

https://dspace.org/

e-Behaviour, personality and academic performance analysis for intervention

Seota, Serepu Bill-William

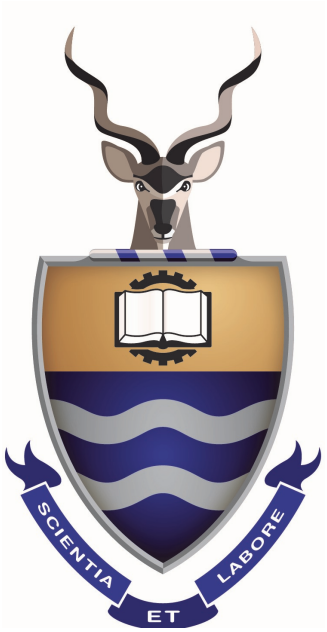
2021
https://hdl.handle.net/10539/32249

Downloaded from a DSpace repository. https://dspace.org/

e-Behaviour, Personality and Academic Performance Analysis for Intervention

Author:
Serepu Bill-William SEOTA
(1043377)

Supervisor:
Dr Richard KLEIN and
Prof Terence VAN ZYL



SCHOOL OF COMPUTER SCIENCE AND APPLIED MATHEMATICS

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Ethics Clearance Protocol Number: H19/06/36

June 28, 2021

Declaration of Authorship

I, Serepu Bill-William SEOTA (1043377), declare that this dissertation titled, "e-Behaviour, Personality and Academic Performance Analysis for Intervention" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 28 June 2021

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Abstract

Faculty of Science
School of Computer Science and Applied Mathematics

Master of Science

e-Behaviour, Personality and Academic Performance Analysis for Intervention

by Serepu Bill-William SEOTA (1043377)

Post-secondary educational institutions have a concerningly low rate of completion. Problems include a lack of student support, a student's social, cultural and economic background, and inability to adapt to the curriculum. This study provides insight into the relationship between e-Behaviour, personality and performance. Student performance analysis involves data modelling that enables the formulation of hypotheses and insights about student behaviour and personality. While the automation of such data pipelines is efficient and economical, it still requires a thorough analysis of data and the results obtained from its usage. To achieve the exploratory data analysis and prediction objectives of this research, the algorithms used are Regression Analysis, Decision-Tree, Support Vector Machine, k-Means clustering, k-Nearest Neighbours, and the Long Short-Term Memory. Our procedures provide methodology to timeously identify students who are likely to become at risk of poor academic performance. For the education sector, this study is valuable because it presents an approach to examining the extrinsic influences – e-Behaviour and personality – on performance. We extract online behaviours as proxies to Extraversion and Conscientiousness, which have been proven to correlate with academic performance. The proxies of Extraversion and Conscientiousness yield significant ($p < 0.05$) population correlation coefficients for the personality traits against grade: 0.846 for Extraversion and 0.319 for Conscientiousness. Using engineered e-Behaviour and personality features, we obtained a classification accuracy (κ) of students at risk of 0.51. Lastly, we design an intervention process that supplements existing performance analysis and intervention methods.

Acknowledgements

I would like to thank my supervisors, Dr Richard Klein and Prof Terence van Zyl, for their support and insightful guidance and comments during my research. To my family: thank you for your support and encouragement throughout my studies.

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

1 Introduction 2
- 1.1 Research Motivation . . . 3
- 1.2 Research Questions . . . 3
- 1.3 Research Aims and Objectives . . . 3
- 1.4 Assumptions on Data Credibility . . . 4
- 1.5 Contributions of Research . . . 4
- 1.6 Research Outline . . . 5

2 Review of Literature 6
- 2.1 Bourdieu's Three Forms of Capital and Student Success . . . 6
  - 2.1.1 Bourdieu's Three Forms of Capital – Definition . . . 6
  - 2.1.2 Social Capital . . . 7
  - 2.1.3 Cultural Capital . . . 7
  - 2.1.4 Economic Capital . . . 8
- 2.2 Five-Factor Personality Model . . . 8
- 2.3 Related Work . . . 9
  - 2.3.1 Log File Studies for User behaviour . . . 9
    - 2.3.1.1 Guidelines for Dealing with Log Data . . . 10
    - 2.3.1.2 Problems Faced when Dealing with Log Data . . . 10
  - 2.3.2 Machine Learning in Education . . . 10
    - 2.3.2.1 Student Performance Analysis and Prediction . . . 10
    - 2.3.2.2 Personalised Education . . . 11

3 Methodology Overview 13
- 3.1 Data Usage and Experimental Design . . . 13
- 3.2 The Four Criteria – Feature Engineering and Selection . . . 14
- 3.3 Interpretable Models and Algorithms . . . 14
- 3.4 Outline and Usage of Data . . . 15
  - 3.4.1 Description of Dataset – Moodle Data . . . 15
  - 3.4.2 Data and Model Flow . . . 16
- 3.5 Encoding Performance . . . 17
  - 3.5.1 Behaviour, Personality and Performance . . . 18
    - 3.5.1.1 Importance and Choice of Personality Traits: . . . 18

3.5.1.2 Encoding Personality Traits: . . . 18
3.6 Algorithms for e-Behaviour, Personality and Performance Analysis . 19
3.6.1 Decision Tree Classifier . . . 19
3.6.1.1 Architecture . . . 19
3.6.1.2 Gini Impurity Index – Decision Factors . . . 19
3.6.2 Ordinary Least Squares Linear Regression Analysis . . . 20
3.6.3 Validity of OLS Regression Models . . . 21
3.6.4 Time Series Analysis . . . 22
3.6.4.1 Autocorrelation . . . 22
3.7 Long Short-Term Memory . . . 23
3.7.1 LSTM Problem Design . . . 25
3.8 Evaluation Metrics for Student Risk Classification . . . 26
3.8.1 The Overall Accuracy of a Model . . . 27
3.9 Contribution to Existing Evaluation Systems . . . 28

4 Student Background and performance **31**
4.1 Background Features and Data Split . . . 31
4.1.1 Classifying a Student based on Background Data . . . 33
4.2 Conclusion . . . 34

5 Forum Activity and performance **35**
5.1 An Overview of Independent and Dependent Variables . . . 36
5.2 Extraversion and Grade . . . 38
5.2.1 Results and Discussion: Extraversion . . . 39
5.3 Academic-groups – Discussion and Collaboration-groups . . . 40
5.3.1 Student Discussions . . . 40
5.3.2 Student Collaboration-groups . . . 46
5.3.3 Academic-group Size Constraints . . . 50
5.3.4 Results and Discussion: *Discussions* and *Collaboration-groups* 50
5.3.4.1 Academic-Groups and Social Capital . . . 53
5.4 Conclusion . . . 54

6 e-Behaviour, Personality, Performance and Intervention **55**
6.1 Choice of Resampling Periods for Login Patterns . . . 55
6.1.1 Performance Analysis: The Imbalance of Outcomes . . . 57
6.1.2 Conscientiousness and Grade . . . 58
6.1.2.1 Results and Discussion: $(G(s), C(s))$ . . . 60
6.2 Behaviour-Personality Model . . . 61
6.2.1 B-PM: Model Design . . . 62
6.2.2 Results and Discussion: B-PM . . . 62
6.2.3 Results and Discussion: BM . . . 67
6.3 Accuracy and Timeliness of Intervention . . . 68
6.4 BM and the Trade-off Waterfall . . . 69
6.4.1 Practical Benefits and Limitations of the Trade-off Waterfall 69
6.5 Classifying a Student Using Trade-in Waterfall and Prediction Sequence . . . 70
6.6 Cohort-Specific Discussion on Intervention . . . 71
6.7 Conclusion . . . 71

7 Conclusion **72**
7.1 Limitations and Future Work . . . 73
7.2 Alternative Formulations of Personalities . . . 74
7.3 Completeness of Data . . . 74
7.4 Recommendations and Future Work . . . 75
7.4.1 Student Relationships . . . 75
7.4.2 e-Behaviour and Personality . . . 75

Appendices **76**

A Appendix **77**
A.1 Background Features Descriptions . . . 90
A.2 OLS Assumptions: Forum Activity and Performance . . . 91
A.2.1 Relationship between Grades and Crude Post Count . . . 91
A.2.2 Relationship between Grades and Number of Posts – OLS Assumptions . . . 91
A.2.2.1 Autocorrelation of Residuals . . . 92
A.2.2.2 Student Discussions – OLS Assumptions . . . 92
A.2.2.3 Residual-Normality Test . . . 92
A.2.2.4 Autocorrelation of Residuals . . . 92
A.2.2.5 Student Collaboration-groups – OLS Assumptions and kNN Code . . . 93
A.3 OLS Assumptions: E-Behaviour, Personality and Performance . . . 94
A.3.1 Average Number of Logins, Conscientiousness and Grade – OLS Assumptions . . . 94
A.3.2 Residual-Normality Test . . . 94
A.3.3 Autocorrelation of Residuals . . . 94

Bibliography **96**

List of Figures

3.1 End-to-end Framework – Behaviour and Performance Analysis . . . 17
3.2 The LSTM unit . . . 26

4.1 Pearson Correlation Coefficients of the Chosen Features . . . 33
4.2 Importance of the Chosen Background Features . . . 34

5.1 Frequency Distribution of Student Post Frequencies . . . 36
5.2 Frequency Distribution of Student Grades . . . 37
5.3 Crude Post Count against Student Grade . . . 38
5.4 Extraversion-level Grade against Extraversion-level . . . 39
5.5 Random Student's Grades against Discussion's Grade Averages . . . 42
5.6 Word Frequency per Discussion Cluster . . . 44
5.7 Clusters – Random-Student's Grades against Discussion's Mean Grade 45
5.8 Random Student's Random Grades against Discussion Grade Averages 47
5.9 Random-Student's Grades against Collaboration-group's Grade Averages . . . 49

6.1 Correlogram for the 114-Day Period . . . 56
6.2 Correlogram for the 17-weekPeriod . . . 56
6.3 Weekly Login Activity per Outcome Group . . . 59
6.4 Cumulative Grade Distribution of Cohort . . . 59
6.5 Average Number of Weekly Active Days against Grade . . . 60
6.6 Grade Distribution of the Safe but Flagged Students . . . 66
6.7 Grade Distribution of the At-risk but Ignored Students . . . 66
6.8 κ for predictions up to Each Week . . . 68
6.9 Trade-off Between Timeliness and Accuracy . . . 69

A.1 Table of LMS Behaviour From Literature Review (1 of 3) . . . 78
A.2 Table of LMS Behaviour From Literature Review (2 of 3) . . . 79
A.3 Table of LMS Behaviour From Literature Review (3 of 3) . . . 80
A.4 Crude Post Count Violation of Autocorrelated Residuals . . . 91
A.5 Autocorrelation of Residuals for Grade against Post Frequency Group 92
A.6 Autocorrelation of Residuals for Grade against Discussion . . . 93
A.7 Autocorrelation of Residuals for Grade against Collaboration-group 93
A.8 kNN Algorithm to Nearest Neighbours of a . . . 94
A.9 Autocorrelation of Residuals for Grade against Estimated Grade . . . 95
A.10 Random-Grade Generation for Students . . . 95

List of Tables

2.1 Facets of OCEAN Personality Traits. Source: Lim (2020) 9

3.1 Moodle Logins and Forum Post Number of Entries 16
3.2 Data Table Summary 16
3.3 Cohen's Kappa Interpretation 28
3.4 Comparison of Evaluation Systems 29

4.1 Background Data Feature Count Per Phase of Transformation 31
4.2 Background Data Features and Labels after RFE 32
4.3 Confusion Matrix and Summary of Background-Grade Test Set Results 33

5.1 Raw Forum Data 36
5.2 Forum Table 36
5.3 Forum Table Features 37
5.4 Input Table – Extraversion-level Grade against Extraversion-level . 39
5.5 OLS Regression Summary – Extraversion-level Grade against Extraversion-level 39
5.6 Discussion Table – Random Student's Grades against Discussion's Grade Averages 41
5.7 OLS Regression Random-Student's Grades against Discussion's Grade Averages 42
5.8 Sample: Random-Grade Generation for Students 46
5.9 OLS Regression Random-Student's Grades against Discussion's Grade Averages 46
5.10 Sample Table of Discussion Participation 47
5.12 OLS Regression Summary Random-Student's Grades against Collaboration-group's Grade Averages 49
5.11 Collaboration-groups and Grades 49
5.13 Discussion and Collaboration-group Differences 50
5.14 Academic-group Comparison: Discussion and Collaboration-group 51
5.15 Summary Statistics Comparison of Academic-group Categories . . 53

6.1 Sample of the Moodle Login Data Log 57
6.2 Login Sequence, $\{L(s)_t\}$ for each Student per Week 57
6.3 Number of Students per Outcome Group 58
6.4 OLS Regression Summary – Average Number of Weekly Active Days against Grade 60
6.5 Oversampling for the Login Table Train Set 62
6.6 B-PM and BM LSTM Hyperparameter Configuration 63
6.7 B-PM and BM Hyperparameter Alternatives 63
6.8 B-PM Training Input and Output Summary 64

6.9 Alternative B-PM Input and Output Summary 64
6.10 Confusion Matrix and Summary of B-PM Test Set Results 65
6.11 BM Input and Output Summary 67
6.12 Confusion Matrix and Summary of BM Test Set Results 67

A.1 Moodle Database Tables 81
A.2 Background Table Features 84
A.3 Moodle Database Tables 88

Nomenclature

α	Level of Significance
β_0	Slope Coefficient
β_1	Intercept
$\hat{y}$	Estimated y Value
κ	Cohen's Kappa Coefficient (Cohen, 1960)
$\mathbb{E}[\cdot]$	Expected Value Operator
$n(\cdot)$	Cardinality Operator
$p(x,y)$	Probability of a rejected null hypothesis for an $x-y$ relationship
$r(x,y)$	Pearson's Correlation Coefficient between variables x and y
r_k	Autocorrelation Coefficient at lag k
DTC	Decision Tree Classifier
LSTM	Long Short-Term Memory
NN	Neural Network
OLS	Ordinary Least Squares

Audience: Although this report contains applications from computer science, mathematics and social psychology, it is aimed at members who would like to survey a framework of methods for improving student performance. The methodology used in this report cites concepts derived for the ultimate usage within computer science literature. However, the usefulness of the methodology and analyses is extendable to the above domains and their intersections.

Chapter 1

Introduction

Post-secondary educational institutions have a concerningly low rate of completion (Ajoodha et al., 2020). According to News24Wire (2015), between 50% and 60% of South African students unenroll within their first year of studies. A study by John (2013) on attrition rates showed that 46% of students in face-to-face institutions between 2005 and 2010 withdrew from their degrees. The proportion for distance-learners was 68%. Richiţeanu-Năstase and Stăiculescu (2018) identify several problems that contribute to the low student retention rates. The identified problems include a lack of support for student, the student's background, and the student's inability to adapt to the curriculum.

The evaluation and analysis of factors affecting a student's performance result from the need to improve his grades. We define a student's performance as his grade at the end of his study programme. We sometimes refer to performance as *risk* or *risk of failure*, since an increase in performance results in a lowered risk of failure. The *e-Behaviour* of a student is defined as:

> "a pattern of engagement with a Learning Management System (LMS)"

and *personality* adopts the definition by Wright and Taylor (1970):

> "[...] the relatively stable and enduring aspects of individuals which distinguish them from other people and form the basis of our predictions concerning their future behaviour."

LMS is a system that tracks online activity for learning and development programs. *Behaviour* refers to behaviour defined in our cited literature.

Traditional approaches to revealing relationships between student behaviour, personality and performance include questionnaires, surveys and interviews. Respondents' biases can compromise the accuracy of responses (Heppner et al., 2015; Stone, 2000; Northrup et al., 1997). Furthermore, it has proven difficult to measure the reliability of an opinion (Fellegi, 1973), especially for each individual in a population. We address the self-reporting problems by using unobtrusive and automated approaches that measure how a student behaves rather than how he thinks he behaves. For instance, instead of asking, 'In how many weekly online discussions do you participate?', whose answer may only be approximate, we instead obtain the exact number of discussions from an LMS register. The models developed in this research use quantitative metrics to proxy behaviour and personality traits that are traditionally obtained from surveys. These metrics were used to draw correlations with and

predict student performance. From e-Behaviour and personality, we modelled an intervention framework that supplements current student intervention systems.

1.1 Research Motivation

Recently, teaching methods have adopted online channels to engage with students. Online LMS systems will likely remain part of teaching, and monitoring e-Behaviour is an efficient, low-cost method of measuring engagement and increasing performance on online-learning platforms. The need for the methodology of this research arises from the following factors that affect student performance:

- Existing structures are not 100 per cent effective at identifying students who are likely not to complete their programmes. University counsellors can be timely informed of students who display behaviour that correlates with or predicts poor performance.
- Students who perform poorly in their academics are most vulnerable to academic distress. This phenomenon may be cyclic, making poor performance due to distress a self-fulfilling process (Pomerantz et al., 2002).
- Identifying personalities and patterns in behaviour helps stakeholders understand where to focus behavioural analyses for performance improvement.
- Automated methods for student flagging allow us to collect, analyse and model data for each student and are scalable across the entire cohort. This scalability enables more efficiency than manual methods.

1.2 Research Questions

The below three questions were used as guidelines for the research:

1. What are the relationships between a student's performance, e-Behaviour and personality? (Chapters 4, 5 and 6)
2. What feature engineering techniques and analyses can be developed to inform hypotheses about a student's e-Behaviour, personality and performance? (Chapters 5 and 6)
3. To what extent can an automated model supplement existing academic intervention programmes? (Chapters 5 and 6)

1.3 Research Aims and Objectives

Our research aims to develop a methodology that makes it easier to produce future research on the intersection between computer science, education and psychology. We achieve this aim by revealing e-Behaviours and personality traits that are inputs to models, and then analysing model outputs and their practical implications. The following objectives guide the research in responding to the research questions:

- Identify behaviours and personalities that correlate with student risk by reviewing literature
- Use literature studies from Computer Science, Psychology and Education as a basis on which to formulate e-Behaviours and personalities
- Measure relationships between e-Behaviour, personality and risk
- Measure the predictive power of e-Behaviour and personality over risk
- Provide an intervention methodology

1.4 Assumptions on Data Credibility

The LMS is not the only online source of academic interaction for students. There exist multiple online forum platforms and instant messaging services which students use to supplement their academic needs. For instance, a student would not need to visit the LMS to download an assignment outline if he has already received it from a source outside of the LMS. The ability of the LMS data to measure academic engagement hinges on three assumptions for each student:

1. We have sufficient data to model his LMS engagement.
2. The LMS is his primary source of digital academic information.
3. His level of engagement with academic content within the LMS is proportional to his level of engagement outside of the LMS.

Even without assumptions 2 and 3 above, the metrics we develop are valid e-Behaviour and personality measures. This is because e-Behaviour and personality are defined based on LMS engagement alone; e-Behaviour and personality need not measure *all* academic engagement.

1.5 Contributions of Research

The studies in this dissertation add to the existing methodology of identifying factors related to student performance analysis and prediction.

The University's academic and student academic support staff members have established systems of identifying the likelihood of students completing a programme. The three systems listed below are aimed at understanding student performance as a part of the University's student support programmes:

1. Questionnaires
2. Observing a student's grades over time for that programme, and
3. One-on-one consultations by a counsellor or lecturer with the student.

e-Behaviour Models have some advantages over using grades or systems 1 and 3 above. These advantages are shown in Table 3.4, later in Chapter 3. e-Behaviour model-development is transferrable from this dissertation into other research and includes the following steps:

1. Taking inventory of datasets to pursue the objectives outlined in Section 1.3:
   - Economic and Cultural Background data as model inputs
   - Forum data as inputs
   - Login data as inputs
   - Grades data as target variables
2. Engineering features
3. Extracting patterns and classifying students
4. Providing an intervention framework for a cohort, based on this study's methodology

An e-Behaviour model can serve as the continuously-proactive component of existing student performance evaluation and intervention systems.

1.6 Research Outline

The following is an outline of the chapters and sections that respond to the research objectives:

1. Review of literature – Chapter 2
2. Contribution of this research to existing performance evaluation systems – Chapter 3
3. Correlation between and predictive power of background, e-Behaviour and personality against performance

   (a) Social and Economic Background and Performance – Chapter 4
   (b) Forum e-Behaviour, personality and performance – Chapter 5
   (c) Login e-Behaviour, personality and performance – Chapter 6

   Chapter 4 quantifies the predictive ability of student background data against their Outcomes. Data on student background was used as a benchmark to our study since the background of a student has been known to correlate with performance (see the Review of Literature in Chapter 2 for citations).

4. Explanation of an intervention mechanism that uses e-Behaviour, personality and performance – Sections 6.3 and 6.4

Chapter 2

Review of Literature

In this chapter, literary frameworks that describe factors affecting academic performance are discussed. Furthermore, existing studies on the aggregation and analysis of student behaviour, personality and academic performance are surveyed. This chapter includes a discussion on the methodology used and the findings across computer science, education and psychology. Figures A.1 to A.3 in the appendix show a table with a summarised set of the most relevant and recent works of literature that inspired our methodology. The table shows each paper's author, research aims and questions, behaviour metrics and measurements of academic risk.

2.1 Bourdieu's Three Forms of Capital and Student Success

Bourdieu's Three Forms of Capital is a framework which suggests that economic, cultural and social capital that can be leveraged by an individual regulates his level of success. We use this framework to support our investigation of the economic, cultural and social capital that a student has available to him as each form of capital relates to his academic performance.

2.1.1 Bourdieu's Three Forms of Capital – Definition

Dauter (2016) defines economic sociology as

> " [...] the application of sociological concepts and methods to analysis of the production, distribution, exchange, and consumption of goods and services. "

Economic sociology has been used extensively by Bourdieu and Richardson (1986), who argue that an individual's possession of three forms of capital regulates his social positions and ability to access goods and services. The three forms of capital are

- economic capital,
- cultural capital, and
- social capital.

The Three Forms of Capital can be considered essential to a student obtaining good grades and acquiring the services he needs to improve his grades. We refer to our proxies for economic and cultural capital as the *background* of a student.

2.1.2 Social Capital

Bourdieu and Richardson (1986) define social capital as:

> "the aggregate of the actual or potential resources which are linked to the possession of a durable network of more or less institutionalised relationships of mutual acquaintance and recognition"

The above definition describes social capital as a resource that is available between people as a result of their relaitonships. An individual may accrue social capital by being part of relationships.

Carpiano (2006) uses the framework by Bourdieu and Richardson (1986) to build onto the theory of social capital. Carpiano (2006) categorises the social capital available to individuals into four types, namely,

- social support,
- social leverage,
- informal social control, and
- community organisation participation.

The above four types of social capital are available to students, who also form relationships for social or academic purposes. Hallinan and Smith (1989) refer to these intra-cohort groups as *social networks* or *cliques*. The common saying, *show me your friends and I will show you your future*, is commonly used to describe the relationship between an individual's affiliates and his results. In this research, these results are referred to as his *Grade* or his *Outcome*. The hypothesis that a student has access to some *social capital* has been validated to various extents by Hallinan and Smith (1989). This hypothesis is adopted for studying student behaviour in this cohort.

A limitation with the social capital frameworks by Bourdieu and Richardson (1986), Carpiano (2006) and Song (2011) is that they provide no standard measure social capital. The definition of social capital leaves no room for a well-defined metric. In our research, a student's social network is evidence of his social capital, and is called his *Academic-group*. His Academic-group is defined in terms of Discussions and Collaboration-groups in Section 5.3. In an academic setting, a student's *quality* of resources social capital can be defined in term of the aggregate grades of his Academic-group. The relationships between Academic-groups and Grades is modelled throughout Chapter 5.

2.1.3 Cultural Capital

According to Hayes (1997), cultural capital is a set of non-economic factors that influence academic success, such as family background, social class and commitments to education, and do not include social capital. Bourdieu and Richardson (1986) categorises cultural capital into three forms, namely:

1. institutionalised cultural capital (highest degree of education),
2. embodied cultural capital (values, skills, knowledge, tastes), and
3. objectified cultural capital (possession of cultural goods).

In this research, the metrics we develop in Chapter 4 are proxies of 1 and 2. Smith and White (2015) found that success in obtaining a degree relates strongly to gender and ethnicity. Caldas and Bankston (1997) found that students' cultural capital affect their performance.

2.1.4 Economic Capital

Bourdieu and Richardson (1986) defines economic capital as material assets that are 'immediately and directly convertible into money'. In turn, an individual's monetary leverage can be converted into cultural and social capital (Bourdieu & Richardson, 1986).

Bourdieu and Richardson (1986) recognise that an individual can increase his social and cultural capital by making use of his economic capital. An individual who leverages his economic capital can obtain more resources to improve his cultural capital. For instance, an individual can improve his cultural capital through in improvement in his position in society. By investing in formal or informal education beyond the classroom, a student may increase his knowledge and the amount of cultural capital that is available to him. Fan (2014) observed that a student's quality and level of education was affected by his cultural and economic capital.

Chapter 4 reveals the relationships between student background (cultural and economic capital) and academic performance.

2.2 Five-Factor Personality Model

The expression of certain personality traits correlates significantly with academic performance (Poropat, 2009). The Five-Factor or OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism) model, often referred to as the Big-Five personality model, is a psychological framework expanded by the work of Costa and McCrae (1985). OCEAN has been used by several researchers to uncover the personality-performance relationships and has become a framework for standardised personality assessment. Personality-performance research has also been conducted in research by Poropat (2009), Furnham et al. (2013), Ciorbea and Pasarica (2013), Kumari (2014), and Morris and Fritz (2015). Section 3.5.1 describes how personality proxies were developed.

NEO PI-R (Revised Neuroticism-Extraversion-Openness Personality Inventory) (Costa Jr & McCrae, 2008) explains each OCEAN personality trait's six facets, which are summarised in Table 2.1.

For Conscientiousness, these facets are competence, order, dutifulness, achievement-striving, self-discipline, and deliberation. Activity, assertiveness, excitement-seeking, gregariousness, positive emotion and warmth are the facets of Extraversion. The rest of the

TABLE 2.1: Facets of OCEAN Personality Traits. Source: Lim (2020)

Trait	**Facets**
Openness	Curiosity, imagination, creativity, openness to new experiences
Conscientiousness	Competence, order, dutifulness, achievement-striving, self-discipline, deliberation
Extraversion	Activity, assertiveness, excitement-seeking, gregariousness, positive emotion, warmth
Agreeableness	Trust, straightforwardness, altruism, compliance, modesty, sympathy, empathy
Neuroticism	Anxiety, hostility, stress, self-consciousness, vulnerability, erraticity in mood

2.3 Related Work

2.3.1 Log File Studies for User behaviour

A log file contains a record of data transfers, among other information (logs), between a server and other computers. A log file collates data on the communications or transactions that typically contain the time, user identifier, and the transaction's description. Log file data has been used in studies of online information retrieval systems for over 40 years and has expanded since the use and development of the Internet (A. Peters, 1993). A widely used benefit of keeping log files is reporting and diagnosing problems with transactions to better a system's processes.

Reports that are generated by log file data are discreet and unbiased. As a result, there has been a continuous exploration of how log files can be used more effectively than for quantitative reporting and identifying system downtime. As an example of such a high-level application, Yu and Apps (2000) explore the features and functionality that make electronic journals more useful to journal readers. Their system showed that log files could be used to extract variables, which can be applied to determine user interaction with their journals. The study provides a recommendation to the process of log file data extraction and analysis; namely, data collection planning, data collection processing for computerised analysis, determining the validity of metrics and deriving statistics from the data.

Anitha and Isakki (2016) conducted a survey on predicting user behaviour based on log files, and their aim was to *discover the web user behaviour of students and faculty through web server log files*. The model used in this study included data preprocessing, Association Rule Mining[1], predicting and classifying user behaviour. Singh and

[1]Association Rule Mining is an algorithm used to discover relationships between variables in a dataset

Jain (2014) present algorithms for identifying cyber-criminal behaviour and subsequent prediction of illegal behaviour through website log files. The survey identifies that clustering techniques can distinguish cyber-criminals from regular users, while Neural Networks (NNs), Genetic Algorithms and Decision Trees can be used to trace criminal patterns.

2.3.1.1 Guidelines for Dealing with Log Data

Yu and Apps (2000) found that there are verbatim techniques for analysing log files, which will help our research. One such technique is keeping a file that stores original identification information, especially for sensitive data. A second guideline provided was to cluster groups of users depending on their frequency of interaction. To aggregate their data, Yu and Apps (2000) found that not all central tendency measures are appropriate to all user groups – some groups favour the use of median and range, and others favour the mean and standard deviation.

2.3.1.2 Problems Faced when Dealing with Log Data

Log files introduce two obstacles that should be addressed:

- Poor connectivity and denied requests can cause a user to make multiple requests to a server. These requests are recorded as separate transactions; for example, a student may log in twice after losing connectivity. These multiple transactions should be consolidated into a single event when collecting data from an LMS.
- There may be user groups that behave differently from one another. Yu and Apps (2000) suggest exploring each user group's distributions before modelling for different groups.

2.3.2 Machine Learning in Education

2.3.2.1 Student Performance Analysis and Prediction

The data used for student performance analysis can take on various forms – a time series representation of their behaviour is one alternative. Shanmugarajeshwari and Lawrance (2016) use attributes like student name, roll number, previous semester marks, attendance, assignment date, seminar performance, lab work and gender. These features were modelled without a time component. Shanmugarajeshwari and Lawrance (2016) aimed to evaluate student performance and create a warning system for student performance improvement. The study uses Naive Bayes and Decision Tree algorithms to 'identify students who need special consideration'.

Wilson and Shrock (2001) use 12 predictors, including mathematical background, previous computer experience, previous programming experience and gender. They obtained a coefficient of determination (R^2) of 0.44 between these 12 predictors and student performance. Evans and Simkin (1989) observed an R^2 value of 0.23 between a student's performance in homework assignments and his proficiency in a computer information systems programme. Evans and Simkin (1989) observed an

R^2 of 0.21 between performance in student's mid-term examination and proficiency. Fowler and Glorfeld (1981) fit a regression model over a student's most recent Grade Point Average, university admission scores and age to predict whether each student has a low or high *aptitude* for a computing course. The authors claim that their logistic classifier would obtain an accuracy of 75% in practice.

Samrit and Thomas (2017) developed a system that recommends elective subjects to students based on their performance. Ashenafi et al. (2015) use a student's pre-examination performance to predict the students' final examination scores. Poh and Smythe (2014) use previous grades and 100 exogenous variables (non-assessment data from surveys) as features to predict performance in an examination. Their regression model was compared to a *best guess*, defined as the mean value of the student's previous grades. The best guess produced a $\pm 17.15\%$ (percentage points) error margin. By contrast, their regression model produced an error margin of $\pm 12.74\%$. Their results indicate that a forecast based on the average of previous grades can be improved.

Yang and Li (2018) analyse a student's academic progress and predict his performance. The authors use a student's prior performance and other student's prior performance as input to a Neural Network. By grouping students into 'Art' and 'Science' subject areas, their model achieved an average prediction error of 4.02 grade points out of a possible 150 points.

2.3.2.2 Personalised Education

Ángel Agudo-Peregrina et al. (2014) found a significant relationship between online-course interaction and online-course performance, but an insignificant relationship for face-to-face courses. Several works by Ciolacu et al. (2018) and Ciolacu et al. (2017) study the use of artificial intelligence in education. Ciolacu et al. (2018) and Ciolacu et al. (2017) seek to define the future of education using *Education 4.0*, defined as education that responds to the needs of *Industry 4.0*. Ciolacu et al. (2017) propose a method that predicts student grades before their final examinations. The approach in their research embraces the use of wearable devices such as augmented reality glasses, biometric, security, health and emotion sensors. Ciolacu et al. (2018) and Ciolacu et al. (2017) review themes of text semantic analysis, electronic assessments for examinations and smart data sensors for student performance evaluation. They propose that Education 4.0 would introduce platforms for LMSs such as flagging systems to caution students at risk of failing. Ciolacu et al. (2018) use cross-sectional data as predictors, namely semester, course of study, number of clicks per month and quiz engagement over each month. Their study uses a NN to learn the click data and online questionnaire results, and a separate NN was trained on each course. The trained NNs were used to generate the forecasts to detect students at risk.

Park et al. (2015) cluster blended learning courses using online behaviour data. They argue that educational institutions have not made full use of LMSs by exploring potentially useful analytics contained in LMS logs. Most blended learning courses use basic LMS functionality rather than using the LMS to provide a user experience beyond accessing course content (Dahlstrom et al., 2013). The approaches used by

Park et al. (2015) investigate how educational services can be made more affordable and effective by leveraging an LMS for blended learning. Their clustering approach partitions courses into four clusters based on student engagement to identify patterns in the data. The blended learning courses are clustered based on the students' behaviour towards the courses. Online forum engagement informs how their courses can be improved.

Chapter 3

Methodology Overview

The previous chapter reviewed studies on student behaviour, personality and performance analysis. This chapter contains the formulation of metrics and the descriptions of algorithms used for this research. We close the chapter with a comparison of our methodology to existing literature's methodology.

3.1 Data Usage and Experimental Design

This dissertation filed for a study ethics application that was approved by the Human Research Ethics Committee at the University of the Witwatersrand (the University). The ethics application included measures imposed on the research methodology to ensure the protection of the student identities and their data's security. This dissertation's application clearance certificate protocol number is H19/06/36.

The relevance of the interpretations of analyses contained in each section was intended to span various educational settings. The exact magnitudes of results reported under the methodology used may not be identical to results from another tertiary institution under the same methodology. This study's results should be taken as a proxy for any other outcome obtainable through our methodology. Thus, obtaining a particular result was a secondary objective to:

- following a sound methodology to obtain the result,
- showing the methodology's ability to test a hypothesis, and
- explaining how a result applies to the University's context.

An empirical methodology based on

- Economic, Cultural and Social Capital theory,
- Computer Science literature, and
- Psychology, and Education literature,

together with collections of data (described in Section 3.4), are used as inputs to regression, predictive and time-series models to identify relationships between e-Behaviour, personality and performance. Section 3.4.2 contains a description of the data flow and models used to formulate e-Behaviour and personality in order to reveal their relationship with performance.

3.2 The Four Criteria – Feature Engineering and Selection

Python's *Tsfresh* package (Christ et al., 2018) contains 63 *classes* of extractable features, and the number of features features extracted can exceed 1 000, depending on the raw set of features. Due to the large possible number of features, *four criteria* were considered when selecting and engineering features for this research's objectives. The *four criteria* pertain to each feature's:

- *Interpretability* – measures how easily a behaviour or personality can be understood from the feature;
- *Generalisability* to other contexts – measures the extent to which the feature and its associated behaviour can be abstracted into other studies;
- *Relevance* to this study's objectives – measures the extent to which the feature is appropriate for measuring the behaviour or personality that it intends to measure;
- *Actionability* – the practical value within the context of the studies in this dissertation.

The algorithms used in this study are parametric – they do not extract explicit features – and the quality of their output depends on the input features. Consulting literature, including the NEO PI-R (Revised Neuroticism-Extraversion-Openness Personality Inventory) framework (Costa Jr & McCrae, 2008) ensured generalisability. This framework is discussed in Section 3.5.1.1. Interpretability and generalisability enable this research to be used outside our context. Relevance and actionability ensure the practical usage of the features within our study's context.

Given the need for our research's practical usage in education, only features whose foundations are based on the *four criteria* and cited theory were engineered. In turn, the features chosen were more intuitive than those obtained from arbitrary feature-extraction methods. If a highly interpretable feature was favoured over a less interpretable one, explicit mention is made about the trade-off between the above *four criteria*. The interpretability of each model's features and algorithms is discussed after the model's formulation.

3.3 Interpretable Models and Algorithms

Aspects of interpretability include fidelity (how well an explanation approximates a model) and simulatability (a user's ability to run a model on a given input) (Molnar et al., 2020). Except for the LSTM, the results of this dissertation's algorithms are mathematically or visually interpretable. Research conducted on the interpretability of deep learning algorithms has made advancements. Samek et al. (2017) calculate the sensitivity of their LSTM's predictions to changes in input (sensitivity test). In

addition to sensitivity tests, this research analysed the LSTM predictions by drawing distributions based on the LSTM classification labels (false positives and false negatives).

The best use case of this research is for practitioners in education to:

1. identify with the features and
2. make clear how our research methodology and analyses can be used in practice.

Satisfying the *four criteria* and using interpretable models aided our understanding of each problem's context and the analyses required. Practitioners can achieve the *best use case* by using the *four criteria* and adapting the concepts cited and methodology developed.

3.4 Outline and Usage of Data

Four data tables were constructed from the following sources:

1. Background data (Excel table from an SQL export) from the University's Business Intelligence department – used to construct the Background Table;
2. 2018 and 2019 Moodle database (SQL dump file) from the Faculty of Science's Moodle Administrator – used to construct the Forum and Logins Tables;
3. Grades data (Excel table from an SQL export) from the University's Business Intelligence department – used to construct the Grades Table.

3.4.1 Description of Dataset – Moodle Data

In the Moodle database, the Users (Student) table consisted of the 'MoodleID' and 'StudentID' columns used as keys to join the rest of the data. StudentID joined the Users table with the Grades Table (with 1 133 students), reducing the Users table's size to 1 133 students. The resulting table has three columns ('MoodleID', 'StudentID' and 'Grade').

The Logins Table results from merging the table of student grades (GT) and the Moodle Dataset (2019 Data) and the Forum Table (2018 and 2019 data).

The Moodle database has 153 tables, each containing logs on a student's activity. The table names are included in Table A.1 in Appendix A. We chose the most relevant set of tables that would enable the abstraction of behaviours and the Conscientiousness and Extraversion personality traits. The choice of these personality traits is discussed in Section 3.5.1.1. Secondary to the tables' relevance, tables that contained more activity logs were preferred to those with fewer logs. We assumed that a table registering more student activity contains more information and gives greater insight into the students' behaviours than a table that logs less student activity. The relevant tables that also contained the most data are the Forum Posts and Login activity tables, whose number of rows are summarised in Table 3.1. The final dataset thus contained four raw tables, shown in Table 3.2.

TABLE 3.1: Moodle Logins and Forum Post Number of Entries

Table Name	**No. Rows**
mdl_logstore_standard_log	467 743
mdl_forum_posts	2 378

TABLE 3.2: Data Table Summary

Table	**Can merge with**
Background – BT	GT
Forum – FT	LT, GT
Login – LT	FT, GT
Grades – GT	BT, FT, LT

3.4.2 Data and Model Flow

This flow of data used to develop the e-Behaviour Models is shown in Figure 3.1. We use *e-Behaviour Models* as a general term that refers to:

- models based on behaviour alone,
- models based on behaviour and personality, or
- both of the above.

The figure illustrates the flow of data into the models, and will guide the methodology of each experiment. Four data sources are queried, which are merged with the Grades Table and processed into model-ready form. As a final step to each model, the results of the experiments are analysed and discussed.

Each data table described in Table 3.2 has different features. However, the features common to all the tables were:

- *student*, a unique field used as a primary key to our SQL joins for each table in Table 3.2, and
- *Outcome*, *Grade* or both as a label to each Table.

Each data table shown in Table 3.2 was processed and transformed before being fit to a model. The algorithm used for transformation and the reason for the transformation in the context of the experiment's aim is detailed. The merging process includes joining the input tables to the experiment with labels (Either Grade – a continuous variable or Outcome – binary variable) and is followed by a combination of feature extraction and feature engineering methods. The resulting *Pandas Dataframe* is then ingested into a set of algorithms to produce a model. The model's results are analysed and discussed at the end of each experiment.

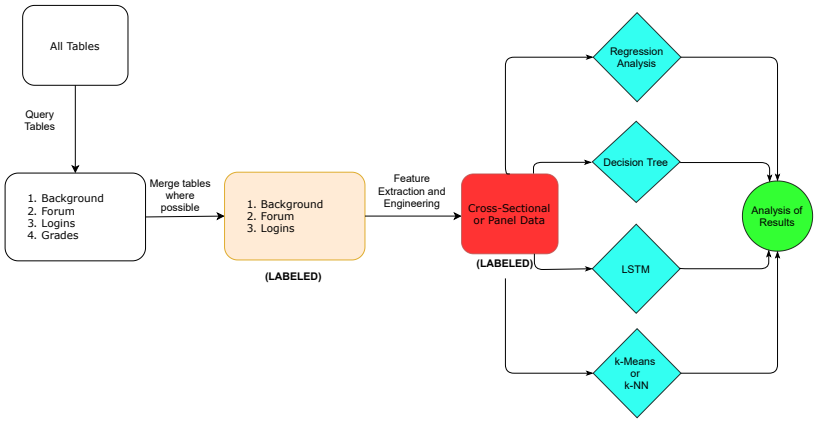


FIGURE 3.1: End-to-end Framework used for Behaviour and Performance Analysis

FT, LT, GT (as shown in Table 3.2), and an appropriate model are used to produce a methodology that suits each experiment's objective, so the algorithms used to fit each dataset will vary slightly. However, the evaluation structure of each classification and regression experiment is kept consistent.

3.5 Encoding Performance

Two measures of performance or *academic output* were constructed from the Grades Table, namely:

1. *Grade*: A continuous label that indicates the mean of a student's performance across all modules taken. This label is continuous and ranges between 0.00 and 100.00. See limitations of this encoding in Section 7.1.

2. *Outcome*: A binary label that indicates whether a student obtained below 51 *Grade* points (*At-risk*) or at least 51 *Grade* points (*Safe*). That is, Outcome is taken to measure the degree of *risk-of-failure*. Note that a *fail* is considered any grade below 50 Grade points. However, the boundary of 51 provides a buffer that allows the models to reveal students who were close to failing (At-risk). Therefore, a student need not fail for him to be considered at risk.

   Each Outcome is associated with a *Safety Score* classification label. Correct classifications would assign a *Flagged* Safety Score for an At-risk student, and an *Ignored* Safety Score for a student with a Safe Outcome.

Grade is used throughout Chapter 5 as a label to the Forum activity experiments, Extraversion (Section 5.2) and Conscientiousness (Section 6.1.2).

Outcome is used as a label to the classification models in Chapters 4 and 6.

3.5.1 Behaviour, Personality and Performance

3.5.1.1 Importance and Choice of Personality Traits:

As discussed in Section 3.4, the Moodle database contains several tables that each provides different information.[1] The Moodle database tables were ordered by the number of entries that they contained. We checked each table's appropriateness in modelling any of the OCEAN traits (see Section 2.2). Extraversion and Conscientiousness were traits linked closest to the information in the tables and were chosen for this study. By comparison, Openness, Agreeableness and Neuroticism were more complex to model, given the available data and the lack of validation of a link to academic performance within literature.

Our data linked closest to the dutifulness facet of Conscientiousness. Dutifulness is defined as the characteristic of being motivated by a sense of duty (Merriam-Webster). Wilt and Revelle (2009) define Extraversion as the 'disposition to engage in social behaviour', which links significantly to the gregariousness facet. Gregariousness is defined as the 'tendency for human beings to enjoy the company of others and to want to associate with them in social activities' (Association, 2020). Alternative formulations of each trait were considered and are described in Section 7.2. We used quantitative proxies to model Conscientiousness (Dutifulness) and Extraversion (Gregariousness).

3.5.1.2 Encoding Personality Traits:

Wright and Taylor (1970) define personality as

> 'the relatively stable and enduring aspects of individuals which distinguish them from other people and form the basis of our predictions concerning their future behaviour'.

According to Ajzen (2005) and Campbell (1963), human behaviour can be explained by reference to stable underlying dispositions, or personality. Therefore, Extraversion and Conscientiousness were modelled as single-valued averages that do not vary through time. Our choice to encode personality traits as unvarying values is based on theory by Wright and Taylor (1970), Ajzen (2005), Campbell (1963), and Hemakumara and Ruslan (2018).

Challenges against encoding personality traits: The above definitions of personality and their link to behaviour may cause a belief that personality and behaviour should be measured identically. However, to understand the separate correlations between either personality and performance or e-Behaviour and performance, it is important to encode an individual's *stable aspects* (personality traits that are less

[1] See AppendixA

likely to change) differently from his *changing* e-Behaviour. As a result, personality metrics are aggregated while e-Behaviour is modelled to vary through time.

3.6 Algorithms for e-Behaviour, Personality and Performance Analysis

3.6.1 Decision Tree Classifier

The Decision Tree Classifier (DTC) is a supervised learning algorithm that iteratively assesses conditions on the values of features in a data set to perform classification. DTC breaks down a decision-making process into a collection of simpler decisions, providing classifications that are easier to interpret than other statistical and machine learning models (Safavian & Landgrebe, 1991).

3.6.1.1 Architecture

DTC is assembled from a root node, edges, internal nodes and leaf nodes. At the root node, DTC conducts a test on each observation's value. Based on its value, the root node assigns a resolution represented by an edge, which the observation then traverses. At the end of the traversed edge is an internal node. An example of a node's test is 'Gender?', and an example of an edge is 'Female'. This decision process continues through the rest of the internal nodes until the tree reaches a leaf node, where a classification is made. See Mitchell (1997) for details on the DTC architecture.

3.6.1.2 Gini Impurity Index – Decision Factors

During prediction, an observation is predicted as part of a class after being checked through a series of conditions. An optimal decision tree results in an *optimal split*. An optimal split is achieved when each leaf node has the fewest possible train-set misclassifications (lowest impurity), and the tree has not been overfitted. Entropy and Gini Impurity Index are two commonly used metrics for impurity. The Gini Impurity Index (Gini) measures the relative frequency that a randomly chosen element from that set would be mislabelled. A Gini score greater than zero describes a node that contains samples belonging to different classes. Raileanu and Stoffel (2004) suggest that the difference between Entropy and Gini is trivial. This research uses Gini, which is interpretable.

Gini Calculation:

The Gini value decreases as a traversal is made down the tree towards its leaf nodes. The decrease happens as each internal node's condition aims to separate the classes according to a criterion that results in more homogeneous separations and higher accuracy in the training data. However, as with other predictive models, a high training-set accuracy is generated at the risk of overfitting. A larger tree (with more edges and branches) is more likely to overfit than a smaller tree and may result in a Gini of 0 at

the tree's leaf nodes. A Gini of 0 represents the minimum probability of misclassification over the training set but may result in weak generalisability over the test set. Therefore, smaller trees are preferred to larger trees (Mitchell, 1997).

The Gini impurity index is calculated using the formula:

$$Gini = 1 - \sum_i p_i^2 \tag{3.1}$$

where p_i is the probability of class i.

Khalaf et al. (2018) model DTCs on survey questions and answers that cover health, social activity and relationships of students to predict their academic performance. Topîrceanu and Grosseck (2017) and Kolo et al. (2015) provide literature in educational data mining and advocate for the use of the DTC due to its low complexity (with a run time of $O(m \times n \times log(n))$ and high interpretability. In Section 4.1.1, the DTC is used to select student economic and social capital features and predict the student Outcomes.

3.6.2 Ordinary Least Squares Linear Regression Analysis

Ordinary Least Squares (OLS) Linear Regression is a statistical model that estimates the linear relationship between one or more independent variables (regressors) and a dependent variable (regressand) (Gujarati & Porter, 2009). Throughout this research, only one independent variable was used per regression model. Using one independent variable per model isolates the effect of each variable on Grade. A Regression model with one independent variable is called a Simple OLS Regression model. Each estimated or predicted value, $\hat{y}_i$, derived from the line of best-fit shown in Equation 3.3, can be determined by

$$\hat{y}_i = \beta_0 x_i + \beta_1 + \epsilon_i, \tag{3.2}$$

where $\hat{y}_i$ is the predicted value of the i^{th} independent variable, x_i. β_0 is slope coefficient of the model, representing the average marginal change in $\hat{y}_i$ for a unit increase in x_i. β_1, the intercept of the model, represents the expected value of $\hat{y}_i$ when $x_i = 0$. $\epsilon_i \in \mathbb{R}$ is the residual term.

Every observed value, y_i, has an associated estimate or prediction value, $\hat{y}_i$. The line of best-fit,

$$\hat{y} = \beta_0 x + \beta_1, \tag{3.3}$$

is obtained by minimising the sum of the squares in the difference between the observed and predicted values of the dependent variable,

$$\sum (y_i - \hat{y}_i)^2 = (y_i - (\beta_0 x_i + \beta_1))^2. \tag{3.4}$$

The (linear) correlation coefficient between x and y is represented by r or $r(x,y)$. $r(x,y)$ measures the extent to which the independent variable, x is correlated with the dependent variable, y. That is, $r(x,y)$ measures the degree of closeness of all points,

(x, y), to the line of best-fit, $\hat{y} = \beta_0 x + \beta_1$. The correlation coefficient lies between -1 and 1, where a r of 1 or -1 means that the change in y is directly proportional to the increase in x. In that case, x and y are said to be perfectly correlated. That is,

$$r = 1 \implies \frac{y_i - y_{i+1}}{(x_i + 1) - x_i} = c, \tag{3.5}$$

for all values of i where x_i and y_i are defined, and where $c \in \mathbb{R}$. r is computed by

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{3.6}$$

where $\bar{x}$ represents the mean average of independent variable x and $\bar{y}$ represents the mean average of dependent variable, y.

The p associated with β_0 shows the probability of a hypothetical value, β_0^*, having an absolute value, $|\beta_0^*|$, that is at least as high as the observed $|\beta_0|$ by chance. The level of significance, α, is used as a threshold for a permissible p. In the domain relating to e-Behaviour, personality and performance, the common α used is 0.05, or a 5% level of significance. In our regression models, β_0 is accompanied by a $(1 - \alpha = 95\%)$ confidence interval, $\beta_0 \pm v$. Suppose $p < \alpha = 0.05$ for β_0. This means that from 100 experiments on similar sample distributions, fewer than 5 experiments would produce a $|\beta_0^*|$ value that lies outside of $\beta_0 \pm v$. Such a result means that the regressor and regressand have a statistically significant correlation different from zero (Gujarati & Porter, 2009). Statistical insignificance may indicate that x on its own does not yield reliable estimates, $\hat{y}$.

Statistical significance is important in the analysis of a student cohort's behaviour since statistical significance confirms the existence of a statistical relationship. Empirical significance refers to the magnitude of β_0 (Gujarati & Porter, 2009) and is also a measure of the model's practical value. One can be more confident in practical decisions if the relationship is not generated by chance (if the relationship is statistically significant). This *chance* is measured by the p.

3.6.3 Validity of OLS Regression Models

The data given in a model must satisfy five OLS regression assumptions (Gujarati & Porter, 2009), namely:

1. Normality of model residuals. The residual for each point is given by $y_i - \hat{y}_i$. $s^2 + k^2$ is computed for the residuals, where s is the z-score returned by the test for skewness and k is the z-score returned by the test for kurtosis.
2. Residual Independence or lack of Autocorrelation in Residuals.
3. Linearity in Parameters.
4. Homoscedasticity of Residuals.
5. Zero Conditional Mean.
6. No Multicollinearity in Independent Variables.

A linear relationship that violates the OLS assumptions is not fit for an OLS model. Therefore, we constructed only OLS relationships that satisfy the assumptions. Mention is made about experiments where the OLS assumptions are violated. All experiments which satisfy the OLS assumptions show the verification of the Normality of model residuals and Residual Independence assumptions in Appendix A. Furthermore, Linearity in Parameters, Homoscedasticity of Residuals and Zero Conditional Mean were verified for all Regression experiments whose results were analysed. The No Multicollinearity assumption was not verified since all OLS regression experiments are Simple. Proof of the validity of each experiment is included in Appendix A. See Gujarati and Porter (2009) for further details on the formulation of the OLS Regression model.

In this dissertation, all regression analyses are accompanied by regression plots.

3.6.4 Time Series Analysis

A time-series variable is a sequence of observations that takes on values ordered by time. In this report, we refer to e-Behaviour time-series as e-Behaviour sequences. Time-series analysis is useful in finding patterns to detect anomalous behaviour (Shumway & Stoffer, 2011). Observations through time may be correlated. Such observations are said to be autocorrelated – autocorrelation is a property of a time series that forms a foundation for time-series analysis (Little & Wei, 2013).

3.6.4.1 Autocorrelation

If linear *correlation* is taken to measure the extent of a linear relationship between two variables, *autocorrelation* measures the linear correlation between a time-series variable's values through time (Hyndman & Athanasopoulos, 2018).

The *Autocorrelation coefficient* r_k measures the extent of linear relationship, where k represents the number of observations lying between the points being correlated. For instance, r_1 would measure the correlation between the pair of values x_t and x_{t-1}, and r_2 measures the correlation between x_t and x_{t-2}. Formally,

$$r_k = \frac{\sum_{t=k+1}^{T} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T} (x_t - \bar{x})^2}, \tag{3.7}$$

where T represents the index of the last point of the time-series variable and $\bar{x}$ represents the mean average of time-series variable x. $r_k \in [-1, 1]$, where $r_k = 1$ indicates perfectly positive linear relationship and $r_k = -1$ indicates a perfectly negative linear relationship (where previous high values are followed by low values or vice versa). If $r_k = 0$, then there is no linear relationship between the values through time. As an example, consider a log file showing student logins for a website. A high correlation coefficient would indicate consistency in the student's activity. By contrast, inconsistent activity is indicated by a coefficient that is close to 0. A negative

correlation coefficient would indicate oscillatory behaviour about the mean. That is, high activity generally precedes low activity.

Throughout Chapters 5 and 6, r is used to identify behaviours that correlate significantly with performance, while r_k is used to determine the resampling periods for the series of student's logins in Section 6.1.

3.7 Long Short-Term Memory

The Long-Short Term Memory algorithm (LSTM) is designed to exploit data's temporal structure since LSTM is designed to model sequences for prediction (Liu & Sullivan, 2019). The LSTM has been used in studies that range from predicting weather-induced background radiation fluctuation by Liu and Sullivan (2019), to human motion classification and recognition by Wang et al. (2019).

The backpropagation through time algorithm computes the error, $\mathbf{E}_t$, at every time step, t, and then computes the total error. The LSTM's parameters are updated to minimise the total error $\frac{\partial \mathbf{E}}{\partial \mathbf{W}}$ with respect to a weight parameter $\mathbf{W}$:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}_t}{\partial \mathbf{W}}. \tag{3.8}$$

Letting $\mathbf{y_t}$ represent the output at time t, $\mathbf{h_t}$ represent the hidden state at time t and applying the chain rule to the Recurrent Neural Network model, the total error in equation 3.8 becomes:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}}{\partial \mathbf{y}_t} \frac{\partial \mathbf{y}_t}{\partial \mathbf{h}_t} \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} \frac{\partial \mathbf{h}_k}{\partial \mathbf{W}}, \tag{3.9}$$

where $\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k}$ involves a product of Jacobian matrices:

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \frac{\partial \mathbf{h}_{t-1}}{\partial \mathbf{h}_{t-2}} \cdots \frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{h}_k}. \tag{3.10}$$

Equation 3.10 illustrates the problem of vanishing gradients in equation 3.8; when the gradient becomes progressively smaller as k increases, the parameter updates become insignificant.

LSTMs are an architecture of Recurrent Neural Networks (RNNs). Bengio et al. (2012) suggest that RNNs are challenging to train because of the vanishing error gradient problem. The following section stipulates how the LSTM's architecture mitigates the vanishing error gradient issue through LSTM cells that maintain a state $\mathbf{c}_t$ at every iteration t. The cell state $\mathbf{c}_t$ serves to *remember* and propagate cell outputs between time steps. Each cell state then allows for temporal information to become available in the next time step, adding greater context to the inputs $\mathbf{x}_t$ that follow.

The activation $\mathbf{h}_t$ of an LSTM unit is:

$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t), \tag{3.11}$$

where

$$\mathbf{o}_t = \sigma\big(\mathbf{W}_{xo}\mathbf{X}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o\big), \tag{3.12}$$

is an output gate that mitigates the amount of content in the memory to expose to the following time step and $\sigma : \mathbb{R} \to (0,1)$ is the logistic sigmoid function.

Given new memory content,

$$\mathbf{i}_t \tanh\big(\mathbf{W}_{xc}\mathbf{X}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c\big), \tag{3.13}$$

where $\mathbf{i}_t$ represents the degree to which new memory is added to the memory cell, and is specified by an input gate

$$\mathbf{i}_t = \sigma\big(\mathbf{W}_{xi}\mathbf{X}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\big), \tag{3.14}$$

the cell state,

$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tanh\big(\mathbf{W}_{xc}\mathbf{X}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c\big), \tag{3.15}$$

can be updated by taking into account the previous cell state $\mathbf{c}_{t-1}$ and a term defined by the forget gate,

$$\mathbf{f}_t = \sigma\big(\mathbf{W}_{xf}\mathbf{X}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f\big). \tag{3.16}$$

Consolidating equations 3.11 to 3.16, the system of equations that describe each LSTM unit given by

$$\mathbf{f}_t = \sigma\big(\mathbf{W}_{xf}\mathbf{X}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f\big), \tag{3.17}$$
$$\mathbf{i}_t = \sigma\big(\mathbf{W}_{xi}\mathbf{X}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\big), \tag{3.18}$$
$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tanh\big(\mathbf{W}_{xc}\mathbf{X}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c\big), \tag{3.19}$$
$$\mathbf{o}_t = \sigma\big(\mathbf{W}_{xo}\mathbf{X}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o\big), \tag{3.20}$$
$$\text{and} \tag{3.21}$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t). \tag{3.22}$$

Let B denote the input batch size (number of time stamps per input chunk), H denote the LSTM hidden state capacity, and D represent the dimensions of the inputs to the LSTM. Then, in equations 3.17 through 3.22:

$$\mathbf{x}_t, \mathbf{h}_{t-1} \in \mathbb{R}^{BxD}, \tag{3.23}$$

$$\mathbf{f}_t, \mathbf{i}_t, \mathbf{c}_t, \mathbf{o}_t, \mathbf{h}_t \in \mathbb{R}^{BxH}, \tag{3.24}$$

$$\mathbf{W}_{xf}, \mathbf{W}_{xi}, \mathbf{W}_{xc}, \mathbf{W}_{xo} \in \mathbb{R}^{DxH}, \tag{3.25}$$

$$\mathbf{W}_{hf}, \mathbf{W}_{hi}, \mathbf{W}_{hc}, \mathbf{W}_{ho} \in \mathbb{R}^{H^2}, \text{and} \tag{3.26}$$

$$\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_c, \mathbf{b}_o \in \mathbb{R}^{BxH}. \tag{3.27}$$

For an illustration, refer to Figure 3.2. The LSTM has four main gates that respond to the values of four functions determined by $\mathbf{f}, \mathbf{i}, \mathbf{c}$ and $\mathbf{o}$, represented in equations 3.17 through 3.20. With the input data matrix $\mathbf{x}_t$ (data vector if $B = 1$) concatenated with previous output matrix $\mathbf{h}_{t-1}$ (vector if $B = 1$), the flow of inputs and outputs from the various gates described in the LSTM equations interact as follows:

1. $\mathbf{h}_{t-1}$ and $\mathbf{X}_t$ are fed into the gate (or function) $\mathbf{f}$, where the output $\mathbf{f}_t$ lies in the open interval $(0, 1)$. $\mathbf{f}_t$ then interacts with previous cell state $\mathbf{c}_{t-1}$ through element-wise multiplication $\bigotimes$, thus $\mathbf{c}_{t-1}$ holds an interim cell state, $\mathbf{f}_t\mathbf{c}_{t-1}$. At this stage, $\mathbf{f}_t\mathbf{c}_{t-1}$ represents a state that has forgotten some previous cell state data in $\mathbf{c}_{t-1}$ that was captured as unimportant (note that importance is regulated by weight coefficients that are trained and stored in their respective weight matrices).

2. Whereas the forget gate $\mathbf{f}_t$ focuses on regulating the extent to which previous data is forgotten, the input gate $\mathbf{i}_t$ focuses on adding new data, scaled by its importance, or extent to which data should be added from the matrix comprised of $\mathbf{h}_{t-1}$ and $\mathbf{X}_t$.

3. The *tanh* gate obtains $\mathbf{h}_{t-1}$ and $\mathbf{X}_t$, but uses the hyperbolic tangent *tanh* function to compute its outputs (between -1 and 1).

4. The result given by *tanh*, and $\mathbf{i}_t$ are then multiplied element-wise and further added ($\bigoplus$) to $\mathbf{f}_t\mathbf{c}_{t-1}$, giving $\mathbf{c}_t$, shown in equation 3.19.

5. The output gate $\mathbf{o}_t$ decides what values to output, given $\mathbf{h}_{t-1}$ and $\mathbf{X}_t$, and also computes its exposure to the following cell state based on trained importance.

6. Finally, the values of the cell state, $\mathbf{c}_t$, are passed through a *tanh* function and multiplied by the output gate result, $\mathbf{o}_t$, such that the LSTM unit keeps only the output that it accounts for as important in $\mathbf{h}_t$, described by equation 3.22.

3.7.1 LSTM Problem Design

Let B (different from the B above) represent a $n \times T$ matrix containing the n number of e-Behaviour sequences of all students. T is the length of all each student's e-Behaviour sequence. Let $B(s)$ represent a $1 \times T$ variable representing the e-Behaviour sequence of student s, and let $B(s)_t$ be a scalar representing the value of $B(s)$ at time t. The LSTM learns the interdependencies between variables B and

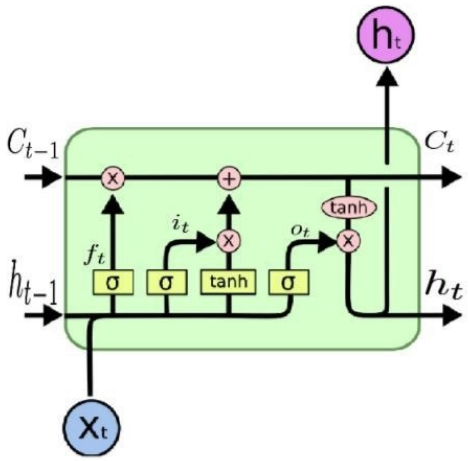


FIGURE 3.2: The LSTM unit keeps a cell state throughout its operations, which serves as input in the next time step. It also outputs $\mathbf{h}_t$, which supplements the input $\mathbf{X}_t$ in the following time step. From Olah (2015).

$B(s)$ with the aim of classifying the risk (or determining the Safety Score) of s given the values of B and $B(s)$. That is, B and $B(s)$ are predictors of the Safety Score *(classification)* of s. Without loss of generality; this framework is used to predict the Safety Score of all students in Chapters 4 and 6.

3.8 Evaluation Metrics for Student Risk Classification

The Result Summary below is used as an evaluation template for the classification problems in Chapters 4 and 6.

Results Summary

		Safety Score (Prediction)		
		Flagged	**Ignored**	**Total**
Outcome (True Label)	**At-risk**	a	b	$a+b$
	Safe	c	d	$c+d$
	Total	$a+c$	$b+d$	$N=a+b+c+d$

Outcome	**Precision**	**Recall**
At-risk	$\frac{a}{a+c}$	$\frac{a}{a+b}$
Safe	$\frac{d}{d+b}$	$\frac{d}{d+c}$

The best-case scenario (where the e-Behaviour model obtains a 100% accuracy) occurs where all students with an At-risk Outcome label are Flagged, and all students with a Safe Outcome label receive an Ignored Safety Score.

Given an Outcome, At-risk, precision measures the proportion of students who were correctly Flagged as At-risk, a, against the total number of Flagged students. Precision is calculated as $\frac{a}{a+c}$, where c represents the number of students who should have been Ignored as Safe. Recall measures the proportion of students who were correctly Flagged as At-risk, a, against all At-risk observations, $\frac{a}{a+b}$. The same calculation generalises to the Safe Outcome.

It is essential to know the most important metrics to measure when evaluating a classifier's performance. Consider the Results Summary above. A perfectly accurate model results in a b and c equal to 0. The precision and recall scores would be 1 for both the At-risk and Safe Outcomes. None of our models achieves perfect accuracy – they make trade-offs regarding precision and recall. For classifying all students who are at risk of failing, a model with higher precision-recall scores for the At-risk Outcome is preferred over one with precision-recall scores for the Safe Outcome. Furthermore, maximising the recall of the At-risk Outcome, $\frac{a}{a+b}$ (where the classifier recalls all students who are at risk) is preferred to maximising either the precision of the At-risk Outcome or the precision-recall scores of Safe students. Recall-maximisation will likely cause a low precision for the At-risk Outcome class (a high c). In such a case, however, no student who is At-risk will have been incorrectly Ignored.

3.8.1 The Overall Accuracy of a Model

While precision and recall are important metrics to measure a binary classifier's performance, they represent four different views of accuracy that must be analysed separately (precision and recall for At-risk and Safe Outcomes). When evaluating a model or accuracy, it is useful to obtain a single metric. A widely-used *accuracy* measure calculates the ratio between the correctly classified number of observations and the total number of observations. This *accuracy* is a good measure for balanced data, not for imbalanced data. For instance, if a test dataset contains 100 observations with an At-risk:Safe split of 10:90, a classifier can obtain an accuracy of 90% by classifying all students as Safe. An accuracy measure that combines the harmonic mean of precision and recall of either class is the *f-1 score*, whose effectiveness is surveyed by Hand and Christen (2018). The f-1 score produces two metrics (one for each Outcome) and does not concisely summarise the model's accuracy. By contrast, *Cohen's Kappa*, κ (Cohen, 1960) is a metric that captures accuracy with a single value. The formula,

$$\kappa = (p_o - p_e)/(1 - p_e), \tag{3.28}$$

measures the *agreement* between the *predicted* Safety Score and the *true* Outcome. Landis and Koch (1977) suggest using the scale in Table 3.3 to interpret the significance of κ values. In Identity 3.28, the observed accuracy (ratio between correctly classified number of students and total students), p_o, is adjusted for the expected *accuracy* when the classifier assigns a label randomly, p_e. In the example above, $p_e = 0.90$ and $p_o = 0.90$, giving a κ value of 0.00, or *no agreement* between the

Outcomes and the Safety Scores assigned by the classifier. κ is thus more representative of a model's performance than the *accuracy* commonly used for data with balanced labels. *Chance* is an event that occurs when a classifier fails to fit an optimised objective function or has not learned anything from the data. In the above example, the κ of 0.00 signifies that the classifier performs no better than chance.

TABLE 3.3: Cohen's Kappa Interpretation

κ	Level of Agreement
< 0.00	Worse than chance
0.00 – 0.20	*Slight* agreement
0.21 – 0.40	*Fair* agreement
0.41 – 0.60	*Moderate* agreement
0.61 – 0.80	*Substantial* agreement
0.81 – 1.00	*Near-perfect* agreement

3.9 Contribution to Existing Evaluation Systems

Chapters 4 to 6 aim is to engineer features and frameworks that are based on principles contained in literature. Before each feature's construction, we mention how its construction metric is a valid representation of the personality or behaviour that it intends to capture. Furthermore, this study adds to the existing methodology of identifying factors related to student performance. Researchers and academic personnel can develop a methodology using similar principles for their own studies.

The University's academic and student support staff members have access to three common systems of identifying the likelihood of students completing a programme, namely:

1. Questionnaires
2. Observing a student's grades over time for that programme, and
3. One-on-one consultations by a counsellor or lecturer with the student.

It has been shown that student grades correlate with their future grades – Evans and Simkin (1989), Fowler and Glorfeld (1981), and Poh and Smythe (2014) prove that prior performance can be a reliable measure for future performance. While e-Behaviour Models[2] do not guarantee a similar reliability, e-Behaviour Models have some advantages over using grades or methods 1 and 3. These advantages are shown in Table 3.4.

System 1 above helps programme coordinators understand the general sentiment that students have towards a programme and lecturer. This system of surveying lends

[2]In this chapter, e-Behaviour Models refers to models that fit students' LMS behaviour and / or personality to a model. In Chapter 6, distinction is made between behaviour-based and behaviour-personality-based models.

itself to two limitations in efficacy. Firstly, the questionnaire is not offered throughout the teaching period. Secondly, the questionnaire is anonymous, so lecturers and staff can not link each response to its respondent. Therefore, there is no reasonably easy way of establishing which students are struggling with their programmes. By their high-touch nature, systems 2 and 3 are usually not anonymous and voluntarily conducted upon consent. The advantage of these systems is that they give a detailed response to a student's feeling towards his programme and are thus corrective (can help resolve poor performance). The disadvantages are that one-to-one consultations are voluntary, are retroactive (the consultation is only conducted after a student has been identified as *at risk*) rather than proactive, and are often not conducted at scale (for each student in a programme). The lack of continuity in gauging student performance gives rise to the proposal for using e-Behaviour Models. e-Behaviour model development is transferrable from this dissertation into other research and includes the following steps:

1. Taking inventory of datasets for the pursual of objectives outlined in Section 1.3:
    - Background data as model inputs
    - Forum data as inputs
    - Login data as inputs
    - Grades data as target variables
2. Engineering features
3. Extracting patterns and classifying students
4. Providing an intervention framework for a cohort, based on this study's methodology

If an e-Behaviour Model borne out of the above framework were used in practice, then the model can serve as the continuously proactive component of existing student performance evaluation and intervention systems. Table 3.4 shows the advantages and disadvantages of e-Behaviour Models and performance Evaluation systems 1 to 3.

TABLE 3.4: Comparison of Evaluation Systems

System	Continuously Proactive	Corrective	Easily Feasible at Scale	Reliable
Questionnaires	✗	✗	✓	✗
Previous Grades	✗	✗	✓	✓
Consultation	✗	✓	✗	✓
e-Behaviour Models	✓	✗	✓	To be shown

Table 3.4 suggests that e-Behaviour Models are the only system of evaluation that is continuously proactive (The continuity of the system depends on the urgency of

intervention and periodicity of evaluation). Therefore, all of the systems are proactive to the extent that they provide a timely response or intervention that an institution's students require. Behaviour, however, can be monitored at any point in time, hence the term *continuously* proactive.

FT, LT, GT (as shown in Table 3.2), and an appropriate model are used to produce a methodology that suits each experiment's objective, so the algorithms used to fit each dataset will vary slightly. However, the evaluation structure of each classification and regression experiment is kept consistent.

Chapter 4

Student Background and performance

This chapter quantifies the predictive ability of student Background data against their Outcomes. Data on student Background was used as a benchmark to our study since the Background of a student has been known to correlate with performance (see Chapter 2). Later in this chapter, the pair-wise r-values between each Background feature and Outcome is shown in Figure 4.1.

4.1 Background Features and Data Split

Refer to Table 4.1. The raw Background Dataset (Table A.2) consists of 4 748 students and 176 features on which experiments in this chapter were conducted. These features captured answers by the student upon registration and data collected throughout their study – for instance, their high-school's facilities, high-school subjects, age and city of residence. Table 4.1 shows a summary of the features after each phase of transformation. The full dataset of 176 features (Table A.2) is included in Appendix A.

TABLE 4.1: Background Data Feature Count Per Phase of Transformation

Transformation Phase	Categorical Features	Total Features
Before One-hot	169	176
After One-hot	6 616	6 623
After RFE	5	5

The 169 categorical features were one-hot encoded, extending the number of features from 176 to 6 623. *Scikit-Learn*'s Recursive Feature Elimination algorithm (RFE) was used to reduce the 6 623 feature set's dimensionality. The optimiser used was a Decision Tree. RFE involves filtering through features with the lowest ranking of importance with respect to Outcome, through the following procedure (Guyon et al., 2002):

1. optimise the Decision Tree weights with respect to its objective function on a set of features, F

2. compute the ranking of importance for the features in F using the Decision Tree optimiser
3. prune the features with the lowest rankings from F
4. repeat 1-3 on the pruned set until the specified number of features is reached

The RFE process produced five Background Features, explained with the Grade and Outcome variables in Table 4.2. The features' full descriptions is under Section A.1

TABLE 4.2: Background Data Features and Labels after RFE

Feature	**Description**	**Type**	**Values**
Quintile	To which of the five categories a student's high-school belongs under the South African Government schools standards; a 6 indicates private high-schools	Categorical	1 - 6
Gauteng Province	Whether a student completed their ultimate year of high school at a school in *GP* (Gauteng Province)	Binary	No, Yes
Gender	Whether the student is female or male	Binary	Female, Male
Financial Assistance	Whether a student received financial aid from the National Student Financial Aid Scheme	Binary	No, Yes
Township School	Whether a student's high school was situated in a township area	Binary	No, Yes
Grade (label)	Grade points out of 100 obtained, as defined in Section 3.5 on Encoding performance	Continuous	0.0-100.0
Outcome (label)	Risk of the student based on their Grade, as defined in Section 3.5 on Encoding performance	Binary	At-Risk, Safe

Figure 4.1 shows the linear correlation between the five Background predictor variables chosen by RFE and the Grade target variable. Decision Trees pick attributes based on the largest Gini Index at each node. Therefore, Decision Trees are not susceptible to multicollinearity. However, it is essential to understand what linear correlations exist between each chosen feature and student Grade. The Pearson Correlation matrix is shown in Figure 4.1.

Quintile (r = 0.170) has the strongest linear correlation with Grade, followed by Township School ($r = -0.140$). The Background data's correlations suggest that students from higher Quintile high-schools generally performed better than students from lower Quintile schools. Students from Township schools performed worse than students from other schools.

Understanding that a relationship exists between the chosen features and Grade shows that these features can inform the student's Grade and Outcome. Section 4.1.1 below shows the results of the Decision Tree Classifier used to classify the student Outcomes based on the five chosen predictor variables.

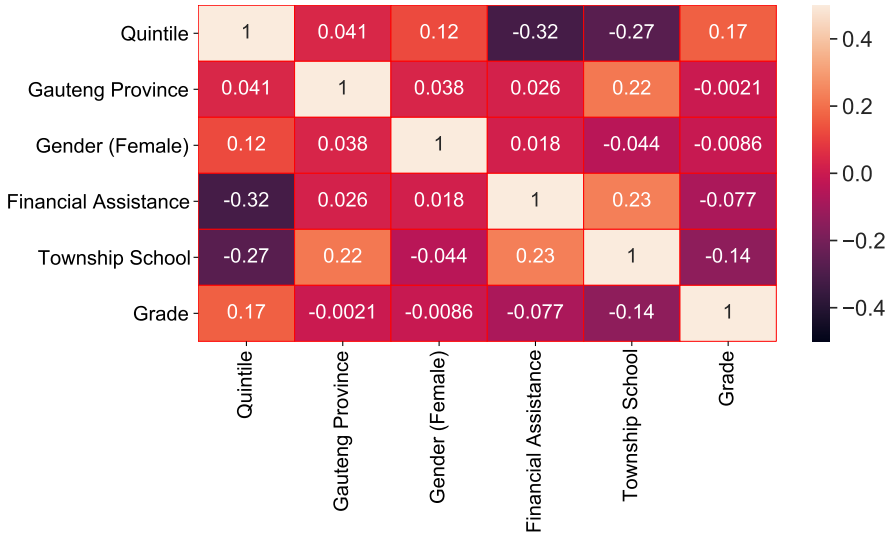


FIGURE 4.1: Pearson Correlation Coefficients of the Chosen Features. Quintile and Township School have the highest correlation with Grade

TABLE 4.3: Confusion Matrix and Summary of Background-Grade Test Set Results

		Safety Score (Prediction)		
		Flagged	**Ignored**	**Total**
Outcome (True Label)	**At-risk**	107	157	264
	Safe	153	533	686
	Total	260	690	950

Outcome	**Precision**	**Recall**
At-risk	0.41	0.41
Safe	0.77	0.78

$\kappa = 0.18$

4.1.1 Classifying a Student based on Background Data

The classifier used was the Decision Tree Classifier. The train-set contained 3 798, while the test-set contained 950 students. The train-test split was stratified by the Outcome of the students. For the Decision Tree, a grid search on the train-set suggested a maximum tree depth of 6 and 8 maximum leaves. A summary of results is presented in Table 4.3.

Results and Discussion: Background Data

Figure 4.2 shows the relative importance of the five chosen features, based on the Decision Tree model whose results are presented above. The most important feature

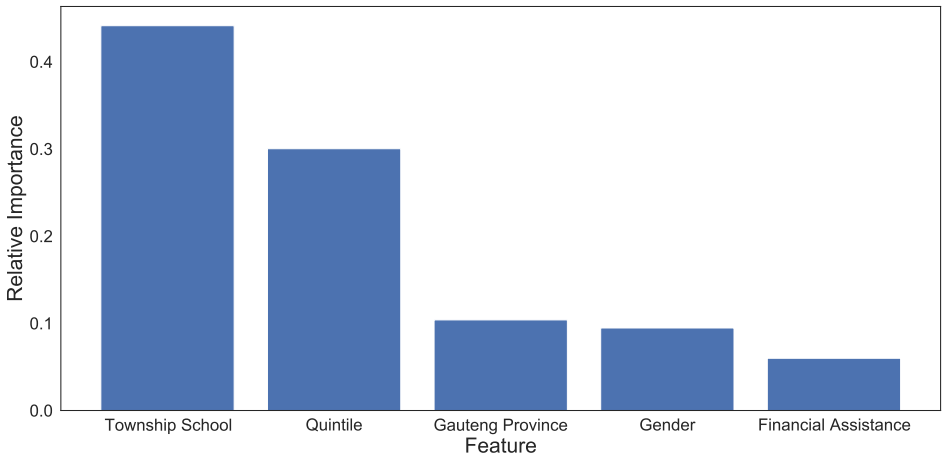


FIGURE 4.2: Importance of the Chosen Background Features

in classifying the students was the Township School feature, followed by the Quintile feature. The remaining features are comparatively less important determinants to a student's Outcome. Excluding the Gauteng Province, Gender, and Financial Assistance features results in a Decision Tree that scores a κ of 0.13.

Refer to Table 4.3. 640 out of the 950 test observations were classified correctly, producing a κ of 0.18 (slight agreement) between the Safety Score and the Outcome. The precision score for the *Flag* students suggests that 107 of the 260 Flagged students were correctly Flagged; the remaining 153 were meant to be Ignored.

The Support Vector Machine algorithm, with a linear classification boundary, produced a κ of 0.16. Lastly, a Random Forest Classifier produced a κ of 0.18.

4.2 Conclusion

This chapter showed that a student's Background is linked to and has predictive power over his academic outcome. We demonstrated that Background data has variables that can be used as input variables to models that predict student risk. In this case, a Decision Tree, Support Vector Machine, and Random Forest produced results superior to a random classifier's. Despite a better-than-random accuracy, both the precision and recall scores for the *At-risk* Outcome group was 0.41. Therefore, the DTC was weaker at classifying students at risk than at classifying Safe students.

A student's Background does not indicate his risk at different points throughout his study programme (see Section 3.9). Also, a student's Background, such as his current or previous economic capital, is more difficult to change than his

- e-Behaviour, which he can improve, and
- social capital, which he can accumulate

at any point during his studies. Chapters 5 and 6 contain analyses and discussions on the predictive power of student e-Behaviour.

Chapter 5

Forum Activity and performance

Chapter 4 extended literature's theory that a student's Background can inform his performance. While the above argument is credible, a student's Background is recorded only once.[1] Therefore, the relationship between Background and performance can only be measured once. We seek to understand how students' performances can be improved upon and improve the likelihood of timeous intervention. Therefore, we must account for factors that may affect student performance *during* their programmes. In this research, these factors are e-Behaviours and expressions of personalities. Both factors were measured through LMS engagement.

Chapter 2 introduced the correlation between e-Behaviours, personality and academic performance in previous literature. In this chapter, forum engagement patterns are examined to uncover the relationship between forum engagement and student Grade. This chapter also shows the formulation of the Extraversion trait (in Section 5.2) that are used in Chapter 6. Extraversion is used as an extension of the e-Behaviour model introduced in Section 3.9.

The use of *Post* logs over *Read* logs

A *post* log shows that the student has shown some interest in the topic. On the other hand, the LMS logs a *Read* if a student enters a discussion, whether or not he engages with or even reads its contents. Hence, *Post* logs present an inherent advantage in filtering out the possibly-passive reads from students and will be used as a measure of forum activity in this chapter.

The Moodle Forum data Table (5.1) contains four fields. The *Discussion* enumerates each discussion on the LMS, while *userid* is the student's unique identifying number in the Moodle database. *Created* indicates the time, in Unix Time format, that the forum posts were created.[2] *Message* shows the content of the messages posted by students.

After joining the table in Table 5.1 with the Grades Table, the resulting table contained 1 133 students whose forum data make up the Forum Table. The final Forum Table is shown in Table 5.2, which is transformed before each experiment.

[1] Although Background may be recorded many times, it is often without any significant changes across the cohort

[2] See Misja (2021)

TABLE 5.1: Raw Forum Data

Discussion	userid	Created	Message
390	6629	1559147973	Thank You for the links we ap...
397	6526	1559148633	THANKS!
444	6490	1559152233	Wow!! Thanks a lot it was!!
468	6522	1559152473	Okay then thanks :)
600	6522	1559152713	There are many resources you...

TABLE 5.2: Forum Table

Student	Discussion	Grade	Message	Time	Created
35	99	49.25	Hi guysDoes any ...	18-08-0919:...	1.533+09
35	241	49.25	hi, I think Pseudo-...	18-10-0220:...	1.538+09
902	309	78.75	thank you for that.	18-11-0609:...	1.541+09
902	327	78.75	yes please.	18-11-2120:...	1.542+09
349	109	69.75	Hi all!I am runni...	18-08-1213:...	1.534+09

5.1 An Overview of Independent and Dependent Variables

As a starting point, we examined the distributions of the student number of posts and Grades, shown respectively in Figures 5.1 and 5.2.

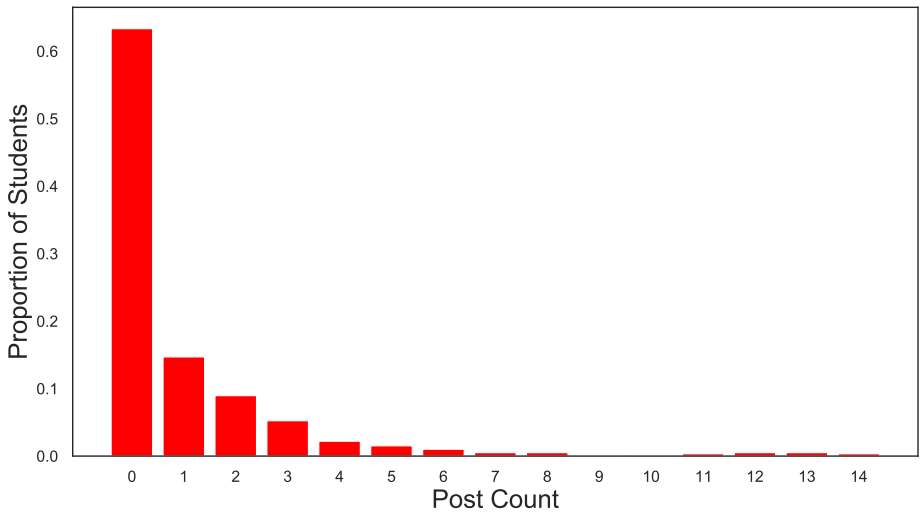


FIGURE 5.1: Frequency Distribution of Student Post Frequencies. The majority of students do not post to forums. The proportion of student posts is inversely proportional to the number of posts.

From Figure 5.1, observe that 63% of students who were given access to the forum did not write any posts. The distribution also shows that the number of students who

post more frequently decreases with the number of posts.

Figure 5.2 shows the frequency distribution of Grades, which is unimodal, has a mean of 59.92 and a median of 59.00 Grade points.

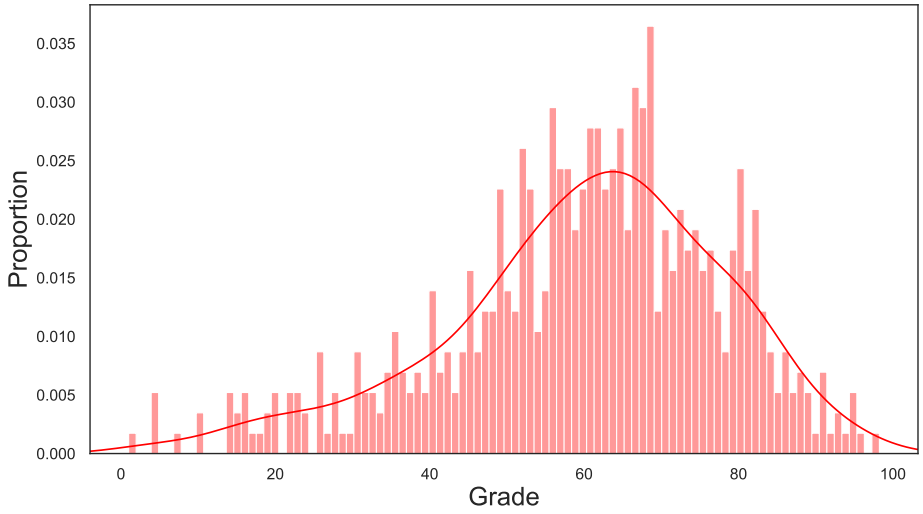


FIGURE 5.2: Frequency Distribution of Student Grades. Mean: 59.92. Median: 59.00.

The meaning of the *Discussion*, *Message* and *Time* independent variables used in this chapter is explained in Table 5.3.

TABLE 5.3: Forum Table Features

Feature	**Description**	**Type**	**Possible Values**
Discussion	Discussion Number. Messages that begin a topic and are posted as responses are assigned the same discussion number.	Categorical	0 – 337
Message	Contents of each forum post.	String	-
Time	Extracted from the *Created* variable. Indicates the time at which each *message* was posted.	yyyy-mm-dd	2018-01-05 to 2019-01-05
Grade (label)	Number of points out of 100 (Section 3.5).	Number	0.0 – 100.0

The four variables in Table 5.3 are used to engineer other features in this chapter. Due to the limited feature set, we did not apply feature selection. The following section draws a relationship between Grades and Extraversion.

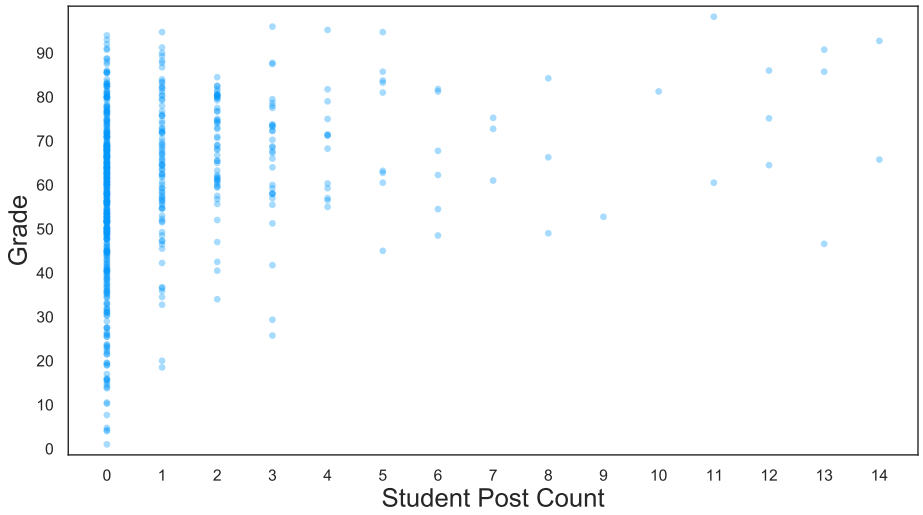


FIGURE 5.3: Crude Post Count against Student Grade. There is a positive relationship between Post Count and Grade that is not suitable for a linear OLS fit.

5.2 Extraversion and Grade

In this section, the linear correlation between the students' Grades and levels of Extraversion is examined. The definition of social capital in Section 2.1 suggests that Extraversion or gregariousness can improve an individual's ability to accumulate social capital, which is correlated with academic performance. We use this argument as a reason to quantify the direct relationship between a student's Extraversion and performance, without the need to consider the level of his social capital. We later examine the relationship between social capital and performance in Section 5.3. A way to model the level of social interaction or gregariousness is by capturing the number of forum posts that an individual contributes to forum discussions. Hence, we chose the student's post count as a quantitative proxy for his *level* of Extraversion. The choice of modelling Extraversion in this way is based on the definition of gregariousness as well as the interpretability and actionability of this proxy (See Section 3.2).

Each student was placed in an Extraversion-level group, E, representing the number of posts that he contributed. Each level, E, was then assigned a mean Grade, G_E, computed by averaging the Grades of all students in E. A linear regression plot of G_E against E is shown in Figure 5.4. Table 5.4 shows the input table to the OLS model, whose results are shown in Table 5.5 and Figure 5.4. Placing the students in Extraversion-levels satisfied the OLS assumptions. While regressing each student's post count against his Grade produced statistically significant results, the data's distribution violated OLS assumptions. Hence the transformation by placing each student into an Extraversion-level. Figure 5.3 shows the Crude Post Count against the Grade of each student. The Residual Normality, Independence, and Homoscedasticity assumptions were violated by the OLS model fitted on the data in Figure 5.3.

TABLE 5.4: Input Table – Extraversion-level Grade against Extraversion-level

E	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
$\mathbf{G_E}$	55.8	64.4	68.3	66.1	69.3	73.3	66.0	69.7	71.4	70.2	81.3	79.4	75.2	74.4	79.3

TABLE 5.5: OLS Regression Summary – Extraversion-level Grade against Extraversion-level

Linear Equation: $\hat{G}_E = \beta_0 E + \beta_1$

$= 1.269E + 62.422$

Feature	**Coeff.**	**r**	***p*-value**	**Coeff. 95% CI**
E	1.269	0.846	0.000	[0.771, 1.767]
Intercept	62.422		0.000	[58.354, 66.491]

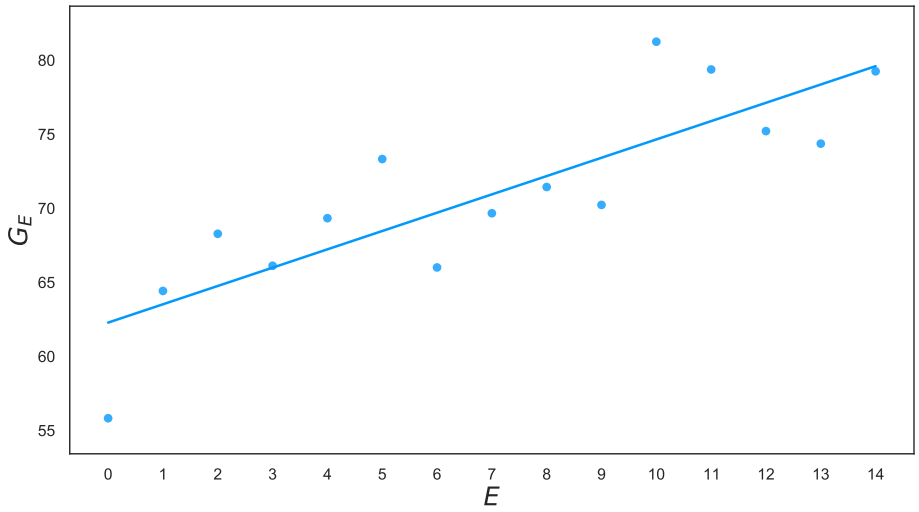


FIGURE 5.4: Extraversion-level Grade against Extraversion-level. $r = 0.846$, $p = 0.000$

5.2.1 Results and Discussion: Extraversion

Figure 5.4 shows that students in higher Extraversion-levels tend to achieve higher Grades, on average. Although the increase in the mean Grade with Extraversion-level is apparent from the line of best-fit, this claim is confirmed by the OLS Regression model's output. The fit described in OLS Summary Table 5.5 shows a linear relationship, $\hat{G}_E = 1.269E + 62.422$, The p-values of 0.000 signify that $\hat{G}_E = 1.269f + 62.422$ is not a relationship by chance. Also, the high r of 0.846 signifies that G_E moves closely with E and can be inferred from E with a 95% confidence that $\beta_0 \in [0.771, 1.767]$.

The positive β_0 coefficient of 1.269 signifies that the average Grade of students in higher Extraversion-levels is higher than the average Grade of students in lower

Extraversion-levels. While one more post than the last may not result in an additional 1.269 points to a student's Grade record, the average Grade of students who contributed to discussions more frequently, in general, was higher than the Grades of students who posted less often. Although this model accounts for the observed effect on Grade of only one independent variable, Extraversion, the probability (p-value) of Extraversion having no relationship with Grade is 0. This shows a statistical significance of Extraversion as a regressor against student Grade.

Post Count serves as a suitable proxy for Extraversion that will be used as a trait of the e-Behaviour model in Section 6.2.

The following section contains two experiments that examine the relationship between Academic-group and Grade. Each Academic-group is a subset of the cohort.

5.3 Academic-groups – Discussion and Collaboration-groups

In modelling Academic-groups, it is useful for us to understand:

1. how to model the patterns that form Academic-groups
2. the difference between the average Grades of the modelled Academic-groups, and
3. the extent to which a random student's Grade correlates with the rest of the students' Grades in his Academic-group

The above points are resolved in two sections:

1. Section 5.3.1 relates the Grades of *Discussions* and Grades of its constituent students. The Discussion relationships are modelled using an OLS Regression model, while
2. Section 5.3.2 relates the Grades of *Collaboration-groups* and the Grades of students in each Collaboration-group. The Collaboration-group relationships are modelled using the kNN and OLS Regression algorithms.

Both Sections 5.3.1 and 5.3.2 make use of the Student, Discussion, Message, Time and Grade variables that are shown in Table 5.2.

5.3.1 Student Discussions

This section analyses the relationships between the Grades of students within each Discussion. Here, a Discussion (group), d_i, is defined as any discussion that contains more than two students created on the Moodle LMS. A Discussion that has fewer than three students is not considered a Discussion by our definition. Linear OLS assumptions for Discussions containing only two or more students do not hold. Section 5.3.3 shows the reasons.

TABLE 5.6: Discussion Table – Random Student's Grades against Discussion's Grade Averages

$\mathbf{d_i}$	$\mathbf{s_i}$	$\mathbb{E}[\mathbf{Gd_i}]$	$\mathbf{Gd_i(s_i)}$
0	23	83.08	92.75
1	728	56.81	59.25
2	833	49.75	48.50
⋮	⋮	⋮	⋮
336	79	75.35	90.75
337	15	74.87	70.25

Let

- $D = \{d_i\}_{i=1}^{k} = \{d_1, d_2, \ldots, d_k\}$ be the set of all Discussions,
- s_j represent any student who participates in discussion d_i,
- s_i represent a selected random student who participates in discussion d_i,
- $\mathbb{E}[Gd_i]$ represent the mean Grade of Discussion d_i, and
- $G(s_j)$ is the grade of student s_j.

$$\mathbb{E}[Gd_i] = \frac{1}{n(d_i) - 1} \sum_{s_j \neq s_i} G(s_j), \tag{5.1}$$

where k represents the number of Discussions in D and $n(d_i)$ denotes the number of students in d_i. This section measures the correlation between $Gd_i(s_i)$ and $\mathbb{E}[Gd_i]$ by following Algorithm 1:

Algorithm 1: Correlation Between Mean Discussion Grade and Student Grade

Result: $\hat{G}d_i(s_i) = \beta_0 \mathbb{E}[Gd_i] + \beta_1$

foreach $d_i \in D$ **do**
    Select s_i
    Obtain $Gd_i(s_i)$ from s_i
    Compute $\mathbb{E}[Gd_i]$
    Plot $(Gd_i(s_i), \mathbb{E}[Gd_i])$
end

In Figure 5.5, each point, $(\mathbb{E}[Gd_i], Gd_i(s_i))$, relates $\mathbb{E}[Gd_i]$, the mean Grade of a *unique* Discussion, d_i along the x-axis, to $Gd_i(s_i)$ the randomly-selected student's Grade, along the y-axis. Extraversion-levels are ordinal, with each level indicating the number of posts in that level. Therefore, an E of 1 is a *lower* Extraversion-level than an E of 2. Note that there are 36 observations because only Discussions that satisfied the criteria of having more than two students were considered – Section 5.3.3

TABLE 5.7: OLS Regression Random-Student's Grades against Discussion's Grade Averages

$$\text{Linear Equation: } \hat{Gd}_i(s_i) = \beta_0 \mathbb{E}[Gd_i] + \beta_1$$
$$= 0.528 \mathbb{E}[Gd_i] + 35.607$$

Feature	**Coeff.**	**r**	***p*-value**	**Coeff. 95% CI**
$\mathbb{E}[Gd_i]$	0.528	0.421	0.011	[0.131, 0.925]
Intercept	35.607		0.014	[7.789, 63.425]

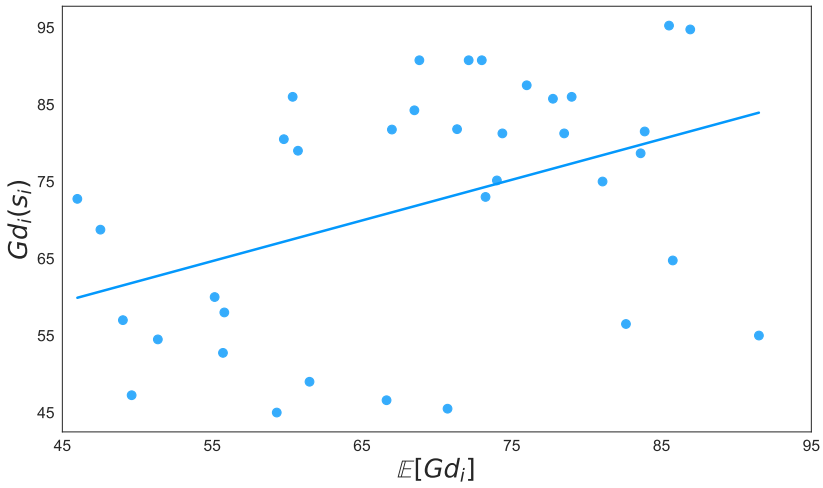


FIGURE 5.5: Random Student's Grades against Discussion's Grade Averages – Positive Correlation.

explains the reasons for the Discussion size limitation. Figure 5.5 shows the regression line fitting the $\mathbb{E}[Gd_i]$ and $Gd_i(s_i)$ values in Table 5.6, while Table 5.7 shows the corresponding OLS Regression Summary. The linear equation of the line of best-fit is given by $\hat{Gd}_i(s_i) = 0.528\mathbb{E}[Gd_i] + 35.607$. The β_0 coefficient of $\mathbb{E}[Gd_i]$ and its statistical significance (p = 0.011) indicates that the a marginal increase in $\mathbb{E}[Gd_i]$ of 1 Grade point corresponds to an average increase of 0.528 in $Gd_i(s_i)$. The $r(\mathbb{E}[Gd_i], Gd_i(s_i))$ of 0.421 suggests that there is a strong correlation between the mean Grade of a Discussion ($\mathbb{E}[Gd_i]$) and the Grade of a student, ($Gd_i(s_i)$), chosen at random, who participated in that Discussion. This correlation also holds for any other set of randomly-selected students. Section 5.3.4 contains a detailed analysis of the above results, along with its empirical meaning and comparison with Collaboration-groups.

The OLS assumption tests (Sections A.2.2.2) show that an OLS model is appropriate

for modelling the relationship between student Discussion data and Grade. In addition to the linear relationship between $\mathbb{E}[Gd_i]$ and $Gd_i(s_i)$, Figure 5.5 shows the presence of two clusters. Analyses on these clusters are discussed in the next section on the Formation of Clusters.

The Formation of Clusters

A k-means algorithm partitions the data into two clusters: one cluster in the lower-left and another in the plot's upper-right quadrant. We call these clusters the *Blue* Discussion Cluster and *Green* Discussion Cluster, colour-coded in Figure 5.7. This figure shows the same points that appear in Figure 5.5 without the regression line. This section investigates a probable cause of the separation between the Grades of these clusters by examining the discussions' contents within each cluster.

Each word's usage in the Discussions that form the two clusters reveals different sentiment in each cluster's Discussion messages. Each word's frequency was divided by the number of words in all Discussions. The resulting quotient is each word's usage frequency relative to the usage frequency of other words across all Discussions.

Figure 5.6 shows a heatmap that compares the frequency (as a percentage of all words mentioned in the cluster) of the top 10 most frequent words per Discussion Cluster.

The value in each cell represents the number of times, as a proportion of the number of words (not unique) in the Discussion Cluster. For instance, the word 'thank' appears 2.6 times for every 100 words written in the Blue Cluster and 0.64 times per 100 words in the Green Cluster. The word 'use' appeared with the highest frequency of all words in the Green Cluster, followed by 'class', 'function' and 'value'. The meaning or context of certain words may provide additional insight into the reason for the diction in each cluster. However, words such as 'thank' and 'helpful' may be taken, without context, to represent gratitude within the Blue Cluster.

On the other hand, the nature of the sentiments in the words 'class', 'function' and 'value' seem to relate specifically to a discussion topic in the Green Cluster. While the most frequently used words in each cluster do not provide information of the context, a look into the content of the messages containing the most frequent words may give some context if the aim is to analyse student sentiments further. The top 50 words and their relative frequencies are shown in Figure A.3 of Appendix A.

The difference between the Blue and Green Clusters' distributions of $Gd_i(s_i)$ and $\mathbb{E}[Gd_i]$ suggest that the Blue and Green clusters can be appropriately named the *Low-achieving* Discussion Cluster and *High-achieving* Discussion Cluster, respectively. The high presence of the word 'thank' in the Blue Cluster suggests that the purpose of a message from the *Low-achieving* Cluster is more likely than any other message in the cluster to be a message of gratitude. That is, students posted a resource, such as a 'wikipedia' or 'khanacademy' link that drew the 2.6% 'thank's in response to the supposedly-*helpful* resources. By contrast, the number of 'thank' words in the *High-achieving* Cluster appears less often. The *High-achieving* discussions appear to be more example-based, with a lower percentage of words showing gratitude. The contrast in word-usage leads to a hypothesis that the two clusters

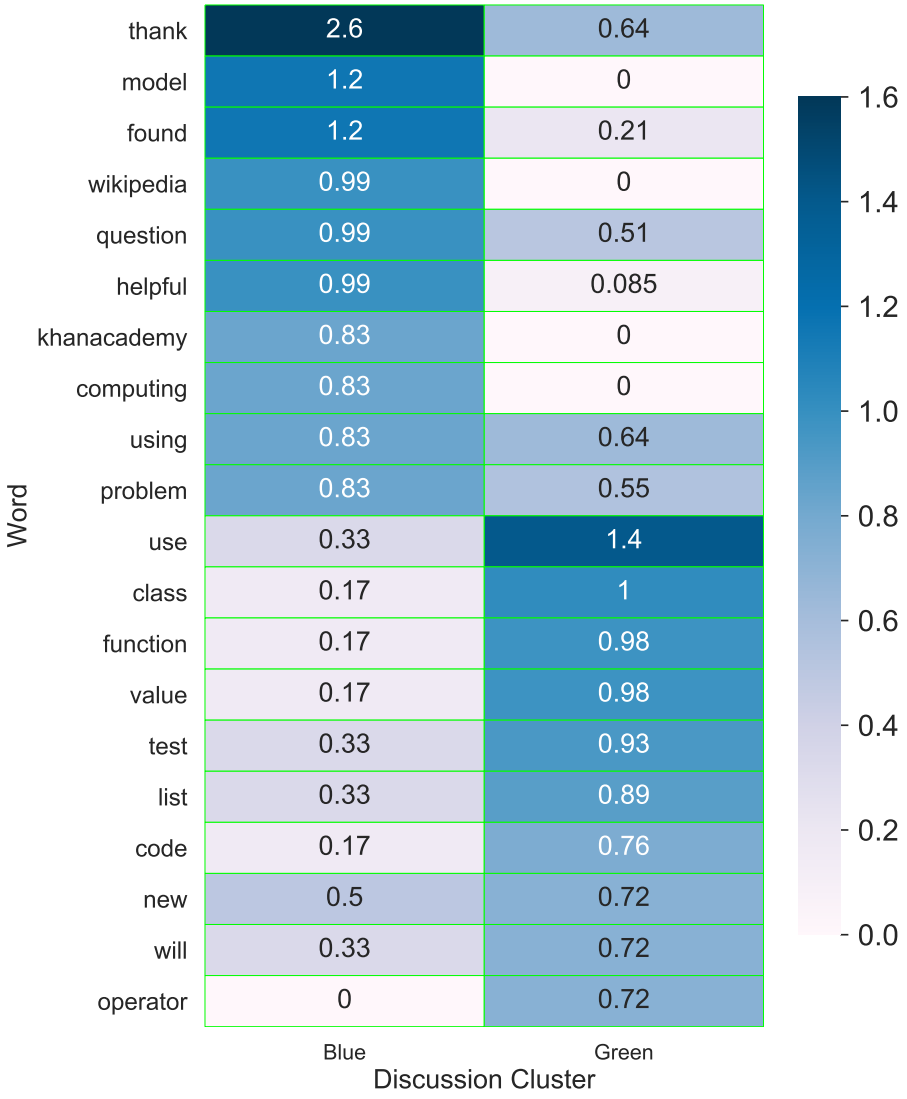


FIGURE 5.6: Word Frequency per Discussion Cluster

tend to discuss different topics, and the *Low-achieving* Cluster seems to contain students who sought help more than students in the High-achieving Cluster. Also, fewer words in the *Low-achieving* Cluster showed topic-appropriateness as opposed to the *High-achieving* Cluster. The analysis in this section shows that the content of a discussion matters; in this case, the users in the *Low-achieving* Discussion Cluster use words that appear to show gratitude for the help received from others. By contrast, words in the *High-achieving* Discussion Cluster appear to contribute examples that explain content more clearly. Furthermore, there are 24 more Discussions created and 19 more students in the *High-achieving* cluster than its *Low-achieving* counterpart, which may illustrate a greater level of interaction among students in the former cluster; a possible reason for the performance superiority of the *High-achieving* Cluster.

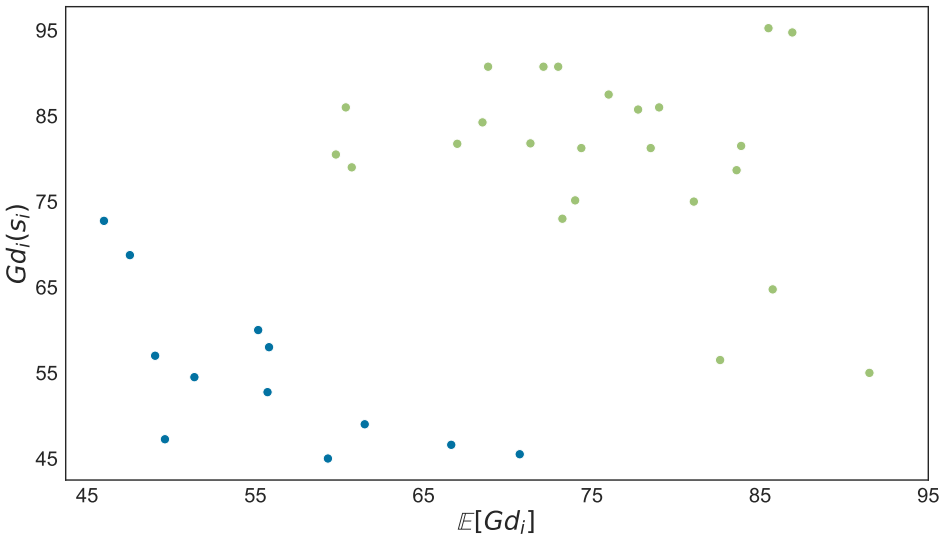


FIGURE 5.7: Clusters – Random-Student's Grades against Discussion's Mean Grade

Model Evaluation – Discussions and the Formation of Clusters

Given that the data in studies under sections 5.2 and 5.3.2 passed the OLS assumptions tests, their linear fit's p-value is sufficient verification against chance. The p-value of the linear relationship given by $\hat{Gd}_i(s_i) = 0.528\mathbb{E}[Gd_i] + 35.607$ tests for the probability of a linear relationship, however, it does not test for the probability for the formation of clusters in the $(\mathbb{E}[Gd_i], Gd_i(s_i))$ scatter plot (Figure 5.5). This section aims to verify whether these clusters occured by chance, through

1. assigning each student a randomly generated Grade, $Gd_i(s_i)^*$, sampled from a distribution that is identical[3] to that of $Gd_i(s_i)$.

2. observing the new linear relationship between $Gd_i(s_i)^*$ and $\mathbb{E}[Gd_i]^*$, and

3. investigating the presence of clusters in the plot relating $Gd_i(s_i)^*$ and $\mathbb{E}[Gd_i]^*$.

The presence of clusters in the fabricated $Gd_i(s_i)^*$, $\mathbb{E}[Gd_i]^*$ relationship will inform whether the presence of clusters in the true $Gd_i(s_i)$, $\mathbb{E}[Gd_i]$ relationship was accidental. For each s_i, Table 5.8 contains the true Grade, $Gd_i(s_i)$, alongside the randomly-assigned Grade[4], $Gd_i(s_i)^*$. The low, statistically insignificant r-value of 0.015 between $Gd_i(s_i)$ and $Gd_i(s_i)^*$ assures that the random Grade allocated to each student in the control sample, $Gd_i(s_i)^*$, is not similar to his true Grade, $Gd_i(s_i)$.

[3] A normal distribution with the same mean, median, standard deviation, minimum and maximum values.

[4] See Figure A.10 for the programme that generated the random grades

TABLE 5.8: Sample: Random-Grade Generation for Students

s_i	**True Grade** $Gd_i(s_i)$	**Random Grade** $Gd_i(s_i)^*$
1	49.25	37.53
2	69.75	59.21
3	89.33	54.01
...	...	...
1131	81.80	56.56
1132	66.28	47.76
1133	78.66	55.07

TABLE 5.9: OLS Regression Random-Student's Grades against Discussion's Grade Averages

$$\text{Linear Equation: } \hat{G}d_i(s_i)^* = \beta_0 \mathbb{E}[Gd_i]^* + \beta_1$$
$$= 0.157 \mathbb{E}[Gd_i]^* + 54.260$$

Feature	**Coeff.**	**r**	***p*-value**	**Coeff. 95% CI**
$\mathbb{E}[Gd_i]^*$	0.157	0.105	0.570	[-0.401, 0.716]
Intercept	54.260		0.005	[17.477, 91.043]

The linear relationship between $Gd_i(s_i)^*$ and $\mathbb{E}[Gd_i]^*$ is shown in Table 5.9. β_0's coefficient's high p-value of 0.570 suggests insufficient evidence against the absence of a linear OLS relationship between $Gd_i^*(s_i)$ and $\mathbb{E}[Gd_i]^*$. Furthermore, the 95% confidence interval lower- and upper-bounds of contain 0. Therefore, if the clustering and linear equation, $\hat{G}d_i(s_i) = 0.528\mathbb{E}[Gd_i] + 35.607$, observed from the true Grade relationship were by chance, then there should be no reason for $\hat{G}d_i(s_i)^* = 0.157\mathbb{E}[Gd_i]^* + 54.260$ to not generate a clustering as similar to Figure 5.7's, or at least produce a statistically-significant $\mathbb{E}[Gd_i]^*$ slope coefficient. Therefore, there is sufficient evidence that the true Grade clustering between $Gd_i(s_i)$ and $\mathbb{E}[Gd_i]$ was not by chance.

5.3.2 Student Collaboration-groups

The previous section showed the Discussion variant of the Academic-group formulation. This section illustrates an alternative method to formulating an Academic Group, namely, the Collaboration-group method. We correlate the Grades of students within each Collaboration-group with the mean Grade of each Collaboration-group.

The Forum Table (5.2) was transformed into Table 5.10 below, which shows discussion participation per student. Each column, d_i, represents a discussion. 1 represents

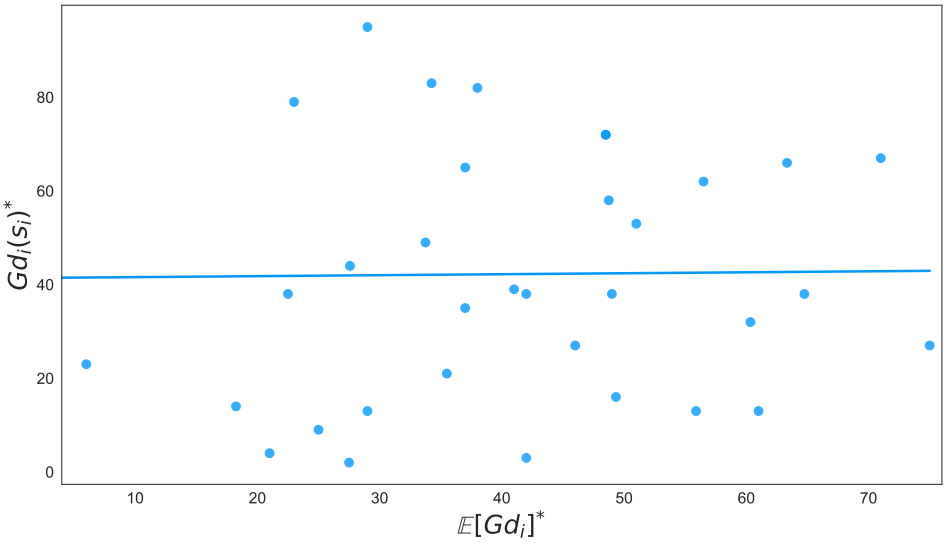


FIGURE 5.8: Random Student's Random Grades against Discussion Grade Averages. $r = 0.105$, $p = 0.570$.

that the student participated in discussion d_i, while 0 shows that he did not participate in d_i.

TABLE 5.10: Sample Table of Discussion Participation

s	$\mathbf{d_0}$	$\mathbf{d_1}$	$\mathbf{d_2}$	$\mathbf{d_3}$	$\mathbf{d_4}$	...	$\mathbf{d_{337}}$
1	0	0	0	0	0	...	1
2	1	1	0	0	0	...	0
3	0	0	1	1	0	...	0
...	...	...	...	...	...	...	...
1131	0	1	0	1	0	...	0
1132	1	0	0	0	0	...	0
1133	0	1	0	0	0	...	1

Let

- $C = \{c_i\}_{i=1}^{k} = \{c_1, c_2, \ldots, c_k\}$ be the set of all Collaboration-groups.
- h_i be the *Host* of c_i
- $H = \{h_i\}_{i=1}^{k} = \{h_1, h_2, \ldots, h_k\}$ be the set of all Hosts, one for each Collaboration-group.

A Collaboration-group, c_i, that is *hosted by* student, h_i, is defined as the group of more than two students with whom h_i shares at least one discussion. Any group that has two or fewer students is not considered a Collaboration-group by our definition;

OLS relationships analogous to those in this section do not hold for groups containing only two or more students. The reasons are presented under Section 5.3.3.

h_i may host a maximum of one Collaboration-group. Let

- $\mathbb{E}[Gc_i]$ represent a the mean Grade of c_i, and
- $\mathbb{E}[Gc_i]_i(h_i)$ represent the Grade of h_i,

where $\mathbb{E}[Gc_i]$ represents the mean Grade of c_i which excludes $Gc_i(h_i)$, as in the case with $\mathbb{E}[Gd_i]$ and $Gd_i(s_i)$ in Equation 5.1.

The kNN algorithm was used to compute the Collaboration-group for each student, using the **Collaboration-group policy** specified in the below paragraph. The kNN algorithm responsible for the policy's construction is shown in Figure A.8 of Appendix A. By this policy, not all students fit the qualify to host a Collaboration-group.

We design the conditions necessary to define the Collaboration-group policy; let:

- h_* be a *candidate* Host of a Collaboration-group,
- c_* represent the Collaboration-group to be hosted by student, h_*,
- $n(c_*)$ represent the number of students in c_*,
- s_1, s_2 and s_3 be any three students in the cohort, and
- h_i represent a (qualified) Host to his (unique) Collaboration-group, c_i

Collaboration-group Policy:

- c_* becomes a Collaboration-group, c_i, if and only if $n(c_*) > 2$ students

Equivalently,

- If h_* shares a discussion with s_1, s_2 and s_3, then h_* *qualifies* as a Host, h_i, and $c_i = \{s_1, s_2, s_3\}$
- If $n(c_*) \leq 2$ students, then h_* remains a candidate until he shares a discussion with at least one more member.

Table 5.11 shows a sample set of the Hosts, $\mathbf{h_i}$, and their Collaboration-groups, $\mathbf{c_i}$. Each entry in column $\mathbf{c_i}$ is a set of indices that represent students in c_i, while column $\mathbf{Gc_i}(\mathbf{h_i})$ shows the Grades of the Hosts. $\mathbb{E}[\mathbf{Gc_i}]$ represents the mean Grades of each c_i. Table 5.12 and Figure 5.9 show the OLS regression results of the fit between $Gc_i(h_i)$ and $\mathbb{E}[Gc_i]$.

TABLE 5.12: OLS Regression Summary Random-Student's Grades against Collaboration-group's Grade Averages

Linear Equation: $\hat{Gc}_i(h_i) = \beta_0 \mathbb{E}[Gc_i] + \beta 1$
$= 0.984 \mathbb{E}[Gc_i] + 5.175$

Feature	**Coeff.**	**r**	***p*-value**	**Coeff. 95% CI**
$\mathbb{E}[Gc_i]$	0.984	0.479	0.004	[0.334, 1.663]
Intercept	5.175		0.797	[-35.501, 0.975]

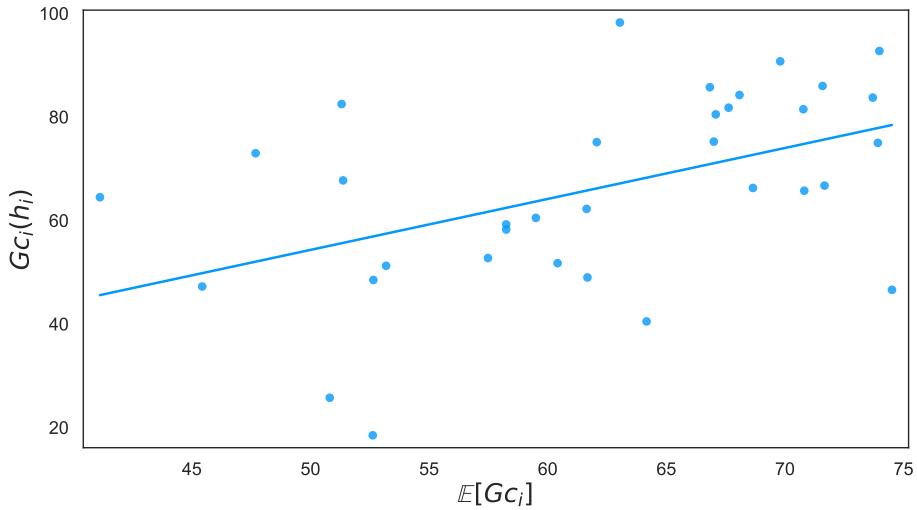


FIGURE 5.9: Random-Student's Grades against Collaboration-group's Grade Averages. $r = 0.479$, $p = 0.004$.

TABLE 5.11: Collaboration-groups and Grades

$\mathbf{h_i}$	$\mathbf{c_i}$	$\mathbf{Gc_i(h_i)}$	$\mathbb{E}[\mathbf{Gc_i}]$
1	**{5, 48, 3, 138}**	73.00	47.68
2	**{119, 172, 199}**	81.80	67.62
3	**{40, 35, 20, 16, 51}**	90.75	69.80
4	**{90, 200, 28, 33, 94, 142, 42, 101, 84}**	49.00	62.08
5	**{81, 209, 143, 206, 12, 150}**	98.25	63.04
6	**{142, 33, 28, 42}**	59.25	58.25
7	**{65, 190, 107, 8, 173}**	46.60	74.51

Refer to Table 5.12. The coefficient of $\mathbb{E}[Gc_i]$ shows that the a marginal increase in $\mathbb{E}[Gc_i]$ of 1 Grade point corresponds to an estimated increase of 0.984 in $Gc_i(h_i)$.

The $r(\mathbb{E}[Gc_i], Gc_i(h_i))$ of 0.479 suggests that there is a strong correlation between the average Grade of a Collaboration-group – $\mathbb{E}[Gc_i]$ – and the Grade of its Host –$Gc_i(h_i)$. The Simple OLS Regression assumptions were used to validate the reliability of the results. See Section A.2.2.5 of the Appendix (A). Section 5.3.4 contains a further analysis of the Collaboration-Group results, along with its empirical meaning and comparison with the Discussion method.

5.3.3 Academic-group Size Constraints

Each Discussion and Collaboration-group were constrained to a minimum size of 3. When the sizes were reduced to 2, all linear relationships between the student's Grade and the group's average Grade collapsed and were statistically insignificant. Furthermore, residuals, $Gd_i(s_i) - \hat{G}d_i(s_i)$ (for Discussion-Grade relationships) and $Gc_i(h_i) - \hat{G}c(h)$ (for Collaboration-group-Grade relationships) are not normally distributed for the group sizes of 2.

Sections 5.3.1 and 5.3.2 on Student Discussion and Collaboration-groups proved the presence of a positive linear trend between the groups with Grades. The next section compares the two methods.

5.3.4 Results and Discussion: *Discussions* and *Collaboration-groups*

The differences, advantages, and disadvantages of each Academic-group formulation are discussed in this section. This section ends with each formulation's practical significance. The similarities and differences in the Academic-group formulations are summarised in Table 5.13.

TABLE 5.13: Discussion and Collaboration-group Differences

ATTRIBUTE	DISCUSSION	COLLABORATION-GROUP
1. Composition	Participants of a forum discussion	Set of students with whom a stud--ent shares at least one discussion
2. Min (Max) St--udents per group	3 (No limit)	3 (No Limit)
3. Group limit per student	No Limit to the number of Discussions he can belong to	A student may host a maximum of 1 Collaboration-group
4. Correlation Drawn	Estimated Grade of a Discussion Participant as a function of the Mean Grade of the Discussion: $\hat{G}d_i(s_i) = \beta_0\mathbb{E}[Gd_i] + \beta_1$	Estimated Grade of a Host a as function of the Mean Grade of his Collaboration-group: $\hat{G}c_i(h_i) = \beta_0\mathbb{E}[Gc_i] + \beta_1$

The differences highlighted in Table 5.13 can be used as a guideline for choosing a method to use in practice. We compare the efficacy and implications of using either method of revealing Academic-groups. Differences in attributes 1, 2 and 3 are

TABLE 5.14: Academic-group Comparison: Discussion and Collaboration-group

	Target Var.	**Indep. Var.**	β_0	r	**p-value**	**95% CI**
Discussion	$Gd_i(s_i)$	$\mathbb{E}[Gd_i]$	0.528	0.421	0.011	[0.131, 0.925]
Collaboration	$Gc_i(h_i)$	$\mathbb{E}[Gc_i]$	0.984	0.479	0.004	[0.334, 1.633]

structural differences that result from the formulations of either group type. Attribute 4 makes a comparison between each formulation's linear correlation with Grade.

Magnitude of Correlation and Practical Significance

Table 5.14 compares the slope coefficients (β_0), r, p and confidence intervals between the Discussion-Grade and Collaboration-group-Grade relationships. This table addresses the **Correlation Drawn** attribute of Table 5.13, taken directly from the OLS models in Tables 5.7 and 5.12. Beginning with the comparison of slope coefficients (β_0), observe that the magnitude of the slope coefficient for $\mathbb{E}[Gc_i]$ is greater than the slope coefficient of $\mathbb{E}[Gd_i]$. The larger coefficient suggests that $Gc_i(h_i)$ moves by a greater magnitude with $\mathbb{E}[Gc_i]$ than $Gd_i(s_i)$ does with $\mathbb{E}[Gd_i]$. If causality could be proven, we would say that a student's Collaboration-group has a larger *influence* on his Grade than his Discussion does. $r(\mathbb{E}[Gc_i], Gc_i(h_i))$ of 0.479, and its lower p-value of 0.004, suggests that: $\hat{Gc}_i(h_i) = 0.984\mathbb{E}[Gc_i] + 5.175$ is a more reliable model to infer a student's grade, $Gc_i(h_i)$ as compared to $\hat{Gd}_i(s_i) = 0.528\mathbb{E}[Gd_i] + 35.607$ to infer a student's Grade, $Gd_i(s_i)$.

Furthermore, the Discussion method includes the limitation that a student may belong to more than one Discussion. By contrast, he may belong to only one Collaboration-group. To illustrate the implications of this difference, suppose that the same student s_m participated in Discussion d_1 and d_2. Suppose $\mathbb{E}Gd_1 = 50$, $\mathbb{E}Gd_2 = 60$ and $\mathbb{E}Gd_3 = 70$. Therefore,

$$\hat{Gd}_1(s_m) = 0.528(50) + 35.607 = 62.01, \tag{5.2}$$

$$\hat{Gd}_2(s_m) = 0.528(60) + 35.607 = 67.29, \tag{5.3}$$

and

$$\hat{Gd}_3(s_m) = 0.528(70) + 35.607 = 72.57, \tag{5.4}$$

are three *different* estimates for the Grade of s_m; each estimate is based on each on the three Discussions in which he participated. As a result, using the Discussion method to infer a single student's Grades is not practical. By contrast, Collaboration-groups produce a single Grade prediction, $\hat{Gc}(h)$ for a student h_i. The design of

a Collaboration-group is slighlty more involved, using kNN and OLS regression, however, its relationships with Grade make the Collaboration-group method a more reliable inference model.

A Deeper Look into Student Grades and Level of Interaction

The Forum data reveals that the median membership per discussion is 2 students. The median number of posts per student is 1. As a result, the expected number of opinions or views that a student can ponder (including his own) after entering a random discussion is $2 \times 1 = 2$ opinions. By simply taking part in a discussion, a student obtains 1 extra opinion, on average. At worst, the extra opinion will not add value to the student's understanding of a topic. However, he has the option to reinforce his knowledge or spark his curiosity about a discussion's topic by participating. To illustrate the effect of being part of an Academic-group, we consider five Academic-categories that extend the original Academic-groups. Both Academic-groups (Discussions – Category 4 – and Collaboration-groups – Category 5 – are also included as Academic-categories). Recall that an Academic-group is either a Discussion or Collaboration-group. By contrast, an Academic-category is described by one of the five categories in this list which includes both Academic-groups. The Non-Participating, Participating and Groupless categories were added to show the difference between Grades of students who do not participate in forums (Non-Participating) and those who do (Participating). Formally, a Student may:

1. not participate in forums (Non-Participating),
2. participate in forums (Participating),
3. not be part of any Academic-group (Participating and Groupless),
4. be part of a Discussion (Participating and Discusser),
5. be part of a Collaboration-group (Participating and Collaborator).

The Non-Participating category contains students appearing in none of the other categories. The Groupless, Discusser and Collaborator categories are subsets of the Participating category. Table 5.15 shows the measures of central tendency and dispersion of Grades within each Academic-category.

For ease of visualisation, a colour scale runs across each row in Table 5.15, where green indicates a higher value and red shows a lower value. The Non-Participating category Grades were the lowest across all central tendency measures except for the standard deviation. The Discusser Academic-category scores highest across four of the seven measures of central tendency, including the mean and median. That is, a student who merely belongs to a Discussion is likely to achieve a higher score than a student in any of the other categories. Note, however, that the pair-wise t-test for the difference in the means between Groupless, Discusser and Collaborator categories indicates that the difference in their mean Grades is not statistically significant – the above three categories' Grades were all sampled from the Participating category.

TABLE 5.15: Summary Statistics Comparison of Academic-group Categories

Measure	**Non Part.**	**Part.**	**Grou-pless**	**Disc-usser**	**Collab-orator**
Mean	55.83	67.49	67.22	70.50	66.16
Std. Dev.	**18.44**	**14.36**	**13.30**	**15.61**	**18.81**
Min	1.00	18.50	20.00	45.00	18.50
25^{th} Perc.	45.23	59.25	59.50	56.50	52.00
50^{th} Perc.	58.00	68.25	68.25	75.00	66.52
75^{th} Perc.	68.25	78.75	76.75	81.75	81.73
Max	94.00	98.25	96.00	95.25	98.25

However, since the average Grades of Participating and Non-Participating categories are statistically different from each other, the results presented in the table confirm that students who participated in at least one discussion (Participating students) completed their programmes with a better Grade than students who did not participate in any discussions.

5.3.4.1 Academic-Groups and Social Capital

The Extraversion-Grade relationship is linked to the social science concept of *social capital*, which was used as a theoretical basis for our formation of Academic-groups. Romero et al. (2008) obtained a classification accuracy of 60% for the expected grade category of a student (Fail, Pass, Good, Excellent) against his LMS behaviour. The features used are shown in row 1 of Figure A.3, among which is the number of messages sent to a forum (Extraversion-level).

Bhandari and Yasunobu (2009) and other research do not illustrate the quantitative effect of social capital. However, the authors cite that 'an individual who creates and maintains social capital subsequently gains advantage from it'. The quantification of the perceived effect of social capital was illustrated by the correlation between the Grade of a student in an Academic-group, $Gc_i(h_i)$, and the average Grade of the group, $\mathbb{E}[Gc_i]$. $Gc_i(h_i)$ responded with a statistically-significant increase of 0.984 to a $\mathbb{E}[Gc_i]$ increase of 1.

Despite showing different insights and patterns, both the Discussion and Collaboration-group methods show that the *quality* of a student's academic output (Grades) is associated with the quality of academic output of his social capital. As stated in Section 2.1, a student may choose to *leverage* his social capital (that is available to all students in a cohort) by becoming part of an Academic-group.

Our Academic-group and Grade relationship, findings like Romero et al. (2008)'s and the above statement by Bhandari and Yasunobu (2009) provide evidence for the positive relationship between student success and the accumulation of social capital. This chapter's work has contributed to the theory that:

> *The quality of a student's social capital is the quality of his Academic group's performance.*

5.4 Conclusion

This chapter discussed the usage of Forum Post data. Forum Posts were used to draw relationships between student Extraversion-levels, Academic-groups and Grades. The relationships revealed that the number of posts is a statistically-significant independent variable for student Grades.

This chapter used quantitative techniques to reveal high-level word sentiments in Discussion clusters. Although this report aims to address quantitative relationships, it was suggested that qualitative procedures can be used to analyse word sentiment between Discussion clusters in greater depth.

Two methods of designing Academic-groups were explored: Discussions and Collaboration-groups. Each of the Academic-groups' merits and problems were shown. Statistical correlations between Grades and Academic-groups were drawn, and comparisons were made between the groups. Both formulations of Academic-groups suggest that a student's Grade resembles the Grade of his Academic-group.

Chapter 6 explains the use of a model that predicts a student's Outcome. In the model, Extraversion is a predictor in addition to Conscientiousness and Login e-Behaviour.

Chapter 6

e-Behaviour, Personality, Performance and Intervention

Chapter 2 briefly introduced the importance of considering a student's personality as an influence on his academic performance. Chapter 5 showed the formulation of a student's Extraversion-level, which will be used in this chapter, alongside his Conscientiousness-level. This chapter relates the Login e-Behaviour, personality and performance of students with Grades and Outcome[1]. Section 6.2 quantifies the change in the Behaviour Model's predictive power after augmenting Conscientiousness-level and Extraversion-level. Section 6.5 concludes this dissertation's methodology with a demonstration of an intervention mechanism based on student e-Behaviour and Safety Scores[2].

The Moodle LMS generates Login logs every second. Therefore, we can choose to resample the Login data logs over any period longer than a second. The section below compares the use of daily-resampled data with weekly-resampled data.

6.1 Choice of Resampling Periods for Login Patterns

The University delivers lectures, tutorials (and, at times, assignments) over weekly cycles. This section compares the practicality of daily aggregations with the practicality of weekly login aggregates. Let L_{day} represent the sequence of the mode number of daily logins[3] and L_{week} be the sequence of the mode number of weekly logins across the cohort. To verify whether the student Login e-Behaviours express a weekly cycle pattern, we plotted correlograms to identify the autocorrelation values at different lags for L_{day} and L_{week}. Visualising the correlogram helps determine a suitable Login data frequency as input to the time-series data for the LSTM-based behaviour and personality models in Sections 6.2.2 and 6.2.3.

Figure 6.1 shows the 114-day autocorrelation function correlogram for L_{day}. The figure shows that the L_{day} pattern's first timestep correlates significantly at time lags 1, 2 and 3, showing autocorrelation values of 0.121, 0.115 and 0.08, respectively. This autocorrelation pattern drops into a level of insignificance at lag 4 and 5, with a value of 0.018 and 0.052, respectively. There appears to be autocorrelation at later lags. However, this autocorrelation is without a visible pattern.

[1] Grades and Outcome defined in Section 3.5

[2] Safety Score is defined as the model's prediction of a student's Outcome in Section 3.5

[3] Number of logins by most people on each day

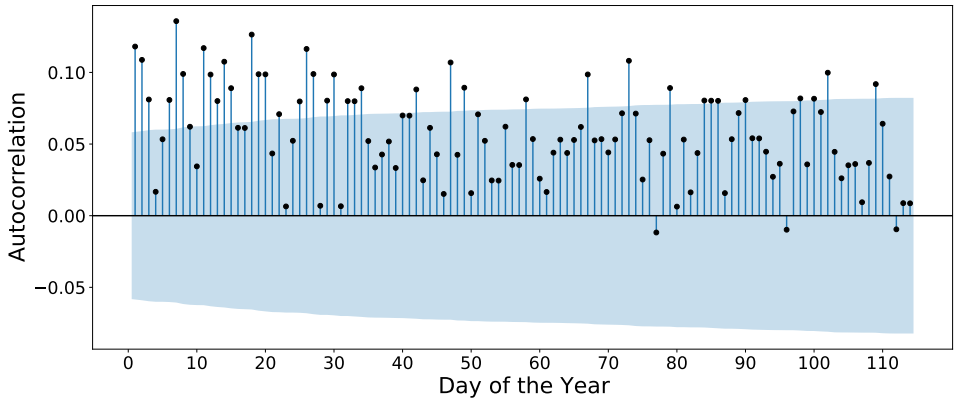


FIGURE 6.1: Correlogram for the 114-Day Period. No significant autocorrelation after day 3.

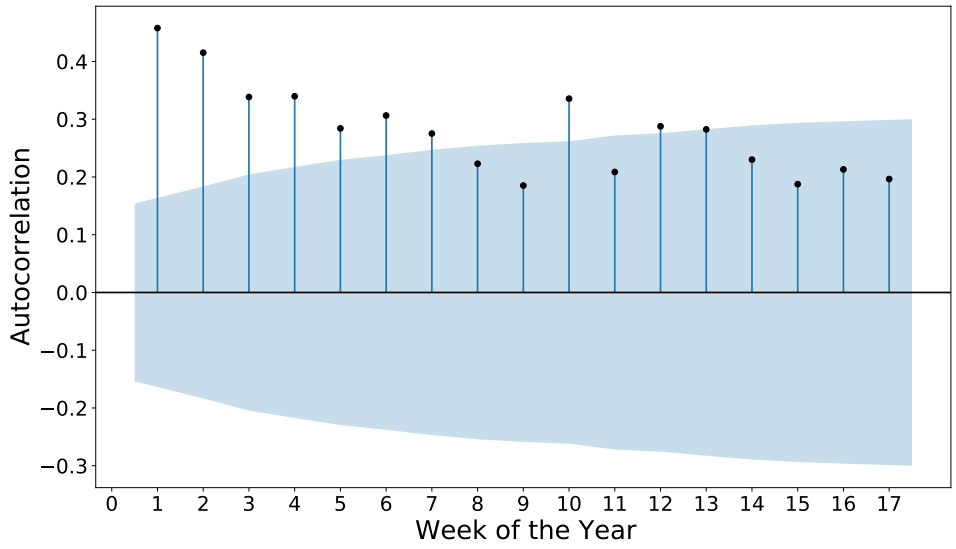


FIGURE 6.2: The Correlogram for the 17-weekPeriod. Significant autocorrelation up to lag 7.

Figure 6.2 shows the resampled 17-week correlogram for L_{week}. The L_{week} autocorrelation values show a more visible autocorrelation pattern, especially at shorter lags, with a decreasing autocorrelation pattern between lags 1 and 5. Furthermore, the significant autocorrelation values for L_{week} extends to lag 7. Finally, the autocorrelation values between shorter lags – lag 1 to lag 7 – range between 0.47 and 0.28 and are statistically significant at a 95% confidence interval. These values are higher than the statistically significant autocorrelation values for L_{day}, ranging between 0.121 and 0.08 between lags 1 and 3. Since time-series-based prediction models learn order dependence on sequential data, a model performs better when the order dependence is predictable.[4]

The autocorrelation values and weekly academic assessment delivery suggest that a weekly-resampled Login Sequence represents a student's Login e-Behaviour than a

[4]See Section 3.7 for details on the LSTM architecture

daily-resampled Login Sequence. Therefore, the following sections use the weekly-resampled Login Sequence to model e-Behaviour and the Conscientiousness trait.

The term *Login Sequence*, written $\{L(s)_t\}$, represents the sequence containing each week's tally of days where student s logged in at least once. A week consists of 7 days. t indicates that each value in $\{L(s)_t\}$ is indexed by a time variable t. The upper bound for t is 17, since our period of study spans 17 weeks. In this chapter, $\{L(s)_t\}$ is shorthand for $\{L(s)_t\}_{t=1}^{17}$. Formally,

$$\{L(s)_t\} = \{L(s)_t\}_{t=1}^{17} = \{L(s)_1, L(s)_2, \ldots, L(s)_{17}\}, \tag{6.1}$$

where $L(s)_t$ is his number of active days during week t.
It follows that $L(s)_t \in \{0, 1, \ldots, 7\}$.

Table 6.1 is a sample of the Login data from the Moodle database. This study used logs belonging to the 1 133 students who are also present in the Grades Table. The Login Data was transformed into a time sequence, then resampled into the weekly Login Sequence, $\{L(s)_t\}$, for each student. $\{L(s)_t\}$ for each student is shown in Table 6.2. Each entry, (s,t), shows the number of active days that student s engaged in during week t.

TABLE 6.1: Sample of the Moodle Login Data Log

student	**timecreated**	**eventname**
108	1531402131	\core\event\user_loggedin
33	1531403377	\core\event\user_loggedin
568	1531414845	\core\event\user_loggedin
568	1531414845	\core\event\user_loggedin
...	...	...
835	1531415179	\core\event\user_loggedin
1083	1531416473	\core\event\user_loggedin

TABLE 6.2: Login Sequence, $\{L(s)_t\}$ for each Student per Week

	Week (t)						
Student (s)	**1**	**2**	**3**	...	**15**	**16**	**17**
1	0	0	1	...	3	0	0
2	0	2	4	...	1	2	0
3	0	1	3	...	2	0	1
...	...	...	...	...	...	...	...
1 132	0	2	6	...	2	4	0
1 133	0	1	1	...	0	3	0

6.1.1 Performance Analysis: The Imbalance of Outcomes

This section is a preliminary analysis of student performance and shows the need for balancing the dataset by student Outcomes. Table 6.3 summarises the number of

Students per Outcome group. Figure 6.3, shows the percentage of students active per week, summarised by Outcome group. To understand the cohort-wide behaviour per Outcome group, observe the bars from Week 1. Only 67.68% of the students who ended up At-risk of failing the year logged in once or more during week 1, while the percentage of active students who were Safe is 83.17% over the same week. The percentages of weekly active students for the At-risk Outcome group fall consistently below those for the Safe group, which means that the At-risk students were less active over the 17 weeks. From Figure 6.3, we can infer that different Outcomes are associated with different behaviours. By training the LSTM on $\{L(s)_t\}$ for all students and associated Outcomes, the LSTM considers how the login pattern of student s compares with other students' login patterns and their associated outcomes. Thus, each student's weekly Login activity sequence, $\{L(s)_t\}$, a suitable input feature to the Behaviour-Personality model.

TABLE 6.3: Number of Students per Outcome Group

Outcome Group	Number of Students
At-risk	309
Safe	824
Total	1 133

Figure 6.4 shows the cumulative distribution curve of the student Grades[5]. This curve shows the proportion[6] (y-axis) of students falling below each Grade (x-axis). The solid green line through the x-intercept shows the classification boundary (Grade ≥ 51), while the solid line through the y-intercept shows that only 27% of students obtained a Grade of 51.00 or fewer points and were therefore *At-risk*. The remaining 73% of the students were *Safe*. In a balanced dataset, we would expect 50% (not 27%) of students to be At-risk, and the remaining 50% to be Safe. Observe the intersection shown by the red set of dashed lines: more than half of the cohort students (Proportion > 50%) obtained a Grade greater than 61.50. Table 6.3 and Figures 6.3 and 6.4 show the need to balance the train set of our models by Outcome. Section 6.1.2 describes the formulation of the Conscientiousness from Login patterns, and how Conscientiousness-level relates to Outcome.

6.1.2 Conscientiousness and Grade

Section 3.5.1.1 explained the facets that describe each personality trait. Our model of Conscientiousness relates closely with dutifulness. Barrick et al. (1993) and Campbell (1990) theorise that Conscientiousness is linked to an individual's choice to expend a level of effort. Therefore, modelling dutifulness requires a formulation that captures the average number of logins per week that the student made throughout the programme. This model of Conscientiousness-level captures the facet of dutifulness and the choice to expend effort. Let $C(s)$ be a variable that represents the

[5]The cumulative frequency could have been used. However, the cumulative distribution normalises the frequencies so that the y-axis shows the proportion, not the number of students

[6]Also referred to as empirical probability

Conscientiousness-level of student s. $C(s)$ is modelled as the average number of logins over the period spanning a student's active weeks. For each student, $C(s)$ is formulated as:

$$C(s) = \frac{\sum_{t}^{17} L(s)_t}{W(s)}, \tag{6.2}$$

where $W(s)$ is the number of weeks spanned between the student's first and last login.

The reason for modelling $C(s)$ as the *average* number of active days per week instead of the *total* number of logins over the period is that the *average* normalises the data. Averaging reduces biases caused by differences in the number of days per cohort, per subject and programme, that students are expected to log in.

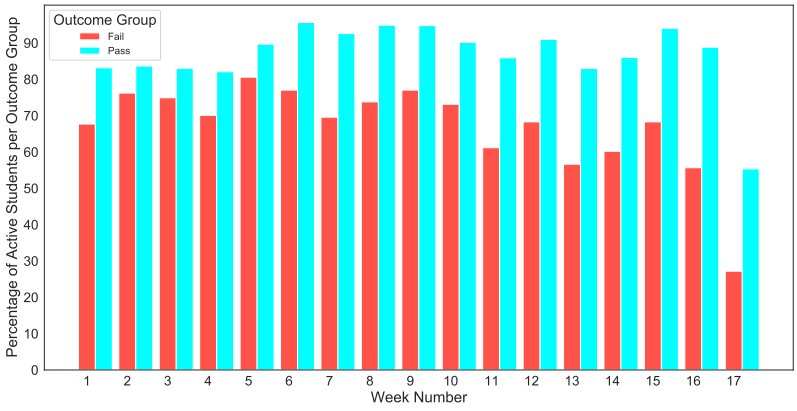


FIGURE 6.3: Weekly Login Activity per Outcome Group

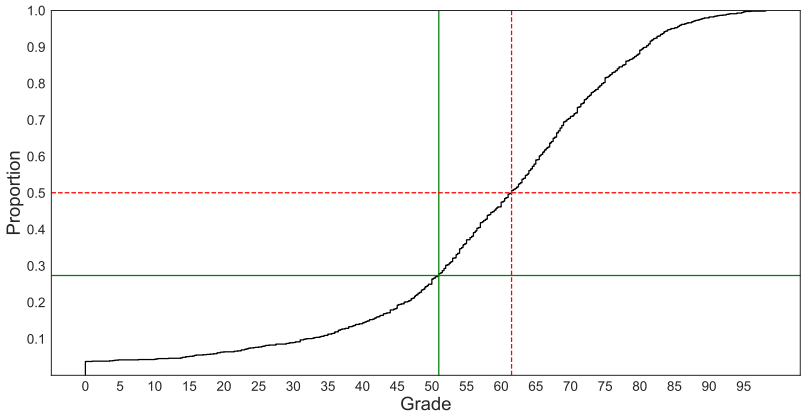


FIGURE 6.4: Cumulative Grade Distribution of Cohort, emphasising the imbalance of student Outcomes. Only 27% of the students were at risk (Grade = 51.00, Proportion = 0.27).

TABLE 6.4: OLS Regression Summary – Average Number of Weekly Active Days against Grade

Linear Equation: $\hat{G}(s) = \beta_0 C(s) + \beta 1$
$= 5.988C(s) + 39.829$

Feature	**Coeff.**	**r**	***p*-value**	**Coeff. 95% CI**
$C(s)$	5.988	0.319	0.000	[4.129, 7.847]
Intercept	39.829		0.000	[34.426, 45.232]

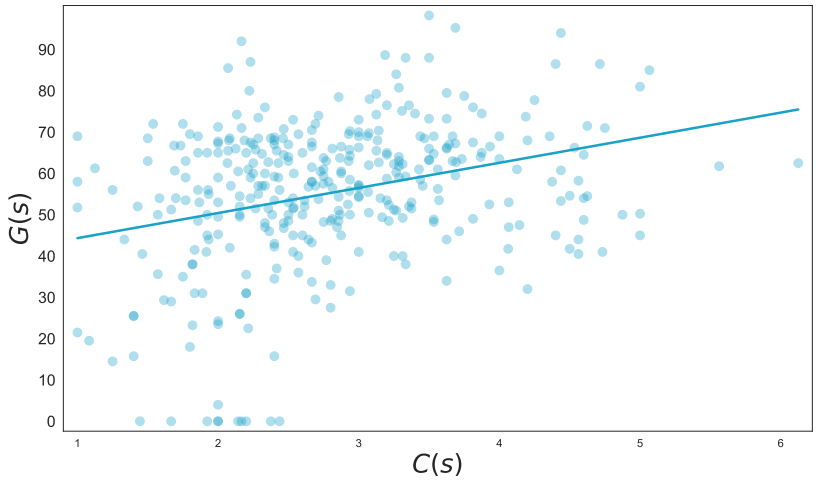


FIGURE 6.5: Average Number of Weekly Active Days against Grade. $r = 0.319$, $p = 0.000$.

The OLS Model, whose results are summarised in Table 6.4, regresses the Conscientiousness-level, $C(s)$, against Grade, $G(s)$. For ease of visualisation, the Summary's Regression Plot shows a random subset of 350 students in Figure 6.5.

6.1.2.1 Results and Discussion: $(G(s), C(s))$

The β_0 coefficient of $C(s)$ indicates that an increase of 1 in Conscientiousness-level is associated with an increase of 5.988 in Grade. Out of the 102 students who ended up at risk of failing their programmes (Grade < 51), 76 had a Conscientiousness-level below 3. The statistically-significant positive correlation between $C(s)$ and $G(s)$ shows that $C(s)$ is a suitable predictor of a student's Outcome.

Hung and Zhang (2009) presented a comparable finding; students who accessed course materials 18.5 times or more throughout their programmes obtained a grade of 77.92 out of 100 or higher, while students who accessed course materials more than 44.5 times obtained a grade of 89.62 or higher.

The assumption tests for the OLS regression between $C(s)$ and $G(s)$ is shown in Section A.3. The tests confirm that an OLS fit is appropriate for regressing $G(s)$ against

$C(s)$. The following section details the use of the LSTM to predict student Outcomes based on their weekly Login Sequence, Extraversion- and Conscientiousness-levels.

6.2 Behaviour-Personality Model

The Behaviour-Personality model (B-PM) consists of two components: the behavioural component is the Login Sequences of students, while the personality component augments the Login Sequences. The traits that make up the personality component are the Extraversion- and Conscientiousness-levels engineered in Sections 5.2 and 6.1.2, respectively. For each student s, there are two Approaches to engineer the Login Sequence ($\{L(s)_t\}$), Extraversion-level ($E(s)$), and Conscientiousness-level ($C(s)$) as model inputs. Either:

1. by augmenting $\{E(s)_t\}_{t=1}^{17}$ and $\{C(s)\}_{t=1}^{17}$ as *sequences* of the *same* values that run parallel to $\{L(s)_t\}_{t=1}^{17}$ through time t. This augmentation forms a 3×17 *input array of sequences*,

$$[\{L(s)_t\}, \{E(s)_t\}, \{C(s)_t\}],$$

   where

   - $\{L(s)_t\}_{t=1}^{17}$ is a sequence of values that *vary* through time t,
   - $\{E(s)_t\}_{t=1}^{17}$ is a sequence of the *same* value through time t, so that $E(s)_t = E(s)_{t-1}$ for all Whole Numbers $t \in [2, 17]$, and
   - $\{C(s)_t\}_{t=1}^{17}$ is a sequence of the *same* value through time t, so that $C(s)_t = C(s)_{t-1}$ for all Whole Numbers $t \in [2, 17]$

   (see Table 6.8), or

2. by augmenting $E(s)$ and $C(s)$ as single values that extend $\{L(s)_t\}$ by two. This augmentation forms a 1×19 *array*,

$$[\{L(s)_t\}, E(s), C(s)],$$

   where

   - $\{L(s)_t\}$ is a sequence of values that vary through time t,
   - $E(s)$ is a *single* value, and
   - $C(s)$ is a *single* value

   (see Table 6.9).

Approach 1 produced a better κ score. Leontjeva and Kuzovkin (2016) guided our usage of Approach 1. B-PM is thus used under Approach 1:
The B-PM **input** for each student is the *array of sequences* –
$[\{L(s)_t\}, \{E(s)_t\}, \{C(s)_t\}]$.
The B-PM **output** for each student is a Safety Score –
Flagged for At-risk students, and *Ignored* for Safe students.
Table 6.8 summarises the B-PM **input** and **output** structures per student.

TABLE 6.5: Oversampling for the Login Table Train Set

Train Subset	**Value**
At-risk – Original	260
Oversampling Factor	$\times 2.703$
At-risk – Oversampled	**= 703**
Safe	+ 703
Train Set – Balanced	**= 1 406**

6.2.1 B-PM: Model Design

The 1 133 observations were split into a train set:test set ratio of 963:170. The Outcome labels stratified the train and test sets. For the train set, the At-risk:Safe student ratio is 260:703, or approximately 1:2.703. The number of samples in the train set was oversampled by a factor of 2.703, resulting in a balanced 703:703 At-risk:Safe train set ratio. Oversampling to balance the data was performed by duplicating the At-risk observations[7]. As a result, the train set consisted of 1 406 students. Table 6.5 shows a summary of the train set oversampling process. The LSTM configuration of the Behaviour-Personality Model is summarised in Table 6.6. B-PM's data configuration is outlined in Table 6.8.

The LSTM for the B-PM was configured with 1 input, 1 hidden, 1 activation and 1 output layer. The Adaptive Moment Estimation (also called *Adam*)[8] optimiser yielded superior results, as compared to Stochastic Gradient Descent's (SGD) results. Both the Hyperbolic Tangent (Tanh) and Sigmoid functions were tried as activation functions, but the results produced by Tanh were inferior to Sigmoid's results. Although the feature shape for each student is 1×19, the input shape into the LSTM is $1\,406 \times 19$, since there are 1 406 students in the oversampled train set. The LSTM's validation set ratio was configured to 20% of the 1 406 observations.

In total, 10 000 experiments were conducted using each combination of hyperparameters shown in Table 6.7. Experiments were terminated if the validation loss did not change over a 100-epoch interval. In Table 6.7, the configuration of B-PM and BM yielding the highest accuracy (measured by κ) is represented by the entries highlighted in green.

6.2.2 Results and Discussion: B-PM

In this section, the predictive power of B-PM against student Outcomes is shown. Table 6.10 shows B-PM's results.

The κ of 0.51 shows a moderate agreement between B-PM's predicted Safety Scores and true student Outcomes. B-PM's precision for the At-risk Outcome group shows that out of the 39 Flagged students, 27 were Flagged correctly (since they ended up at

[7] The SMOTE (Chawla et al., 2002) was also used, without a significant performance improvement (κ increase of 0.032)

[8] See Kingma and Ba (2014) for details on the Adaptive Moment Estimation optimiser

TABLE 6.6: B-PM and BM LSTM Hyperparameter Configuration

Hyperparameter	**Value**
No. of Input Layers	1
No. of Hidden Layers	1
No. of Hidden Units in Hidden State Vector	550
Epochs	500
Batch Size	15
Optimiser	Adaptive Moment Estimation
Learning Rate	0.0004
Loss Function	Binary Cross Entropy
No. of Activation Layers	1
Activation Function	Sigmoid
No. of Output Layers	1
Input Shape (**B-PM Approach 1**)	963×19
Input Shape (**B-PM Approach 2**)	$963 \times 3 \times 17$
Input Shape (**BM**)	963×17
Output Shape	1×1

TABLE 6.7: B-PM and BM Hyperparameter Alternatives

Hyperparameter	**Value 1**	**Value 2**	**Value 3**	**Value 4**	**Value 5**
Hidden Layers	1	2			
Hidden Units	500	550	600	650	700
Epochs	450	500	550	600	650
Batch Size	6	9	12	15	18
Optimiser	Adam	SGD			
Learning Rate	0.04	0.004	0.0045	0.0004	0.0001
Activation Function	ReLu	Sigmoid	Tanh	Softmax	

risk of failing). 12 out of the 39 Flagged students were not meant to be Flagged. The At-risk recall indicates that out of 46 At-risk students, 27 were correctly Flagged, and the remaining 19 were incorrectly Ignored.

B-PM performed better at classifying Safe students than at classifying At-risk students: only 12 out of 124 Safe students were incorrectly Flagged, and 19 out of 131 Ignored students were wrongly Ignored.

TABLE 6.8: B-PM Training Input and Output Summary

Feature	**Shape**	**Type**	**Example Value**
$\{L(s)_t\}$ Login Sequence	(1×17)	A Sequence of Whole Numbers	$[3, 7, \ldots, 0]$
$\{E(s)_t\}$ Extraversion[9] Sequence	(1×17)	A Sequence of Whole Numbers	$[8, 8, \ldots, 8]$
$\{C(s)_t\}$ Conscientiousness[10] Sequence	(1×17)	A Sequence of Real Numbers	$[1.1, 1.1, \ldots, 1.1]$
Input: $[\{L(s)_t\}, \{E(s)_t\}, \{C(s)_t\}]$	(3×17)	An *Array of Sequences* of Real Numbers	$[[3, 7, \ldots, 0]$ $[8, 8, \ldots, 8]$ $[1.1, 1.1, \ldots, 1.1]]$
Output: Safety Score = $\{Flagged, Ignored\}$	(1×1)	Binary	Ignored

TABLE 6.9: Alternative B-PM Input and Output Summary

Feature	**Shape**	**Type**	**Example Value**
$\{L(s)_t\}$ Login Sequence	(1×17)	A Sequence of Whole Numbers	$[3, 7, \ldots, 0]$
$\{E(s)_t\}$ Extraversion-level	(1×1)	A Whole number	8
$\{C(s)_t\}$ Conscientiousness-level	(1×1)	A Real Number	1.1
Input: $[\{L(s)_t\}, E(s), C(s)]$	(1×19)	A Sequence of Real Numbers	$[3, 7, \ldots, 0, 8, 1.1]$
Output: Safety Score = $\{Flagged, Ignored\}$	(1×1)	Binary	Flagged

The confusion matrix in Table 6.10 partitions the students into four Classification groups, namely,

1. At-risk and Flagged (Correctly Classified),
2. Safe and Ignored (Correctly Classified),
3. Safe but Flagged (Misclassified),
4. At-risk but Ignored (Misclassified),

Figure 6.6 shows the Grade Distribution of the Safe but Flagged students, while Figure 6.7 shows the Grade Distribution of the At-risk but Ignored students. Recall that the Outcomes (target variables for the LSTM) were constructed based on the student Grades: At-Risk students obtained a Grade below 51 Grade Points, while

TABLE 6.10: Confusion Matrix and Summary of B-PM Test Set Results

		Safety Score (Prediction)		
		Flagged	**Ignored**	**Total**
Outcome (True Label)	**At-risk**	27	19	46
	Safe	12	112	124
	Total	39	131	**170**

Outcome	**Precision**	**Recall**
At-risk	0.69	0.59
Safe	0.85	0.90

$\kappa =$	0.51

Safe students obtained a Grade of 51 Points and above. The purpose of the Grade distributions is to show each student's Grades in the Misclassified groups and how far from a correct classification their Grades were. That is, Figures 6.6 and 6.7 show the distance from the classification boundary (51 Grade points) that the students in the Misclassified groups lie.

Note that only the Misclassified groups' Grade distributions are evaluated. The Correctly Classified group Grade distributions are not considered since their Grades lie on the correct side of the Grade points scale, respectively (On the left of the 51-point boundary for the At-risk and Flagged and on the right of the boundary for the Safe and Ignored group).

Refer to Figure 6.6 which shows the Grade distribution of the Safe but Flagged students. The Safe but Flagged students are those whose behaviour was fit as *Flagged*, despite being not At-risk. The density curve peaks at 67.25 Grade Points – 16.25 points away from the classification boundary.

Refer to Figure 6.7, showing the Grade distribution of the At-risk but Ignored students. According to B-PM, these students demonstrated a *Safe* pattern in their Login e-Behaviour and were thus Ignored. 19 out of the 46 students who should have been Flagged were incorrectly Ignored. 15 of these 19 falsely Ignored students achieved Grades greater than 40.00 points. Although 5 out of the 19 students (26%) achieved Grade points of greater than 49.00 points and may have passed their programmes[11], the threshold of 51 provides a buffer that allows B-PM to reveal students who were *at risk* of failing. The number of students in the At-risk but Ignored group is seven more than those in the Safe but Flagged group. The At-risk but Ignored students' Grade density peaks at 47.5 Grade points, which is only 3.50 Grade points from the boundary.

Comparing the Grade distributions of the two Misclassified groups reveals that although the Safe but Flagged student behaviour may have been modelled as *risky*,

[11] Depending on faculty rules

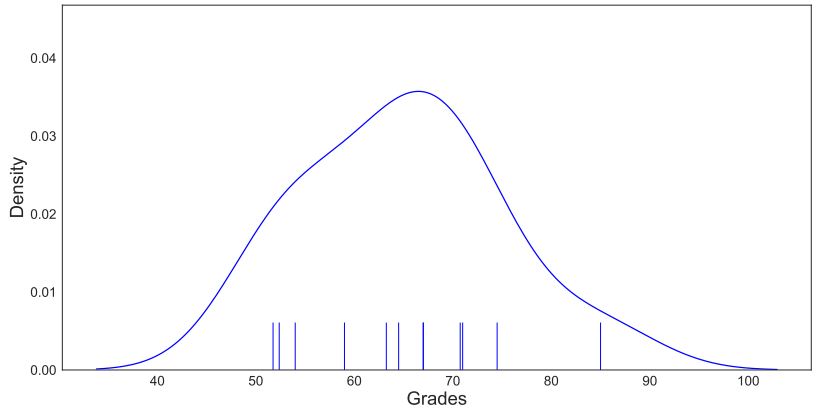


FIGURE 6.6: Grade Distribution of the Safe but Flagged Students

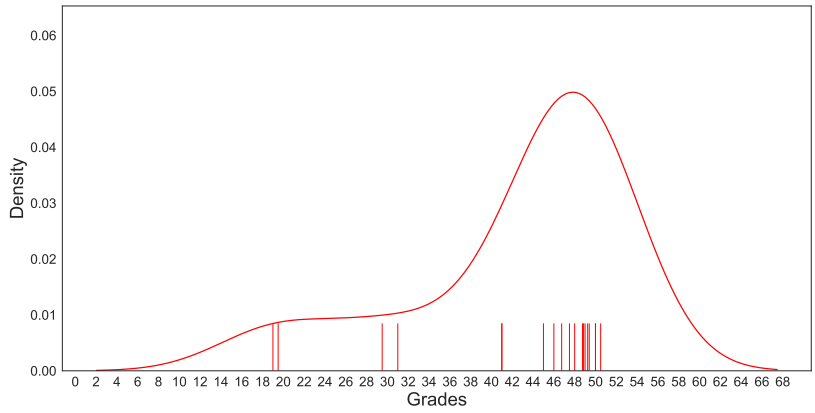


FIGURE 6.7: Grade Distribution of the At-risk but Ignored Students. 26% of the students achieved within 2 grade points of the classification boundary (Grade = 52).

their Grades were an average (median) distance of 14.75 Grade points from the classification boundary. Contrastingly, the Grades of the At-risk but Ignored group were a median distance of 3.50 Grade points from the classification boundary. The difference between these medians shows that students in the Safe but Flagged group were far from being *at risk of failure*. By comparison, the At-risk but Ignored had Grades close to a *Safe* Outcome, which corresponds to their behaviour that B-PM modelled as *Safe*. Increasing the classification boundary to 55.00 Grade points yielded a κ of 0.42, while a 60.00 Grade Point boundary produced a κ of 0.30. In both the above cases, shifting the boundary produced worse accuracies than B-PM's κ of 0.51.

The following section reports on the results of a modified model of B-PM outlined in Table 6.9 – a model without the $\{E(s)_t\}$ and $\{C(s)_t\}$ personality Sequences.

TABLE 6.11: BM Input and Output Summary

Feature	**Shape**	**Type**	**Example Value**
Input: $\{L(s)_t\}$	(1×17)	A Sequence of Whole numbers	$[3, 7, \ldots, 0]$
Output: Safety Score = $\{Flagged, Ignored\}$	(1×1)	Binary	Flagged

TABLE 6.12: Confusion Matrix and Summary of BM Test Set Results

Safety Score

Outcome		**Flagged**	**Ignored**	**Total**
	At-risk	27 [27]	19 [19]	46
	Safe	22 [12]	102 [112]	124
	Total	49 [39]	121 [131]	**170**

Outcome	**Precision**	**Recall**
At-risk	0.55 (0.69)	0.59 (0.59)
Safe	0.84 (0.85)	0.82 (0.90)

$\kappa =$	0.40
$(\kappa =$	0.51)

6.2.3 Results and Discussion: BM

The previous section outlined the Behaviour-Personality Model, which used both the e-Behaviour and personality components of a student's LMS data. In this section, B-PM is compared to an alternative. The comparison helps determine the change in the accuracy of B-PM after removing its personality components. This resulting model is called the Behaviour Model (BM); the only difference between BM and B-PM is that BM has only one input Sequence, $\{L(s)_t\}$. Table 6.11 shows the Input and Output structures of BM.

Table 6.12 shows BM's results. For reference, the comparable B-PM results are shown in brackets. The At-risk recall of BM equals the At-risk recall of B-PM, meaning that BM correctly Flagged as many At-risk students as B-PM did. The lower $\kappa = 0.40$ score of BM results from the lower precision and recall scores of the Safe Outcome and a lower precision score for the At-risk Outcome. All other model evaluation metrics recorded a higher score for B-PM. The result of the higher precision and recall scores of B-PM is a higher κ score of 0.51, making B-PM a better overall classifier of student risk.

While we only showed B-PM and BM predictions for the end of the 17 weeks, the models also produced predictions at the end of each week. Flagging students at risk

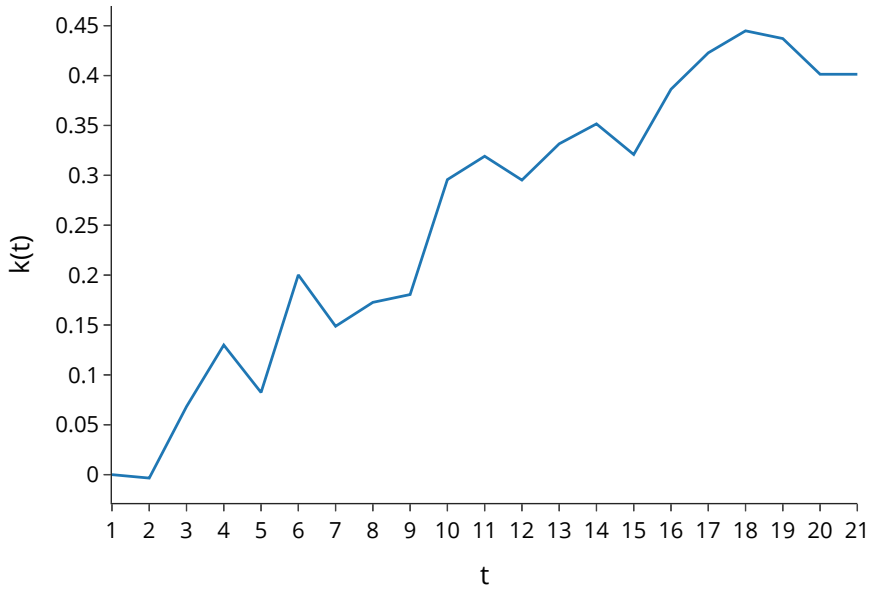


FIGURE 6.8: κ for predictions up to Each Week

earlier may be more beneficial to a student and an institution's stakeholders, since early-flagging allows more time for interventions. The following section reports on the trade-off between timeliness and accuracy.

6.3 Accuracy and Timeliness of Intervention

Each coordinate in Figure 6.8 shows the κ value for BM between weeks 1 and 21. An extra four-week period extends the 17 weeks in the original BM model to include Login e-Behaviour just before the start of the examination period (week 21). At each week, t, a prediction is made based on Login e-Behaviour between week 1 and t. $\kappa(t)$ is the κ computed for the model's prediction accuracy up to week t. The increase in κ between the beginning and the end of the period confirms that the accuracy of BM generally increases when provided with more $\{L(s)_t\}$ data from all students. Predictions at week 18 yield the highest accuracy, with a $\kappa(18)$ of 0.444. While the $\kappa(18)$ is greater than the $\kappa(17)$ of 0.423, $\kappa(18)$ is registered a week after the $\kappa(17)$. The trade-off between the model's accuracy and the intervention's timeliness is a κ of +0.021 in exchange for a week's delay in intervention.

Figure 6.9 is a Trade-off Waterfall that is drawn from the same values in Figure 6.8. The chart presents a more explicit visual to interpret the trade-off between the *benefit of intervention timeliness* and the *cost of intervention success*. Intervention success is determined by the model's accuracy. A red bar represents a decrease in accuracy from one week to the next, and a blue bar indicates an increase. The length of a bar indicates the magnitude of change in accuracy, $\kappa(t)$.

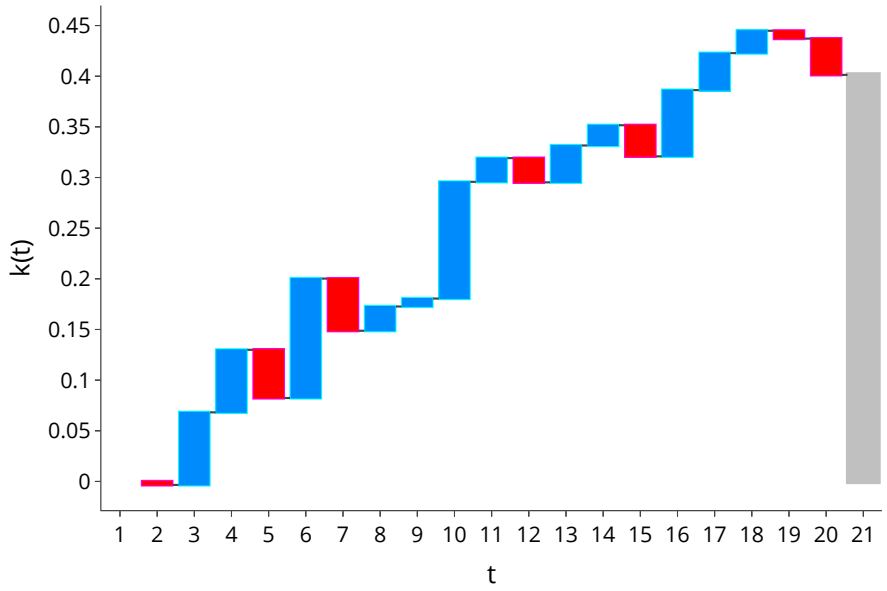


FIGURE 6.9: Trade-off Between Timeliness and Accuracy

6.4 BM and the Trade-off Waterfall

The trade-off between the *benefit of intervention timeliness* and the *cost of intervention success* can help identify whether there are patterns over each year that can inform users about the *optimal* week (t^*) to intervene. If the current year is 2019, then t^* for 2019 (t^*_{2019}) can be determined by either one or a combination of the following factors:

1. t^*_{2019} is chosen to be the t in 2018 that yielded the maximum value of $\kappa(t)$ in 2018.

2. t^*_{2019} is based on *exogenous* considerations determined by the institution's stakeholders. Examples of exogenous considerations are the urgency required for intervention and resources required to make interventions.

For instance, the 2018 Trade-off Waterfall was not available to this study. So the $t^*_{2019} = 17$ used in this cohort's B-PM and BM was based only on the exogenous consideration that interventions should be made by week 17.

6.4.1 Practical Benefits and Limitations of the Trade-off Waterfall

The Trade-off Waterfall is computed after the Outcome of the students is made available. It does not show the week that produces the highest accuracy in real-time, and in some weeks, the trade-off between accuracy and timeliness is not positive. For example, observe that $\kappa(15) < \kappa(14)$. In exchange for delayed intervention, BM produces a worse κ score from $\{L(s)_t\}$, which would not have made the delay

worthwhile. A similar observation is made for the delay between weeks 18 and 19. Therefore, in real-time, there is no way to infer the optimal week to make predictions and interventions. Instead, the Trade-off Waterfall indicates that

1. the general trade-off is that $\kappa(t)$ increases as t increases, and
2. the trade-off peaks at some point, and in this case, three weeks before the examination period at $t = 18$. Therefore, it may not be worth waiting for the start of an examination period (such as $t = 21$) before conducting interventions.[12] For example, the Trade-off Waterfall shows that after $t = 18$, there is no benefit of waiting for an extra one, two or even three weeks to intervene because $\kappa(19), \kappa(20), \kappa(21) < \kappa(18)$.

6.5 Classifying a Student Using Trade-in Waterfall and Prediction Sequence

This section gives a summary of how a student was classified. The following procedure may be used to classify new students across any cohort using:

1. the cohort's BM.
2. the cohort's Trade-in Waterfall,
3. each student's $\{L(s)_t\}$,
4. each student's Prediction Sequence, $\{P(s)_t\}$. The derivation for $\{P(s)_t\}$ is shown in the next paragraph.

Through the explanation, a student, s, will be used to proxy an arbitrary student. Let

- $\{L(s)_t\}$ represent the Login Sequence of Student s, and
- $\{P(s)_t\}$ represent the Prediction Sequence of Student s

If the cohort's B-PM is used, then $[\{L(s)_t\}_{t=1}^{W}, \{E(s)_t\}_{t=1}^{W}, \{C(s)_t\}_{t=1}^{W}]$ replaces $\{L(s)_t\}_{t=1}^{W}$.

Derivation of $\{P(s)_t\}$ for each student, s: At each week, W a student generates a Login Sequence, $\{L(s)_t\}_{t=1}^{W}$. Similarly, at each W, BM makes a prediction from the sequence from $\{L(s)_t\}_{t=1}^{W}$.

The Prediction Sequence, $\{P(s)_t\}_{t=1}^{W}$, is derived from $\{L(s)_t\}_{t=1}^{W}$. It follows that the Prediction Sequence, $\{P(s)_t\}_{t=1}^{W+1}$, is derived from $\{L(s)_t\}_{t=1}^{W+1}$, for any $W = \{1, 2, \ldots, 17\}$.

To classify a student as Safe or At-risk, the Safety Score $P(s)_{t^*}$ (Safety Score at chosen t^*) may be used in conjunction with the Trade-in Waterfall under the following procedure. Suppose the aim is to classify a student during the Year 2020. Given a trained 2019 BM model:

[12] Given the login data of this cohort. Different cohorts and different datasets from those presented in this report may produce different peak-periods

1. Determine t^* for 2020 using the guideline in Section 6.4, and either
2. Flag s as At-risk if $P(s)_{t^*} < 0.5$, or
3. Ignore s as Safe if $P(s)_{t^*} \geq 0.5$.

In this cohort, 0.5 was the chosen critical Safety Score; 0.5 is the threshold for mapping the LSTM predictions, where LSTM predictions below 0.5 were mapped to a 0 (At-risk), and LSTM predictions of 0.5 and above were mapped to a 1 (Safe).

6.6 Cohort-Specific Discussion on Intervention

Conducting the analysis using the four weeks before examinations helped determine the extent to which a model can predict if a student is at risk, given only a subset of a student's e-Behaviour. The model's *extent* is given by $\kappa(17)$. The 17-week period used in this chapter, however small, gives plenty of time for the intervention required from a student whose status is Flagged based on the models. Considering the entire year's time series may result in a more accurate model. However, such a model's intervention would be untimely and would compromise the usefulness of e-Behaviour and personality and performance analysis in practice.

6.7 Conclusion

This chapter started with the formulation of the $\{C(s)_t\}$ (Conscientiousness-level) personality component for the Behaviour-Personality Model. We explained the importance of using a weekly-resampled Login Sequence as input to BM. The Behaviour-Personality Model extended the behaviour Model with $\{C(s)_t\}$ and $\{E(s)_t\}$ personality trait values and produced a superior κ of 0.51, compared to the behaviour Model's κ of 0.40. An analysis of B-PM's results was provided, and it was shown that B-PM classifies Safe students more accurately than it classifies At-risk students. An intervention framework was demonstrated through the cohort's BM, the cohort's Trade-in Waterfall, each student's Login Sequence and each student's Prediction Sequence. The caveats of using the intervention framework were provided.

Chapter 7

Conclusion

Our cited literature included sources on a student Background, behaviour, personality, and these factors' relationship with student performance. The main difference between our methodology and methodology from previous literature is that we engineered features from an LMS system. We used these features to engineer e-Behaviours and personality traits that served as input to our models. We then analysed the model outputs and their practical implications. The results demonstrated that a student's background has lower predictive power over academic performance than a student's e-Behaviour and personality does. We found that modelling student behaviours and personality traits requires considering how accurately our proposed e-Behaviour and personality proxies model *true* behaviours and personality traits – we based our models on definitions found in previous literature.

Our aims guided us in developing a methodology that makes it easier to study the intersection between e-Behaviour, personality and performance. We achieved the aims by:

- using Bourdieu's Three Forms of Capital to model social, economic and cultural capital, and
- using the Big Five Personality traits to model e-Behaviour and personality

from student Background and LMS engagement data.

Bourdieu's Three Forms of Capital were modelled in the following ways:

- Economic Capital – modelled by the *Financial Assistance* feature,
- Cultural Capital – modelled by the *Quintile, Gauteng Province* and *Township School* features, and
- Social Capital – modelled by Academic-groups.

Bourdieu and Richardson (1986) argues that Cultural, Economic and Social Capital regulate the level of success attainable by individual. The correlation values for Financial Assistance, Quintile and Township School provides evidence for the authors' argument. Cultural and Economic Capital, combined with the Gender feature, performed better than chance at predicting student performance. A student's quality of Social Capital available to him also correlates positively with his academic performance.

We used two of the Big Five personality traits previously found to correlate strongly with performance – Conscientiousness and Extraversion. Conscientiousness and Extraversion each showed significant correlations with performance. With these personality features, our e-Behaviour classifier achieved better accuracy than without the personality features. The works cited do not discuss how changes in student behaviour relate to changes in their performance. We constructed a temporal e-Behaviour model that showed an increase in accuracy over time. This e-Behaviour model can be used to flag students at risk at any point throughout their study programmes.

Practical Value of Findings

The analyses in this research are practically useful if they can inform or influence student behaviour. Firstly, a student may find useful the linear relationship between academic performance and Extraversion in Section 5.2. The empirical evidence that a higher Extraversion-level is associated with a better academic performance may encourage students to engage in forums more frequently. This evidence may encourage him to engage with the academic content more thoroughly so that he contributes meaningfully to discussions. Secondly, Section 5.3 shows that the performance of a student is congruent with the performance of his Academic-group. The above result is a further motive to action a student into leveraging his social capital by engaging in forums more frequently, since engaging more frequently increases his chances of being in an Academic-group (see Sections 5.3.1 and 5.3.2).

Changes in student e-Behaviour and performance may stem from confounding factors that we did not capture, such as a student's state of distress. While e-Behaviour itself may not be the root cause of poor student performance, e-Behaviour can reveal signs of a distressed student. Therefore, any flags made by the e-Behaviour models can serve as an early warning mechanism for the confounding factors that are causing a student's *poor* behaviour.

7.1 Limitations and Future Work

This research is a study on the methodology that guides the use of algorithms in academic performance analysis, rather than a study on the efficiency and improvement of the algorithms themselves.

Our results are likely to differ across contexts since different data and algorithm configurations can generate several different model outputs. The results obtained serve only as proxies for the possible outputs in academic performance research.

In the domain of LMS system user engagement, there are no formal definitions and standards, analogues or equivalent metrics that proxy a student's e-Behaviour and personality from LMS data. We modelled features as well-understood traits to further the studies of relationships between behaviour, personality and academic performance.

An unknown in all model outcomes was the presence of causality. For instance, whether e-Behaviour has an *effect* on performance is not known. Although the

methodology followed aimed to set up conditions for inference, diction such as *tend to correspond with*, and *have relationships with*, instead of *causes*, showed sensitivity to all likelihood of effects from confounding variables.

In encoding performance (student Grade), we used uniform importance across all modules. We did so despite some students' modules accounting for a higher proportion of points towards obtaining a qualification from the University. We did not have access to the relative weighting of each module for each student, and therefore did not account for the differences in module weightings.

7.2 Alternative Formulations of Personalities

An approach to capture various facets of Extraversion and Conscientiousness was attempted. For instance, the *orderliness* facet of Conscientiousness required a metric that models routine or consistency of engagement. Orderliness was modelled by computing the sum of the squared deviations, SS, from the mean number of logins for each student. However, the regression model that correlated Grades with SS violated the normality-of-residuals test for normality. Thus, the test for a relationship between SS and Grade was inappropriate under a linear regression model.

7.3 Completeness of Data

The discussion on Background data in Chapter 4 validated the assumptions that existed behind the usage of Background information as a predictor for performance. However, only 104 students from 1 133 had corresponding Background data from the 2019 subset of the cohort, so Background features were not augmented to the e-Behaviour models.

7.4 Recommendations and Future Work

Models with different aims from those presented in this research can be designed to include extra features. Such features may be believed to confound relationships with final grades, such as background features. Grades from pre-examination assessments can also be used as predictors of overall student performance. The prevalence of personality traits could also be measured by traditional surveys instead of automated systems.

7.4.1 Student Relationships

Academic-groups were used as models of student relationships. An avenue of future work can explore modelling student relationships through a network distance matrix – an $n \times n$ matrix, D, showing the pair-wise number of common discussions between each student. The entry in $D_{i,j}$ could be taken to represent the *degree of closeness* between student i and student j. Our methodology modelled student groups from previous literature on social capital.

7.4.2 e-Behaviour and Personality

We used data from Learning Management Systems to model e-Behaviour and personality. Instead of using Learning Management System engagement data, datasets that capture physical behaviour can be used for students whose curriculum requires in-person attendance. Physical behaviour, such as a student's library usage or tutorial attendance, can be combined with e-Behaviours and personality to obtain a more comprehensive view of each student.

The Extraversion and Conscientiousness personality traits were modelled. Future work can focus on modelling Openness, Agreeableness and Neuroticism. A student's Openness, Agreeableness and Neuroticism can be modelled through observing patterns in his text. If a physical behaviour dataset is used, a student's Openness can be modelled by observing how often he visits a library that he has never visited. Furthermore, a student's (academic) Openness can be modelled by extracting sentiments from his forum posts or responses to open questions in online examinations.

We mentioned that NEO PI-R is used as an instrument that assesses the prevalence of personality traits through questionnaires. Similarly, finding reliable e-Behaviour and personality proxies can establish a framework for standard behaviour and personality measures in online learning environments.

Appendices

Appendix A

Appendix

	Literature	Questions and Aims	Measurement of Behaviour	Risk Metric
1	**Gray & Perkins** *(Utilizing early engagement and machine learning to predict student outcomes)*	1) To evaluate several candidate machine learning methods using derived metric measuring student attendance/engagement to produce predictions of student outcome 2) To discuss related issues, including student motivation and potential interventions tutors may wish to undertake. 3) To define a descriptive statistic for student attendance and applies modern machine learning tools and techniques to create a predictive model	1) Student Attendance - Engagament Metric	Grades (Categorised): 1) Pass 2) Fail - Can't Progress 3) Fail Conditional - Supplementary exam 4) Repeat year 5) Repeat Semester
2	**Talavera and Gaudioso** *(Mining Student data to characterize similar behaviour groups in unstructured collaboration spaces)*	Explore the use of data mining to LMS and build analytical models summarizing interaction patterns for insight. To detect patterns of interaction and relate to performance	1) No. sessions started 2) No. entrances to the course chat 3) No. messages sent to the course chat 4) No. messages sent to the course forum 5) No. files in the file storage area of the course 6) Activated alerts in the forum of the course 7) Created forum 8) Published a web presentation 9) Added bookmarks 10) Sent a email to the whole course 11) Registered in other courses 12) No. messages sent to other forums 13) No. entrances to other chats 14) No. messages sent to other chats 15) No. files in other file storage areas 16) Activated alerts in other forums 17) No. static course pages visited 18) No. threads started in the course forum 19) No. threads replied by other students in the course forum 20) No. threads finished in the course forum	Grades: 1) Pass 2) Fail

FIGURE A.1: Table of LMS Behaviour From Literature Review (1 of 3)

	Literature	Questions and Aims	Measurement of Behaviour	Risk Metric
3	**Romero, Lopez, Luna, Ventura** *(Predicting student final performance from participation in online discussion forums)* ***	1) What Data Mining Algos and techniques are best for [redicting student perfomrance from participation in online forums? 2) What attributes are the best predictors? Are all attributes/ vars about forum usage relevant or can reasonable accuracy be obtained on a subset? 3) What kinds of messages are the best predictors? 4) Can we make early predictions with accuracy? Is it necessary to wait until the end of the course or can we do it before the end of the course??	1) Messages no. 2) Threads no. 3) Words no. 4) Sentences no. 5) Reads no. 6) Time mins 7) AvgScoreMsg (0-3 relevance and quality 8) Centrality (based on sent msgs) 9) Prestige (metric - based on received msgs)	Grades: 1) Pass 2) Fail
4	**Hung and Zhang (Revealing Online Learning Behaviors and Activity Patterns and Making Predictions with Data Mining Techniques in Online Teaching)** ****	1) What are the typical online learning behaviors of undergraduate students (Taiwan)? 2) What are the typical patterns of online learning behaviors of undergraduate students? 3) What are the most important predictive indicators for learning outcomes of undergraduate students in an online learning environment?	1) ID: User ID 2) LoginFre: Total frequency of LMS logins 3) LastLog: When was the last time logged into LMS 4) ClassFre: Total frequency of accessing course materials 5) LastClass: When was the last time accessed course materials 6) NoPosting: Total number of bulletin board messages posted 7) DisFre: Total number of synchronous discussions attended 8) ReadHr: Hours spent reading bulletin board messages 9) ReadMegs: Total number of bulletin board messages read	Grades: 0%-100%

FIGURE A.2: Table of LMS Behaviour From Literature Review (2 of 3)

	Literature	Questions and Aims	Measurement of Behaviour	Risk Metric
5	**Romero, Ventura, Espejo, Hervas** *(Data Mining Algorithms to Classify Students)*	Compare data mining techniques for classifying students based on both their usage data in a web-based course and the final marks obtained.	1) Course: Identification number of the course. 2) n_assigment: Number of assignments done. 3) n_quiz: Number of quizzes taken. 4) n_quiz_a: Number of quizzes passed. 5) n_quiz_s: Number of quizzes failed. 6) n_posts: Number of messages sent to the forum. 7) n_read: Number or messages read on the forum. 8) total_time_assignment: Total time used on assignments. 9) total_time_quiz: Total time used on quizzes. 10) total_time_forum: Total time used on forum. 11) Mark: Final mark the student obtained in the course.	Grades: 1) Pass 2) Good 3) Excellent
6	**Asif, Merceron, Ali, Haider** *(Analyzing undergraduate students' perfirmance using educational data mining)*	Aim: to analyze the performance of students pursuing a 4-year Bachelor degree programme in the discipline of Information Technology Questions: 1. Can we predict students' performance with a reasonable accuracy at an early stage of the degree programme using marks only? 2. Can we identify courses that can serve as indicators of a good or low performance at the end of the degree? 3. Can we identify typical progressions of students' performance during their studies and relate them with the indicator courses?	1) Admission marks 2) First-year marks 3) Second-year marks	Grades: A (100-90) - E (59-50)
7	**Cerezo, Sanchez-Santillan, Paule-Ruiz, Nunez** *(Students' LMS interaction patterns and their relationship with achievement: A case study in higher education)*	1) Are there different patterns of effort and procrastination behavior among students when they learn through LMS in an authentic context? 2) To what extent the variables selected from the Moodle records to configure the patterns are equally suitable for the configuration of those clusters? 3) Whether effort and procrastination patterns, measured on the basis of the Moodle logs, are related to final marks?	1) Time tasks 2) Time theory 3) Time forums 4) Words forums 5) Relevant acrions 6) Time "hand in"(Effort Time spent working, Procrastination)	FOUR GROUPS: Two Task-Oriented Groups (socially or individually focused) and Two Non-Task Oriented Groups (procrastinators or non-procrastinators)

FIGURE A.3: Table of LMS Behaviour From Literature Review (3 of 3)

TABLE A.1: Moodle Database Tables

Order	NAME	ROWS
1	mdl_question_attempt_step_data	1584013
2	mdl_grade_grades_history	1581282
3	mdl_question_attempt_steps	935784
4	mdl_question_attempts	336707
5	mdl_grade_grades	328429
6	mdl_course_modules_completion	191770
7	mdl_assign_user_mapping	124968
8	mdl_assign_submission	97694
9	mdl_assign_grades	80804
10	mdl_assignsubmission_file	78720
11	mdl_assignfeedback_witsoj	51503
12	**mdl_logstore_standard_log**	**467743**
13	mdl_question_usages	28900
14	mdl_quiz_grades	24255
15	mdl_grade_items_history	19186
16	mdl_user_enrolments	17444
17	mdl_moodleoverflow_read	17360
18	mdl_user_lastaccess	16937
19	mdl_course_completions	14912
20	mdl_forum_read	12269
21	mdl_question_answers	11565
22	mdl_assign_plugin_config	10300
23	mdl_qtype_stack_deployed_seeds	9235
24	mdl_assign_user_flags	7350
25	mdl_user	6684
26	mdl_question	5670
27	mdl_course_modules	4623
28	mdl_quiz_slots	4440
29	mdl_qtype_stack_qtest_inputs	3492
30	mdl_qtype_stack_qtest_expected	3368
31	mdl_qtype_stack_qtests	3222
32	mdl_quiz_overview_regrades	3162
33	mdl_qtype_stack_prt_nodes	2579
34	**mdl_forum_posts**	**2378**
35	mdl_assignfeedback_file	2170
36	mdl_moodleoverflow_ratings	1897
37	mdl_assignsubmission_onlinetext	1896
38	mdl_qtype_multichoice_options	1859
39	mdl_grade_items	1712
40	mdl_qtype_shortanswer_options	1573
41	mdl_moodleoverflow_posts	1558
42	mdl_qtype_stack_cas_cache	1322
43	mdl_grade_categories_history	1286

Continued on next page

Table A.1 – continued from previous page

Order	NAME	ROWS
44	mdl_qtype_stack_inputs	1271
45	mdl_qtype_stack_prts	1233
46	mdl_question_categories	1186
47	mdl_qtype_stack_options	1166
48	mdl_forum_discussions	1022
49	mdl_course_sections	949
50	mdl_forum_subscriptions	906
51	mdl_moodleoverflow_discuss_subs	892
52	mdl_grading_areas	889
53	mdl_assign	738
54	mdl_quiz_feedback	697
55	mdl_assignfeedback_comments	676
56	mdl_question_numerical	662
57	mdl_question_numerical_options	632
58	mdl_moodleoverflow_discussions	464
59	mdl_question_hints	403
60	mdl_quiz	393
61	mdl_quiz_sections	393
62	mdl_user_devices	380
63	mdl_qtype_match_subquestions	371
64	mdl_enrol	362
65	mdl_forum_discussion_subs	351
66	mdl_forum_queue	329
67	mdl_assignment	319
68	mdl_course_format_options	251
69	mdl_question_multianswer	239
70	mdl_log_display	198
71	mdl_grade_categories	192
72	mdl_forum	175
73	mdl_qtype_ddimageortext_drags	134
74	mdl_question_dataset_items	123
75	mdl_course	122
76	mdl_quiz_overrides	120
77	mdl_assign_overrides	116
78	mdl_qtype_ddimageortext_drops	106
79	mdl_qtype_essay_options	99
80	mdl_question_truefalse	85
81	mdl_qtype_match_options	54
82	mdl_assignment_upgrade	35
83	mdl_moodleoverflow	28
84	mdl_course_categories	21
85	mdl_qtype_ddimageortext	20
86	mdl_question_dataset_definitions	18
87	mdl_question_datasets	18

Continued on next page

Table A.1 – continued from previous page

Order	NAME	ROWS
88	mdl_grade_settings	17
89	mdl_moodleoverflow_subscriptions	17
90	mdl_assignfeedback_witsoj_mkr	10
91	mdl_moodleoverflow_tracking	9
92	mdl_question_gapselect	8
93	mdl_quiz_reports	5
94	mdl_question_calculated	4
95	mdl_question_ddwtos	4
96	mdl_course_completion_defaults	2
97	mdl_forum_digests	2
98	mdl_forum_track_prefs	2
99	mdl_question_calculated_options	2
100	mdl_assignfeedback_editpdf_annot	0
101	mdl_assignfeedback_editpdf_cmnt	0
102	mdl_assignfeedback_editpdf_queue	0
103	mdl_assignfeedback_editpdf_quick	0
104	mdl_assignfeedback_witsoj_langs	0
105	mdl_assignment_submissions	0
106	mdl_course_completion_aggr_methd	0
107	mdl_course_completion_crit_compl	0
108	mdl_course_completion_criteria	0
109	mdl_course_published	0
110	mdl_course_request	0
111	mdl_enrol_flatfile	0
112	mdl_enrol_lti_lti2_consumer	0
113	mdl_enrol_lti_lti2_context	0
114	mdl_enrol_lti_lti2_nonce	0
115	mdl_enrol_lti_lti2_resource_link	0
116	mdl_enrol_lti_lti2_share_key	0
117	mdl_enrol_lti_lti2_tool_proxy	0
118	mdl_enrol_lti_lti2_user_result	0
119	mdl_enrol_lti_tool_consumer_map	0
120	mdl_enrol_lti_tools	0
121	mdl_enrol_lti_users	0
122	mdl_enrol_paypal	0
123	mdl_grade_import_newitem	0
124	mdl_grade_import_values	0
125	mdl_grade_letters	0
126	mdl_grade_outcomes	0
127	mdl_grade_outcomes_courses	0
128	mdl_grade_outcomes_history	0
129	mdl_grading_definitions	0
130	mdl_grading_instances	0
131	mdl_gradingform_guide_comments	0

Continued on next page

Table A.1 – continued from previous page

Order	NAME	ROWS
132	mdl_gradingform_guide_criteria	0
133	mdl_gradingform_guide_fillings	0
134	mdl_gradingform_rubric_criteria	0
135	mdl_gradingform_rubric_fillings	0
136	mdl_gradingform_rubric_levels	0
137	mdl_log	0
138	mdl_log_queries	0
139	mdl_qtype_ddmarker	0
140	mdl_qtype_ddmarker_drags	0
141	mdl_qtype_ddmarker_drops	0
142	mdl_qtype_randomsamatch_options	0
143	mdl_qtype_stack_qtest_results	0
144	mdl_question_numerical_units	0
145	mdl_question_response_analysis	0
146	mdl_question_response_count	0
147	mdl_question_statistics	0
148	mdl_quiz_slot_tags	0
149	mdl_quiz_statistics	0
150	mdl_user_info_category	0
151	mdl_user_info_data	0
152	mdl_user_info_field	0

TABLE A.2: Background Table Features

Variable
GENDER_CODE
ST_RACE_CODE
ENROLLED_AGE
HP_PROVINCE
HP_CITY
HP_ST_POST_CODE
PROGRAM_CODE
PROGRAM_TITLE
PROGRAM_CAT
PROGRAM_TYPE_DESC
PROGRAM_TYPE
ALIGNED_PROGRAM_CODE
CREDIT_POINTS_REQUIRED
FACULTY_NAME
AREA
CALENDAR_INST_YEAR
NEW_RETURN

Continued on next page

Table A.2 – continued from previous page

Variable
SPECIAL_GROUP
YOS
ATTENDANCE_TYPE_CD
FT_PT_DESC
EXT_NONEXT_DESC
ATTENDANCE_DESCRIPTION
ADMISSION_RATING
PSFT_COUNTRY_CODE
ST_NATIONALITY_SHT_NM
VISA_TYPE_CODE
VISA_TYPE_GRP
INTERN_STUDENT_STATUS
RESCH_AREA_CD
RESCH_AREA_DESC
RESCH_AREA_IND
IN_RESIDENCE
RESIDENCE_CODE
RESIDENCE_DESCRIPTION
ENROLLED
PROCEED
MINIMUM_REQ_NOT_MET
RETURNED_YOS
QUALIFIED
DECISION_PENDING
NON_DEGREE_PURPOSE
NO_DECISION
FAILED
INTERMISSION
TRANSFERRED
CANCELLED
FINAL_YEAR_STUD_IND
READMIT_PROGRESS_OUTCOME
READMIT_IND
READMIT_BY_WHOM
READMIT_TYPE_DESC
NEW_TO_PROGRAM_FLAG
CONV_TO_HIGHER_DEGREE_FLAG
DISABILITY_STATUS
PROG_LAPSED_FLAG
PROG_DROPOUTS_FLAG
CONVTD_FRM_LOWER_DEG_FLAG
ACAD_CAREER
NEW_TO_UNIV
REQ_COMPLETION_STATUS

Continued on next page

Table A.2 – continued from previous page

Variable
YR_OF_ENTRY
MATRIC_PROVINCE
PROVINCE_DESC
SIMS_PROVINCE
PROVINCE_COUNTRY
MIN_ENRL_DATE
PROG_ENRL_DATA_CHNG_FLAG
NSFAS_BURSARY_FLAG
STAFF_BURSARY_FLAG
FB_OFFERED_AMT
FB_ACCEPTED_AMT
FB_PAID_AMT
NSFAS_OFF_AMT
NSFAS_ACCP_AMT
NSFAS_PAID_AMT
EXT_BUR_OFF_AMT
EXT_BUR_ACCP_AMT
EXT_BUR_PAID_AMT
UG_SCHL_ACCP_AMT
UG_SCHL_PAID_AMT
PG_SCHL_OFF_AMT
PG_SCHL_ACCP_AMT
PG_SCHL_PAID_AMT
INT_BUR_OFF_AMT
INT_BUR_ACCP_AMT
INT_BUR_PAID_AMT
CLRD_AMT
PAYMENTS
UG_SCHL_OFF_AMT
RESI
AVERAGE_MARKS
REG_ONLINE
FORML_EXT_YR
REGISTRATION_STATUS
OFFICIAL_LANGUAGE_DSCR
HOW_PAID_SCH_FEES
SCH_LOCATION_DSCR
PART_TIME_WORK
ON_OFF_CAMPUS_ACC_DSCR
FAM_UNIV_ATTENDANCE
OWN_PHONE
MATHEMATICS
PHYSICAL_LIFE_SCI
LANGUAGES

Continued on next page

Table A.2 – continued from previous page

Variable
SOCIAL_SCIENCES
BUSINESS_ACCOUNTING
ACCOMODATION_CATERING_DSCR
TRAVEL_TIME_MRNG_DSCR
TUIT_SINGLE_BOTH_PRNTS_DSCR
TUIT_SINGLE_PARENT_DSCR
TUIT_EMPLYMNT_FATHER_DSCR
TUIT_OCCUPATION_FATHER_DSCR
TUIT_FEEPERC_FATHER_DSCR
TUIT_NUM_DPNDNTS_FATHER_DSCR
TUIT_EMPLYMNT_MOTHER_DSCR
TUIT_OCCUPATION_MOTHER_DSCR
TUIT_FEEPERC_MOTHER_DSCR
TUIT_EMPLYMNT_GRDIAN_DSCR
TUIT_OCCUPATION_GRDIAN_DSCR
TUIT_NUM_DPNDNTS_GRDIAN
TUIT_EMPLYMNT_SPOUSE_DSCR
TUIT_OCCUPATION_SPOUSE_DSCR
TUIT_OCCUPATION_SPOUSE_OTHER
TUIT_FEEPERC_NSFAS_DSCR
TUIT_FEEPERC_TRUSTFUND
TUIT_FEEPERC_TRUSTFUND_DSCR
TUIT_FEEPERC_GOVSCHLRSHP
TUIT_FEEPERC_GOVSCHLRSHP_DSCR
TUIT_FEEPERC_EDUINSUREPOL
TUIT_FEEPERC_EDUINSRPOL_DSCR
TUIT_FEEPERC_MERITSCHRSHP
TUIT_FEEPERC_MERITSCHSHP_DSCR
TUIT_FEEPERC_MERIT_TRUST
TUIT_FEEPERC_MERIT_TRUST_DSCR
TUIT_FEEPERC_LOAN
TUIT_FEEPERC_LOAN_DSCR
TUIT_FEEPERC_SELFFUND
TUIT_FEEPERC_SELFFUND_DSCR
TUIT_FEEPERC_BURSARY
TUIT_FEEPERC_BURSARY_DSCR
L_EXP_OCCUPATION_FATHER_DSCR
L_EXP_NUM_DPNDNTS_GRDIAN_DSCR
L_EXP_FEEPERC_NSFAS
L_EXP_FEEPERC_NSFAS_DSCR
L_EXP_FEEPERC_TRUSTFUND
L_EXP_FEEPERC_TRUSTFUND_DSCR
L_EXP_FEEPERC_GOVSCHLRSHP
L_EXP_FEEPERC_GOVSCHLRSHP_DSCR

Continued on next page

Table A.2 – continued from previous page

Variable
L_EXP_FEEPERC_EDUINSUREPOL
L_EXP_FEEPERC_EDUINSRPOL_DSCR
L_EXP_FEEPERC_MERITSCHRSHP
L_EXP_FEEPERC_MERITSCHSHP_DSCR
L_EXP_FEEPERC_MERIT_TRUST
L_EXP_FEEPERC_MERIT_TRUST_DSCR
L_EXP_FEEPERC_LOAN
L_EXP_FEEPERC_LOAN_DSCR
L_EXP_FEEPERC_SELFFUND
L_EXP_FEEPERC_SELFFUND_DSCR
L_EXP_FEEPERC_BURSARY
L_EXP_FEEPERC_BURSARY_DSCR
SCH_HAD_COMPUTERS
SCH_HAD_LIBRARY
SCH_HAD_SCIENCE_LABS
SCH_HAD_ELECTRICITY
SCH_HAD_RUNNING_WATER
SCH_HAD_SCHOOL_HALL
SCH_HAD_SPORTS_FIELDS
SCH_HAD_BOARDING_FACILITIES
SCH_HAD_OSIDE_TOILET_NO_WATER
SCH_HAD_STUDENT_DESKS_TABLES
SCH_HAD_TEXT_BOOKS
SCHOOL_CODE
URBAN_RURAL
QUINTILE
SIMS_URBAN_RURAL

TABLE A.3: Moodle Database Tables

Order	Word	Relative Frequency
0	use	0.011893
1	thank	0.010533
2	class	0.008495
3	function	0.008155
4	value	0.008155
5	test	0.008155
6	list	0.007815
7	using	0.006796
8	new	0.006796
9	will	0.006456
10	code	0.006456

Continued on next page

Table A.3 – continued from previous page

Order	Word	Relative Frequency
11	problem	0.006116
12	question	0.006116
13	operator	0.005776
14	need	0.005437
15	temp	0.005437
16	want	0.005097
17	prime	0.004757
18	file	0.004417
19	way	0.004417
20	one	0.004417
21	make	0.004417
22	found	0.004077
23	think	0.004077
24	case	0.004077
25	say	0.004077
26	youtube	0.004077
27	help	0.004077
28	github	0.004077
29	linked list	0.004077
30	number	0.003738
31	linkedlist	0.003738
32	work	0.003738
33	add	0.003738
34	first	0.003398
35	now	0.003398
36	newlist	0.003398
37	video	0.003398
38	another	0.003058
39	try	0.003058
40	pre	0.003058
41	data	0.003058
42	object	0.003058
43	lab	0.003058
44	white space	0.003058
45	terminal	0.002718
46	reference	0.002718
47	see	0.002718
48	course	0.002718
49	example	0.002718

A.1 Background Features Descriptions

1. *Quintile* is a field that indicates which of the five categories a school belongs to under the South African Government schools standard (Which range between 1 and 5, also called *quintile*), and an additional category that indicates private high schools (Quintile label 6). The category is based on the *relative wealth of the school's surrounding communities* "Groundup", 2019.

   Variable type: **Categorical**

   Possible values: **1, 2, 3, 4, 5, 6**.

2. *Gauteng Province* is a variable that indicates whether a student completed their ultimate year of high school at a school in *GP* (Gauteng Province). GP is a province with the highest population (25%) and considered the wealthiest of the nine provinces, as measured by its contribution to South Africa's national income (between 33% and 35% annually) and income per person (Gateway, 2019; OECD, 2020).

   Variable type: **Boolean**

   Possible values: **0,1 = No, Yes**.

3. *Gender*: A Boolean Variable indicating whether the student is female.

   Variable type: **Boolean**

   Possible values: **0,1 = Male, Female**.

4. *State Financial Assistance* is a variable indicating if a student received financial aid from the National National Student Financial Aid Scheme.

   Variable type: **Boolean**

   Possible values: **0,1 = No, Yes**.

5. *Township School* is a variable indicating if a student's high school was situated in a township area. This is defined as an urban area that was demarcated for Black, Coloured and Indian residents of South Africa. Under the the Land Act of 1913 and the Group Areas Act of 1950, each of the above racial groups was segregated into their respective townships for residential purposes, before 1994 (Ladd, 2008)

   Variable type: **Boolean**

   Possible values: **0,1 = No, Yes**.

6. *Grade* takes on two possible values that show whether a student obtained an average grade above 51%, as defined in Section 3.5 on Encoding Performance.

   Variable type: **Boolean**

Possible values: **0,1 = No, Yes**

7. *Grade* is a continuous variable that shows the average grade that a student achieved. We use Grade for the regression experiment, which is used to introduce the linear relationship between Background features of a student and their Grades.

   Variable type: **Continuous**

   Possible values: **0-100**

A.2 OLS Assumptions: Forum Activity and Performance

A.2.1 Relationship between Grades and Crude Post Count

Figure A.4 shows the autocorrelation of the residuals for student Grades against their Crude Post Count. The large autocorrelation values show that the a linear model is an invalid inference model.

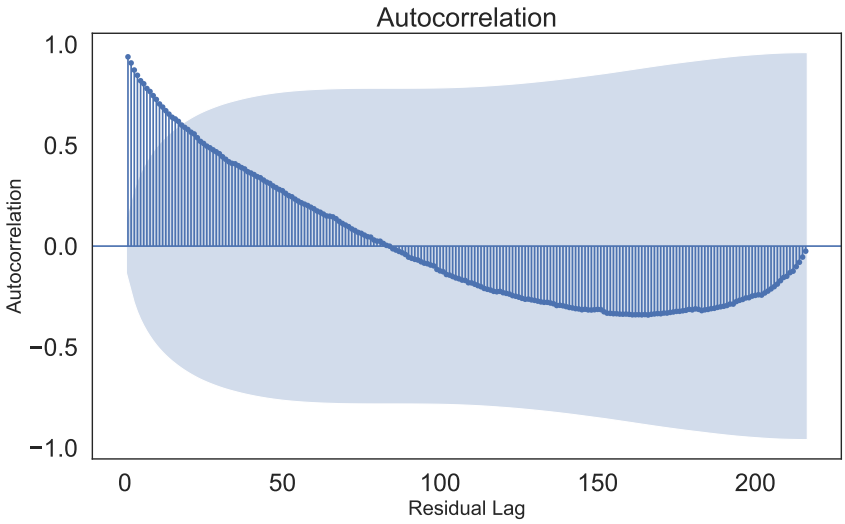


FIGURE A.4: Crude Post Count Violation of Autocorrelated Residuals

A.2.2 Relationship between Grades and Number of Posts – OLS Assumptions

Here, we validate the OLS assumptions for the OLS model data whose results are reported in Table 5.5.

Residual-Normality Test

1. Null hypothesis: That the residuals, $G_f - \hat{G}_f$, are not normally distributed.

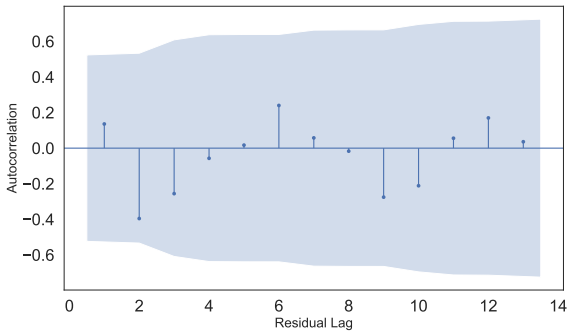


FIGURE A.5: Autocorrelation of Residuals for Grade against Post Frequency Group

2. $s^2 + k^2$ statistic: 1.359,
3. p-value: 0.507
4. Outcome: The insignificant p-value of $s^2 + k^2$ suggests that we fail to reject the null that the residuals are not normally distributed

A.2.2.1 Autocorrelation of Residuals

Figure A.5 shows the autocorrelation of residuals, $G_f - \hat{G}_f$. Since the autocorrelation at all lag levels is not statistically significant at a 95% confidence level, we can reject the null that the residuals are autocorrelated.

A.2.2.2 Student Discussions – OLS Assumptions

Here, the OLS assumptions for the OLS model data whose results are reported in Table 5.7 are validated.

A.2.2.3 Residual-Normality Test

1. Null hypothesis: That the residuals, $\hat{G}d_i(s) - Gd_i(s)$, are not normally distributed.
2. $s^2 + k^2$ statistic: 4.538,
3. p-value: 0.103
4. Outcome: Fail to reject the null that the residuals are not normally distributed

A.2.2.4 Autocorrelation of Residuals

Figure A.6 shows the autocorrelation of residuals, $\hat{G}d_i(s) - Gd_i(s)$. Since the autocorrelation at all lag levels is not statistically significant at a 95% confidence level, we can reject the null that the residuals are autocorrelated.

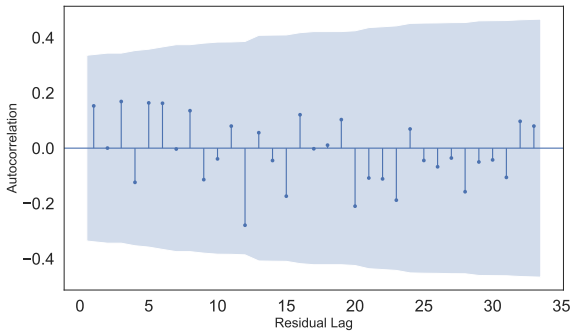


FIGURE A.6: Autocorrelation of Residuals for Grade against Discussion

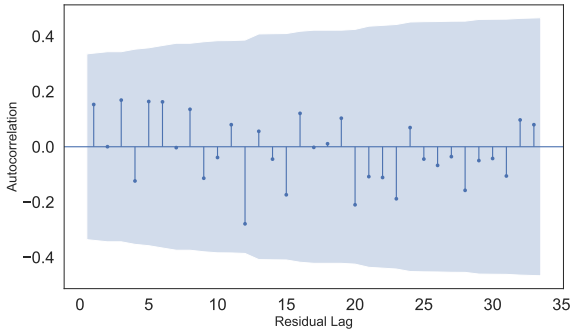


FIGURE A.7: Autocorrelation of Residuals for Grade against Collaboration-group

A.2.2.5 Student Collaboration-groups – OLS Assumptions and *k*NN Code

Here, the assumptions for the OLS model data whose results are reported in Table 5.12 are validated.

Residual-Normality Test

1. Null hypothesis: That the residuals, $Gc(a) - \hat{G}c(a)$, are not normally distributed.
2. $s^2 + k^2$ statistic: 1.359,
3. p-value: 0.507
4. Outcome: The insignificant p-value of $s^2 + k^2$ suggests that we fail to reject the null that the residuals are not normally distributed

Autocorrelation of Residuals

Figure A.7 shows the autocorrelation of residuals, $Gc(a) - \hat{G}c(a)$. Since the autocorrelation at all lag levels is not statistically significant at a 95% confidence level, we can reject the null that the residuals are autocorrelated.

```
from sklearn.neighbors import NearestNeighbors
discNeighs = []
samples = idDiscArr
neigh = NearestNeighbors(n_neighbors=10)
neigh.fit(samples)
allNeighbs = []
for i,s in enumerate(samples):
    #find the nearest neighbours
    sNeighbs = neigh.kneighbors([s])[1]
    #find the true neighbours (buddies) by checking if (but not how many) they have common discussions
    #we append these neighbours in sNeighbsTrue
    sNeighbsTrue = []
    for n in range(len(sNeighbs[0])):
        if sNeighbs[0][n] != i:

            #check in which discussions this neighbour participated;
            #NeighParticipated = array of discs neighbour participated in
            NeighParticipated = np.where(samples[sNeighbs[0][n]]==1)
            #check the intersection of the above participant, sNeighbs[0][n] and s's discussions
            inter = set(np.where(s==1)[0]).intersection(NeighParticipated[0])
            # it doesn't matter how many discussions they have in common; everyone they share a discussion with
            #if they share no discussions, do nothing but if they do (their intersection is the empty set()),
            #then append them to the list of true neighbours.
            #but knn show that a and b are neighbours even if they have no common discussions; so we eliminate those
            if inter != set():
                sNeighbsTrue.append(sNeighbs[0][n])
        else: #this else block is the only difference between this and the next cell
            sNeighbsTrue.append([])
    #the neighbours, in no order and of which the distance is deliberately unknown (unknown; our max is 10 though)
    # append sNeighbsTrue to allNeighbs, which will go back into idDisc as a new column
    # each entry showing an array of nearest neighbours (or at least a reference the array...)
    if sNeighbsTrue != []: #after checking all neighbours, n, of s, append the ones that aren't empty
        allNeighbs.append(sNeighbsTrue)
```

FIGURE A.8: kNN Algorithm to Nearest Neighbours of a

A.3 OLS Assumptions: E-Behaviour, Personality and Performance

A.3.1 Average Number of Logins, Conscientiousness and Grade – OLS Assumptions

Here, we validate the OLS assumptions for the OLS model data whose results are reported in Table 6.4.

A.3.2 Residual-Normality Test

1. Null hypothesis: That the residuals, $G - \hat{G}$, are not normally distributed.
2. $s^2 + k^2$ statistic: 4.801,
3. p-value: 0.091
4. Outcome: Fail to reject the null that the residuals are not normally distributed

A.3.3 Autocorrelation of Residuals

Figure A.9 shows the autocorrelation of residuals, $G - \hat{G}$. Since the autocorralation at all lag levels is not statistically significant at a 95% confidence level, we can reject the null that the residuals are autocorrelated.

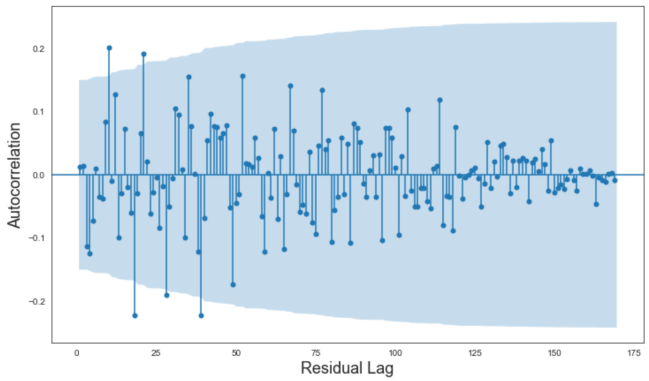


FIGURE A.9: Autocorrelation of Residuals for Grade against Estimated Grade

```
# random grades.
np.random.seed(97)
# min and max values that appear in data
mi, ma = idGrade['SemAvg'].min(), idGrade['SemAvg'].max()
# Random data from normal distribution; parameters are mean, standard deviation and the length of the data
idGrade['SemAvg'] = np.random.normal(idGrade['SemAvg'].mean(),idGrade['SemAvg'].std(), len(idGrade['SemAvg']))
# cap to min and max
idGrade = idGrade[(idGrade['SemAvg']>=mi) & (idGrade['SemAvg']<=ma)]
# end random grades
```

FIGURE A.10: Random-Grade Generation for Students

Bibliography

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Ajoodha, R., Jadhav, A., & Dukhan, S. Forecasting learner attrition for student success at a south african university. In: New York, NY, USA: Association for Computing Machinery, 2020. ISBN: 9781450388474. https://doi.org/10.1145/3410886.3410973.

News24Wire. (2015). South africa's alarming university drop out rate. https://businesstech.co.za/news/trending/87770/south-africas-alarming-university-drop-out-rate/

John, V. (2013). Dropout rate points to lack of support [Accessed: 2020-01-04].

Richiţeanu-Năstase, E.-R., & Stăiculescu, C. (2018). University dropout. causes and solution. *1*, 71–75.

Wright, D., & Taylor, A. (1970). Introducing psychology: An experimental approach.

Heppner, P. P., Wampold, B. E., Owen, J., Wang, K. T., & Thompson, M. N. (2015). *Research design in counseling* (Fourth Edition). Cengage Learning.

Stone, A. A. (2000). *The science of self-report: Implications for research and practice*. Lawrence Erlbaum.

Northrup, D. A., York University (Toronto, O., & for Social Research, I. (1997). *The problem of the self-report in survey research: Working paper*. Institute for Social Research, York University.

Fellegi, I. P. (1973). The evaluation of the accuracy of survey results: Some canadian experiences. *International Statistical Review / Revue Internationale de Statistique*, *41*(1), 1–14.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology*, *94*(2), 396.

Dauter, L. G. (2016). Economic sociology [Accessed: 2020-12-29].

Bourdieu, P., & Richardson, J. G. (1986). The forms of capital.

Carpiano, R. M. (2006). Toward a neighborhood resource-based theory of social capital for health: Can bourdieu and sociology help? *Social science & medicine*, *62*(1), 165–175.

Hallinan, M. T., & Smith, S. S. (1989). Classroom characteristics and student friendship cliques. *Social forces*, *67*(4), 898–919.

Song, L. (2011). Social capital and psychological distress. *Journal of health and social behavior*, *52*(4), 478–492.

Hayes, E. (1997). Elaine hayes on "the forms of capital". https://web.english.upenn.edu/~jenglish/Courses/hayes-pap.html

Smith, E., & White, P. (2015). What makes a successful undergraduate? the relationship between student characteristics, degree subject and academic success at university. *British Educational Research Journal*, *41*(4), 686–708.

Caldas, S. J., & Bankston, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research*, *90*(5), 269–277. https://doi.org/10.1080/00220671.1997.10544583

Fan, J. (2014). The impact of economic capital, social capital and cultural capital: Chinese families' access to educational resources. *Sociology Mind*, *04*, 272–281. https://doi.org/10.4236/sm.2014.44028

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, *135*(2), 322.

Costa, P. T., & McCrae, R. R. (1985). *The neo personality inventory*. Psychological Assessment Resources Odessa, FL.

Furnham, A., Nuygards, S., & Chamorro-Premuzic, T. (2013). Personality, assessment methods and academic performance. *Instructional Science*, *41*(5), 975–987.

Ciorbea, I., & Pasarica, F. (2013). The study of the relationship between personality and academic performance. *Procedia-Social and Behavioral Sciences*, *78*, 400–404.

Kumari, B. (2014). The correlation of personality traits and academic performance: A review of literature. *IOSR Journal of Humanities and Social Science*, *19*, 15–18.

Morris, P. E., & Fritz, C. O. (2015). Conscientiousness and procrastination predict academic coursework marks rather than examination performance. *Learning and Individual Differences*, *39*, 193–198.

Costa Jr, P. T., & McCrae, R. R. (2008). *The revised neo personality inventory (neo-pi-r)*. Sage Publications, Inc.

Lim, A. (2020). The big five personality traits.

A. Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech*, *11*, 41–66. https://doi.org/10.1108/eb047884

Yu, L., & Apps, A. (2000). Studying e-journal user behavior using log files: The experience of superjournal. *Library Information Science Research*, *22*(3), 311 –338. https://doi.org/https://doi.org/10.1016/S0740-8188(99)00058-4

Anitha, V., & Isakki, P. A survey on predicting user behavior based on web server log files in a web usage mining. In: *2016 international conference on computing technologies and intelligent data engineering (icctide'16)*. 2016, 1–4.

Singh, A. P., & Jain, R. (2014). A survey on different phases of web usage mining for anomaly user behavior investigation. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *3*(3).

Shanmugarajeshwari, V., & Lawrance, R. Analysis of students' performance evaluation using classification techniques. In: *2016 international conference on computing technologies and intelligent data engineering (icctide'16)*. 2016, 1–7. https://doi.org/10.1109/ICCTIDE.2016.7725375.

Wilson, B. C., & Shrock, S. Contributing to success in an introductory computer science course: A study of twelve factors. In: SIGCSE '01. Charlotte, North Carolina, USA: Association for Computing Machinery, 2001, 184–188. ISBN: 1581133294. https://doi.org/10.1145/364447.364581.

Evans, G. E., & Simkin, M. G. (1989). What best predicts computer proficiency? *Commun. ACM*, *32*(11), 1322–1327. https://doi.org/10.1145/68814.68817

Fowler, G. C., & Glorfeld, L. W. (1981). Predicting aptitude in introductory computing: A classification model. *AEDS Journal*, *14*(2), 96–109.

Samrit, N. B., & Thomas, A. (2017). A recommendation system for prediction of elective subjects. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *5*(4), 36–43.

Ashenafi, M. M., Riccardi, G., & Ronchetti, M. Predicting students' final exam scores from their course activities. In: *2015 ieee frontiers in education conference (fie)*. IEEE. 2015, 1–9.

Poh, N., & Smythe, I. To what extend can we predict students' performance? a case study in colleges in south africa. In: *2014 ieee symposium on computational intelligence and data mining (cidm)*. 2014, 416–421. https://doi.org/10.1109/CIDM.2014.7008698.

Yang, F., & Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, *123*, 97–108.

Ángel Agudo-Peregrina, Iglesias-Pradas, S., Ángel Conde-González, M., & Ángel Hernández-García. (2014). Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning. *Computers in Human Behavior*, *31*, 542 –550.

Ciolacu, M., Tehrani, A. F., Binder, L., & Svasta, P. M. Education 4.0 - artificial intelligence assisted higher education: Early recognition system with machine learning to support students' success. In: *2018 ieee 24th international symposium for design and technology in electronic packaging (siitme)*. 2018, 23–30.

Ciolacu, M., Tehrani, A. F., Beer, R., & Popp, H. (2017). Education 4.0 — fostering student's performance with machine learning methods. *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 438–443.

Park, Y., Yu, J. H., & Jo, I.-H. (2015). Clustering blended learning courses by online behavior data case study in a korean higher education institute. *The Internet and Higher Education*, *29*.

Dahlstrom, E., Walker, J., & Dziuban, C. (2013). *Ecar study of undergraduate students and information technology* (tech. rep.). 2013.

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, *307*, 72 –77. https://doi.org/https://doi.org/10.1016/j.neucom.2018.03.067

Molnar, C., Casalicchio, G., & Bischl, B. (2020, October). Interpretable machine learning – a brief history, state-of-the-art and challenges.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

Merriam-Webster. Dutiful [Accessed: 2020-11-29].

Wilt, J., & Revelle, W. (2009, January). Extraversion.

Association, A. P. (2020). Apa dictionary of psychology – gregariousness [Accessed: 2020-12-29].

Ajzen, I. (2005). *Attitudes, personality, and behavior*. McGraw-Hill Education (UK).

Campbell, D. T. (1963). Social attitudes and other acquired behavioral dispositions. *Psychology: A study of a science. study ii. empirical substructure and relations with other sciences. volume 6. investigations of man as socius: Their place in psychology and the social sciences.* (94–172). McGraw-Hill. https://doi.org/10.1037/10590-003

Hemakumara, G., & Ruslan, R. (2018). Spatial behaviour modelling of unauthorised housing in colombo, sri lanka. *25*, 91–107. https://doi.org/10.21315/kajh2018.25.2.5

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660–674.

Mitchell, T. M. (1997). *Machine learning* (1st ed.). McGraw-Hill, Inc.

Raileanu, L., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*, 77–93.

Khalaf, A., Hashim, A., & Akeel, W. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*, 26–31.

Topîrceanu, A., & Grosseck, G. (2017). Decision tree learning used for the classification of student archetypes in online courses [Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France]. *Procedia Computer Science*, *112*, 51 –60.

Kolo, K. D., Adepoju, S. A., & Alhassan, J. K. (2015). A decision tree approach for predicting students academic performance. *International Journal of Education and Management Engineering*, *5*(5), 12.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics*. Douglas Reiner.

Shumway, R. H., & Stoffer, D. S. Time series regression and exploratory data analysis. In: *Time series analysis and its applications*. Springer, 2011, pp. 47–82.

Little, T. D., & Wei, W. W. (2013, October). Time series analysis. oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898-e-022

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice, 2nd edition [Accessed: 2019-04-01]. OTexts.com/fpp2

Liu, Z., & Sullivan, C. J. (2019). Prediction of weather induced background radiation fluctuation with recurrent neural networks. *Radiation Physics and Chemistry*, *155*, 275 –280.

Wang, M., Zhang, Y. D., & Cui, G. (2019). Human motion recognition exploiting radar with stacked recurrent neural network. *Digital Signal Processing*, *87*, 125 –131.

Bengio, Y., Boulanger-Lewandowski, N., & Pascanu, R. (2012). Advances in optimizing recurrent networks. *CoRR*, *abs/1212.0901*. arxiv.org/abs/1212.0901

Olah, C. (2015). Understanding lstm networks [Accessed: 2019-04-19].

Hand, D., & Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, *28*(3), 539–547.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389–422.

Misja. (2021). Epoch unix timestamp conversion tools [Accessed: 2019-05-03].

Romero, C., Ventura, S., Espejo, P., & Martínez, C. Data mining algorithms to classify students. In: 2008, January, 8–17.

Bhandari, H., & Yasunobu, K. (2009). What is social capital? a comprehensive review of the concept. *Asian Journal of Social Science*, *37*, 480–510. https://doi.org/10.1163/156853109X436847

Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of applied psychology*, *78*(5), 715.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology.

Hung, J.-L., & Zhang, K. (2009). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, *4*.

Leontjeva, A., & Kuzovkin, I. Combining static and dynamic features for multivariate sequence classification. In: 2016, October, 21–30. https://doi.org/10.1109/DSAA.2016.10.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Groundup. (2019).

Gateway, S. A. (2019). The nine provinces of south africa [Accessed: 2020-09-10].

OECD. (2020). Regional income per capita [Accessed: 2021-01-04].

Ladd, B. (2008). Townships. *University of Kwa-Zulu Natal. International Encyclopedia of Social Sciences, 2nd Edition. Published By: Macmillan Reference USA*, 405–406.