value of 2 and an aggregated teachers' age of 31-50 years were observed in these clusters. The teachers' peer ranking of 2 is anomalous and could mean that the L2Ms might not be conducting proper assessment of teachers with respect to their colleagues. The schools in these clusters had an average of 4-10 teachers and an average class size of 1-15 students. The process variables of test report issuance which signifies the issuance of cluster reports to the district centre and pre-mentoring

status which shows the education status of cluster before the mentoring process began were seen to be Good and Bad respectively in the clusters achieving bad results.

Table 8.6: Itemsets from parallel coordinates plot – Cluster outcome = Bad

Position	Itemsets	Comments
1	Average teacher peer ranking = 2	Average teacher ranking of 2 as identified by peers on a scale of 1-4 (4 being the worst)
	Average L2M age = 25-35 years	Average L2M age is 42 years
2	Average L2M experience = Less than 5 years	Average L2M experience is 6 years.
	Mode L2M professional qualification = B.Ed.	Relatively lower professional qualification. All levels = {B.Ed., M.Ed.}
	Cluster test report issuance = Good	The status of cluster report that is to be issued to the district centre being Good among {Bad, Good}
3	Average teacher age = 31-50 years	Average teacher age is 46 years
	Mode class size in cluster = 1-15	Average class size is 21 students
	Average number of teachers in schools = 4-10	An average of 5 teachers are present at a school
	Cluster pre-mentoring status = Bad	The status of education in cluster before the mentoring process began being Bad among {Bad, Good}
4	Cluster result = Bad	Cluster result being Bad among {Good, Average, Bad}

8.1.4. What distinguished Good from Bad clusters? A comparison of rules obtained from experiments 1 and 2 at the macro level of analytics did not show any variables which differed in both experiments. The rules obtained for good cluster outcome suggest that good school result, good percentage of present students, good attendance of staff, and the presence of 6-10 Elementary School teachers in the schools relate to good overall cluster result. Alternatively, the bad cluster results were linked to L2M characteristics of L2Ms being younger (25-35 years), less experienced (less than 5 years), and having a relatively lower qualification of B.Ed. The schools mostly had 4-10 total teachers and had a class size of 1-15 students in clusters that secured bad results. In addition, middle-aged teachers (31-50 years) were employed in these clusters with an average peer ranking of 2 on a scale of 1-4 (1 being the best). Again, the peer ranking of 2 in clusters with bad results point towards a violation of fidelity, since bad overall results were obtained in the clusters despite the good peer ranking value for teachers.

8.1.5. Conclusion. The experiments conducted for cluster outcome analysis gave rules with good quality measures of support, confidence and lift indicating strong association between the antecedents and consequent. Table 8.7 shows a summary of these experiments. The obtained rules had relatively higher support values than the

values obtained in micro- and meso-level experiments and they were applicable to 4-7 clusters out of the entire 59 clusters.

Table 8.7: Summary of rules for cluster outcome analysis

Cluster outcome	Good				Bad			
No. of rules	2				5			
Quality measure	Min	Max	Mean	StDev	Min	Max	Mean	StDev
Support	0.102	0.12	0.11	0.0085	0.068	0.068	0.068	0
Lift	3.16	3.68	3.42	0.26	8.43	8.43	8.43	0
Confidence	0.857	1.0	0.93	0.07	1.0	1.0	1.0	0

8.2. Summary

The macro-level analysis on education data was performed in this chapter. The unit of cluster was operational at this level. The educational basket that was used at the macro-level was described. Then, cluster outcome analysis was performed with the help of rule template discussed in Chapter 4. The aim of this analysis was to distinguish clusters having good results from the ones who got bad results. The resulting patterns were examined for interestingness by means of different visualization plots.

Chapter 9 . Discussion, Conclusion, and Future Directions

The purpose of this thesis was to determine if the association mining using Apriori algorithm an appropriate technique to obtain a description of large-scale educational data characterized by high volume, high velocity, low variety and veracity. The goal was to discover relationships between different variables of sparse educational data and to answer educational questions at different levels of analytics for policy-making and decision-making at regional or state level.

A novel approach to relationship mining was used by creating education baskets or education data subsets at each level of analysis. The active variables pertaining to the learning outcomes at each level of analysis were found and explored. This approach of modelling educational data at different levels worked by integrating lower-level variables into higher-level analysis. By using this methodology of rule discovery on multi-tier educational data, interesting associations were produced especially at the meso and macro levels of analyses. The grade and teacher characteristics that were found to be associated with higher-level educational units like clusters' and schools' outcomes were interesting, since these rules inform about what should be controlled or investigated at the micro-level with the help of which good results are obtained on large-scale regional or state level.

The macro-level results depicted a strong relation with the school result and attendance of students and teaching staff in schools' variables. Certain L2M features and process variables were also observed to be impacting the cluster results at this level. Class size and school characteristics variables like number of teachers and ESTs that relate to the overall cluster result were also brought out at this level.

The meso-level results indicated strong relationship between the schools' process variables and their results. In addition, the qualification, training, age, experience, and designation of L2Ms played an important role in the data collection and monitoring process and impacted the meso-level results because L2Ms were responsible for the training and mentoring of teachers. Finally, the micro-level indicator of teachers' workload, academic qualification, and peer ranking also associated with the school outcome at the meso-level. The features of schools with different sizes were

also studied and the relationship of various school outcomes with different students' enrolment was examined.

The micro-level results were obtained by conducting experiments for teachers and subject outcomes. The class size, distribution of boys and girls in a classroom, attendance of students in a class, and teacher characteristics like age, experience, designation, qualifications, trainings etc. were studied that can be controlled or monitored to achieve good results. The various training areas that were assigned to different kinds of teachers were examined with respect to different teacher results.

9.1. Quality of Rules

From the experiments performed at micro, meso, and macro levels of analyses, groups of teachers, schools, and geographical clusters were obtained to which a certain rule was applicable. Table 9.1 presents a summary of rules obtained at various levels by the achieved coverage, and mean support, lift, and confidence values.

At the micro level, groups of 5-32 teachers out of the total 2,613 teachers were obtained, hence the low support values. However, as we moved up the level of analysis, a better coverage of schools was obtained of 7-81 schools out of the total 1,391 schools. Finally, at the macro level, the support further increased and better coverage of 4-7 clusters out of the total 59 clusters was obtained. This improvement in the objective goodness of rules with respect to support at the higher levels was achievable due to the aggregation of variables from the lower level. At the higher levels, the sparse data became more aggregated and thus gave rules with better coverage and higher support values.

The high confidence of 85% was used for all the experiments, due to which the mean confidence is high for all the experiments. Finally, higher lift was observed in some of the experiments like Teacher outcome = Good, Subject English outcome = Bad, School outcome = Bad, and Cluster outcome = Bad, deeming the resulting rules as strong association rules.

Table 9.1: Summary of rules obtained at each level of analysis

Level of Analytics	Number of Transactions	Template	Experiment Number	Number of rules	Coverage	Support				Lift				Confidence			
						Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev
Micro	2613	Teacher outcome	1	1	5 teachers	0.0019	0.0019	0.0019	-	14.76	14.76	14.76	-	1.00	1.00	1.00	-
			2	18	20-32 teachers	0.0073	0.011	0.008738	0.00099	1.71	1.92	1.79	0.061	0.852	0.95	0.89	0.03
		Subject outcome	1	3	6 teachers	0.002296	0.002296	0.002296	0	6.153	6.153	6.153	0	0.857	0.857	0.8571	0
			2	1	6 teachers	0.002296	0.002296	0.002296	-	15.45	15.45	15.45	-	0.857	0.857	0.8571	-
			3	1	6 teachers	0.002296	0.002296	0.002296	-	5.989	5.989	5.989	-	0.857	0.857	0.8571	-
			4	0	-	-	-	-	-	-	-	-	-	-	-	-	-
			5	6	6-9 teachers	0.002296	0.0031	0.002424	0.00028	4.93	5.76	5.10	0.3	0.857	1.00	0.8862	0.052
			6	0	-	-	-	-	-	-	-	-	-	-	-	-	-
			7	4	6-8 teachers	0.002296	0.00268	0.002488	0.000019	5.83	5.95	5.893	0.061	0.857	0.875	0.8661	0.0089
		8	0	-	-	-	-	-	-	-	-	-	-	-	-	-	
		9	1	9 teachers	0.003062	0.003062	0.003062	-	5.836	5.836	5.836	-	0.889	0.889	0.889	-	
Teacher training	1	21	15-28 teachers	0.0038	0.0092	0.0056	0.0017	1.7	2.0	1.849	0.096	0.85	1.0	0.919	0.047		
Meso	1391	School outcome	1	10	8-11 schools	0.005	0.0072	0.0057	0.0006	4.91	10.0	5.27	0.3	0.889	1.0	0.95	0.054
			2	14	7-11 schools	0.0043	0.0072	0.0054	0.0008	8.96	10.46	9.5	0.52	0.86	1.0	0.91	0.049
		School size	1	299	8-81 schools	0.0036	0.049	0.0066	0.0043	1.39	9.15	1.6	0.67	0.85	1.0	0.92	0.061
Macro	59	Cluster outcome	1	2	7 clusters	0.102	0.12	0.11	0.0085	3.16	3.68	3.42	0.26	0.857	1.0	0.93	0.07
			2	5	4 clusters	0.068	0.068	0.068	0	8.43	8.43	8.43	0	1.0	1.0	1.0	0

9.2. Limitations

This research had various limitations in terms of availability of data, analysis of obtained rules, and application of constraints and rule templates. These limitations are detailed below:

9.2.1. Educational data characteristic limitations. The available educational data lacked certain variables at the micro-level like students' personal information, financial status and occupation of families of enrolled students, and teachers' gender. The availability of these variables would help to gain further insight into the micro-level analysis. For example, the teachers' gender relation to class ratio, students' grades relation to their financial backgrounds can be studied. Furthermore, no cluster environment variables were available due to which macro-level analysis were performed by studying only the process variables of the cluster with respect to the cluster outcome.

The educational dataset was missing a large percentage of values for most of the variables. The missing values especially in outcome variables impacted the quality of association rules and gave very low support rules which were not representative of the entire dataset but were applicable to small groups of teachers, schools, and clusters.

Certain variables had values such that they were translated to highly-skewed distributions due to which most of the rules were dominated by the frequently occurring values of these variables. For example, the distance of schools from the cluster centre being near, school mentoring completion being good, and cooperation of head teacher being good were the variables which did not add to the interestingness of the rules since they were occurring in all kinds of relations.

9.2.2. Formulation of rule templates. The rule templates should be developed by the educational stakeholders who are familiar with the nature of the data and can ask real questions that are helpful in gaining insight into various educational problems. These rule templates will help to achieve better and more interesting rules.

9.2.3. Application of constraints. A trial-and-error method were used to determine the minimum support that was used to mine the resulting association rules. However, in most of the experiments this approach resulted in very low support (mostly

less than 1% of transactions) which resulted in rules that were not applicable to most of the transactions in the dataset.

Furthermore, due to the applied constraints on the left- or right-hand-side of the rules, any particularly interesting rule was not introduced (rules that are interesting but could not be thought of by the stakeholders). For example, the strange but interesting rule of *Diaper* \Rightarrow *Beer* that is famous in the supermarket transactions. However, constraining the antecedent or consequent helped in the study of rules in a structured manner which would have been difficult otherwise due to the large number of resulting rules.

9.2.4. Analysis of rules. In this research, the rules that were analysed were only those that were highlighted by the visualization plots. This approach can result in missing interesting rules which have slightly lower support or lift values. In addition, the interesting rules were sought only with respect to the support, confidence, and lift parameters and did not employ other evaluation parameters.

9.3. Future Directions

In future, the meso and micro analysis can be performed for selected clusters and schools, and patterns can be extracted that are useful for each cluster and/or school only, and can also be analysed together to answer questions like:

- How does the performance of the students in one school relates to other students in different schools of the same geographical location?
- How does the performance of the students in one location relates to other students in a different location?
- How do the teacher variables in one school and location relate to teacher variables in different schools and locations?
- Which teachers in a school need training for different subjects?
- Are the learning outcomes achieved in all locations? If yes, then to what extent?
- How do the learning outcomes achieved in one location vary as opposed to other locations?

The longitudinal analysis of educational data for each of the months can be performed to track the achievement and deviation (if any) of learning outcomes in all

clusters. This approach to rule mining can be scaled by using optimized versions of Apriori or other algorithms like FP-Growth [15] which offer less computational complexity and better run-time performance. The time taken by the algorithm to find the largest set of 299 rules was 0.31 seconds (worst-case) on the CPU Intel Pentium N3540 with a RAM of 4 GB and 64-bit Windows 8.1 operating system, therefore, with current size of data Apriori algorithm worked fine but for larger datasets more efficient algorithms can be used. In addition, a reporting or discussion tool to state the most interesting rules to educational experts can be developed, as suggested in [65]. Finally, a subjective analysis of the obtained rules can be performed by domain experts in the field of education in developing countries to convert the obtained associations into actionable educational reforms.

References

- [1] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann, 2011, pp. 3-40.
- [2] R. S. Baker. "Educational data mining: An advance for intelligent systems in education." *IEEE Intelligent Systems*, vol. 29, no. 3, pp. 78–82, May 2014.
- [3] S. B. Shum. (2012, Nov.). "Learning analytics." *UNESCO IITE*. [Online]. Available: <http://iite.unesco.org/publications/3214711/> [Oct.17, 2017].
- [4] A. J. Bowers. "Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis." *Practical Assessment Research and Evaluation*, vol. 15, pp. 1–18, May 2010.
- [5] M. Tobin, P. Lietz, D. Nugroho, R. Vivekanandan, and T. Nyamkhuu. (2015, Sep.). "Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region." *Melbourne: ACER and Bangkok: UNESCO*. [Online]. Available: <https://research.acer.edu.au> [Nov. 10, 2017].
- [6] "The four V's of big data." Internet: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>, [Apr. 30, 2018].
- [7] D. Laney. "3D data management: Controlling data volume, velocity and variety." *META Group Research Note*, vol. 6, no. 70, Feb. 2001.
- [8] J. Brownlee. "Supervised and unsupervised machine learning algorithms." Internet: <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>, Mar. 16, 2016 [Nov. 12, 2017].
- [9] P.-N. Tan, M. Steinbach, and V. Kumar. "Data mining cluster analysis: basic concepts and algorithms," in *Introduction to Data Mining*, 1st ed. Ed. Boston: Pearson Addison Wesley, 2005, pp. 487-568.
- [10] J. T. Chi, E. C. Chi, and R. G. Baraniuk. "k-POD: A method for k-means clustering of missing data." *The American Statistician*, vol. 70, no. 1, pp. 91–99, Jan. 2016.
- [11] L. Ertöz, M. Steinbach, and V. Kumar. "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *Proceedings of the 2003 SIAM International Conference on Data Mining*, 2003, pp. 47–58.
- [12] D. Napoleon and S. Pavalakodi. "A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set." *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, Jan. 2011.

- [13] Z. Huang. “Extensions to the k-means algorithm for clustering large data sets with categorical values.” *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.
- [14] A. Peña-Ayala. *Educational Data Mining: Applications and Trends*. Springer International Publishing, 2013, pp. 9-53.
- [15] P.-N. Tan, M. Steinbach, and V. Kumar. “Association analysis: basic concepts and algorithms,” in *Introduction to Data Mining*, 1st ed. Ed. Boston: Pearson Addison Wesley, 2005, pp. 327-414.
- [16] N. R. Mabroukeh and C. I. Ezeife. “A taxonomy of sequential pattern mining algorithms.” *ACM Computing Surveys (CSUR)*, vol. 43, no. 1, p. 3, Nov. 2010.
- [17] A. Hero. “Correlation mining in massive data.” Internet: <https://pdfs.semanticscholar.org>, Apr. 11, 2013 [Jan. 15, 2018].
- [18] R. Scheines. “Causal data mining.” Internet: <https://www.slideshare.net/Tommy96/causal-data-mining>, May 10, 2010 [Jan. 15, 2018].
- [19] C. C. Aggarwal, C. Procopiuc, and P. S. Yu. “Finding localized associations in market basket data.” *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 51–62, Jan. 2002.
- [20] A. Loraine Charlet and A. Kumar. “Market basket analysis for a supermarket based on frequent itemset mining.” *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 257, Sep. 2012.
- [21] J. Dongre, G. L. Prajapati, and S. Tokekar. “The role of Apriori algorithm for finding the association rules in data mining,” in *2014 IEEE International Conference on International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 2014, pp. 657–660.
- [22] C. Zhang and J. Ruan. “A modified Apriori algorithm with its application in instituting cross-selling strategies of the retail industry,” in *2009 IEEE International Conference on Electronic Commerce and Business Intelligence ECBI*, 2009, pp. 515–518.
- [23] N. Elgendy and A. Elragal. “Big data analytics: a literature review paper,” in *Springer Industrial Conference on Data Mining*, 2014, pp. 214–227.
- [24] S. Borkar and K. Rajeswari. “Predicting students academic performance using education data mining.” *International Journal of Computer Science and Mobile Computing (IJCSMC)*, vol. 2, no. 7, pp. 273–279, Jul. 2013.
- [25] Y. Kurniawan and E. Halim. “Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application

- and benefits),” in *2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 2013, pp. 98–103.
- [26] B. Şen. “Predicting and analyzing secondary education placement-test scores: A data mining approach.” *Expert Systems with Applications*, vol. 39, no. 10, pp. 9468–9476, Aug. 2012.
- [27] Q. Yang and Y. Hu. “Application of improved Apriori algorithm on educational information,” in *2011 IEEE 5th International Conference on Genetic and Evolutionary Computing (ICGEC)*, 2011, pp. 330–332.
- [28] T. Pradhan, S. R. Mishra, and V. K. Jain. “An effective way to achieve excellence in research based learning using association rules,” in *2014 IEEE International Conference on Data Mining and Intelligent Computing (ICDMIC)*, 2014, pp. 1–4.
- [29] P. Bhandari, C. Withana, A. Alsadoon, and A. Elchouemi. “Enhanced Apriori algorithm model in course suggestion system,” in *2015 IEEE International Conference and Workshop on Computing and Communication (IEMCON)*, 2015, pp. 1-5.
- [30] C. Romero and S. Ventura. “Educational data mining: A review of the state of the art.” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
- [31] C. Romero, S. Ventura, A. Zafra, and P. De Bra. “Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems.” *Computers & Education*, vol. 53, no. 3, pp. 828–840, Nov. 2009.
- [32] M. Yu-Ling. “The research on courses correlation based on the intelligent education frame.” *DEStech Transactions on Social Science, Education and Human Science*, 2016.
- [33] F. Liu, S. Zhang, J. Ge, F. Lu, and J. Zou. “Agricultural major courses recommendation using Apriori algorithm applied in China Open University system,” in *2016 IEEE 9th International Symposium on Computational Intelligence and Design (ISCID)*, 2016, pp. 442–446.
- [34] B. Mahatthanachai, H. Ninsonti, and N. Tantranont. “A study of factors influency student dropout rate using data mining.” *The Golden Teak: Humanity and Social Science*, vol. 22, no. 4, pp. 46–55, 2016.
- [35] H. Maniar and S. Khanna. “A predictive student performance analytics scheme using auto-adjust Apriori algorithm.” *International Journal of Computer Applications*, vol. 157, no. 6, Jan. 2017.
- [36] H. Zhang, T. Huang, Z. Lv, S. Liu, and Z. Zhou. “MCRS: A course recommendation system for MOOCs.” *Multimedia Tools and Applications*, vol. 77, no.6, pp. 7051-7069, Mar. 2018.

- [37] M. Zaharia *et al.* “Apache Spark: A unified engine for big data processing.” *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, Oct. 2016.
- [38] T. White. *Hadoop: The Definitive Guide*. O’Reilly Media, Inc., 2009.
- [39] A. Merceron and K. Yacef. “Interestingness measures for associations rules in educational data.” *Educational Data Mining*, vol. 8, pp. 57–66, Jun. 2008.
- [40] A. Buldu and K. Üçgün. “Data mining application on students’ data.” *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 5251–5259, Jan. 2010.
- [41] S. Parack, Z. Zahid, and F. Merchant. “Application of data mining in educational databases for predicting academic trends and patterns,” in *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, 2012, pp. 1–4.
- [42] M. Matetic, M. B. Bakaric, and S. Sisovic. “Association rule mining and visualization of introductory programming course activities,” in *ACM Proceedings of the 16th International Conference on Computer Systems and Technologies*, 2015, pp. 374–381.
- [43] H. Wang, X. Hao, W. Jiao, and X. Jia. “Causal association analysis algorithm for MOOC learning behavior and learning effect,” in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 2016, pp. 202–206.
- [44] J. Deng, J. Hu, H. Chi, and J. Wu. “An Apriori-based approach for teaching evaluation,” in *2010 2nd International Symposium on Information Engineering and Electronic Commerce (IEEC)*, 2010, pp. 1–4.
- [45] C.-L. Mao, S.-L. Zou, and J.-H. Yin. “Educational evaluation based on Apriori-Gen algorithm.” *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 10, pp. 6555–6564, Sep. 2017.
- [46] S. Ahmed, R. Paul, and A. S. M. L. Hoque. “Knowledge discovery from academic data using association rule mining,” in *2014 IEEE 17th International Conference on Computer and Information Technology (ICCIT)*, 2014, pp. 314–319.
- [47] N. D. Guerrero and S. C. Ambat. “Mining students’ performance in Cisco networking academy program using Apriori algorithm.” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 11, Nov. 2016.
- [48] M. Srilekshmi, S. Sindhumol, S. Chatterjee, and K. Bijlani. “Learning analytics to identify students at-risk in MOOCs,” in *2016 IEEE Eighth International Conference on Technology for Education (T4E)*, 2016, pp. 194–199.

- [49] K. Mouri, C. Yin, F. Okubo, A. Shimada, and H. Ogata. “Profiling high-achieving students for e-book-based learning analytics,” in *CrossLAK*, 2016, pp. 5–9.
- [50] “BookLooper | KCCS.” Internet: <http://www.kccs.co.jp/ict/cloud-booklooper/>, [Nov. 24, 2017].
- [51] S. Ahmed, A. S. M. L. Hoque, M. Hasan, R. Tasmin, D. M. Abdullah, and A. Tabassum. “Discovering knowledge regarding academic profile of students pursuing graduate studies in world’s top universities,” in *IEEE International Workshop on Computational Intelligence IWCI*, 2016, pp. 120–125.
- [52] Y. Teng, L. Zhang, Y. Tian, and X. Li. “A novel FAHP based book recommendation method by fusing Apriori rule mining,” in *2015 IEEE 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2015, pp. 237–243.
- [53] C. Romero, J. R. Romero, J. M. Luna, and S. Ventura. “Mining rare association rules from e-learning data,” in *3rd International Conference on Educational Data Mining*, 2010, pp. 171–180.
- [54] C.-T. Chen and K.-Y. Chang. “A study on the rare factors exploration of learning effectiveness by using fuzzy data mining.” *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 13, no. 6, pp. 2235–2253, Jun. 2017.
- [55] C.-H. Chen, T.-P. Hong, and V. S. Tseng. “A cluster-based fuzzy-genetic mining approach for association rules and membership functions,” in *IEEE International Conference on Fuzzy Systems*, 2006, pp. 1411–1416.
- [56] R. Dash, R. L. Paramguru, and R. Dash. “Comparative analysis of supervised and unsupervised discretization techniques.” *International Journal of Advances in Science and Technology*, vol. 2, no. 3, pp. 29–37, 2011.
- [57] “Market basket analysis - identifying products and content that go well together,” Internet: <https://discourse.snowplowanalytics.com/t/market-basket-analysis-identifying-products-and-content-that-go-well-together/1132> [Dec. 25, 2017].
- [58] Y. Zhao. “Association rule mining with R.” Internet: <http://www.rdatamining.com/docs/association-rule-mining-with-r>, Oct. 7, 2016 [Nov. 17, 2017].
- [59] K. McGarry. “A survey of interestingness measures for knowledge discovery.” *The Knowledge Engineering Review*, vol. 20, no. 1, pp. 39–61, Mar. 2005.
- [60] M. Hahsler and S. Chellubhoina. *arulesViz: Visualizing Association Rules and Frequent Itemsets*. 2012.

- [61] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.
- [62] A. A. Dragulescu. *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files*. 2014.
- [63] F. Eibe, M. Hall, I. Witten, and J. Pal. “The WEKA workbench.” *Data Mining: Practical Machine Learning Tools and Techniques*, Burlington: Morgan Kaufmann, 2016.
- [64] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik. *arules: Mining Association Rules and Frequent Itemsets*. 2017.
- [65] C. Angeli, S. Howard, J. Ma, J. Yang, and P. A. Kirschner. “Data mining in educational technology classroom research: Can it make a contribution?” *Computers & Education*, vol. 113, pp. 226-242, Oct. 2017.
- [66] R. Agrawal and R. Srikant. “Fast algorithms for mining association rules.” in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, vol. 1215, pp. 487–499, Sep. 1994.
- [67] H. Wickham, R. Francois, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*. 2017.

Appendix A: Apriori Algorithm

Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers). Each transaction is seen as a set of items (an itemset). Given a threshold C , the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database. Apriori uses a “bottom-up” approach, where frequent subsets are extended one item at a time (candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate itemsets of length k from itemsets of length $k - 1$.

The pseudo code for the algorithm is given below in Fig. A.1 for a transaction database T , and a support threshold of ϵ . C_k is the candidate set for level k . At each step, the algorithm is assumed to generate the candidate sets from the large itemsets of the preceding level, heeding the downward closure lemma. $count[c]$ accesses a field of the data structure that represents candidate set c , which is initially assumed to be zero.

```

Apriori( $T, \epsilon$ )
   $L_1 \leftarrow \{\text{large 1 - itemsets}\}$ 
   $k \leftarrow 2$ 
  while  $L_{k-1} \neq \emptyset$ 
     $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}$ 
    for transactions  $t \in T$ 
       $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$ 
      for candidates  $c \in C_t$ 
         $count[c] \leftarrow count[c] + 1$ 
       $L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\}$ 
       $k \leftarrow k + 1$ 
  return  $\bigcup_k L_k$ 

```

Fig. A.1: Apriori algorithm [66]

Appendix B: ETL for “Teacher”

There are two kinds of attributes for each of the educational unit of ‘Cluster’, ‘School’, ‘Teacher’, ‘L2M’, and ‘Grade’ entities. The first is the identity attributes comprising of ID and Name, and the second one is the characteristic attributes comprising of features and outputs of the respective educational unit.

Data Extraction as Relational table

For the extraction of class ‘Teacher’, a Visual Basic for Applications (VBA) script was written to extract all the attributes of Teacher in the form of relational data. The data derivation was performed at this stage to obtain derived attributes like Teacher’s age from date of birth, Teacher’s experience from the date of joining, etc. The output data was saved to an MS Excel worksheet and it consisted of the following attributes:

- Teacher ID
- Teacher Name
- Cluster ID
- School ID
- L2M ID
- Teacher designation
- Teacher workload per week
- Teacher academic qualification
- Teacher professional qualification
- Teacher age
- Teacher experience
- Teacher training duration
- Recommended teacher training
- Subject team of teacher
- Level of teacher identified by peers
- Teacher result

After the extraction of this class, the table was loaded into RStudio. The load and transform steps were performed using R version 3.4.1.

The `xlsx` [62] package's `read.xlsx` function was used to load this data from MS Excel to R.

```
> teacher<- read.xlsx("C:\\~\\output.xlsx", sheetName="Teacher")
```

Data Cleansing:

The numeric teacher attributes which had out-of-range values and missing values coded as 0 were cleansed and these values were changed to missing values.

```
> teacher$Teacher.Age[teacher$Teacher.Age<0 | teacher$Teacher.Age>100]  
> ]<-NA
```

Attributes Consistency:

The teacher attributes were modified for consistency. For example, the degrees B.ED. and BS.ED. were changed to B.Ed.

```
> teacher$Teacher.Prof.Qual.<-ifelse(teacher$Teacher.Prof.Qual.=="B  
.ED.", "B. Ed.", as.character(teacher$Teacher.Prof.Qual.))
```

```
> teacher$Teacher.Prof.Qual.<-ifelse(teacher$Teacher.Prof.Qual.=="B  
S.ED.", "B. Ed.", as.character(teacher$Teacher.Prof.Qual.))
```

Attributes Discretization

The attributes discretization was performed for numeric attributes such as Teacher workload per week using the fixed interval binning. The *discretize* method in package *arules* [64] was used for this kind of binning.

```
> teacher$Teacher.workload.per.week<-discretize(teacher$Teacher.workl  
oad.per.week, method="fixed", categories = c(-Inf, 20, 40, Inf), labels=c("1-20 hours", "21-40 hours", "More than 40 hours"))
```

Formulation of Educational Basket for Micro analysis

The teacher table was merged with "Grade" table to prepare the data for micro analytics using the *left_join* command from *dplyr* [67].

```
> micro_teacher_outcome<-left_join(teacher, class)
```

Attributes Selection:

The identity attributes were removed from the micro-level basket by using the *select* function from package *dplyr* and using the column number of the attributes that were to be removed.

```
> micro_teacher_outcome<-select(micro_teacher_outcome, -c(1:4))
```

Appendix C: Rules Generation for Teacher Outcome Analysis

The template discussed in Section 4.5.1.1. for Teacher outcome analysis is used in this Appendix for rule generation.

Conversion of Basket to Transactions:

The micro-level basket containing the teacher attributes was converted to transactions using *arules* [64].

```
> t_micro_teacher<- as(mi_micro_teacher_outcome, "transactions")
```

Rules for Teacher outcome = Good/Bad

The following statement was used to obtain rules for teachers who obtained Good results.

```
> r_micro_teacher_good<- apriori (t_micro_teacher, parameter=list(supp=0.0015, maxlen=4, conf=0.85), appearance=list(rhs="Teacher.result=Good", default="lhs"))
```

In the above statement, the following parameters were used:

- *supp*= 0.015; value was obtained by decreasing the support from 0.1 until useful rules were obtained.
- *conf*= 0.85; to mine high-confidence rules
- *maxlen* = 4; to obtain rules with a maximum of 4 itemsets in each rule.

The appearance parameter was provided according to the template specifying RHS as Teacher result = Good, and all the other variables on the LHS.

Similarly, the following statement was used to mine rules for teachers who got bad results.

```
> r_micro_teacher_bad<- apriori (t_micro_teacher, parameter=list(supp=0.007, maxlen=4, conf=0.85), appearance=list(rhs="Teacher.result=Bad", default="lhs"))
```

Removing redundant rules

There are often many redundant rules generated by the Apriori algorithm. These rules are super rules to rules with higher lift and add no extra knowledge. The redundant rules were removed by:

```
> r_mi cro_teacher_bad<- r_mi cro_teacher_bad[!is.redundant(r_mi cro_tea  
her_bad) ]
```

Sorting rules

The obtained rules can be sorted with respect to support, confidence, and lift. For example,

```
> r_mi cro_teacher_bad<- sort(r_mi cro_teacher_bad, by="l i f t")
```

Displaying rules

The following command is used to view the sorted rules alongwith their quality measures.

```
> i n s p e c t (r_mi cro_teacher_bad)
```

Visualizing rules

The resulting rules were visualized for interestingness using the *arulesViz* [60] package by the following commands:

1. Scatter plot:

```
> p l o t l y _ a r u l e s (r_mi cro_teacher_bad)
```

2. Matrix plot:

```
> p l o t (r_mi cro_teacher_bad, method="m a t r i x")
```

3. Parallel coordinates plot

```
> p l o t (r_mi cro_teacher_bad, method="p a r a c o o r d")
```

Vita

Tasneem Yousuf was born in 1990, in Karachi, Pakistan. She did her Matriculation in 2005 from Little Folk's Secondary School and her Intermediate in 2007 from Sir Syed Government Girls College, in Karachi. In 2011, she completed her B.E. in Computer and Information Systems Engineering from the NED University of Engineering and Technology, in Karachi. From 2012 to 2013, she worked in the capacity of Business Analyst at eDev Technologies Inc.

In 2014, Ms. Tasneem joined the Master of Science in Computer Engineering program at the American University of Sharjah as a Graduate Teaching Assistant. During her master's study, she co-authored 3 papers which were presented in international conferences.