# Investigation of Variable Selection Methods for Latent Class Analysis to Identify Patterns of Learner Behaviour

**Chavvi Nihal Chandani**

**BCA(H)**

**A Dissertation**

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

**Master of Science in Computer Science (Intelligent Systems)**

Supervisor: Dr. Arthur White

September 2020

# Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

[redacted]

Chavvi Nihal Chandani

September 7, 2020

# Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

[redacted]

Chavvi Nihal Chandani

September 7, 2020

# Acknowledgments

First of all, I would like to thank my supervisor Dr. Arthur White for providing me with the opportunity and having trust in me to take his research forward. He has been a source of continuous support, guidance and motivation. Successful completion of this research wouldn't have been possible without his direction.

I would also like to thank my parents for continuously motivating me to pursue what I want, without their strong will and support it wouldn't have been possible for me to pursue this degree.

Lastly, I would like to express my gratitude towards my friends and flatmates who stuck together during these tough times and helped each other to shine bright during this pleasant journey of Masters.

Chavvi Nihal Chandani

*University of Dublin, Trinity College*
*September 2020*

# Investigation of Variable Selection Methods for Latent Class Analysis to Identify Patterns of Learner Behaviour

Chavvi Nihal Chandani, Master of Science in Computer Science
University of Dublin, Trinity College, 2020

Supervisor: Dr. Arthur White

In this paper, the associations of early-stage business students with learning assets for an introductory statistics module over a period of four months were analysed using Latent Class Analysis. The examples of conduct of studies, by the various groups give bits of knowledge in relation to which learning assets undergraduates use more frequently. Although varying degrees of face to face participation and online association existed, all the groups neglected to connect with online material in an ideal way. Later on, two variable selection method called Headlong Search Algorithm and Swap Stepwise Selection Algorithm for Latent Class Analysis were explored. These methods compared two models to determine a variables usefulness for clustering, provided the clustering variables are already selected. Implementation of Headlong Search Algorithm resulted in the selection of variables with clustering information, thereby removing 17 variables out of 59 variables. Most of the variables containing information in relation to scheduled online material were dropped, whereas all the variables containing lecture and tutorial attendance were retained. Consequently, led to classification performance improvements and accuracy in the choice of the number of classes. The results from this examination can be further used for the production of various models like Early

Warning Systems, that warns the students at risk presumably during the mid-semester, to give them sufficient time to improve before final examination or it can also be used by educators to build student-specific supplies, which caters to student needs as per their behaviour, etc

# Summary

In this research, the focus is on recognising the different behavioural learning patterns amongst the first-year undergraduate business students taking the Data Analysis for Decision Makers(DADM) module at University College Dublin, in 2013-2014 academic year. A total of 524 students data has been investigated in this research.

This paper explores how five different learning resources such as lectures, tutorials, online material, printable pdfs and blackboard logins are utilised by undergraduate business students over a period of twelve weeks. For this purpose, model based clustering has been employed using Latent Class Analysis. 7 different models have been trained to find the best-suited parameters for further research. An iterative model was created to perform Latent Class Selection, that results in the selection of the optimal number of clusters that outlines the different behavioural patterns amongst each group. Four was selected as the optimal number of latent classes, which presented varying student behaviours such as students with very high motivation during the start of the course, loose interested over the passage of time, next are students with good intentions who maintain high attendance throughout the course, another group also has good intentions but also showed an early interest in online resources when compared to others and the final group was the one that had minimal interaction with all the learning resources.

This modelling framework of Latent Class Analysis does not address the selection of essential variables on its own. It uses all the available variables to form different group

structures. Broadly, removing unnecessary variables and parameters can also improve classification performance and the precision of parameter estimates. Hence, two different variable selection methods have been explored, namely Headlong Search Algorithm and Swap - Stepwise Selection Algorithm, the former resulted in the removal of 17 variables out of 59, thereby improving latent class selection along with improvement in classification performance, whereas further analysis regarding the later outcomes will be considered for future work, as it was computationally expensive due to time and device constraints.

The implementation can be viewed at:

https://github.com/ChavviChandani/ModelBasedClusterigAndVariableSelection

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and Background

## 1.1 Introduction

The rebuilding of undergraduate business degree programs at the School of Business in University School Dublin (UCD) Ireland in 2011 distinguished quantitative and expository aptitudes as fundamental, to all-encompassing training of business studies. Clear decision-making practices based upon data analysis were considered as an essential skill for future business pioneers. The audit included a conference with key partners and recognized three program pillars, i.e., business in society, innovation and enterprise, and personal development planning. Thus, during this survey, the Data Analysis for Decision Makers(DADM) module was structured to create business undergraduates' logical and analytical abilities by acquainting them with standard statistical practices. To underpin their knowledge, online resources were designed and to deepen student learning and critical skills development, lectures and tutorials were used [1].

In this research, the focus is on identifying the behavioural learning patterns in the ecosystem of undergraduate business students taking the Data Analysis for Decision Makers(DADM) module in the 2013-2014 academic year at University College Dublin.

This paper examines how teaching aids, such as online resources and head-to-head lectures and tutorials, are used by undergraduate first-semester business students, as well as it identifies if each of these resources over the period of twelve weeks were essential in identifying learner pattern behaviour. This has been implemented on the data set of 524 students each with 59 variables. Hence, for this purpose, a variable selection method has been applied for the selection of the most useful variables in detecting the group structure in the data.

Latent Class Analysis is used to reveal clusters in multivariate categorical or binary data. It models the data as a finite mixture of distributions, each one corresponding to a class or group or cluster. It has an underlying statistical model that makes it possible to discover the number of classes using a model selection method.

The modelling framework of Latent Class Analysis does not address the selection of essential variables on its own. It uses all the available variables to form different group structures. There exist multiple reasons why the selection of variables for Latent Class Analysis is beneficial, one of them being that it assists in better understanding of the model, also, it makes it easy to fit a model with more number of classes when compared to clustering with all the variables. Broadly, removing unnecessary variables and parameters can also improve classification performance and the precision of parameter estimates.

Therefore, during this research, variable selection methods have been examined to extract variable to be used for clustering. One of them is based on Dean and Raftery (2010) [2] method for variable selection in model-based clustering of discrete variables. The strategy behind this is, the comparison of two different models and then allows the discarding of those variables with no group information and those variables carrying the same information as the already selected ones, given the clustering variables already selected. This is followed by enforcement of Headlong Search Algorithm, based on Badsberg (1992) [3], for exploring the space of possible models. Not only this but it also easily selects the variables along with the number of classes in the model. The other method is based on Fop, Smart and Murphy (2017) [4] method were two different models are compared again and the variables are selected using Swap-Stepwise selection algorithm.

***Research Question:*** In general, the data suggested low levels of lecture attendance and lack of engagement with online material. However, as the module is delivered to a large group of students, it would not be right to quickly narrow down this conclusion for all students. Hence, a further study is done to find out if these claims are valid for all students or just for certain groups of students. The two primary research questions addressed here are:

- What are the identifiable behavioural patterns of learning objects used?

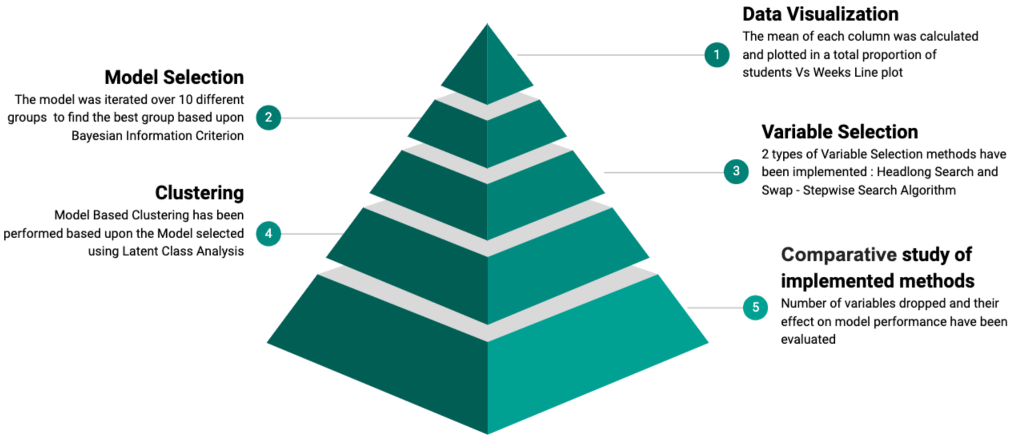

Data Visualization
The mean of each column was calculated and plotted in a total proportion of students Vs Weeks Line plot
1
Model Selection
The model was iterated over 10 different groups to find the best group based upon Bayesian Information Criterion
2
Variable Selection
2 types of Variable Selection methods have been implemented : Headlong Search and Swap - Stepwise Search Algorithm
3
Clustering
Model Based Clustering has been performed based upon the Model selected using Latent Class Analysis
4
Comparative study of implemented methods
Number of variables dropped and their effect on model performance have been evaluated
5

Figure 1.1: Research Overview

- Are all these variables essential for identifying behavioural patterns? Does each of these 59 variables provide essential clustering information?

The entire project overview can been seen in Figure 1.1.

## 1.2 Literature Review

The academic performance of students is likely to increase if they are not only made aware of their learning styles but are also supported with learning material that incorporates their individual learning styles [5]. The behaviour of the students taking online classes has been closely observed by learning management systems to infer an automatic learning style using the Felder-Silverman Learning Styles Model. The Felder-Silverman model is based on the supposition that students have preferences in terms of the way they receive and process information. This approach was used on a set of 127 students, and yielded good results, thereby proving that the proposed approach was suitable for identifying learning styles.

As per [1], there is a high possibility that the positive interaction between the students and the learning objects can result in effective learning. The engagement of the biotic components, i.e students with biotic, i.e lecturers and tutors and abiotic components, i.e eLearning with the content of the learning ecosystem has been examined. The cause of students detachment from the lectures might be a result of the acces-

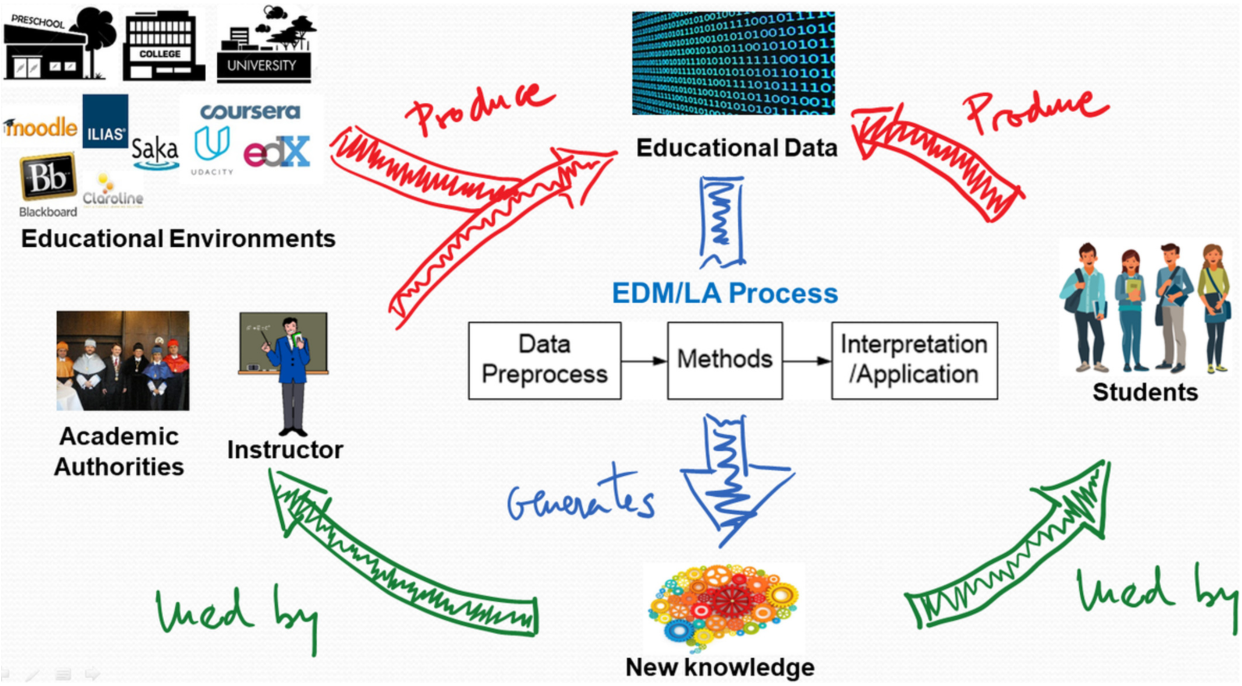


Figure 1.2: Educational Data Mining and Learning Analytics Process

sibility of eLearning resources [6]. [7] states that teaching efficiently with technology needs continuous creation, maintenance and re-establishment of a dynamic equilibrium amongst these three components. Student interaction with technology can be readily understood using Learning Analytics and Educational Data Mining techniques. Learning Analytics remodels the way, impacts and outcomes in learning environments are measured. It allows users to develop new ways of obtaining excellence in teaching and learning as well as providing students with new information to make the best choice about their education. These techniques contribute to furnishing experimental proofs regarding what students do with the learning resources (See Figure [8]). Hence, Learning Analytics (LA) has been applied to recognize how learners use the newly designed learning objects; do these patterns have any relation with the grades that student receives, and providing suggestions on the type of learning objects that can increase the learning productivity of business students. The results of this approach stated that all the learning objects were used differently by distinct groups of learners.

Cluster analysis automatically looks for gatherings of related observations in a dataset [9]. It helps in placing data objects in the same group that are more similar to each other when compared to those in other groups. It is a powerful data mining tool for the identification of discrete groups in any kind of data. Similarly, Model-based clustering is a well established and popular tool for clustering multivariate data. In this approach, clustering is formulated in a modelling framework and the data generating process is represented through a finite mixture of probability distributions [10]. In general, if the number of groups is known beforehand, then parameters are usually estimated using the EM algorithm [11]. Generally, model selection corresponds to the selection of the number of groups and to accomplish this task, a plethora of methods have been suggested in the literature, the Bayesian Information Criterion [12] being the most popular one. Another popular approach for mixture model selection is the Integrated Complete-Data Likelihood criterion [13], which gives more importance to models with well-separated clusters.

Latent class analysis (LCA) is a type of finite mixture model that helps in determining unobserved subgroups of binary data. It is based upon the local independence assumption stating that the variables are independent within each latent class [14]. Few examples of binary data are, the symptoms observed in people with major depressive disorder [15], disability index recorded by long-term survey [16] or the correct or incorrect answers submitted during an exam [11].

R provides an environment for statistical computing. It features several packages for finite mixture model, some for mixtures of multivariate Gaussian distributions, regression models, etc. Latent Class Analysis can be performed using R with the help of packages like e1071 [17] and in particular poLCA [18] which allows the user to perform inference within a maximum likelihood estimate, frequentist framework. In order to perform inference on model posterior, no dedicated package existed. Hence, Bayes LCA package was developed to perform Latent Class Analysis within a Bayesian paradigm [19]. This package helps its users to include prior information into their analysis upon their requirement. Features like these were outside the scope of frequentist analysis.

In all cases, the model with the higher value of Bayesian Information Criterion concerning a criterion is deemed the better fit to the data [19]. Generally, at first, all the variables are used for modelling. But, in many cases using all the variables disadvantageously increases the model complexity. It is often noted that some variables do not contain any clustering knowledge, thus are of no use in detecting the group structure. Rather, they could be pernicious to clustering. Furthermore, the case where all of the variables contain clustering information can also be problematic.

Along with the increasing number of dimensions comes the curse of dimensionality [20] and including unnecessary variables in the model leads to identifiability problems and over-parameterization [21] [22]. Therefore, resorting to variable selection techniques can expedite model fitting, facilitate the interpretation of the results and lead to better quality data classification. Reducing the set of variables used in the clustering process, even in situations of medium or low dimensionality can be beneficial [23].

There are various methods for variable selection for Latent Class Analysis like Bayesian approach, Penalization approach, etc. In model-based clustering, the distribution of the relevant variable directly depends on the group membership variable. As per distributional postulates on relevant and irrelevant variables different model-based clustering and variable selection approaches can be outlined. Two main assumptions unique to the task of variable selection and mixture model clustering are: Local independence assumption, i.e the relevant variables are conditionally independent within the groups and Global independence assumption which states that the irrelevant variables are independent of relevant clustering variables. In particular, the local independence assumption helps to interpret the modelling of the joint distribution of relevant variables and is especially useful in high-dimensional data frameworks. It is a conventional assumption of the latent class analysis model [14].

For Gaussian mixture models, the statement corresponds to assuming components with diagonal covariance matrices. The global independence assumption implies that the joint distribution of the variables factors into the product of the mixture distribution of the relevant variables and the distribution of the irrelevant variables. The term “global” is employed because the independence statement affects the distribution of all the variables, which is not only specific to the clustering ones. This assumption interprets the modelling of the association between relevant and irrelevant variables [10].

[24] considers full independence of the variables, while a further enhancement of this method is [2] which considers conditional independence of variables. The later has been used in the selection of variables from the DADM Dataset of students. A headlong search algorithm consisting of inclusion and removal steps was introduced by the authors to explore within the model space [3]. The algorithm has the benefit of being more computationally productive than a backward search but has the drawback

of being sensitive to the initialization of the set of clustering variables. To surmount this [25] added a random check step aimed at initializing the estimation algorithm with a large number of random starting values, so as to prevent the problem of incurring in local optima.

A major drawback of [2] framework is the independence assumption between the proposed variable and the set of clustering ones in the model. Given the set of already selected relevant variables, the model does not take into account that the proposed variable could be redundant for clustering based upon the above assumption. Thus results in this approach were not only being capable of dropping variables related to the clustering ones, but not directly related to the group structure itself. To overcome the problem, [4] propose the modelling framework, which has been considered as the second variable selection method in this research. Although the results for the second model couldn't be received due to time and device constraints.

# Chapter 2

# Data Analysis for Decision Makers

Data Analysis for Decision Makers (DADM) is a fundamental subject which needs to be studied by all first-semester business students at UCD. The learning outcomes upon successful completion of this course were :

1. Analysis of quantitative data using common probability distributions and statistical functions.

2. Calculation, analysis and presentation of useful statistical measurements from large-scale data sets.

3. Creation and interpretation of inferential statistical statements about population parameters.

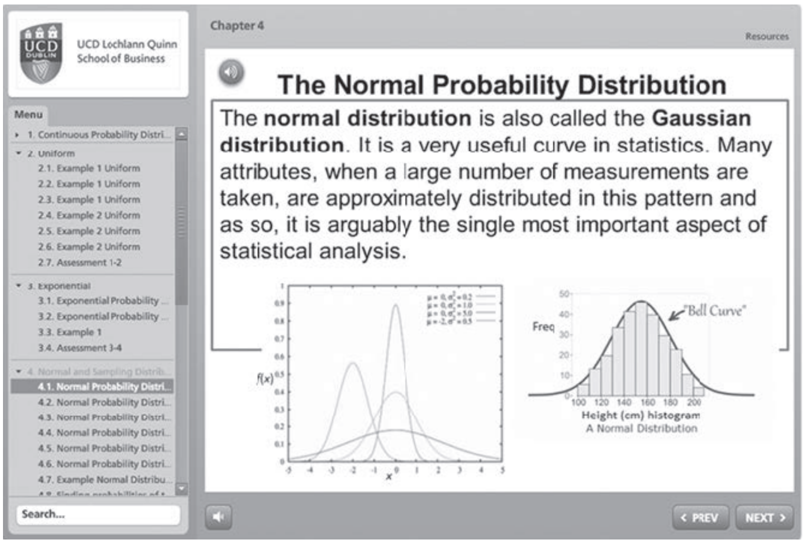
4. Interpretation of the results of data analysis with a view to inform decision making [1].

The students are given to access online learning resources at any given time on the Virtual Learning Environment(VLE) used by UCD, i.e. Blackboard. A blended learning approach is incorporated, where pre-lecture reading and reflection on online materials is complemented by two hours of lectures and one-hour tutorials. It has been noted that on an average around 600 students every year take up this module at UCD. The total students were broken down as follows: each week comprised of 4 lectures with 150 students each and 12 tutorials for 50 students at a time. The Professors motive is to make data analysis concepts in a business context clear to the students and promote

| **Week** | **Lecture content** | **Tutorial content** | **CA task** | **Additional activity** |
|---|---|---|---|---|
| 1 | Intro to data analysis and descriptive statistics | Intro to excel | | Drop-in clinic |
| 2 | More descriptive statistics | Descriptive statistics | Excel work | Drop-in clinic |
| 3 | Basic probability | Graphs and tables | Excel work | Drop-in clinic & industry speaker |
| 4 | Discrete random variable probability distributions | Excel functions | Excel work | Excel training probability distributions |
| 5 | Continuous random variable probability distributions | Excel test | Excel test | Drop-in clinic probability distributions |
| 6 | Normal distributions | Probability exercises | | MSC Hot topic & industry speaker |
| 7 | Sampling theory | MCQ1 | MCQ1 | Drop-in clinic |
| 8 | Confidence intervals | Probability exercises | | MSC Hot topic & industry speaker |
| 9 | Confidence intervals | MCQ2 | MCQ2 | Drop-in clinic |
| 10 | Hypothesis testing | Inferential statistics exercises | | Drop-in clinic & industry speaker |
| 11 | Hypothesis testing | Inferential statistics exercises | Team assignment | |
| 12 | Case study | Revision | | |

Table 2.1: DADM Schedule

active/task-based learning while teaching assistants facilitate a discussion of peer and group work, enquiry and problem-based learning exercises during the tutorials. Apart from this to provide external exposure to the business students, guest speakers from industry, drop-in clinics and hot topic sessions are also arranged by the university's Math Support Centre (MSC) [26]. An abstract schedule of the DADM activities and continuous assessment (CA) tasks at UCD can be seen in Table 2.1 [1].


UCD Lochlann Quinn School of Business
Chapter 4
Resources
Menu
1. Continuous Probability Distri...
2. Uniform
2.1. Example 1 Uniform
2.2. Example 1 Uniform
2.3. Example 1 Uniform
2.4. Example 2 Uniform
2.5. Example 2 Uniform
2.6. Example 2 Uniform
2.7. Assessment 1-2
3. Exponential
3.1. Exponential Probability ...
3.2. Exponential Probability ...
3.3. Example 1
3.4. Assessment 3-4
4. Normal and Sampling Distrib...
4.1. Normal Probability Distri...
4.2. Normal Probability Distri...
4.3. Normal Probability Distri...
4.4. Normal Probability Distri...
4.5. Normal Probability Distri...
4.6. Normal Probability Distri...
4.7. Example Normal Distribu...
Search...
f(x)
x
Freq
"Bell Curve"
Height (cm) histogram
A Normal Distribution
< PREV
NEXT >

The Normal Probability Distribution

The **normal distribution** is also called the **Gaussian distribution**. It is a very useful curve in statistics. Many attributes, when a large number of measurements are taken, are approximately distributed in this pattern and as so, it is arguably the single most important aspect of statistical analysis.

Figure 2.1: Online Learning Portal

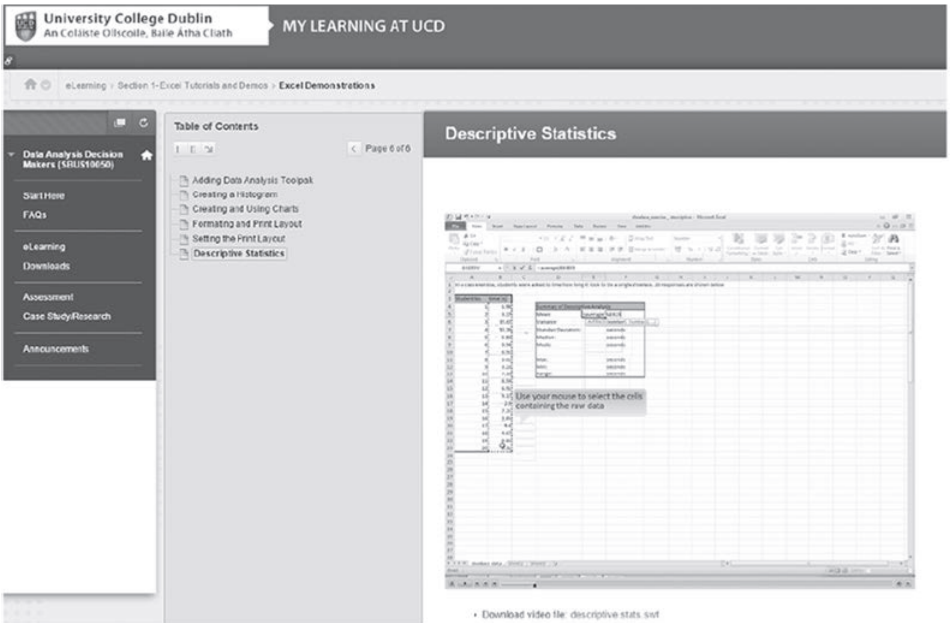
## 2.1 Design of Data Analysis for Decision Makers Learning objects

In the designing of Data Analysis for Decision Makers module the Business School faculty focused on:

- Using technology to help develop conceptual understanding and for analysing data, and
- Using activities to improve learning and engage students during contact time.

A story-based approach using the Articulate software [27] was used by the professors for the creation of online courseware. The layout was designed based upon the design principles in Mayer (2001)[28] as it suggests that Multimedia learning occurs when a learner builds a mental representation from words and pictures that have been presented, and as Mayer and Moreno [29] prove that cognitive load is a central consideration in the design of multimedia instruction. Hence, the course material was structured into sections and chapters. A fragment of Online Learning Portal can be seen in Figure 2.1 [1].

The pages with speaker icon provide options to listen to the explanatory voice-over for that particular section. For example, the voice over in Figure 2.1 says, "Note the


University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath
MY LEARNING AT UCD
eLearning > Section 1-Excel Tutorials and Demos > Excel Demonstrations
Data Analysis Decision Makers (SBUS10050)
Start Here
FAQs
eLearning
Downloads
Assessment
Case Study/Research
Announcements
Table of Contents
Page 6 of 6
Adding Data Analysis Toolpak
Creating a Histogram
Creating and Using Charts
Formatting and Print Layout
Setting the Print Layout
Descriptive Statistics
Descriptive Statistics

Use your mouse to select the cells containing the raw data

- Download video file: descriptive stats.swf

Figure 2.2: A screenshot from MS Excel Demonstration

figure on the left that there are many normal distributions. Which one we are dealing with is determined by the mean and the variance of the distribution. In the figure on the right, we see some data that has been gathered and formed into a histogram. We see that the pattern in the histogram, if we were to throw a blanket over it, almost forms a normal distribution or bell-shaped curve." The online material also consisted of various essential links to YouTube Videos, excel and sampling distribution demonstration for better understanding. A screengrab of MS Excel Demonstration can be seen in Figure 2.2 [1].

This was created using Adobe Captivate, to promote the practice and better understanding to perform Data Analysis in MS Excel. The topics covered during Excel Online Learning were an introduction to spreadsheets, data entry, use of statistical and descriptive statistics functions, and drawing charts and histograms. Students were assigned timely assessments and were expected to complete short exercises before attending the tutorials as they focused on the practical implementation of the theory taught during lectures. Although completion of these short practical exercises triggered automated feedback upon the performance of the student, yet were not graded. Each chapter was designed to include a conclusions/summary section and end with a set of review and more detailed Theory in Practise (TIP) case study questions. The TIP topics were carefully selected so that they are of at most relevance to the early stage business students.

Students were given the option to download Online Learning material as per their convenience. Students could download zipped folders of the full content or a pdf printable version of the text content as per their wish. This was done to facilitate students who were more comfortable to print and bring the hard copy notes to face-to-face sessions as these supplements are more traditional pdf copies of the text of the learning materials.

### 2.1.1 Data Analysis for Decision Makers Assessment Components

The Assessment Component for Data Analysis for Decision Module module was designed very carefully keeping in mind that it should help students to reflect on their interpretations as well as let them derive how these results will enhance business decision making. Hence, the assessments were divided into two types :

- Continuous Assessments and
- Exams

Where Continuous Assessments constituted of 40% of the total marks where other 60% were evaluated using the traditional semester exam system. The Continuous Assessments were further subdivided into various types in order to promote effective learning. The further subdivision of Continuous Assessments can be seen in Table 2.2 [1].

Week 2, 3 and 4 requires continuous submission of MS Excel Tutorial work followed by an MS Excel Test in Week 5 which is graded by Teaching Assistants. It requires the student to create a small spreadsheet, perform some descriptive statistics calculations, and draw a graph. Week 7 and 9 are designated for Open Book Multiple Choice Quizzes (MCQs), where students are allowed to access online resources as well as their own notes. This process was used to develop discovery learning among students. The results for these MCQs are automated by the Virtual Learning Environment used by UCD. It also provides instant feedback on the performance of the test takers. The

| **Week** | **Continuous Assessments Tasks** | **Grade %** |
|---|---|---|
| 2 | MS Excel work | 1 |
| 3 | MS Excel work | 1 |
| 4 | MS Excel work | 1 |
| 5 | MS Excel test | 12 |
| 7 | MCQ 1 | 7.5 |
| 9 | MCQ 2 | 7.5 |
| 11 | Team project | 10 |

Table 2.2: DADM Continuous Assessmnet Schedule

week before the final week is when the students need to submit their Group Projects, which requires them to pen down their understanding, observations and analysis of business datasets in the form of a report. The students are given the freedom to choose the dataset they want to analyse as a team. This final assessment is graded by the Professors as well as checked for Plagiarism using SafeAssign tool embedded into the Virtual Learning Environment.

### 2.1.2 Data Analysis for Decision Makers Student Guide

The DADM student guide reviews the blended learning teaching strategy of the DADM module as well as the schedule of lecture, tutorial, and assessment activities. It was developed to set student expectations from the module. It provided in-depth information about each Continuous Assessments Tasks, university policies on plagiarism, late submission, and the group conduct code of behaviour. It has been noticed that students struggled in transitioning from teacher-led learning approach from their secondary school to independent learning approach in college. Hence, this guide was developed to ease the new process for the students. This guide was made available in Blackboard to all students under the ”Start here” tab, that can be seen in the top left corner of Figure 2.2. All frequently asked questions were also put together into a single file and uploaded on the Blackboard to clarify most of the student doubts, as seen in the top left corner of Figure 2.2 [1].

# Chapter 3

# Data

Data of the students attending the Data Analysis for Decision Makers module at UCD has been used for this research.

## 3.1 Data Collection

First, the Ethics approval was provided by the University Research Ethics Committee to Professor White and Paula. Secondly, the students were informed before the start of the semester about the research study and its objectives. The students' interactions with the online resources were captured upon their logging into the Virtual Learning Environment, i.e.Blackboard. In order to collect true data of the students who attended lectures or tutorials, UCD provided provision for students to swipe their student cards at a barcode scanner which was observed by the lecturer or tutor during respective sessions. UCD uses a smart card system by using barcode scanners that only stores student IDs and no other student information. The students' attendance for Lectures and Tutorials can be seen in blue and orange colour respectively in Figure 3.1. These attendance records were even used by the faculty for the formation of groups for team tasks. This entire process was done by maintaining a high level of data integrity and there is no information about students swiping cards for their friends that the DADM faculty was aware of. Not only this, but the faculty also made sure to enter data manually for students who forgot to bring their identity cards to college, with the hope to deliver accurate results [1].

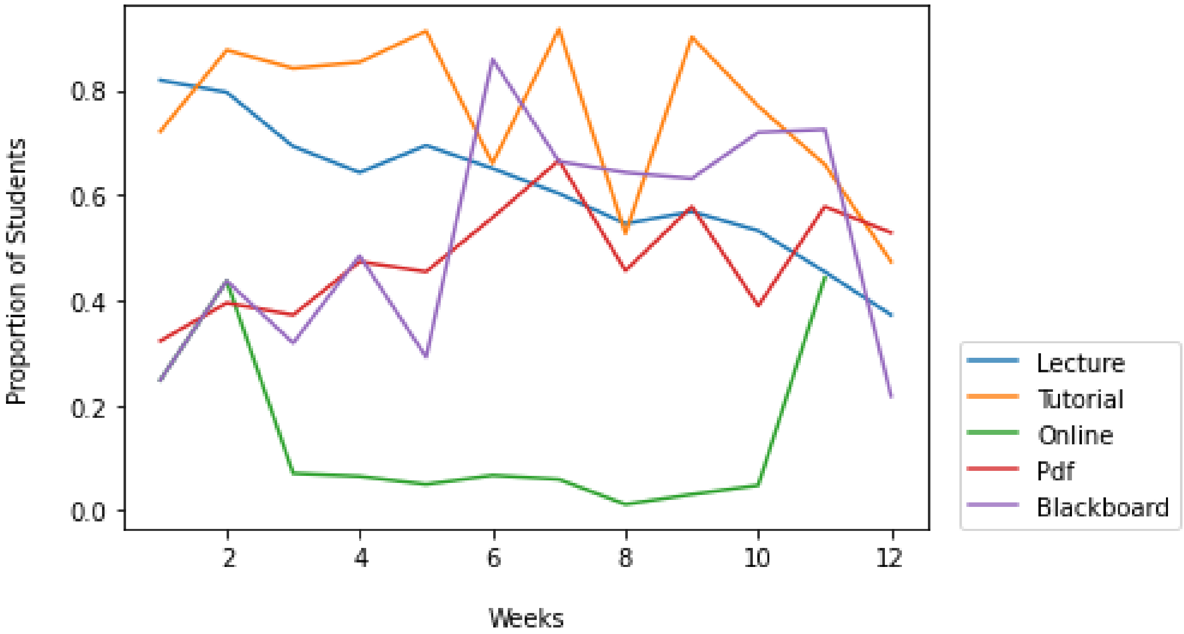


Figure 3.1: Proportion of Students Attending Lectures, Tutorials or Engagement with Online Materials like printable chapters,etc by Week

## 3.2 Dataset

Using the above stated method, a total of 524 student records were collected with regards to five learning objects over the 12 teaching weeks of the DADM module. Figure 3.1 shows very clearly the interaction of the students with different learning resources over the semester of 12 weeks. The details of the various learning resources are given below:

1. The blue line denotes the lecture attendance;

2. The orange line denotes the tutorial attendance;

3. The green line denotes the proportion of students that accessed the online material on the same week when it was taught in the class;

4. The red line denotes the proportion of students who accessed the printable pdf learning materials that could be brought to the class;

5. The purple line denotes the students who accessed any kind of material using Blackboard during that particular week.

This constitutes a total of 59 variables for each student. The data is in the form of 0s or 1s, i.e 1 if the student attended the lecture or tutorial or accessed the online learning materials or 0 otherwise. This kind of data is known as Binary data. For each learning resource data has been recorded for over 12 weeks except for Week 12 for the Online Learning Material because no new learning material was uploaded during the last week of the module [1].

# Chapter 4

# Methodology

## 4.1 Model Based Clustering

Model based clustering is a framework for clustering multivariate discrete data. It is based on the idea that the observed data comes from a population with several classes, groups or subpopulations and models each with its own probability distribution [30, 9]. In this approach, each sub-population is modelled separately and the overall population is modelled as a mixture of these subpopulations, using a finite mixture model. It is used to discover hidden groupings in multivariate categorical data. The general form of the finite mixture model with G sub-populations or groups can be written as :

$$p(X) = \sum_{g=1}^{G} \pi_g p_g(X),$$

where G is the number of groups, $\pi_g$ is the mixture proportion of the population in the $g_{th}$ group and $p_g$ is the probability density function of the $g_{th}$ group. [24] A special case of model-based clustering also exists, in which $p_g$ are from the same parametric density family and this case is well known as Latent Class Analysis.

### 4.1.1 Latent Class Analysis

Latent class analysis was proposed by Lazarsfeld (1950a,b) [31, 32] and Lazarsfeld and Henry (1968) [33] and can be viewed as a special case of model-based clustering, for multivariate discrete data. Hence, the general form of finite mixture model can be rewritten as :

$$p(X_n) = \sum_{g=1}^{G} \tau_g p(X_n|\theta_g),$$

Where, X is a N x M data matrix, each row $X_n$ is the realization of a M-dimensional vector of random variables $X_n = (X_{n1}, ..X_{nm}, ..X_{nM})$, composed of G sub-populations or classes. The two sets of parameters $\tau_g$ and $\theta_g$, underly the model and are known as parameters for Class and Item probability respectively. The parameter $\tau_g$ denotes the prior probability of belonging to group g, where $\tau_g \geq 0$ and $\sum_{g=1}^{G} \tau_g = 1$. The parameter $\theta_{gm}$ denotes the set of parameters for the $g^{th}$ group, such that $X_{im} = 1$, for any $i \in 1, ....., N$, so that $p(X_{im}|\theta_{gm}) = \theta_{gm}^{X_{im}}(1 - \theta_{gm})^{1-X_{im}}$ ,for $X_{im} \in \{0; 1\}$[19].

In latent class analysis, if the class value of an observation is given, it is assumed that the variables are statistically dependent. This is known as local independence. Each variable within each group is then modelled with a multinomial density. The general density of a single variable x given that it is in group g is, therefore,

$$p(X_m|\theta_g) = \Pi_{c=1}^{C_m} \theta_{gmc}^{1\{X_m=c\}},$$

where c= 1,....,$C_m$ are the possible categories values for variable m, $\theta_{gmc}$ is the probability of the variable taking value c given class g, and 1{$X_m$=c} is the indicator function equal to 1 if the variable takes value c, otherwise its 0 [4].

In this case of conditional independence, if we have M variables, their joint group density can be written as a product of their individual group densities as:

$$p(X_m|\theta_g) = \Pi_{m=1}^{M} \Pi_{c=1}^{C_m} \theta_{gmc}^{1\{X_m=c\}};$$

where $1\{X_m = c\}$ is the indicator function equal to 1 if the observation of the $m^{th}$ variable takes value c, otherwise its 0, $\theta_{gmc}$ is the probability of variable m taking value c in group g and $C_m$ is the number of possible values or categories the $m^{th}$ variable can take. The overall density of the finite mixture model, is then a weighted sum of these individual product densities, namely

$$p(X_n) = \sum_{g=1}^{G} \tau_g \Pi_{m=1}^{M} \Pi_{c=1}^{C_m} \theta_{gmc}^{1\{X_m=c\}}.$$

where $0 > \tau_g < 1$, $\forall_g$ and $\sum_{g=1}^{G} \tau_g = 1$.

For a fixed value G, the model parameters $\{\tau_g, \theta_{gmc} : m = 1, ..., M; c = 1, ..., C_m; g = 1, ..., G\}$ can be estimated from the data by maximum likelihood using the EM algorithm or the Newton–Raphson algorithm or a hybrid of the two [34]. A randomly generated starting values are then given to these algorithms because the algorithms are not guaranteed to find a global maximum and are usually fairly dependent on good starting values. Hence, it is a routine to generate several random starting values and use the best solution given by one of these [2].

Thus, because of the underlying statistical model it is possible to determine the number of classes using model selection methods.

More details about the model and the parameter estimation are provided in, Goodman (1974) [35]; Haberman (1979), Clogg (1995) [36], and Hagenaars and McCutcheon (2002) [37].

As in this case, the binary variables for the students data attending DADM module , $\theta_{gmc}$ will represent the probability of being into a certain pattern for each student belonging to class g.

### Selecting the number of Latent Classes / Model Selection

Different LCA models are defined by assigning different values to G. In order to choose the best number of classes for the data we need to choose the best model, this is carried out using an approximation to their Bayes factor. The Bayes factor for comparing model $M_i$ versus model $M_j$ is equal to the ratio of the posterior odds for $M_i$ versus $M_j$ to the prior odds for $M_i$ versus $M_j$ . When the prior model probabilities are equal the Bayes factor reduces to the posterior odds. Hence, the general form for the Bayes factor is:

$$B_{ij} = \frac{p(Y|M_i)}{(Y|M_j)},$$

where $p(Y|M_i)$ is known as the integrated likelihood of model $M_i$ (given data Y ). It is known as the integrated likelihood because it is obtained by integrating over all the model parameters, like the mixture proportions and the group variable probabilities [2].

The ratio of the integrated likelihoods of the two models is the Bayes factor, $B_{i,j}$. The quantity $p(Y|M_i)$ is conveniently approximated using the Bayesian Information

Criterion (BIC), defined by

$$BIC(Y|M_i) = 2log(maximizedlikelihood)(no.ofparameters)log(n),$$

where n is the number of observations [12].

Then twice the logarithm of the Bayes factor is approximately equal to the difference between the BIC values for the two models being compared is:

$$2log(B_{i,j}) \approx BIC(Y|M_i) - BIC(Y|M_j),$$

and if this difference is greater than zero the evidence is in favour of model $M_i$, otherwise in favour of $M_j$ [38].

BIC is proved to be consistent for the choice of the number of components in a mixture model under certain conditions, when all variables are relevant to the grouping [2]. Several arguments in favor of BIC for model selection in mixture models have been given in the literature [4]; see McLachlan and Rathnayake (2014) [39] for a recent review.

A rule of thumb for differences in BIC values is that a difference of less than 2 is viewed as barely worth mentioning, while a difference greater than 10 is seen as constituting strong evidence [38].

By assigning different values to G, for a given number of variables, not all the models specified are identifiable. In fact, a necessary condition to the identifiability of a model with G latent classes can be considered as:

$$\Pi_{m=1}^{M} C_m > (\sum_{m=1}^{M} C_m - M + 1)G,$$

with $C_m$ the number of categories taken by variable $X_m$ [35]. Thus, when selecting the number of classes, hereafter values of G can be considered for which this identifiability condition holds.

## 4.2 Variable Selection Methods for Latent Class Analysis

For the selection of relevant variables for clustering in Latent Class Analysis, Dean and Raftery (2010) [2] introduced a stepwise model comparison approach. At each step of this method the collection of variables are partitioned into three sets :

- $Y^{(clust)}$ : It is the set of variables which are already considered as useful for clustering,
- $Y^{(proposal)}$ : It is the variable(s) that are being considered for inclusion into/exclusion from $Y^{(clust)}$,
- $Y^{(other)}$ : It is the set of all other variables which are not relevant for clustering.

Once this partitioning is done, and unknown clustering memberships $\mathbf{z}$ is known, figuring out the usefulness of $Y_{(prposal)}$ for clustering can be served as a model selection question. The solution to this requires selection between two models $M_1$, which considers that $Y_{(prposal)}$ is not useful and $M_2$, that considers that $Y_{(prposal)}$ is useful. The models $M_1$ and $M_2$ can be stated as:

$$M_1 : p(Y|z) = p(Y^{(clust)}, Y^{(proposal)}, Y^{(other)}|z)$$

$$= p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)})p(Y^{(proposal)})p(Y^{(clust)}|z)$$

,

$$M_2 : p(Y|z) = p(Y^{(clust)}, Y^{(proposal)}, Y^{(other)}|z)$$

$$= p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)})p(Y^{(proposal)}, Y^{(clust)}|z)$$

$$= p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)})p(Y^{(proposal)}|z)p(Y^{(clust)}|z),$$

According to Model $M_1$, when $Y^{(clust)}$ is known and when $Y^{(proposal)}$ is independent of clustering membership defined by unknown clustering variables z, then $Y^{(proposal)}$ doesn't provide any further information about the clustering model. This can be seen in figure 4.1.

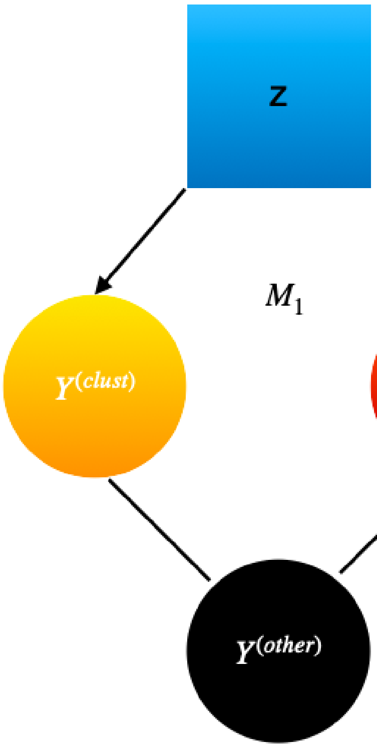


Figure 4.1: $M_1$ states that $Y^{(proposal)}$ is independent of clustering membership z, when $Y^{(clust)}$ variables are already in the model

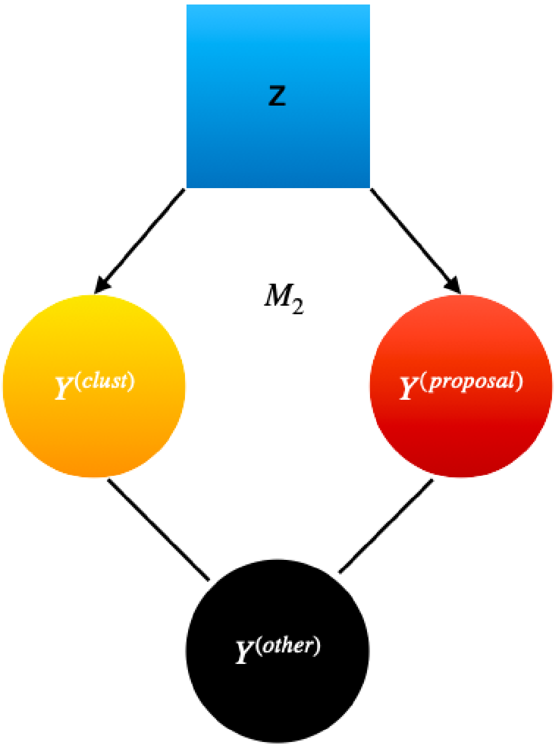


Figure 4.2: $M_2$ states that $Y^{(proposal)}$ is dependent on clustering membership z

Whereas $M_2$ states apart from the clustering information provided by $Y^{(clust)}$, $Y^{(proposal)}$ can also be seen as an beneficial parameter that might provided clustering information beyond $Y^{(clust)}$ variables. This is clearly shown using Figure 4.2.

Both these model are based upon the assumption that $Y^{(other)}$ are conditionally independent of the two other sets, namely $Y^{(clust)}$ and $Y^{(proposal)}$ and belong to the same parametric family.

Both the models $M_1, M_2$ are compared using the approximation to Bayes factor which lets the high dimensional $p(Y^{(other)}|Y^{(clust)}, Y^{(proposal)}$ to cancel from the ratio. If data Y is given, the Bayes factor, $B_{12}$ for two models against each other would be:

$$B_{12} = p(Y|M_1)/p(Y|M_2),$$

where $p(Y|M_k)$ is the integrated likelihood for a model $M_k$, here k = {1,2}.Hence, this can be written as:

$$p(Y|M_k) = \int p(Y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k.$$

Here, $\theta_k$ is the vector-valued parameter of model $M_k$, and $p(\theta_k|M_k)$ is its prior distribution [38].

If integrated likelihood for $M_1$ is considered, $p(Y|M_1) = p(Y^{(clust)}, Y^{(proposal)}, Y^{(other)}|M_1)$, then $M_1$ is stated using three probability distributions namely:

- Latent class model that states $p(Y^{(clust)}|\theta_1, M_1)$
- Distributions like $p(Y^{(proposal)}|\theta_1, M_1)$ and $p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, \theta_1, M_1)$.

The parameter vectors for these three probability distributions is based on the assumption that their prior distributions are independent and is denoted as $\theta_1, \theta_2$ and $\theta_3$ for further calculations. Hence, the integradted likelihood factors can be written as:

$$p(Y|M_1) = p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, M_1)p(Y^{(proposal)}|M_1)p(Y^{(clust)}|M_1),$$

where

$$p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, M_1) = \int p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, \theta_3, M_1)p(\theta_3|M_1)d\theta_3$$

.

Similar results have been obtained for $p(Y^{(proposal)}|M_1)$ and $p(Y^{(clust)}|M_1)$.

Now, if the integrated likelihood for latent class model $M_2$ is considered which is $p(Y^{(proposal)}, Y^{(clust)}|M_2)$ for $(Y^{(proposal)}, Y^{(clust)})$, then we again obtain,

$$p(Y|M_2) = p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, M_2)p(Y^{(proposal)}, Y^{(clust)}|M_2)$$

.

Therefore,

$$p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, M_2) = p(Y^{(other)}|Y^{(proposal)}, Y^{(clust)}, M_1)$$

as the prior distribution of parameter $\theta_3$ is considered to be same under $M_2$ as under $M_1$.

Thus, The Bayes factor, $B_{12}$, for $M_1$ against $M_2$ based on the data Y is given by :

$$B_{12} = \frac{p(Y^{(proposal)}|M_1)p(Y^{(clust)}|M_1)}{p(Y^{(proposal)}, Y^{(clust)}|M_2)},$$

Due to the cancellation of the factors involving the potentially high-dimensional $Y^{(other)}$ the equation has been greatly simplified. Still, the integrated likelihoods are still hard to be decided analytically, and so they are approximated using the BIC approximation [2].

## 4.3 Method 1

### 4.3.1 Headlong Search Algorithm

As introduced in [2] the variables need to be partitioned such that, $Y^{(clust)}$ should consist of enough variables initially so that a latent class model for G > 1 can be identified. If a latent class model is identifiable for G > 1 using all variables, in this case, the largest number of classes that can be identified should be selected to estimate the model. The variance of the probability of the variable across the group is calculated, for each category of each variable. The variances for each variable are added and then variables are ranked according to the sum. The variables that are ranked higher are considered very useful for clustering as they have high between-group variation in probability. Using this ranking, the smallest number of top k variables are selected that are just sufficient enough to identify a latent class model with G > 1. These set of variables are identified as the initial value for $Y^{(clust)}$.

If the above results from the above mentioned process are not received as expected, then the minimum number of variables required identification of the latent class model with G > 1 is calculated. Once, the minimum number of variables are found, then multiple random subsets of variables are generated with the same number of variables that were calculated before. In this case, the initial $Y^{(clust)}$ are selected based upon the variable set that gives the greatest overall average variance of categories' probabilities across the groups.

As soon as the set of variables for $Y^{(clust)}$ have been identified, the inclusion and exclusion steps of the headlong algorithm are followed. These steps, at first require an upper and a lower constant to be set. The upper constant is the value above which the difference in BIC for models $M_2$ and $M_1$ will result in a decision of a variable being included in $Y^{(clust)}$, otherwise will result in a variable being excluded from $Y^{(clust)}$. A default value for upper is 0, which refers to the fact that any positive difference in BIC between models is a measure of a variable's usefulness for clustering and any negative difference is taken as a measure of a variable's lack of usefulness.

On the other hand, the lower constant is the value below which the difference in BIC for the two models will result in a variable being removed from consideration for the rest of the procedure. A difference of lower constant indicates that a variable is unlikely to ever be useful as a clustering variable and needs not to be checked any more. By default, a large negative number such as 100 (which by [2] rule of thumb would constitute strong evidence against) is selected a the initial values of lower constant.

### Inclusion Step

Each variable in $Y^{(other)}$ singly in turn for $Y^{(proposal)}$ is proposed. The difference in BIC for models $M_2$ and $M_1$ is calculated when the current $Y^{(clust)}$ is given . If the variable's BIC difference is:

- between upper and lower, the variable is not included in $Y^{(clust)}$, instead it is returned to the end of the list of variables in $Y^{(other)}$;
- below lower, then the variable is not included in $Y^{(clust)}$ as well as removed from $Y^{(other)}$;
- above upper, in this case the variable should be included in $Y^{(clust)}$ and inclusion step is stopped.

If the end of the list of variables in $Y^{(other)}$ is reached, it marks the end of the inclusion step.

### Exclusion Step

Each variable in $Y^{(clust)}$ singly in turn for $Y^{(proposal)}$ (with the remaining variables in $Y^{(clust)}$ not including current $Y^{(proposal)}$ now defined as $Y^{(clust)}$ in $M_1$ and $M_2$) is proposed. The difference in BIC for models $M_2$ and $M_1$ is calculated. If the variable's BIC difference is:

- between upper and lower constant values, the variable from (the original) $Y^{(clust)}$ is excluded and returned to the end of the list of variables in $Y^{(other)}$ and exclusion step is stopped;
- below lower,then the variable is excluded from (the original) $Y^{(clust)}$ and from $Y^{(other)}$ and the exclusion step is stopped;
- above upper, in this case the variable from (the original) $Y^{(clust)}$ are not excluded.

If the end of the list of variables in $Y^{(clust)}$ is reached, it marks the end of the exclusion step.

Once the headlong algorithm has converged, it stops as $Y^{(clust)}$ remains the same even after consecutive inclusion and exclusion steps [2].

## 4.4 Method 2

### 4.4.1 Swap-stepwise Selection Algorithm

Using swap stepwise algorithm the clustering variables are selected by switching between exclusion, inclusion and swapping steps. In the inclusion step, all the variables in $Y^{(other)}$ are examined in turn to be added to the clustering set, ie $Y^{(clust)}$ . In the exclusion step, all the variables in $Y^{(clust)}$ are examined in turn to be removed from the clustering set. In the inclusion and exclusion step, model $M_1$ is compared against model $M_2$, while in the swapping step, two different configurations of model $M_2$ are compared that differ in the fact that one clustering variable is swapped by one of the non-clustering variables are compared. The grounds for the swap step lies in the postulates of model $M_2$. In model $M_2$ the proposed variable is assumed independent from z conditionally, given the set of already selected variables. Therefore, $Y^{(proposal)}$ is actually allowed to contain some information about the clusters, because if one of the variables of the optimal set for $Y^{(clust)}$ has been discarded during the search, $Y^{(proposal)}$ in such situation may be the best information available. Hence the algorithm could converge to a sub-optimum. Henceforth, to avoid it, two different sets of clustering variables in the swapping step are compared and if a true clustering variable has been removed during this search, then the informative variable is added back to the clustering set. The algorithm looks for the optimal combination of clustering variables and the number of classes at each stage and thus selects the best number of latent classes. The procedure stops when no change has been made to the set $Y^{(clust)}$ after the consecutive exclusion, inclusion and swapping steps.

# Chapter 5

# Implementation & Results

A multivariate binary data with 524 students records, where each student has 59 variables associated with them, which provides information about their lecture and tutorial attendance as well as interactions with online resources like logging in to Blackboard to access any learning data, downloading printable pdfs and accessing lecture notes, over 12 weeks were analyzed.

In order to perform variable selection and model based clustering to investigate the learning behaviours of 524 students taking the DADM module, the prime thing that was required was data visualisation, to understand the data better, and to obtain some valuable hidden information.

To start with this, I first read the csv file using the **readr** package which is one of the fastest ways to read rectangular data. Once the data was read from the csv file, it was converted to a data frame using the **rio** package, which makes appropriations about the file format and consequently applies import functions suitable to that format.

Now, to recognise various patterns form the data, the mean of each column was computed and stored it in a separate data frame. Some format conversions like column renaming, transpose to change the data frame orientation, aligning mean of each educational resource over the 12 weeks with the correct week number and under correct learning resource category, etc was done in order to plot the overall association of the students with the learning resources like lectures, tutorials, and various online materials over a tenure of 12 weeks. These patterns can be distinctly seen in Figure 3.1.

Advancing in this process, Bayesian Latent Class Analysis was applied to the entire student's dataset of 524 rows and 59 columns, by utilising **BayesLCA** library. The BayesLCA fills the gap of performing Latent Class Analysis within a Bayesian criterion. BayesLCA package proposes to [19]:

- Cluster observations into groups.
- Perform inference on the model posterior.
- Report parameter estimates and posterior standard deviations.
- Provide plotting tools to visualize parameter behaviour and assess model performance.
- Provide summary tools to help in model selection and fit assessment.

The three main inferential functions in BayesLCA are:

1. an EM algorithm [40],
2. a Gibbs sampler [41] and,
3. a variational Bayes approximation [42].

Upon importing BayesLCA, two other packages are introduced called, **e1071** and **coda**. Where **e1071** offers functions for latent class analysis, support vector machines, fuzzy clustering, naive Bayes classifier, short-time Fourier transform, shortest path computation, bagged clustering, etc and **coda** provides functions for abstracting and outlining the output from Markov Chain Monte Carlo (MCMC) simulations and performs indicative tests of convergence to the equilibrium distribution of the Markov chain.

In this research, the EM algorithm has been utilised as an inferential function in BayesLCA.

Before proceeding further, the seed was set to 123, to ensure that the same results are obtained each time when the same process is run with the same seed.

The dataset was fit on EM LCA model with a default restart = 5 with a specification to form 6 clusters (See Figure 5.1). This was to determine if that random group number

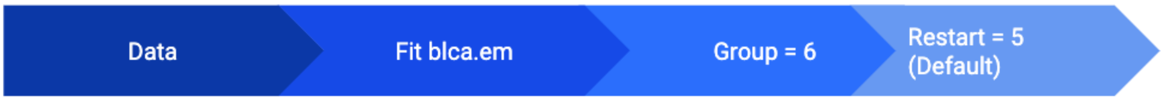


Figure 5.1: Model 1: Six Group Model

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.211 | 0.210 | 0.190 | 0.152 | 0.150 | 0.087 |

Table 5.1: Membership Probabilities, Six Groups Model

could convey any sort of clustering information regarding the student's behaviours and which students belong to them. The Group probability for 6 Group Model can be seen in Table 5.1.

The log-posterior in EM algorithms is increased at each iteration, hence it may converge to only a local maximum or saddle-point. To overcome this problem, [19] suggests restarting the algorithm multiple times from randomly picked starting values, this will allow the set of parameters to easily achieve the highest log-posterior value". Hence, following a similar solution from, restart = 20 was particularised to 6 (randomly selected) group model (See Figure 5.2). With just 20 restarts, the algorithm finds the new maxima at restart 17, as seen in Table 5.2.

The Group probabilities received after implementing Model 2, i.e 6 Group Model with restart = 20 can be seen in Table 5.3.

As per [19] while achieving the global maximum, if a sub-optimal set of estimates are inaccurately classified this possibly would lead to a very different interpretation of the dataset, probably pointing to a flawed investigation. Therefore, in this case, the authors recommend that it's better to run the algorithm multiple times in order to identify the optimal parameters accurately. Thus, as the above algorithm with 20 restarts found its global maxima only once, the group model is run again with 50

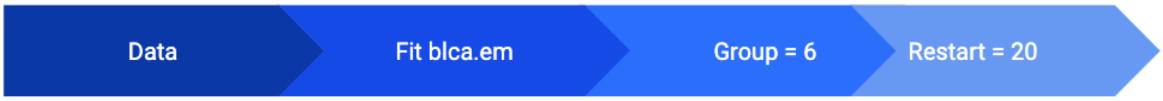


Figure 5.2: Model 2: Six Group Model with 20 restarts

| Restart = 20 |
|---|
| Restart number 1, logpost = -14868.65... |
| Restart number 2, logpost = -14934.02... |
| Restart number 3, logpost = -14899.29... |
| Restart number 4, logpost = -14926.44... |
| Restart number 5, logpost = -14950.08... |
| Restart number 6, logpost = -14914.64... |
| Restart number 7, logpost = -14894.42... |
| Restart number 8, logpost = -14918.59... |
| Restart number 9, logpost = -14878.31... |
| Restart number 10, logpost = -14910.9... |
| Restart number 11, logpost = -14884.37... |
| Restart number 12, logpost = -14924.95... |
| Restart number 13, logpost = -14951.48... |
| Restart number 14, logpost = -14905.96... |
| Restart number 15, logpost = -14895.11... |
| Restart number 16, logpost = -14936.81... |
| New maximum found... Restart number 17, logpost = -14803.52... |
| Restart number 18, logpost = -14844.18... |
| Restart number 19, logpost = -14876.57... |
| Restart number 20, logpost = -14965.64... |

Table 5.2: Log-Posterior Values with Restart = 20

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.230 | 0.229 | 0.176 | 0.164 | 0.138 | 0.064 |

Table 5.3: Membership Probabilities, Six Groups Model with Restart = 20

Data
Fit blca.em
Group = 6
Restart = 50

Figure 5.3: Model 3: Six Group Model with 50 restarts

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.229 | 0.216 | 0.180 | 0.163 | 0.148 | 0.064 |

Table 5.4: Membership Probabilities, Six Groups Model with Restart = 50

restarts to attain the correct optimal parameters (See Figure 5.3).

With 50 restarts the algorithm finds its highest log-posterior value multiple times ensuring non-erroneous research. The Group probability for 6 Group Model with restart = 50 can be seen in Table 5.4.

The log posterior and BIC values for all the three models defined above were compared and the best results were achieved with the third model where the restart was set to 50, both concerning log-posterior as well as BIC. Therefore, model 3 was used for further implementations. Table 5.5 to be referred for results.

There are 2 approaches to specify starting values in BayesLCA :

- single: Using this each unique data point is randomly assigned membership to a single class. It is the default method.
- across: Using this class membership is randomly assigned across groups with respect to a uniform distribution.

Both of these methods mentioned above were implemented on Model 3 and on Model 1 to get understand the contrast they produce in the results.

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| | Restart 5 | Restart 20 | Restart 50 |
| Log-posterior | -14834.60 | -14803.52 | -14802.27 |
| BIC | -31907.51 | -31845.34 | -31842.84 |

Table 5.5: Different models are compared considering log-posterior and BIC as criterion for Six Groups Model

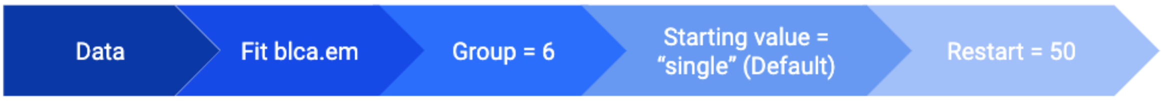


Figure 5.4: Model 4: Six Group Model with 50 restarts and stating value set to "single"

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.230 | 0.218 | 0.175 | 0.168 | 0.146 | 0.064 |

Table 5.6: Membership Probabilities, Six Groups Model with Restart = 50 and start.vals = "single"

Model 3 was fine-tuned with an addition of an extra parameter, where starting value is set to "single" (Figure 5.4). The Group Probabilities for the same can be seen in Table 5.6.

Moving forward, the same Model 3 was re-trained but this time with starting value as "across" (Figure 5.5). The Membership Probabilities achieved by this model can be seen in Table 5.7.

Just to cross-check if there is any need to perform restart along with starting values, 2 different models were trained, considering Model 1 as the base model.

A new model, Model 6 was trained with starting values = "single", keeping Model 1 as the base, this can be seen in Figure 5.6, along with that the membership probabilities for this model can be seen in Table 5.8.

Once again, a similar procedure was repeated on Model 7 by changing the starting value to "across", with Model 1 being the base model. Refer Figure 5.7 for clarity. The Group membership probabilities for Model 7 can be seen in Table 5.9.

Comparison of the log-posterior values for all four models (4,5,6,7) is done. The outcomes in Table 5.10 make is very explicit that accepting the "single" method as

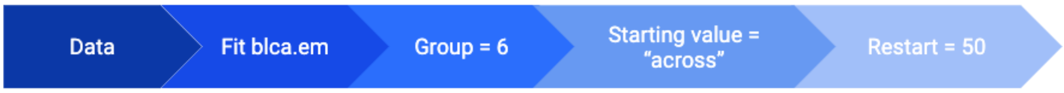


Figure 5.5: Model 5: Six Group Model with 50 restarts and starting value set to "across"

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.221 | 0.207 | 0.178 | 0.177 | 0.153 | 0.064 |

Table 5.7: Membership Probabilities, Six Groups Model with Restart = 50 and start.vals = ”across”

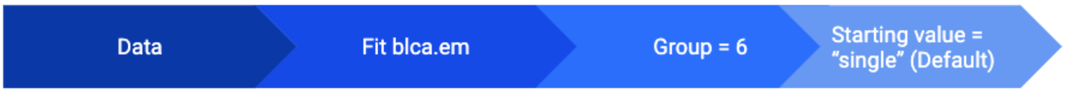


Figure 5.6: Model 6: Six Group Model with starting value set to ”single”

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.229 | 0.198 | 0.181 | 0.179 | 0.147 | 0.066 |

Table 5.8: Membership Probabilities, Six Groups Model with start.vals = ”single”

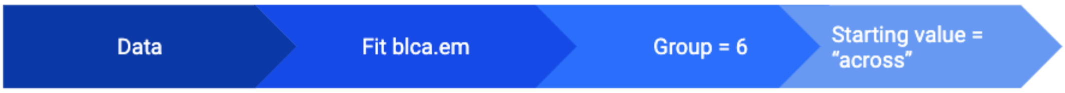


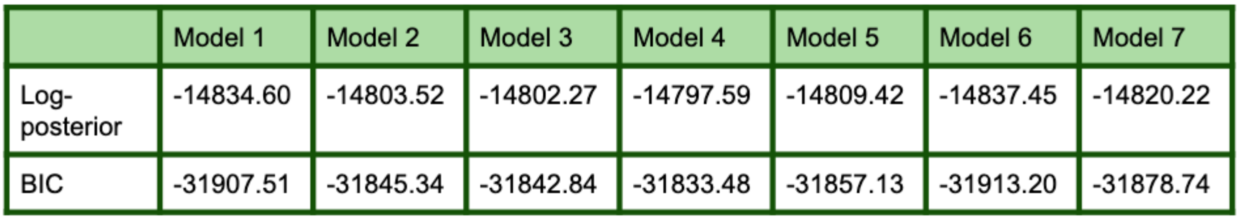
Figure 5.7: Model 7: Six Group Model with starting value set to ”across”

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Membership Probabilities: | 0.212 | 0.208 | 0.199 | 0.155 | 0.148 | 0.079 |

Table 5.9: Membership Probabilities, Six Groups Model with start.vals = ”across”

| | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|
| | start.vals = "single" | start.vals = "across" | start.vals = "single | start.vals = "across" |
| | Restart = 50 | Restart = 50 | | |
| Logpost | -14797.59 | -14809.42 | -14837.45 | -14820.22 |

Table 5.10: Membership Probabilities, Six Groups Model Comparison with and without starting value



| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Log-posterior | -14834.60 | -14803.52 | -14802.27 | -14797.59 | -14809.42 | -14837.45 | -14820.22 |
| BIC | -31907.51 | -31845.34 | -31842.84 | -31833.48 | -31857.13 | -31913.20 | -31878.74 |

Figure 5.8: Comparison of 7 Models

starting value provides a better fit with 50 restarts.

Merely for better transparency on results, the Log posterior and BIC values for all the 7 models were examined (Figure 5.8).

The comparison Figure 5.8 indicates that Model 4 (with starting value as single and restart = 50) appeared to provide the most reliable results.

In accordance to the results obtained above, Model 4 group probabilities are closely looked through and it vividly signifies that group 1,2,3 and 4 provide strong grouping information as the total number of students are almost equally distributed amongst the four groups, with Group 1 comprising of the 23%, Group 2 consists of 21.8%, Group 3 with 17.5% and Group 4 with 16.8% of the total students, although Group 5 and 6 fail to provide strong grouping information, as only 14.6% and 6.4% of the cumulative students fall into these groups respectively. Consequently, examining this it might be assumed that, there is a possibility that group 5 and 6 students might fall in either of the groups between 1 and 4.

To support my inference above, I took inspiration from [1] and applied Bayesian Latent Class Analysis to the anonymised data, over a range of numbers of groups:

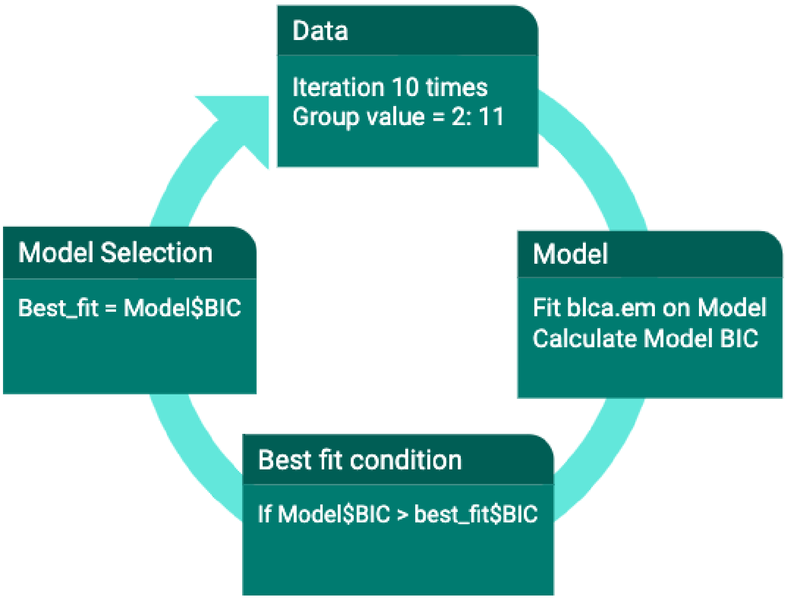


Figure 5.9: Model Selection Process

| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Membership Probabilities: | 0.333 | 0.288 | 0.277 | 0.102 |

Table 5.11: Membership Probabilities, Four Groups Model after Model Selection

$G = \{1, ....., 10\}$ and allowed the Bayesian Information Criterion (BIC) to pick the best model (Figure 5.9).

Figure 5.10 shows the various BIC values obtained by each Group Model amongst the ten iterations.

Using BIC approach, the optimal number of the group was found to be G = 4 as it attains the highest BIC value. Group Probabilities can be seen in Table 5.11.

The completion of the above code took some extra time, which inspired the question of this research ; "Are all these variables essential for clustering?" Does each of these 59 variables provide essential clustering information ?"

This was an essential question that required an answer, to save both time and cost. While researching through various papers I came across [2], taking guidance from which Variable Selection Selection Method for Latent Class Analysis was applied on the dataset of students taking Data Analysis for Decision Makers Module at Business School of University College Dublin. The Headlong search algorithm based upon

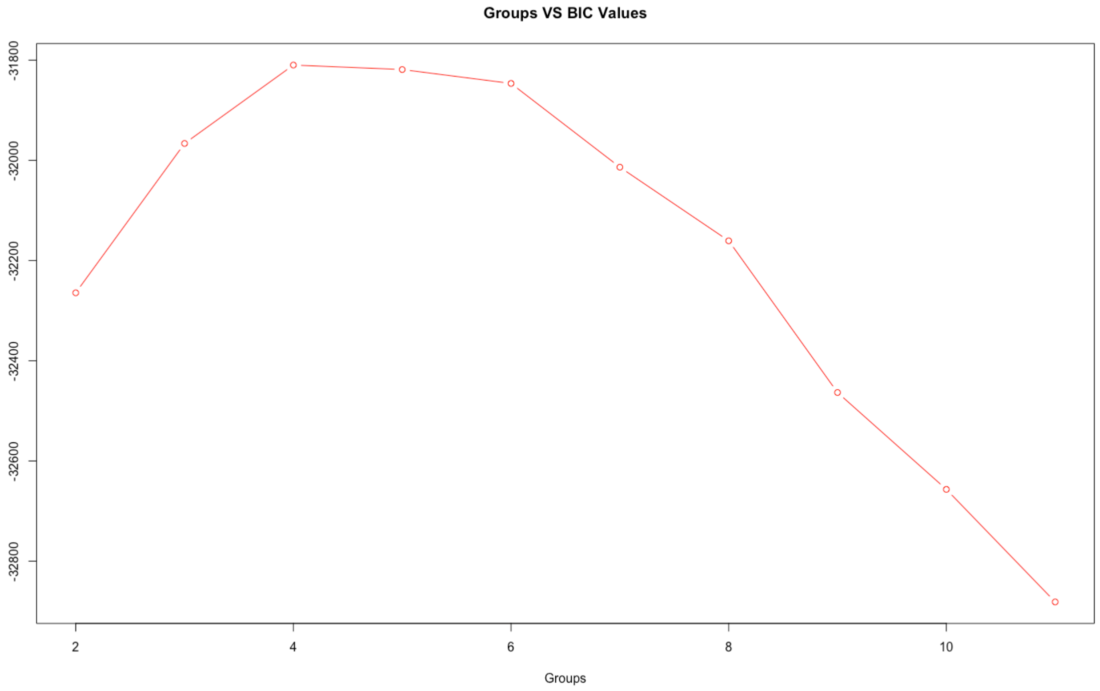


Figure 5.10: Groups VS BIC Values

global independence assumption was fit on the dataset for inclusion and exclusion of variables, with 10 iterations to find the most important variables along with the best fit model (Figure 5.11). For the implementation of this **LCAVarsel** library was used which performs Variable selection for latent class analysis for model-based clustering of multivariate categorical or binary data.

Applying this method to the students data set, achieved 5 Cluster model, see Table 5.12 for membership probabilities.

In order to perform a comparative study, Swap-Stepwise Selection Algorithm was

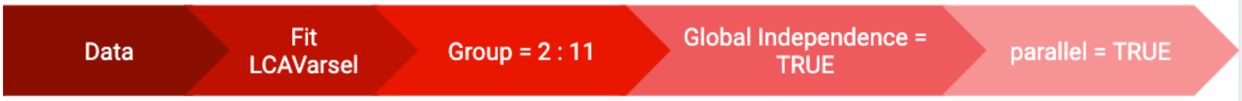


Figure 5.11: Variable Selection using Headlong Search Algorithm

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 |
|---|---|---|---|---|---|
| Membership Probabilities: | 0.228 | 0.061 | 0.240 | 0.281 | 0.189 |

Table 5.12: Membership Probabilities, Five Groups Model using Headlong Search Algorithm

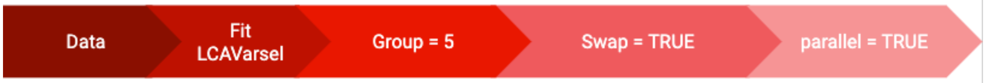


Figure 5.12: Variable Selection using Swap-Stepwise Selection Algorithm

implemented (Figure 5.12)

The results from these implementation have been discussed in the next chapter.

# Chapter 6

# Discussion

[1] stated that majority of DADM students were initial stage students, with 87% entering directly from the Irish secondary school system in the 2013/2014 academic year and most of them had never come across a blended learning environment before taking the DADM module. Therefore, it appears it would be fascinating to examine patterns in such students dataset, as it may provide unbiased results due to the rawness of the students' exposure to such an environment.

Therefore, plotting the entire proportion of learners Vs their attendance (Figure 3.1) in terms of attending lectures or tutorials and accessing various online materials weekly elucidated that:

- lecture attendance of students considerably falls from above 80% to as low as 40%
- tutorial attendance goes through various ups and downs as it witnesses its peaks during Week 5, 7 and 9 which coincides with the Continuous Assessment submission weeks,
- students fail to engage in accessing materials taught on the same week, especially in-between week 3 to 10.
- More than 30% of the students consistently preferred downloading printable pdfs to be brought to the class, although a hike in the number of downloads has been noticed in Week 7, which clashes with the open book based first MCQ Test.

- More than 20% of the students regularly login into Blackboard, aloof of the fact that this percentage went up to as high as 85% in Week 6 and 72% in Week 11 and then dropped back to 20% in Week 12.

The best performing model out of the variation of 7 different six groups models was visualized. The different student behaviour discovered for Six Group Model with 50 restarts along with starting value set to single can be seen in Figure 6.1. Looking at plots for various learning resources, one can clearly state that there are a few Group which completely overlap with each other like for Lecture and Tutorial attendance, group 2, 3 and 5 completely shadow each other for the first 3 weeks and show somewhat similar patterns over the span of twelve weeks. Each of these group maintains a Lecture attendance of above 50% staring from 90%, whereas they keep a record of above 60% Tutorial attendance, while the for the other 3 groups the lecture and tutorial attendance fell as low as below 10%. All of these 6 groups showed similar patterns in engaging with the online materials from Week 3 to Week 10. With Group 3, 4 and 6 showed similar behaviour in accessing the scheduled online material during the first three weeks. Although, each group showed varying behaviour when students behaviour over downloading pdfs to bring to the next lecture was analysed. Moving on to analyzing learners behaviours about accessing any study material on Blackboard, it noticed that Group 3 and Group 5 show at a lot of similarity in the 12 weeks tenure, especially between week 3 to 7.

Studying the results of six group model it appeared like a smaller group model would provide better results as few of the groups were overlapping each other. Henceforth, when the Bayesian Information Criterion was used to find out the most suitable group model, it gave Group = 4, as the best model. See Figure 6.2 for Behaviour of students with different Learning Resources using 4 Group Model.

These four clusters were further drilled to identify learner behavior in detail.

### 6.0.1 Interpreting the Clusters for 4 Group Model

**Group 1 : High Motivation to Low Interest**

Group 1 comprised 33.3% of the total number of students, thereby forming the largest group. Their Lecture attendance considerably falls with the passage of week from

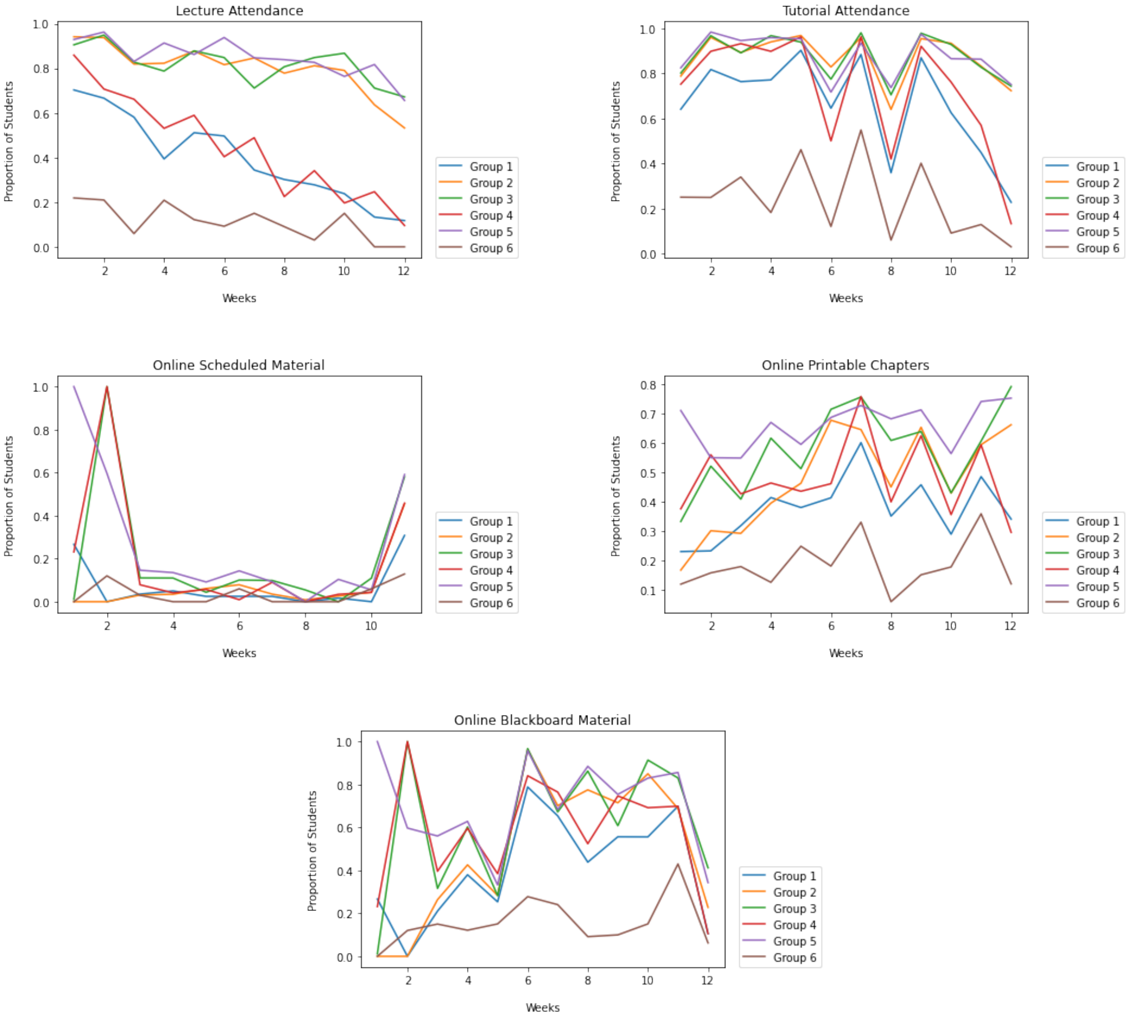


Figure 6.1: Behaviour of students with different Learning Resources using 6 Group Model

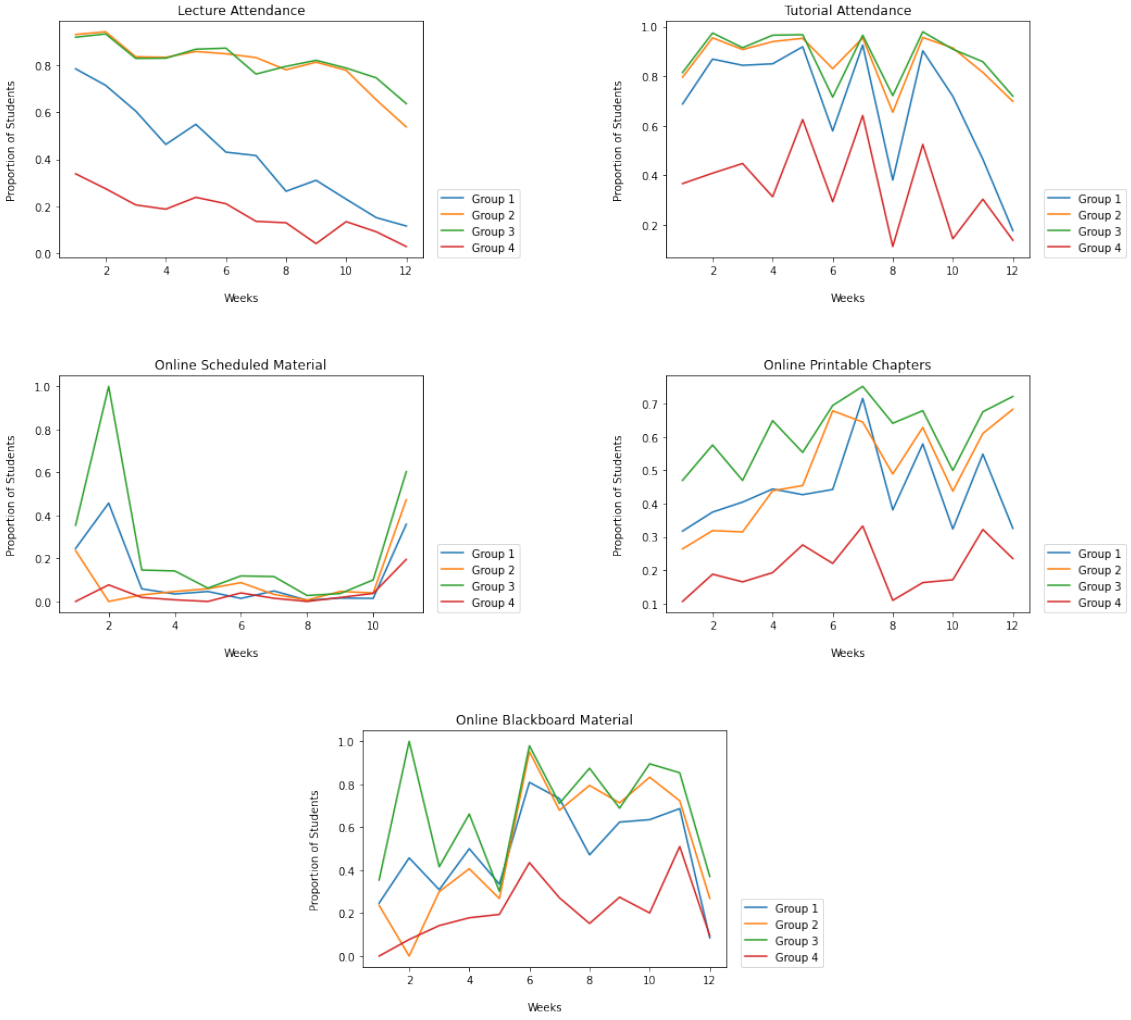


Figure 6.2: Behaviour of students with different Learning Resources using 4 Group Model

above 70% to as low as below 20%, whereas their tutorial attendance witnesses various hikes and lows especially during Week 5, 7 and 9, the students' attendance equals close to 85% or more. Along with the tutorial, 50% or more of these students seem to download printable scheduled material during Week 7, 9 and 11. Group 1 turns out to be the second-highest group in accessing online scheduled material during the initial week, apart from this they eventually started accessing the full online material on Blackboard as well, but mostly after Week 5, the pattern seems to change. Maybe a practical argument for this gathering can be that probably had high motivations during the initial stage, as they manage to maintain above-average attendance in accessing all of the learning resources, yet they struggle to maintain this pattern throughout the course.

### Group 2: Principle Attenders

Group 2 maintains a high attendance proportion throughout the course for nearly all the learning resources on an aggregate. The lecture and tutorial attendance ratio for this group drops with time, yet they maintain an attendance of more than 50% for both the learning resources. Like other groups, this group also fails to access online material resourcefully, although there has been a slight increase in their accessing online scheduled material in between week 2 to 6. After week 5, this group exhibits interest in accessing printable pdfs as well as accessing other resources more often on Blackboard when compared to the first few weeks. This pattern hints that students were more principled in their attendance.

### Group 3: Principle Attenders with early online adopters

Group 3 shows quite similar patterns of behaviour in case of Lecture and Tutorial attendance when compared to Group 2. A bit of similarity can also be noticed between these after Week 5 and Week 7 for accessing any online material on the blackboard and downloading printable chapters respectively. Out of all the four group students, students in this group showed the highest rate of engagement with online materials, including close to 100% of the students accessing online material during Week 2, and maintaining a percentage of above 10 throughout the course. This set of students likely mark early eLearning adopters.

### Group 4: Lack of Engagement

This group manifested the lowest levels of engagement in relation to all learning resources. The highest lecture and tutorial attendance for this set of students is 40% which falls to as low as below 5%. The overall highest attendance, witnessed by this group of students is close to 60% for tutorials during Week 5 and 7. The group starts engaging with printable chapters and any blackboard material after Week 6, which if compared with other groups is still too low. This group possibly demonstrates a lack of engagement with learning resources throughout the course.

The item probabilities against each variable were plotted for the 4 group model (Figure 6.5), where 0 indicated less importance over 1 denoting very high importance. The size of each block was based upon the size of the cluster. Block 1, 2 and 3 can be seen of a similar size whereas Block 4 is the smallest as it contains only 10.2% of the total students' data. This plot also implies variables based upon the colour palette, with an increase in the shade of the colour the importance of the variable decreases. Hence, studying the plot it can be stated that variables from 27 to 34 don't seem to carry useful clustering information. This along with the complexity of previous code lead to the formation of the next research question.

In order to provide an answer to the research question, if all 59 variables were necessary for clustering, Headlong search Algorithm was implemented which resulted in dropping of 17 variables out of 59 variables. Figure 6. shows the list of variables that were dropped by Headlong Search Algorithm. The variable selection method was successful in the removal of variables from 24 to 34, i.e Online Week 3 to 10. Not only this the algorithm received better accuracy by removing other unnecessary variables as well.

The Headlong search algorithm lead to the formation of 5 group model, where Group 4 formed the largest group with 28.1% of the students, followed by Group 3 which comprised of the 24% of the students, with Group 1 being next in line with 22.8 of the students, Group 5 was comparatively small with 18.9% of the students, while Group 2 was less than half of Group 5 with only 6% of the students, turning to be the smallest of all the group.

The interaction of the students with various learning resources after the removal of unnecessary or redundant variables can be seen in Figure 6.5.

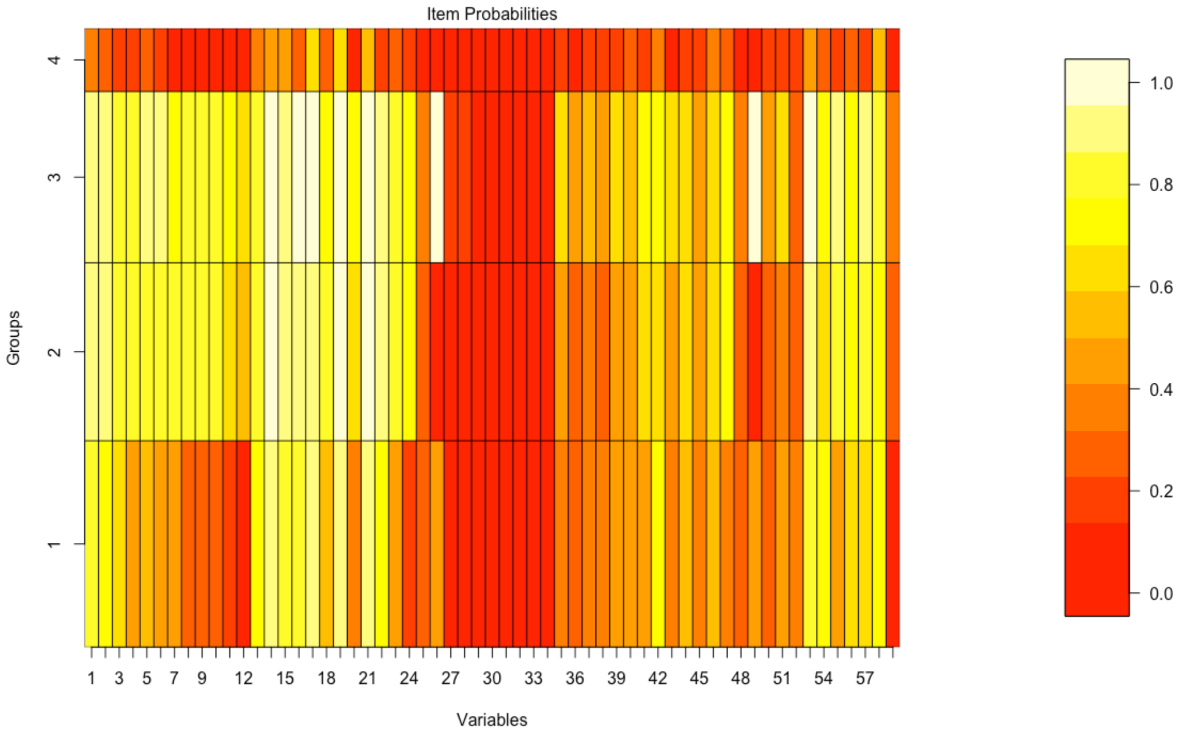
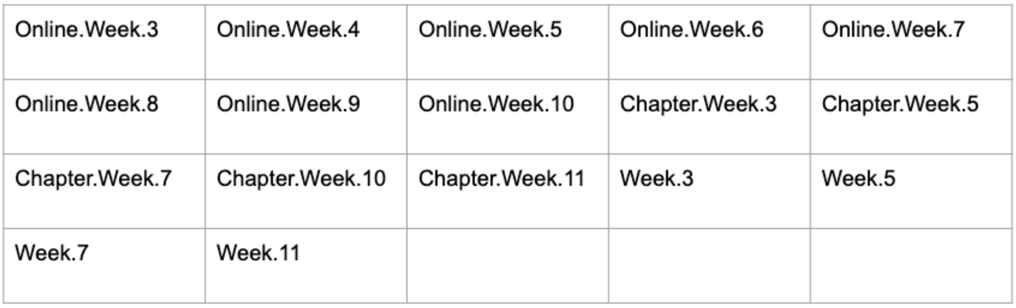
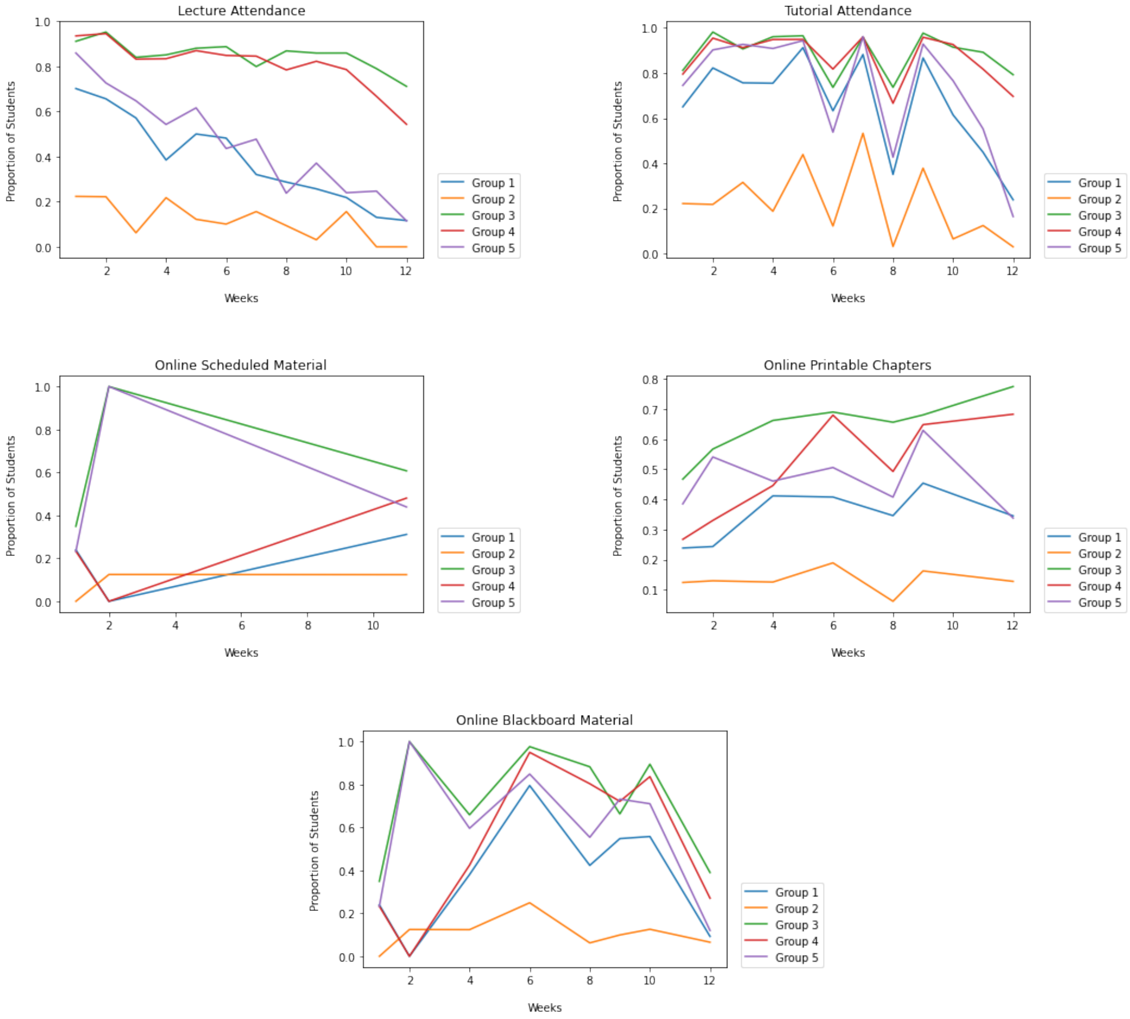


Figure 6.3: Variables VS Groups



| | | | | |
|---|---|---|---|---|
| Online.Week.3 | Online.Week.4 | Online.Week.5 | Online.Week.6 | Online.Week.7 |
| Online.Week.8 | Online.Week.9 | Online.Week.10 | Chapter.Week.3 | Chapter.Week.5 |
| Chapter.Week.7 | Chapter.Week.10 | Chapter.Week.11 | Week.3 | Week.5 |
| Week.7 | Week.11 | | | |

Figure 6.4: Dropped Variables from Headlong Search Algorithm



Lecture Attendance

Tutorial Attendance

Online Scheduled Material

Online Printable Chapters

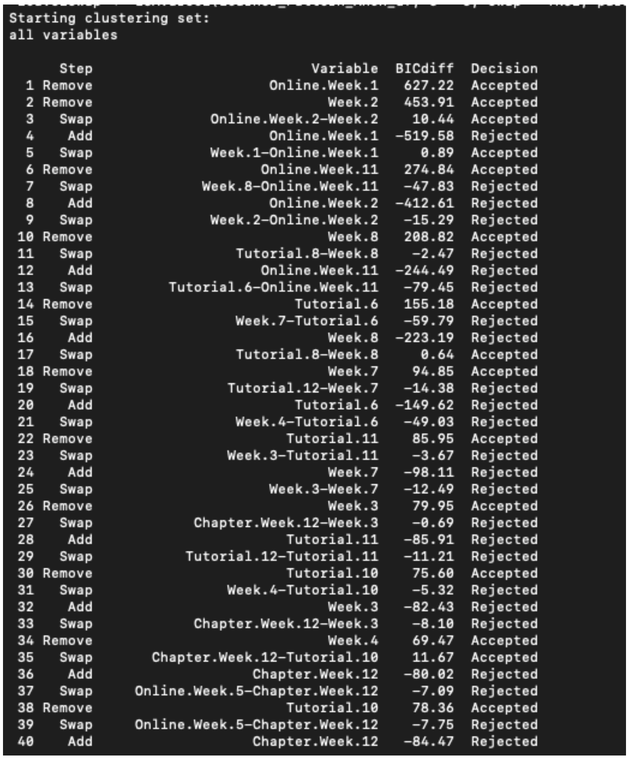
Online Blackboard Material

Figure 6.5: Behaviour of students with different Learning Resources after Variable Selection using Headlong Search Algorithm

There weren't many noticeable changes amongst the groups for Lecture and Tutorial attendance, as no column had been dropped from either of them. It can be deduced from this that all of those variables comprised of the most valuable information for clustering. Moving ahead with the other resources, we can see very clean defined plots in comparison to the previous plots (Figure 6.1 / Figure 6.2) for online scheduled material, printable pdfs or blackboard logins. Different patterns of students behaviour can be identified in accessing online scheduled material, Group 1 and 4 exhibits similar behaviour and so does Group 3 and 5, where it is noticeable that engagement of Group 1 and 4 with scheduled online material increases with time, whereas for Group 3 and 5 it decreases with time, although Group 2 exhibits constant pattern from Week 2 till the end of the course. Similarly, distinct five groups can be noticed when printable pdfs and blackboard logins are monitored. As noted in previous models, Group 3 consistently outperforms all other models in accessing printable pdfs throughout the course. The least involvement of Group 2, with all the learning resources, is evidently visible after variable selection.

In order to perform a comparative study, Swap - Stepwise search algorithm was implemented, unfortunately, results for which couldn't be produced as it was computationally expensive. A small screenshot for the inconclusive code has been added for reference.

This last part of this work would be considered for future work.



```
Starting clustering set:
all variables

      Step                          Variable  BICdiff  Decision
  1 Remove                     Online.Week.1   627.22  Accepted
  2 Remove                            Week.2   453.91  Accepted
  3   Swap              Online.Week.2-Week.2    10.44  Accepted
  4    Add                     Online.Week.1  -519.58  Rejected
  5   Swap              Week.1-Online.Week.1     0.89  Accepted
  6 Remove                    Online.Week.11   274.84  Accepted
  7   Swap             Week.8-Online.Week.11   -47.83  Rejected
  8    Add                     Online.Week.2  -412.61  Rejected
  9   Swap              Week.2-Online.Week.2   -15.29  Rejected
 10 Remove                            Week.8   208.82  Accepted
 11   Swap                 Tutorial.8-Week.8    -2.47  Rejected
 12    Add                    Online.Week.11  -244.49  Rejected
 13   Swap         Tutorial.6-Online.Week.11   -79.45  Rejected
 14 Remove                        Tutorial.6   155.18  Accepted
 15   Swap                 Week.7-Tutorial.6   -59.79  Rejected
 16    Add                            Week.8  -223.19  Rejected
 17   Swap                 Tutorial.8-Week.8     0.64  Accepted
 18 Remove                            Week.7    94.85  Accepted
 19   Swap                Tutorial.12-Week.7   -14.38  Rejected
 20    Add                        Tutorial.6  -149.62  Rejected
 21   Swap                 Week.4-Tutorial.6   -49.03  Rejected
 22 Remove                       Tutorial.11    85.95  Accepted
 23   Swap                Week.3-Tutorial.11    -3.67  Rejected
 24    Add                            Week.7   -98.11  Rejected
 25   Swap                     Week.3-Week.7   -12.49  Rejected
 26 Remove                            Week.3    79.95  Accepted
 27   Swap            Chapter.Week.12-Week.3    -0.69  Rejected
 28    Add                       Tutorial.11   -85.91  Rejected
 29   Swap           Tutorial.12-Tutorial.11   -11.21  Rejected
 30 Remove                       Tutorial.10    75.60  Accepted
 31   Swap                Week.4-Tutorial.10    -5.32  Rejected
 32    Add                            Week.3   -82.43  Rejected
 33   Swap            Chapter.Week.12-Week.3    -8.10  Rejected
 34 Remove                            Week.4    69.47  Accepted
 35   Swap       Chapter.Week.12-Tutorial.10    11.67  Accepted
 36    Add                   Chapter.Week.12   -80.02  Rejected
 37   Swap     Online.Week.5-Chapter.Week.12    -7.09  Rejected
 38 Remove                       Tutorial.10    78.36  Accepted
 39   Swap     Online.Week.5-Chapter.Week.12    -7.75  Rejected
 40    Add                   Chapter.Week.12   -84.47  Rejected
```

Figure 6.6: Screenshot of Swap-Stepwise Search Algorithm implementation

# Chapter 7

# Conclusion

***Research objectives revisited:***

The patterns of behaviours show that some groups adapt well and transition to self-directed learning over the semester. Figure 3.1 shows the students are more enthusiastic about attending lectures during the start of the semester, while the percentage of students that accessed the online materials at the same time, as they were being covered in lectures is consistently low, below 20%, except for weeks 2 and 11. However, if the proportion of students who access online material regardless of whether the content matches the scheduled lecture content are considered, the rate of engagement steadily grows for the most part, by contrast to the lecture and tutorial levels. The same holds for students accessing the printable material. Although, the students' login into Blackboard more often during Week 6 and 11 as they probably mark the Weeks before a test or an examination. Students seem to use printable copies during their MCQ tests, especially during test 1. Although during MCQ2, there was a certain amount of drop-in downloading of pdfs. This is consistent with a lag between the material being covered at lectures and that being accessed by students at a later time, as it possibly indicates a lack of interest in understanding or revisiting topics taught in class. Although students tend to seek help from the tutors mostly during the time of submissions as the number of students attending tutorials just before the submission weeks are relatively high.

After performing variable selection, inspecting the behaviour of students using 5 cluster model, we see that Group 3 consistently accessed all the learning resources, their attendance went up as high as above 95%, while, Group 2 had the lowest levels of activity of the five groups. In terms of online activity, Groups 1 and 5 behaved similarly, except in case of scheduled online material, where engagement of Group 5 and Group 1 with online material increased and decreased with time respectively. Group 4 had high lecture and tutorial attendance as well as eventually starts adopting the blended learning approach by accessing all the online resources with time as seen by Principle attenders with online adopters group in 4 cluster Model. This research hence concludes that the majority of the students eventually start adopting the blended learning approach, and to deduce this results all variables were not necessary. Hence, the variable selection method seems to be effective for this research.

***Learning outcomes:***

This research helped me increase my understanding of Unsupervised learning and helped me gain a deep understanding of various clustering algorithms like Model-Based and Iterative, accompanied by various Variable selection Methods for Latent Class Analysis. Using an R environment was as well as a brand new instance for me, as I had completed various machine learning tasks using Python in past. Hence, Learning and Working with different R libraries was an add on learning experience during this research.

***Future work:***

In future, I would like to use students educational scores data to analyse if these behavioural patterns of the students relate to the levels of attainment in terms of learning outcomes. I would even like to consider some external factors to the ecosystem such as gender and prior educational attainment for the purpose of analysing if this has any connection with the patterns they exhibit. I would even like to continue working on my research for performing a comparative study of the two different variable selection methods, one that was based on the assumption of global independence (Implemented in this paper) and other that is based upon relaxing the independence assumption (Swap-Stepwise Selection algorithm).

The results from this research can be further used by tutors or researchers to ease the transfer of knowledge between the educator and the student, by creation of student specific content, which caters to individual students need or by development of Early Warning Systems, which alerts the students during their mid semesters to improve their performance if necessary.

# Bibliography

[1] P. Carroll and A. White, “Identifying patterns of learner behaviour: what business statistics students do with learning resources,” *INFORMS Transactions on Education*, vol. 18, no. 1, pp. 1–13, 2017.

[2] N. Dean and A. E. Raftery, “Latent class analysis variable selection,” *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 1, p. 11, 2010.

[3] J. H. Badsberg, “Model search in contingency tables by coco,” in *Computational statistics*, pp. 251–256, Springer, 1992.

[4] M. Fop, K. M. Smart, T. B. Murphy, *et al.*, “Variable selection for latent class analysis with application to low back pain diagnosis,” *The Annals of Applied Statistics*, vol. 11, no. 4, pp. 2080–2110, 2017.

[5] S. Graf, T.-C. Liu, *et al.*, “Identifying learning styles in learning management systems by using indications from students’ behaviour,” in *2008 eighth ieee international conference on advanced learning technologies*, pp. 482–486, IEEE, 2008.

[6] N. Gordon, “Flexible pedagogies: Technology-enhanced learning,” *The Higher Education Academy*, pp. 1–24, 2014.

[7] R. Carver, M. Everson, J. Gabrosek, G. Rowell, N. Horton, R. Lock, M. Mocko, A. Rossman, P. Velleman, J. Witmer, *et al.*, “Guidelines for assessment and instruction in statistics education (gaise) college report draft,” 2016.

[8] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[9] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[10] M. Fop, T. B. Murphy, *et al.*, "Variable selection methods for model-based clustering," *Statistics Surveys*, vol. 12, pp. 18–65, 2018.

[11] D. Bartholomew, "Knott (1999)," *Latent variable models and factor analysis.*

[12] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[13] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

[14] C. C. Clogg, "Latent class models for measuring," in *Latent trait and latent class models*, pp. 173–205, Springer, 1988.

[15] E. S. Garrett and S. L. Zeger, "Latent class model diagnosis," *Biometrics*, vol. 56, no. 4, pp. 1055–1067, 2000.

[16] E. A. Erosheva, S. E. Fienberg, and C. Joutard, "Describing disability through individual-level mixture models for multivariate binary data," *The annals of applied statistics*, vol. 1, no. 2, p. 346, 2007.

[17] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang, and C. Lin, "e1071: Misc functions of the department of statistics (e1071), tu wien," *R package version*, vol. 1, no. 3, 2014.

[18] D. A. Linzer, J. B. Lewis, *et al.*, "polca: An r package for polytomous variable latent class analysis," *Journal of statistical software*, vol. 42, no. 10, pp. 1–29, 2011.

[19] A. White and T. B. Murphy, "Bayeslca: An r package for bayesian latent class analysis," *Journal of Statiscal Software*, vol. 61, no. 13, pp. 1–28, 2014.

[20] R. Bellman, "E. 1957. dynamic programming," *Princeton UniversityPress. BellmanDynamic programming1957*, p. 151, 1957.

[21] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[22] C. Bouveyron and C. Brunet-Saumard, "Discriminative variable selection for clustering with the sparse fisher-em algorithm," *Computational Statistics*, vol. 29, no. 3-4, pp. 489–513, 2014.

[23] E. B. Fowlkes, R. Gnanadesikan, and J. R. Kettenring, "Variable selection in clustering," *Journal of classification*, vol. 5, no. 2, pp. 205–228, 1988.

[24] A. E. Raftery and N. Dean, "Variable selection for model-based clustering," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 168–178, 2006.

[25] F. Bartolucci, G. E. Montanari, and S. Pandolfi, "Item selection by latent class-based methods: an application to nursing home evaluation," *Advances in Data Analysis and Classification*, vol. 10, no. 2, pp. 245–262, 2016.

[26] "Ucd_dadm." `https://sisweb.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=MODULE&MODULE=MIS10090`.

[27] J. Smith, "Articulate (2013) articulate storyline.," *Articulate*, vol. 12, 2013.

[28] R. E. Mayer, "Multimedia learning," in *Psychology of learning and motivation*, vol. 41, pp. 85–139, Elsevier, 2002.

[29] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educational psychologist*, vol. 38, no. 1, pp. 43–52, 2003.

[30] G. J. McLachlan and D. Peel, *Finite mixture models.* John Wiley & Sons, 2004.

[31] P. F. Lazarsfeld, "The logical and mathematical foundation of latent structure analysis," *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pp. 362–412, 1950.

[32] P. Lazersfeld, "The interpretation and computation of some latent structures," *Measurement*, 1950.

[33] P. F. Lazarsfeld and N. W. Henry, *Latent structure analysis.* Houghton Mifflin Co., 1968.

[34] G. McLachlan and T. Krishnan, "The em algorithm and extensions, vol. 382 john wiley & sons," *Hoboken, New Jersey.[Google Scholar]*, 2008.

[35] L. A. Goodman, "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, vol. 61, no. 2, pp. 215–231, 1974.

[36] G. Arminger, C. C. Clogg, and M. E. Sobel, *Handbook of statistical modeling for the social and behavioral sciences.* Springer Science & Business Media, 2013.

[37] J. A. Hagenaars and A. L. McCutcheon, *Applied latent class analysis.* Cambridge University Press, 2002.

[38] R. Kass and A. Raftery, "Bayes factors, in journal of the american statistical association," 1995.

[39] G. J. McLachlan and S. Rathnayake, "On the number of components in a gaussian mixture model," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 5, pp. 341–355, 2014.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[41] S. F. Arnold, "18 gibbs sampling," 1993.

[42] J. T. Ormerod and M. P. Wand, "Explaining variational approximations," *The American Statistician*, vol. 64, no. 2, pp. 140–153, 2010.