Leveraging Learning Analytics to Optimize Virtual Learners' Performance

By

Beatrice Chebet Ng'eno

145614

Master of Science in Data Science and Analytics

2024

Leveraging Learning Analytics to Optimize Virtual Learners' Performance

By

Beatrice Chebet Ng'eno

145614

Submitted in Partial fulfillment of the Requirements for the Degree of Master of Science in Data Science and Data Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

JULY 2024

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's Name: Sign: ..... Date: 03/26/2024

Approval

The thesis of **Beatrice Chebet Ng'eno** was reviewed and approved for examination by the following:

Dr. John Olukuru

Institute of Mathematical Sciences,

Strathmore University

8/4/2024

Dr. Kitawi Alfred

School of Humanities & Social Sciences,

Strathmore University

Dedication

I dedicate this dissertation to my family, friends, mentors and classmates who provided unwavering support in my academic journey. Their belief in my abilities have been my driving force towards my determination in pursuing knowledge. This dissertation is a testament to the impact of the values instilled in me throughout this transformative journey.

Acknowledgements

I am thankful to my supervisors, Dr. Alfred Kitawi and Dr. John Olukuru, whose expertise, constructive feedback and commitment to academic excellence have played a pivotal role in shaping the outcomes of this research. I extend my gratitude to my colleagues for their support towards enriching the scope of this research endeavor. I would also like to appreciate Strathmore University for granting research resources that enable the completion of this study.

Abstract

Learning analytics has gained traction globally over the years with many institutions acknowledging its potential to optimize learning and the environments in which learning occurs. The study is structured around three primary objectives aiming to provide a key focus on optimization of virtual learners' academic outcome using learning analytics approaches. Firstly, it aims to identify key indicators that reliably predict students' performance within academic settings. Secondly, it seeks to examine and compare the effectiveness of different algorithms in accurately forecasting students' performance outcomes. Lastly, the research endeavours to develop and deploy a performance prediction and early alert tool utilizing R-Shiny. In this study, the performance of Logistic Regression, Naive Bayes, K-Nearest Neighbors and Support Vector Machine in predicting learners' performance were evaluated. Utilizing 21,216 records from students at the The Open University UK, the results indicated Logistic Regression as the best performing model with a precision rate of 90% and key features encompassed student demographic information and academic history. The findings of this study give invaluable insights to educational institutions on leveraging learning analytic practices for data-driven interventions to optimize and enhance student performance. In conclusion, this study not only provides a tangible solution of students' performance optimization but also contributes to the growing body of knowledge on learning analytic practices that provide solutions which can be incorporated in the education sector.

Keywords: Learning analytics, machine learning, student performance, R-Shiny.

Contents

List of Figures

List of Abbreviations

ROSE	-Random Over-Sampling Examples
SOLAR	-Society of Learning Analytics Research
LMS	-Learning Management Systems
VLE	-Virtual Learning Environment
TEL	-Technology-Enhanced Learning
PLE	-Personal Learning Environment
KNN	-K-Nearest Neighbours
SVM	-Support Vector Machine
EWS	-Early Warning System
STEM	-Science, Technology, Engineering, and Mathematics
EDA	-Exploratory Data Analysis
UI	-User Interface
SNA	-Social Network Analysis
GPA	-Grade Point Average
RMSE	-Root Mean Squared Error
LDA	-Linear Discriminant Analysis
CART	-Classification and Regression Trees

CHAPTER ONE: INTRODUCTION

1.1 Introduction

In this chapter, background, problem statement, the objectives, significance, scope, justification as well as the limitation of the study is presented. The key aspects revolving around the applications of learning analytic practices will be discussed with a focus of the context for the same.

1.2 Background of the study

Society of Learning Analytics Research (SOLAR) (2011), describes learning analytics as capturing learners' data and conducting analysis to optimize learning as well as the environments in which they occur. The availability of data captured in Learning Management Systems (LMS) and Virtual Learning Environments (VLEs), has facilitated the use of learning analytic practices to optimize learning and its environments (Clow,2013). The need to measure learners progress and management has seen many institutions apply learning analytics to inform their decisions (Clow, 2013).

The use of learning analytics practices to aid students at risk of failing has been profound across the globe. Purdue University in the US is known for its Course Signal system that was developed to identify students at risk through the analysis of data captured on the institution's LMS, Blackboard Vista (Arnold & Pistilli, 2012). Purdue University is among many institutions in the US that have reported better performance and retention among its students from data-driven decisions. In the UK, The Open University, which offers most of its courses online, has made major advancements in capturing and analyzing learners' data. The institution has been able to enhance students' performance and reduced churn among the students (Sclater et al.,2016).

Institutions in South Africa leverage learning analytics practices to inform their decisions. According to a report by Lourens and Bleazard (2016), Cape Peninsula University of Technology is among the institutions that apply predictive analytics to identify struggling students and those at risk of churn. Although there are few institutions that leverage learning analytics to optimize learning and its environments in Africa, there are promising advancements towards it due to the substantial adoption of LMS across many institutions (Prinsloo & Kaliisa,2022).

1.3 Problem Statement

Virtual learning environments have provided the ability to increase the scope of learning various content to a wider group of individuals. With the increased coverage, more and more students are able to take up courses of their choice. The ideal in the concept of learning, has always been taking up a course and successfully finishing it as a measure of gained knowledge (Looi et al.,2010). Learning institutions are therefore tasked with ensuring that students' achievement for ideal outcome is attained by investing in best practices informed by learners' data (Clow,2013).

When we look into virtual learning, the experience is quite different from that of physical learning. The online experience may present challenges that affect a learners' course outcome and this has led to the increasing need of performance management and a measurement system as Clow (2013) mentioned. Some of the challenges that may face online learners, as Ferguson (2012) reported, include, lack of motivation and diverting attention. The use of learning analytics to predict learners' course outcome has seen many researchers implement various models. The choice of factors that affect learners' results has also been different. In University Pendidikan Sultan Idris, Malaysia, they utilized age as the only variable determining the performance of a student (Zulfikri et al.,2021). While it is evident that various factors could affect the end results of students, some of these studies chose to apply only a few. Most of the use cases of learning analytics do not have an end-product such as an early alert tool and the few with early alert tools are plug-ins to various LMS. Generally, in Africa the adoption of learning analytics practices to aid in optimizing learning processes and its environments has been very minimal (Prinsloo, 2018) including incorporating early alert tools. We noted that there was no data product for most of the research that were conducted. The aim of this study is to create an early alert tool that incorporates key indicators in predicting learners' performance using data from the Open University's virtual learners. The data product can be utilized in institutions in Africa including the first virtual public university in Kenya (Open University of Kenya) which was recently launched (2023, August). We utilized Open University's virtual learners' data due to its reliability and open accessibility including that no information was captured through informal means.

1.4 Research Objectives

1.4.1 General Objective

The general objective of this study is to develop and deploy a web-based analytic tool that predicts the course outcome of students and alerts those at risk of failing for early intervention.

1.4.2 Specific Objectives

1 To identify key indicators for predicting students' performance.

2 To examine which algorithm performs best in predicting students' performance.

3 To develop and deploy a Performance Prediction and Early Alert Tool using R-Shiny.

1.5 Research Questions

1 What are the significant key indicators for predicting the performance of a student?

2 Which algorithms perform best in predicting students' performance?

3 How can I use the findings from the above research questions to create and implement a performance prediction and early warning tool using R-Shiny?

1.6 Significance of the study

This study provides a focus on the key aspects that affect learners' course outcome. This is vital for institutions to identify struggling students (at- risk of failure) (SOLAR,2011). Just as Clow (2013) reported on the use of data-driven insights to come up with performance management measures, this study will benefit the school managers and policy makers in decision making as far as key indicators that affect students' outcomes are concerned. The school administrators will also be able to utilize the insights from the findings of this study to improve learning processes and its environments for better course outcomes. Through the web-application that sends alerts to students, students can seek instructors' assistance at personal level and consequently improve their grades. The teachers/instructors are able to use insights of this study to come up with teaching strategies that will assist learners improve their course outcome. To the data analytic experts who provide data-driven solutions to various institutions, this study is beneficial to them in as far as the

entire study pipeline is concerned. The study provides a concise description of the process of choosing algorithms and development of an alert tool so as to enable data analytic experts to draw insights and enrich their processes within the learning analytics space. Generally, the contribution of this study is enhancement of learners' success by encouraging early data-driven interventions to aid struggling students just as is in the case of Purdue University in the US. Leveraging learning analytics to inform decisions among stakeholders would in the long-run boost performance of the students (Smith et al., 2012). The findings of this study are therefore beneficial to any institution with an environmental setting similar to that of The Open University.

1.7 Scope of the study

This study focused on predicting learners' performance at The Open University, UK. The data which contain behavioral and performance aspects were collected by the institution on their VLE and were open for research purposes. It consists of 21,216 anonymized students cutting across 22 courses offered at the institution.

1.8 Justification of the study

This study had to be conducted because the need to track struggling students is increasing by-day as more institutions adopt virtual and hybrid learning. The study will help to reveal the factors that are useful to understand learners' performance. As such, institutions may be able to use insights from this study to come up with measures to assist students to improve in their academics (Clow,2013). Also, with the many institutions especially in Africa who have data capturing tools but have not implemented using the data to inform decisions (Prinsloo,2018), this study is conducted to provide a clear use case of the data to improve learners' outcome and thus might be convinced to invest in learning analytic practices.

1.9 Limitation of the study

This study is focused on only the students taking courses off-campus. The variables applied for in this study are limited to the data provided by The Open University.

CHAPTER TWO: RELATED WORKS

2.1 Introduction

This chapter seeks to delve into related works that have been done by other researchers and present their findings on the concept of learning and learning analytics. We seek to contextualize the frameworks that other researchers have implemented and address the issues they try to resolve. Existing gaps were identified after critical review of the literature for this study to address.

2.2 An overview of learning and learning analytics

Learning is defined as increasing the value of students and the change in their character which will shape the future of the individual (Looi et al., 2010). With increasing adoption of technology within the education sector around the globe to substitute or complement traditional approaches of learning, learning has taken a new shift towards Technology- Enhanced Learning (TEL) (Kirkwood et al., 2014). TEL enhances the pedagogical shift from a teacher-constrained learning towards full participation that is student-centered (Looi et al., 2010).

According to Society of Learning Analytics Research (2011), Learning Analytics involves the measurement, capturing, analysis and reporting of learners' data in order to understand and optimize learning and the environments in which they occur. Learning analytics can also be defined as the analysis of students' data and presenting them in a way that can assist to improve quality of learning (Clow, 2013). Big data techniques such as to capture numerous students' data, modeling, understanding and prediction have been applied in the education sector to improve the learning processes and the environments to which they occur (Clow, 2013). To inform their practices, providers of Massively Open Online Courses (MOOCs) such as Coursera and edX, leverage learning analytics.

Several facilitators have led the use of learning analytics to informal educational decision making and practices. One of the major drivers of learning analytics is the availability of data especially for learning that occurs in the online environments, that is, data captured in Virtual Learning Environments (VLEs) and Learning Management System (LMS) (Clow, 2013). Every interaction within VLEs and LMS are recorded and stored. LMS such as Moodle have led to the increased amount of data captured that includes personal data, interaction data and academic information about every student (Ferguson, 2012). Data collected from LMS and are used to understand and

optimize online learning environments (Na, Kew Si, & Zaidatun Tasir, 2017). With metrics such as frequency of interaction with teaching materials as well as the performance in students' assessment, researchers have been able to predict the overall learner's performance (Smith et al., 2012). The second facilitator for learning analytics practices is the need to have performance management, system of measurement and quantification (Clow, 2013). The third driver is the availability of statistical and highly intensive computational tools that can be used to capture and manage high volumes of data through which analysis and interpretation can be easily done (Clow, 2013).

2.3 Contextual Background of Learning Analytics

2.3.1 Global Context

There has been an increased pressure to measure, illustrate and enhance learners' performance. In the USA, the government aims at increasing overall education attainment within its population (Ferguson, 2012). In a report from the U.S. Department of Education (2012), it was noted that higher learning institutions are beginning to leverage data analytics to enhance students' grades and improve retention. One of the key applications that have been successful in the U.S. is predicting learners' performance. Examples of data-driven systems that have been in place to identify students at-risk include; Course Signal System from Purdue University and Moodog System from University of California and University of Alabama (Arnold & Pistilli, 2010; Fiaidhi, 2014, Bienkowski et al., 2012).

In Australia, more and more institutions have started to leverage learning analytics for various purposes. Although many institutions are still at implementation stages of learning analytics some of them have applied it to drive their decisions (Siemens et al., 2013). At the University of Wollongong, through social network analysis of data captured in forums, the institution has been able to record interaction patterns which facilitate the improvement of teaching quality (Sclater et al., 2016). Queensland University of Technology in Brisbane city, tracks behavioral and cognitive indicators from approximately 45,000 students with the intention to improve retention of students. The monitoring and performance reporting programs at the university are reported to be successful due to significant improvement in student retention (Siemens et.al., 2013).

Advancements towards data-driven educational practices have also been profound in the UK. The Open University whose most courses are completed online, has been successful in utilizing learning analytics to enhance student progression, experience and retention (Sclater et al.,2016). The aspects it utilizes include, courses, student_info, assessments among others. Nottingham Trent University, with a capacity of approximately 28,000 students also leverages academic analytics specifically, predictive analytics, to identify struggling students with the level of engagement of each of the students as the indicator of performance. Early intervention is provided to support students (Sclater et al., 2016).

2.3.2 Regional Context

In Africa, implementation of learning analytics has been quite minimal. Although there has been substantial adoption of learning management systems in countries such as Tanzania, Nigeria, Ghana and Zimbabwe, the adoption of learning analytic practices is generally low with internet accessibility as one of the factors attributed to it (Prinsloo & Kaliisa,2022). Mostly the documented institutions are in South Africa (Prinsloo, 2018). University of Cape Town are reported to have developed two MOOCs to aid in understanding and analyzing how people responded to learning designs (Walji et al.,2016). Lourens & Bleazard (2016), reported to have deployed predictive learning analytics for Cape Peninsula University of Technology to aid identify students at-risk and promote retention.

2.3.3 Local Context

In Kenya, there was hardly any documentation on any learning analytics practices that have been implemented though various institutions have invested in learning management systems used to capture students' data. In studies conducted by several researchers such as Ogwoka, Cheruiyot and Okeyo (2015), they only provide the big data practices on students' data but nothing more past that.

2. 4 Theoretical Review

2.4.1 The Learning Analytics Cycle

The Learning Analytics Cycle, is a theoretical guide that enhances the effectiveness of learning analytics projects (Clow, 2012). The cycle consists of four main steps that can be implemented to

be able to come up with actionable insights from learners' data. The cycle, however, must not be necessarily completed as some of the projects like coming up with a report might not end up to the last phase. Fundamentally, for effectiveness of this theoretical framework for learning analytics projects, the cycle must be complete (Clow, 2012).

The first step in this cycle involves the learners, that is the students taking a particular course either in virtual platforms or physically at a given institution. Generating and capturing volumes of data from the students under the various settings is the next goal. The kind of data that can be captured within this stage includes, the students' demographic data, results of assessments they have been tested on, login patterns and clicks made within VLEs/LMS (Clow, 2012).

The third step involves data preprocessing and general analysis to come up with actionable insights. This stage may be characterized by visualizations, interactive dashboards as well as comparison of outcome of some measures put in place etc. (Clow, 2012). Coming up with metrics and analytics might require significant efforts from an analytical team who leverage statistical tools to come up with predictive models, interactive dashboards, recommender systems, early warning systems, social network analysis and so on.

Interventions informed by the above processes complete the learning analytics cycle. Clow(2012), stressed on the need for the metrics to be used to come up with interventions that might aid the learners even if the data did not originate from that specific group of learners. Coming up with dashboards might allow learners to compare each other's academic progress and teachers might as well benefit from this as they can easily compare academic achievements from different cohorts (Clow, 2012). The analytics might also be used to pin-point 'high-risk' learners, that is, those students who are likely to churn and swift measures can be employed to mitigate such scenarios.

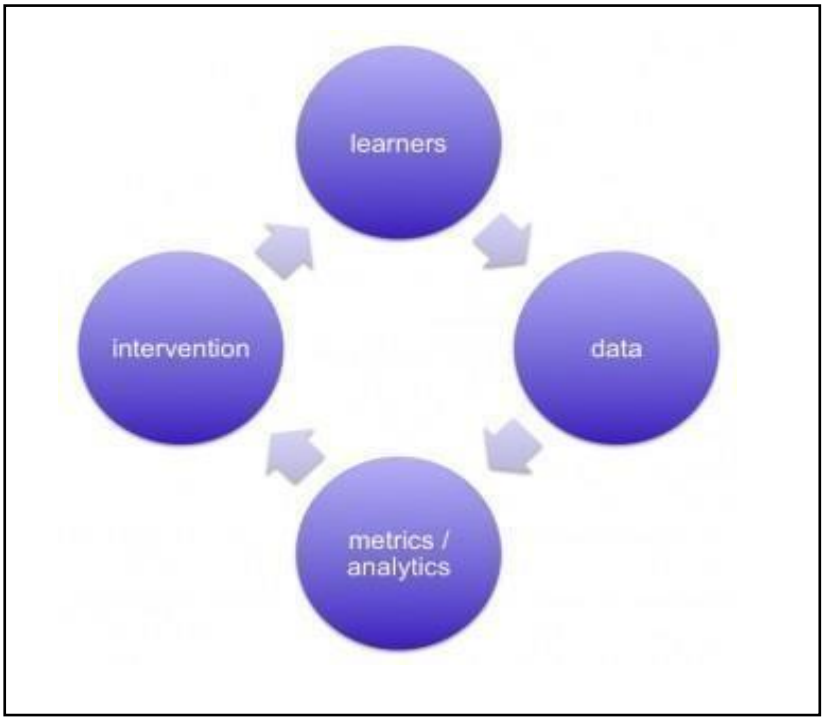


Figure 2.1 The learning analytics cycle, Clow (2012).

2.4.2 Learning Analytics Reference Model

Given the rising need for learning analytics in TEL, Chatti, Dyckhoff, Schroeder and Thüs(2012), proposed a four-dimension model that is a reference model for learning analytics. The dimensions are;

2.4.2.1 What data sources are utilized?

The focus here is on the data and environment, it helps us answer the question on TEL data sources. Educational data sources are categorized into two, that is the centralized systems and distributed learning environments. A good example of a centralized system is an LMS. Distributed data sources capture data beyond that on LMS; an example is from the concept of Personal Learning Environment (PLE) (Chatti et al., 2012).

2.4.2.2 Who benefits from learning analytics?

This dimension helps to answer the question, who are the stakeholders targeted by the analysis of the educational data? The stakeholders who might benefit from learning analytics practices include, students, teachers, administrators from an institution, system designers as well as

researchers. Each of the stakeholder would probably have different goal and expectation from application of learning analytics practices (Chatti et al., 2012)

2.4.2.3 Why is the analysis done?

Focuses on the objective of analyzing a given data. Chatti, Dyckhoff, Schroeder and Thüs(2012), gave a list of possible purposes of applying learning analytics that include, reflection, prediction, recommendation and customizing learning among others. They indicated the need to measure performance using useful metrics for each of the objectives. For this reason, they proposed that before beginning any learning analytics exercise then one must have a clear definition on their Objective/Indicator/Metric (OIM).

2.4.2.4 How is the analysis conducted?

This dimension answers the question on the techniques applied to analyze the collected data (Chatti et al., 2012). Some of the methods outlined include, visualization, Social Network Analysis (SNA), statistics and data mining. Visualizing educational data makes it easier for interpretation by the various stakeholders. According to Chatti, Dyckhoff, Schroeder & Thüs(2012), data mining involves pattern discovery by applying machine learning techniques. As for basic statistics, one can get them from an LMS reporting tool. To support networked learning, a learning analytics technique known as SNA is applied. It contains a graphical representation of nodes representing actors and edges to represent linkage between actors which is informative on the social network of actors. Generally, depending on the why?, then there are various methods to apply in order to design and develop useful tools to aid the stakeholders (Chatti et al.,2012).

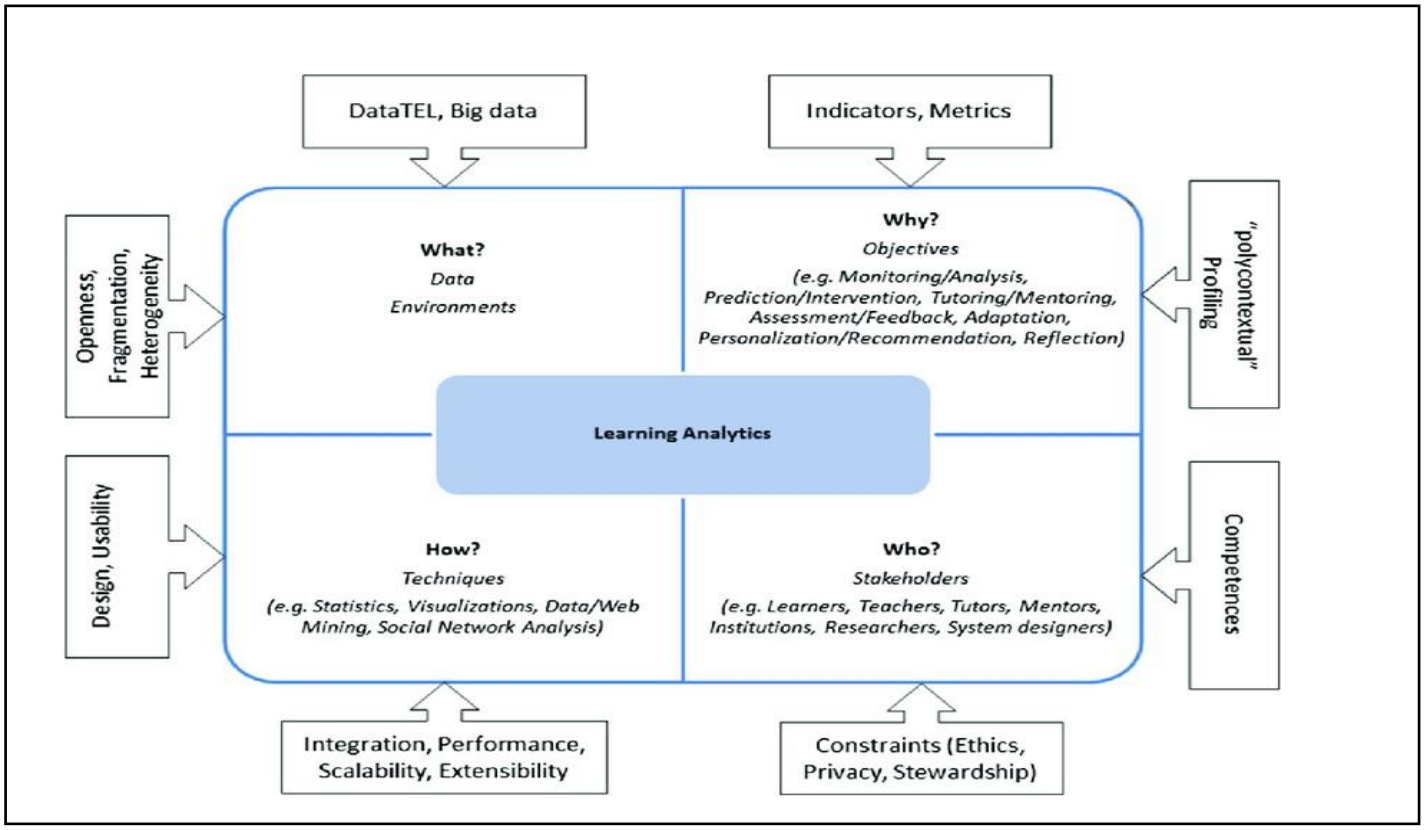


Figure 2.2 Learning Analytics Reference Model, (Chatti et al., 2012).

2.5 Empirical Review

2.5.1 Challenges facing virtual learners

Online learning has benefits attached to it but also, has various challenges that may affect students' performance. Virtual learners may experience lack of motivation, feel isolated due to lack of face-to-face interaction with teachers and peers, they may divert attention in the online space or even experience technical issues (Ferguson, 2012). Some learners are unable to keep up with virtual learning activities due to several factors including multi-tasking and or lack of internet access (Zulfikri et al., 2021). These challenges have seen to have created an impact in the attendance of the learners and institutions such as North Arizona have built a system to monitor the attendance of students and further give them alerts on grades, attendance among other feedback(Picciano,2012). Queensland University of Technology in Brisbane city also tracks behavioral aspects as a result of the challenges learners face to aid in the retention of students. Clow (2012) encourages collection of behavioral data such as login patterns and clicks made within VLEs/LMS and analyzing them to provide actionable insights for better decision making and data-driven interventions.

2.5.2 Specific use of Learning Analytics

According to the Society of Learning Analytics Research (SoLAR), there are various uses of learning analytics in prediction of students' academic success. Some common uses include detecting the learners who are at risk of failing a course or those at risk of churn. In this subsection, look into one of the methodologies that SoLAR indicates which is predictive analytics that involves the analysis of students' data to try and understand the future in terms of performance. Studies have been conducted in various institutions to elucidate the impact of predictive analytics towards the performance of students. In this subsection we will also look into the models applied by various researchers and highlight outcomes, strengths and weaknesses that will be identified.

2.5.2.1 Predictive Analytics using Students' Data

Predictive analytics is an important aspect in many institutions, it allows investing more resources to students at risk (Nghe et al.,2007). Nghe, Janecek and Haddawy (2007), conducted prediction of learners' performance in a comparative approach to try and reveal the accuracy between Decision Trees and Bayesian Network Algorithms. They used datasets from two different institutions, one, Can Tho University based in Vietnam and Asian Institute of Technology based in Thailand. Some of the attributes captured in these datasets include, gender, entry mark, age among others whose importance in predicting Student performance were gauged using Information Gain. In their study, Decision Trees were said to have done a better job as it was more accurate than Bayesian Network Algorithm. They employed the Cross-Validation technique with 10 folds to be able evaluate the accuracy of the prediction for each of the algorithms. The study mainly dwelt on accuracy as a performance evaluation metric, with overall accuracy of 86% for Can Tho University and 74% for Asian Institute of Technology for 3-class prediction say (Fail, Good, Very Good). It was highlighted that the predictions for student performance served different purposes in the institutions. At Can Tho University they applied this to identify students at risk and assist them whereas for Asian Institute of Technology they applied predictive analytics to identify students eligible for scholarships.

In an online setting, a study was conducted in Malaysia at Universiti Pendidikan Sultan Idris which captured students' data from the Faculty of Science and Mathematics (Zulfikri et al.,2021). They applied a linear regression model to predict students' results with the target variable being Grade

Point Average (GPA) and the predictor variable being Age. To evaluate the performance of their model, they leveraged R-squared and Root Mean Squared Error (RMSE). The findings indicated an improvement of GPA in upcoming semesters.

In another study, several classification algorithms were used to predict the course outcome of students taking a course, from data of the previous cohort. Linear Discriminant Analysis (LDA), K- Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian Naïve Bayes, Classification and Regression Trees (CART) as well as Logistic Regression were used (Gull et al., 2020). Historical data containing attributes of the students' assessments that is, Quiz1, Quiz2, Midterm, Project and Lab marks and the target variable as Grade were used. The performance metrics used for model selection were, Accuracy, F1Score, Kappa, Recall and Precision. In the comparative analysis of the performance of each model, LDA was the best where it predicted 49 out of a total of 59 records correctly (Gull et al., 2020).

In a study conducted by Hashim and fellow researchers in November 2020 on predicting student performance using supervised machine learning models, SVM, KNN, Decision Trees, Naïve Bayes, Logistic Regression, Neural Network and Sequential Minimum Optimization were used. They obtained data from the University of Basra and behavioral features, academic history as well as demographic features were applied in the prediction of final course outcome. The Logistic Regression outperformed the other models as a better classifier in predicting final grades (Hashim et al., 2020).

Ogwoka, Cheruiyot and Okeyo (2015), used Decision Trees and KNN to predict the performance of students. The undergraduate students' data was obtained from the learning management system of Technical University of Mombasa-Kenya. Some of the predictor variables used in this research were, gender, assessment marks and attendance. The model evaluation metrics applied were Precision, Recall and F-Measure as well as accuracy.

Mwalumbwe and Mtebe (2017), conducted a study to assess the relationship between LMS usage and the course outcome of students. They captured data from two courses done at Mbeya University of Science and Technology-Tanzania that is, Analysis of Applied Biology Course and Service and Installation course from a total of 171 students. They applied a linear regression model with LMS data as predictor variables and final grade as the target variable. Significance of

predictor variables was gauged using beta values, with general reports that peer interaction, exercises performed and number of forum posts had an impact on course outcome of the students. They reported that the limitation of their study was not qualitatively subjective as it failed to explain why some factors were significant and some were not. They also had issues with data reliability as the courses were blended and suggest that the use of courses that are fully virtual would give a better picture.

2.5.2.2 Early Warning Systems in Education

Akçapınar, Altun and Askar (2019), suggested that the next step in prediction of student academic outcome is to develop an Early Warning System (EWS). One of the most profound systems which was developed at Purdue University in the USA known as Course Signals is used to predict student performance. Arnold and Pistilli (2012), developed Course Signals with data from Purdue's Learning Management System (LMS), Blackboard Vista. They captured the demographic data, academic history, grades as well as the effort from students' interaction with Blackboard Vista. Struggling learners were identified as early as the second week of the semester (Sclater et al., 2016). The predicted outcome is sent to individual emails together with a color on a stoplight (Traffic lights). With this tool the university has seen better performance and overall student retention. It is utilized with more than 140 instructors and for more than 100 courses offered at Purdue University (Arnold & Pistilli, 2012).

Moodog system is another notable early alert tool that tracks the online activities of students at University of California (Fiaidhi, 2014). It allows instructors to view how the learners are interacting with course materials. Moodog sends alert mails to the students on the course materials they need to check and review. This system is also utilized as a course management tool at the University of Alabama (Bienkowski et al., 2012).

Grade Performance System (GPS), from the Northern Arizona University is to provide early feedback on the performance of students (Picciano, 2012). Some of the indicators used include the attendance and assessment marks (Bin Mat et al.,2013). According to Picciano, GPS also serves as a retention system. It provides alerts on grades, attendance among other feedback regarding the academic situation of a student.

Krumm, Waddington, Teasley, and Lonn (2014), designed Student Explorer, an EWS that aggregates data from LMS that provides the data to academic advisors in a program known as STEM (Science, Technology, Engineering and Mathematics) Academy. The STEM Academy is modelled on programs taken at University of Maryland and University of California. The students' data utilized are log-ins and grade data which are analyzed and used to come up with visualizations that display the learners' performance compared to the peers. A three-level classification colored scheme, that is, Engage (red), Explore(yellow) and Encourage(green) was developed to provide academic advisors with a complete picture of the students' data for provision of early intervention (Krumm et al.,2014).

2.7 Gaps in the Research

After a concise review of past studies several gaps were identified. All the studies reviewed in subsection 2.6.1 did not have a data product for their models. Further, an in-depth look at section 2.6.1 revealed several gaps from each study. Nghe, Janecek & Haddawy (2007), dwelt on accuracy as a performance evaluation metric for their models. As for the study conducted at University Pendidikan Sultan Idris, it only had one predictor variable, that is age. Ogwoka, Cheruiyot and Okeyo (2015), applied two classification algorithms which implied it was a comparative study that did not clearly specify the best model. Most of the researchers dwelt on few predictor variables for the prediction of students' performance. We also noted that there was no data product for most of the research that were conducted, which this study aims to address.

2.8 Operationalization of Research Objectives.

To operationalize the objectives of this based on the gaps identified the researcher intends to use The Open University open-source data as it is one of the universities that offers most of its courses online. The reason for utilizing this data is for its reliability as no informal record of data is done which may leave some data out including its open accessibility. I seek to conduct a correlational analysis to figure out which variables are useful in determining the performance of a student. This will aid in the second-tier of the study which is to perform classification to identify the likelihood of a student failing given the set of performance indicators identified from the correlation analysis. The indicators will also come in handy in creating a data product, which is a web-based early performance prediction tool that leverages on the best classification model selected from the

second part of my objective. Operationalizing the research objective will create an impact for students to improve their course outcome.

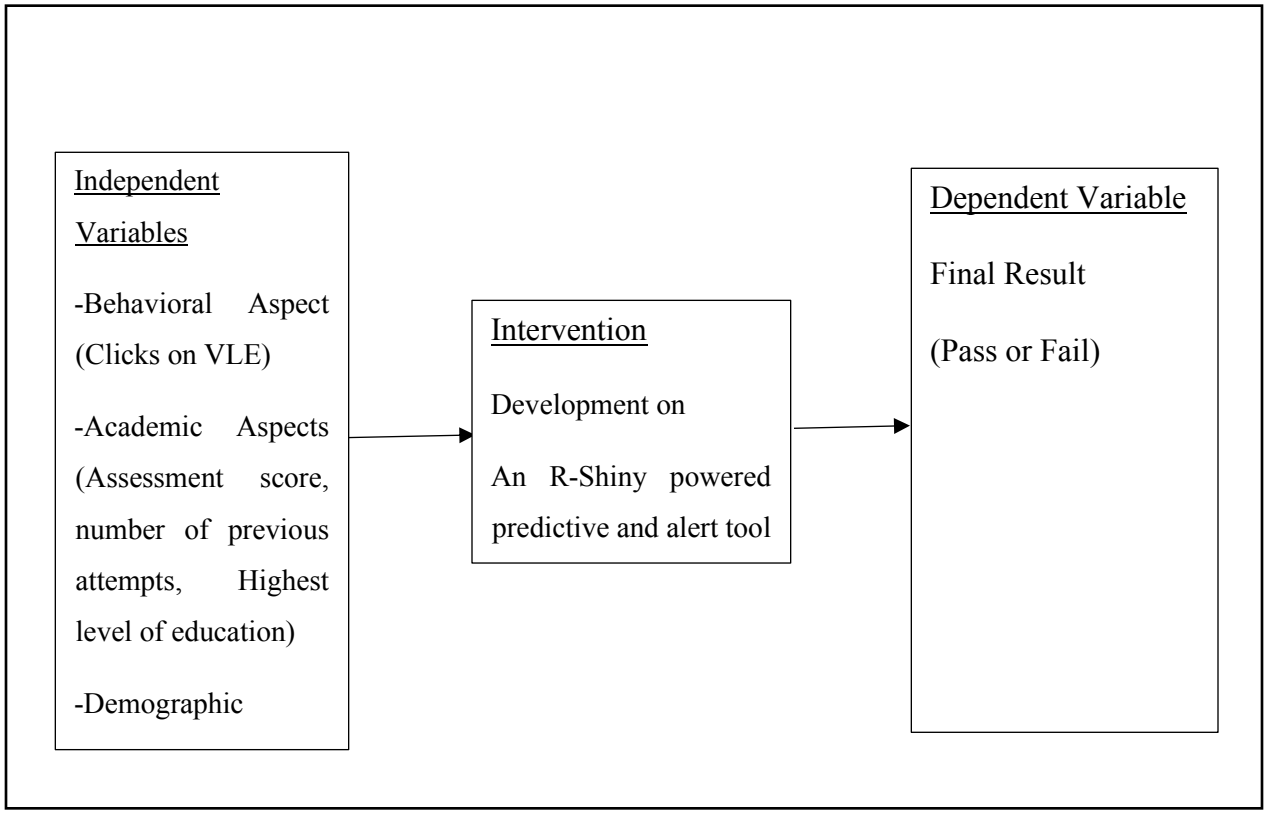
Table 1. Operationalization of Research Objectives.

Research Objective	Aspect	Measure	Test
RO1. To identify key indicators for predicting students' performance.	Key indicators for students' performance (eg.Assessments, student_info)	Student Likelihood to Pass or Fail (Gull et al., 2020)	Correlation analysis (R) [Spearman's]
RO2. To identify which algorithm performs best in predicting students' performance.	Algorithms	Algorithm aspects on Student Likelihood to Pass or Fail (Hashim et al., 2020)	Best Model from Evaluation of: Logistic Regression model Naïve Bayes model. KNN SVM
****RO2 CONTD' EVALUATION**	Model performance evaluation	Classification models evaluation (best of the four models mentioned above chosen based on these metrics) (Hashim et al., 2020)	Precision Recall F1 Score
RO3. To develop and deploy an Early Performance Prediction and Alert Tool.	R-shiny Functions for User-Interface and Server	Students' performance prediction and early alerts (Arnold & Pistilli, 2012)	Early warning input features (significant predictor variables)

Source: (Researcher, 2024)

2.9 Conceptual Framework

This study will draw upon the Learning Analytics Cycle (Clow,2012). The cycle for this study begins with the virtual learners identified as the OU's students. The university provides a schema of every detail contained in their students' data. The data drawn include their performance aspect, demographic aspect and lastly their behavioural aspect which are all indicators of students' course outcome which will be used in this study for prediction of final learners' grade. The performance aspect has the grades of assessments and presentations of both tutor marked assessments and computer marked assessments. The demographic aspect contains the data collected upon student registration such as their age. The behavioural aspect contains the data on each learners' interaction with learning materials provided in the OU's VLE. Given the gap identified in various studies applying a selected few variables, this study expounds on the variables that are utilized from the data. Another core part of Clow's (2012) cycle is the metrics in which this study takes up various measures to select significant input features for predictive analysis. Most of the research works done have indicated the use of several models to predict students' performance but did not explain the reason for applying the said models. SVM, Naive Bayes, Logistic Regression and KNN will be used. The study will take up various performance evaluation metrics to determine a reliable model for the predictive analytics and consequently the server functions of the web-based analytic application. The study also seeks to have an end-product in which several studies lack in the intervention phase. The intervention phase within this study is coming up with the alert tool that effectively predicts learners' outcome and where a student is predicted to fail an email is sent to them to seek assistance from their teachers. The application is to be accessed on the web, unlike most of the early warning systems which are plugged into institutions' LMSs.



Conceptual Framework (Researcher, 2024)

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter provides a focus on the research design, the target population, data collection, data analysis processes as well as the ethical considerations that were applied in this study. In light of the gaps identified in the previous chapter, this chapter seeks to clearly present the steps applicable to the study in order to bridge them.

3.2 Research Design

The study adopts a quantitative approach. Marczyk, DeMatteo and Festinger (2010), described quantitative research design as one that utilizes statistical analysis to obtain findings. Prediction is one of the main goals for this study. A prediction-based study stems from descriptive research which falls under the quantitative research design category. An example of descriptive research is correlational study. Correlational study is to be conducted to uncover the relationship between the dependent and independent variables for this research. From correlational research, a researcher is able to make predictions by applying the predictor variables which are strongly related to the response variable (Marczyk et al., 2010).

3.3 Data Collection and target population

This study utilized secondary data which is open-sourced for research purposes by The Open University, UK. It contains over 20,000 anonymized students' data captured on the institution's VLE. The university provided a schema to better understand the data captured. The data include academic history information, demographic information, registration information and behavioural aspects (clicks on educational materials provided on the VLE). Figure 3.1 is a schema of the dataset that sheds light into the description of the variables that are contained in the data.

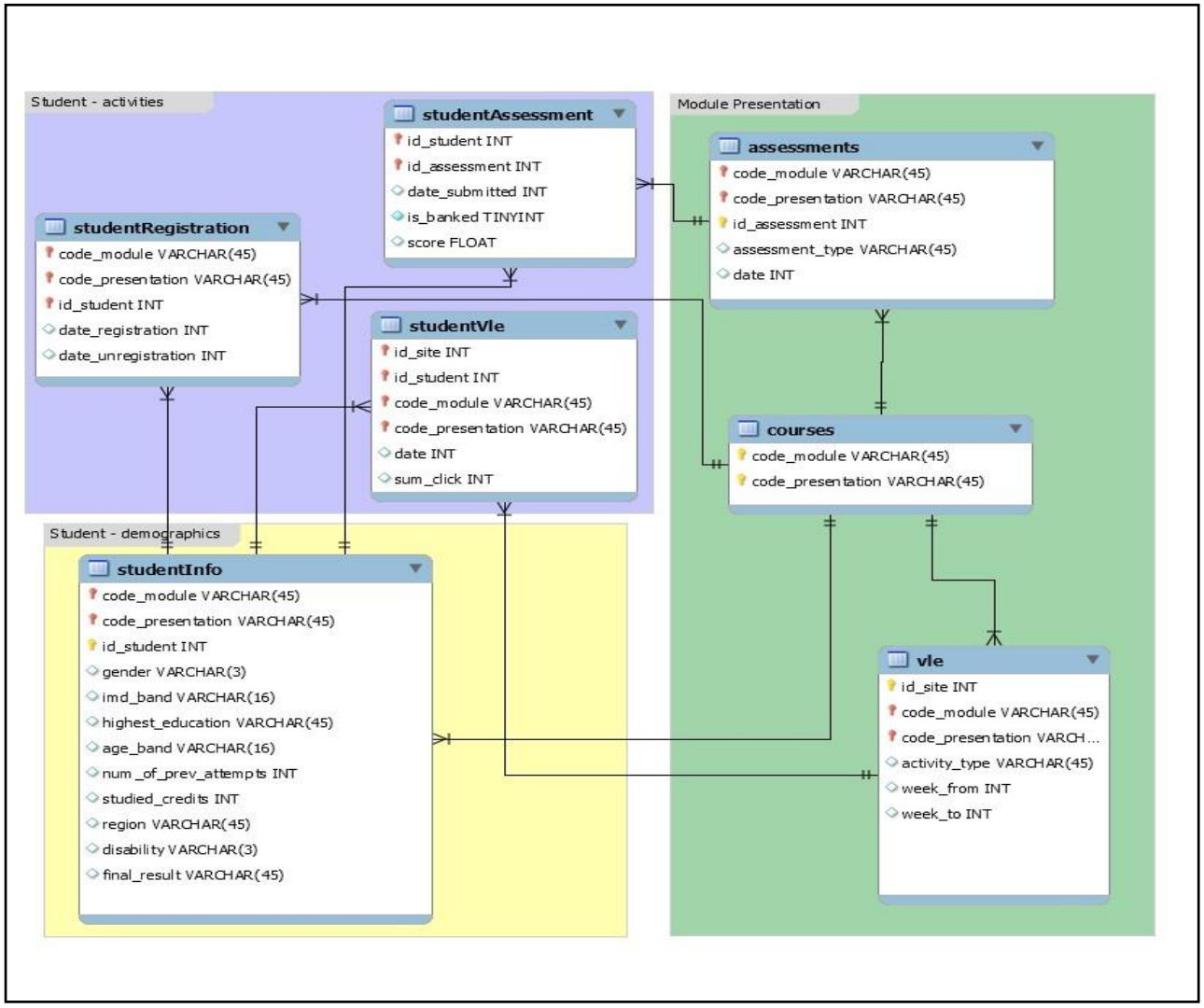


Figure 3.1 The Open University Data Description Schema.

3.4 Data Eye-balling and Manipulation

The journey towards the realization of the solution developed under this study began with data eye-balling and manipulation. This is a crucial step that sheds light to understanding the data in-depth. The datasets provided by the Open University as seen on the schema provided in Figure 3.1 include, “assessments” which contains information about the assessments given, “courses” which contains details on the courses offered, “vle” which contains information about the activities on the VLE, “studentAssessment” which contains the information of each student with regards to the assessments they took, “studentInfo” which contains the demographic information and final outcome of each student, “studentRegistration” which contains the dates of registration for each student, and finally “studentVle” which contains interactive information of each student within the institutions VLE.

From data eye-balling, this study revealed that the dataset containing student information i.e (studentAssessment, studentInfo, studentRegistration, studentVle) were beneficial to the realization of the tangible solution under this study as it focused on students. Each of the datasets were then manipulated individually and later merged using unique identifier i.e "id_student" to come up with a single dataset. The next section will further discuss the data manipulation processes that were undertaken towards the realization of a complete dataset applied in this study.

3.5 Data pre-processing

Data preparation is an important part of analysis since we draw conclusions from it. It involves having treatment options for missing values such as imputing with mean (Marczyk et al., 2010). Data pre-processing enhances the performance of models. For this study, putting various treatment options into consideration, we handled duplicates as well as missing values that might affect the overall representation of the population.

3.5.1 Treatment for missing values

Treatment of missing values was done independently for the following datasets (studentAssessment, studentRegistration). Looking keenly at the studentAssessment 19 students did not attempt to do any assessment, the records were removed since after a careful history look up on the students they finally withdrew from the course. Since the students who withdrew from the course are not part of this study's aim, we opted to remove them. From the data we clearly can equate withdrawal to failure because some students e.g., of id_student= 94961 validates the fact that a student can withdraw from a course and re-do it and actually Pass when they join the course again when it is next offered. As for students missing just a few assessments, the records were imputed with mean scores of their assessments.

As for studentRegistration dataset we noted that data_unregistration, which is a column holding information on the date that a student deregister or rather withdrew from a course, had over 60% of missing values. This was expected because most of the students did not deregister from their courses, we therefore removed this field as it was not significant to the study. As for the date_of registration column, the researcher imputed missing records with the mean since most learners are expected to register for a course around the same time.

3.5.2 Handling duplicates.

The data manipulation in this study also involved handling duplicates in the following datasets (studentInfo, studentAssessment, studentVLE and studentRegistration). We will concisely discuss on what attributed to the duplicates on each dataset and provide the means that were used to handle this.

On the studentInfo dataset we noted a few student records have been duplicated due to some of them withdrawing in some module presentations and coming in for the courses again when they are offered next case id_student=94961. This particular case on student number 94961 is a clear indication that withdrawal does not equate to failure. In this study we focus on the pass or fail labels to classify the performance of the students, therefore we opted to remove all the records in which the students withdrew from a course.

On the studentAssessment we noted having duplicates on the id_student column attributed to the fact that each student takes several assessments and takes. To handle this, we obtained the sum of all the assessments each student took and grouped by the id of the student. We did not use mean of all the assessments the student took because it is highly sensitive to outliers i.e(students who did assessment only once and had really high scores).

On the studentVLE dataset, we noted duplicates that were brought about due to the number of times each student interacts with the materials they access on the Virtual Learning Environment. This was handled by obtaining the total number of clicks and grouping by the id_student number.

On the studentRegistration dataset noted duplicates brought about by students who withdrew and re-registered for the course when it was offered next. The duplicates were removed by filtering out all students who de-registered as this research will only focus on those who completed a course and either passed or failed.

All the datasets were merged using unique identifier (id_student) and for consistency, next step taken was to convert the data types to numerical.

3.6 Feature Transformation

This section provides the practical solutions considered towards feature transformation. Feature Scaling and variable encoding will be discussed.

3.6.1 Variable Encoding

Various techniques were applied in encoding of the categorical features in the complete dataset. For low cardinality variables such as gender we performed one-hot encoding. For variables with high cardinality, we employed ordinal encoding for those that take a particular order i.e (highest_education) and target encoding for features i.e (region).

3.6.2 Feature Scaling

The complete dataset presented features that required scaling. Feature scaling enables us have features redefined to a common scale. To boost the performance of machine learning algorithms applied in this study we used standard scaling technique on R to adjust the range of values in the dataset.

3.6.3 Feature Selection

Feature selection was also carried out so as to address the issue on curse of dimensionality as well as to include statistically significant variables to the model. We applied the use of Elastic Net regularization which boasts of utilizing the full power of both L1(Lasso Regularization) andL2 (Ridge Regularization). This approach facilitated the identification of significant features while handling multicollinearity between them.

3.7 Exploratory data analysis

Exploratory Data Analysis (EDA) involves coming up with visualizations and summaries of sample distributions that allow us to better understand and make inferences about the entire population (Igual & Seguí, 2017). Generally, EDA is conducted to obtain an overall representation of the population from the analysis and examining the central measures of tendency of sample data. The study focused on both univariate and multivariate analysis of the data to check the distribution of data. This is where descriptive and correlational statistics were obtained to aid in the selection of significant variables within the academic history, behavioral and demographic aspects for the

predictive analysis. Descriptive statistics are meant to assist the researchers in data description and establish relationships between variables (Marczyk et al., 2010). We assessed the relationship between the various predictor variables (assessments, demographics and behavioural factors) with the dependent variable (overall outcome) using spearman's correlation coefficient and provided a heatmap depicting the same. The variables which are strongly correlated to performance of a student were utilized in the prediction algorithms and consequently in setting up the User Interface and the Server functions in R-Shiny app.

3.8 Machine learning algorithms

There are several machine learning techniques that can be used to perform prediction-based tasks. This study sequentially looks into four supervised machine learning techniques namely, Naive Bayes, Logistic Regression, KNN and SVM. These models were used to perform classification tasks to predict whether or not a student will fail or pass their course.

3.8.1 Logistic Regression

In Logistic Regression, predictions are made by calculating the chance of an event occurring from the linear combination of one or more explanatory variables (Sperandei, 2014). Typically, a logistic regression applied for this study is represented as;

$$\log\left[\frac{p}{1-p}\right] = \beta 0 + \beta 1X1 + \cdots + \beta nXn$$

where, p represents the probability, X_n represent the predictor variables/input features and β_n represent the coefficients associated to the impact brought about by each explanatory variable.

3.8.2 Naive Bayes.

Naive Bayes is a supervised classification technique in which the idea of Bayes Theorem is applied. It assumes the independence of variables (Mahesh, 2020). The equation used takes the form:

$$P(y|x) = \frac{p(x|y)}{p(x)} * P(y)$$

where;

$P(y/x)$ =posterior probability i.e probability of target given the predictor.

$P(x/y)$ = likelihood i.e the probability of the predictor given the target variable.

$P(y)$ = the prior probability of the predictor

$P(x)$ = the prior probability of the target.

3.8.3 K- Nearest Neighbours

K-Nearest Neighbours is also another powerful algorithm that can be used for classification tasks (Mahesh, 2020). Zhang (2016), tried to explain how the algorithm works by assigning variables to a given class based on two predictor variables. For this study the predictor variables applied were more than two. Generally, two concepts are of importance when applying KNN, that is, the Euclidean Distance and k. Euclidean distance is the distance between unlabelled data points to the labelled data points (Zhang, 2016). It is given by;

$$Distance(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

k is the number of neighbours a researcher decides to use. The goal is to choose the parameter such that we avoid overfitting or underfitting (Zhang, 2016).

3.8.4 Support Vector Machine

The main objective of SVM is creating a hyperplane with the biggest separation between support vectors in the data used in this study. According to Mahesh we aim at maximizing the distance between the different classes so as to minimize the classification error.

The optimal hyperplane is obtained by;

$y = w.x + b$ w represents a normal vector whereas b is some offset.

Figure 3.3 illustrates the concept of SVM as discussed
above.

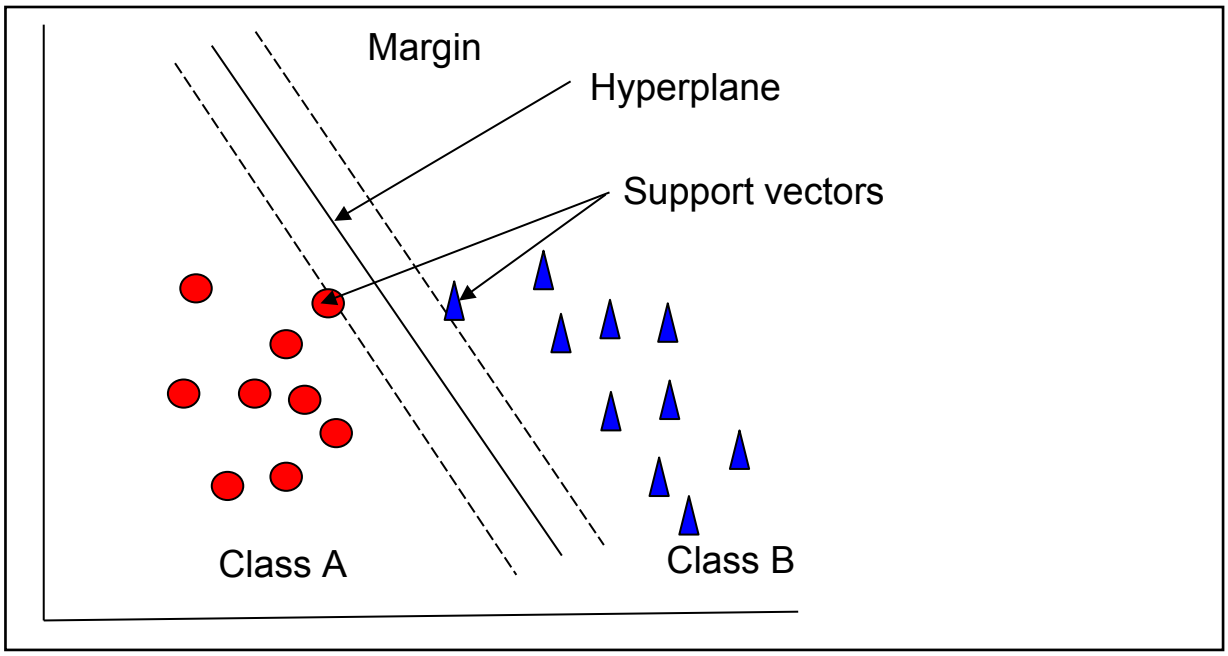


Figure 3.3 Illustration of the concept of Support Vector Machine.

3.9 Handling Class Imbalance

Prior to proceeding to train the machine learning algorithms discussed in the previous section, this study employed ROSE (Random Over-Sampling Examples) technique to generate synthetic examples for the under-represented class for our case, "Fail". This was done to improve the performance of the machine learning algorithms discussed in section 3.8.

3.10 Performance Evaluation

The models applied for this study were subjected to evaluation using the following metrics; F1 - score, Precision and Recall. The choice of the performance evaluation metrics was influenced by the nature of dataset at hand (i.e Imbalanced data) and the goal of optimization in academic domain.

Precision is one of the most applied metrics to evaluate the performance of classification algorithms. It measures the fraction of the positive cases that are correctly predicted from the total positive instances (Hossin & Sulaiman2015). Precision is important for this study because we aim at minimizing false positives i.e instances where students passing are incorrectly classified as failing. We consider precision to ensure that the model is not overly aggressive in classifying instance where a student is predicted to fail when they would have actually passed as this would lead to unnecessary resource allocation and interventions measures. The formula for precision is;

$$Precision = \frac{TP}{TP + FP}$$

Recall, also the true positive rate, is a measure of proportion of positive cases that were correctly classified (Hossin & Sulaiman2015). For this study, the positive cases here are the students at risk of failing. A high recall indicates that the model is effective at capturing the students at risk of failing, which is desirable for this study on optimization where we prioritize catching positive cases for interventions. Mathematically it is represented as;

$$Recall = \frac{TP}{TP + FN}$$

F measure combines both precision and recall, ideally getting the weighted harmonic mean of the two gives the F-measure (Hossin & Sulaiman2015). It combines the trade-offs of recall and precision. Tentatively is obtained by;

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

All the above metrics were used to evaluate all the four classification models and the best model for the classification identified was Logistic Regression with F1 Score of 95%.

3.11 Overall Testing and Diagnostic Approach

The complete dataset was partitioned into training set (70%) and testing set (30%) utilizing the caret package on R. The models were evaluated using the metrics discussed earlier in this section.

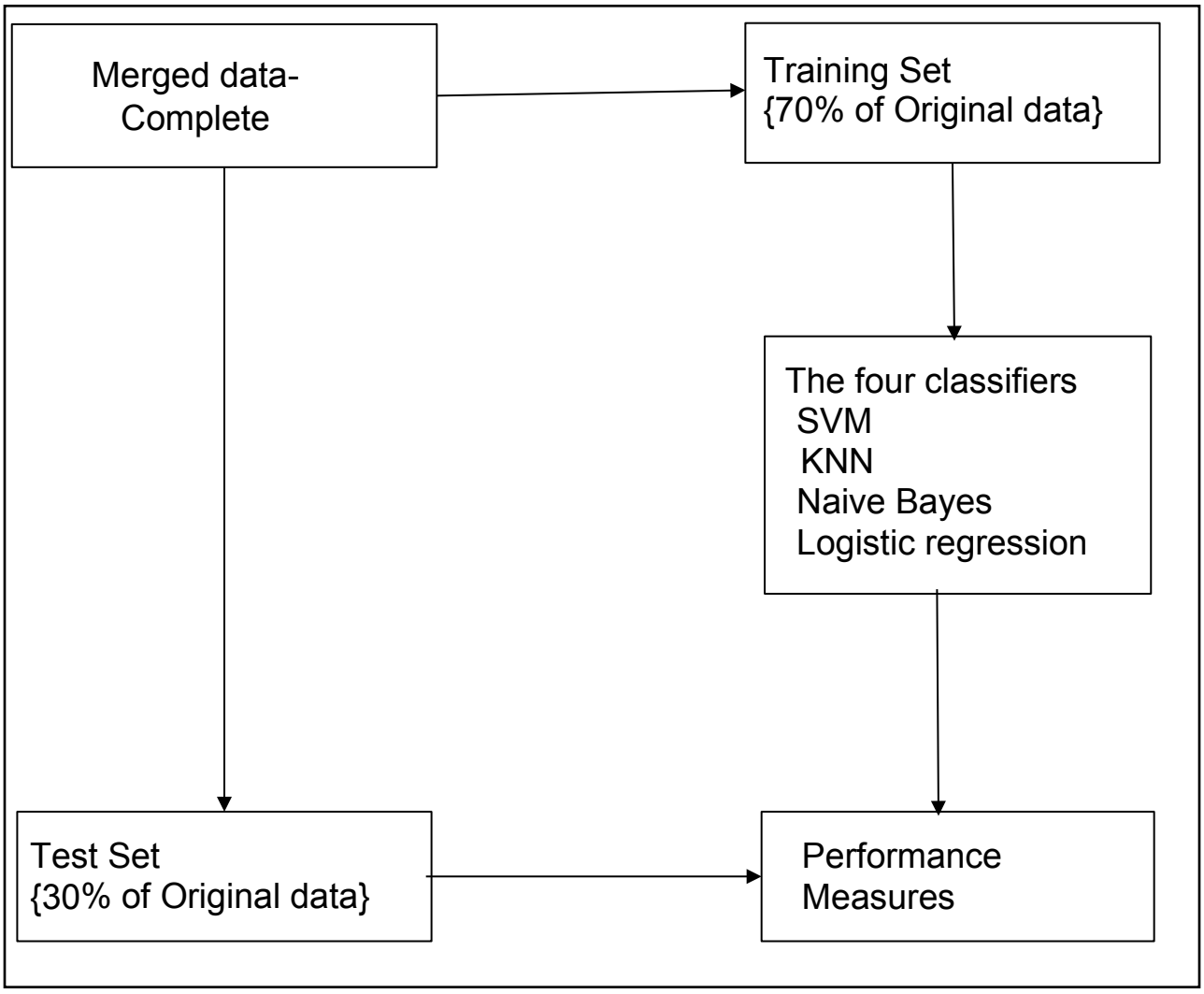


Figure 3.5 Illustration of testing and diagnostic approach (Researcher, 2024).

3.12 Model Deployment

The ultimate goal of the study is to develop a web-based analytic tool using R-Shiny. As documented by R, shiny is a package used to develop interactive web- applications. Applications created using the shiny package have two main parts, the User-Interface (UI) and the Server. The UI functions control what a user can see once they open the application. The server function presents the logic behind the web-based application. Having gone through all the stages of data

analysis and modelling, the best model identified was used to come up with the function code for the server, that is, a code to predict the learner's course outcome given the input features by the user. Once the prediction is made, an email is sent to the learner and gmailr package and email code function within the server are responsible for that.

3.13 Ethical Considerations

Confidentiality is one of the key ethical considerations that researchers are advised to observe (Marczyk et al.,2010). This study maintains confidentiality as it utilizes anonymized data. This helps in mitigating the risk of discomfort among students whose data were being used for research purposes. The data is certified by the Open Data Institute and is open-sourced for research purposes from the Open University site (https://analyse.kmi.open.ac.uk/open_dataset.). The Open University also obtained a CC-BY 4.0 license that authorized sharing its content to the public for research purposes. The institution only requires that researchers must cite the dataset in their work, and we have therefore acknowledged the use of their students' data for this research.

CHAPTER FOUR: SYSTEM DESIGN AND ARCHITECTURE

4.1 Introduction

This chapter provides focus on a detailed machine learning system design and architecture for the operationalization of the solution in this study. System requirements, system design and User Interface design will be discussed under this chapter.

4.2 System Requirements

The requirements of a machine learning system entail the description of the functionality of the system in achieving the goals under this study. In this section, we look into the functional and non-functional requirements that are embedded on the R-Shiny powered solution.

4.2.1 Functional Requirements

In line with the research objectives, the predictive tool requires several key functional specifications. Firstly, it must feature a user-friendly R-Shiny Interface, allowing virtual learners to effortlessly interact with and navigate through the application, facilitating the input of relevant data for analysis. Secondly, real-time data extraction and analysis capabilities are essential, ensuring prompt extraction of learners' input and subsequent analysis using R to generate timely insights. Lastly, the tool must possess alert generation functionality, enabling the execution of predict prompt within the application and the delivery of analysis feedback to learners via email. These functionalities collectively empower learners to engage effectively with the tool, facilitating the implementation of proactive support and intervention strategies to enhance academic performance.

4.2.2 Non-Functional Requirements

In line with the research objectives, the non-functional requirements are crucial considerations for the effectiveness of the predictive tool. Firstly, accessibility is paramount, learners need a stable internet connection to access the web application seamlessly. Secondly, usability is key, necessitating clear and comprehensive instructions for learners to effectively utilize the application. Thirdly, reliability is essential, the application expected to consistently deliver timely insights

throughout its processes. Lastly, security is imperative, requiring the application to robustly safeguard sensitive information pertaining to the virtual learners. Ensuring these non-functional requirements are met is essential for fostering trust, usability, and effectiveness in the tool's implementation and usage within academic settings.

4.3 System Design

In context of this study, the system design refers to the web-app, data and the learning analytics practices providing the ecosystem for the implementation of an R-Shiny powered solution. In this section, we provide a high-level blue-print for the alert tool and the concepts of architectural components and interaction between components will be discussed.

4.3.1 Architectural Components

The primary system components for the R-Shiny powered solution include, the front-end (User-Interface), database and the back-end server. The UI component is developed by R-Shiny with necessary packages i.e shinyjs that allow developers perform Java Script operations to allow seamless user experience. The database component entails the historical records of learners as well as real-time extracted data from the user input in the front-end. The back-end server component entails the function codes that are responsible for executing main goal of the application which is predictive analytics on the dataset and providing timely insights via email. The back-end contains the logic behind the realization of the tangible solution under this study.

4.3.3 Interaction Between System Components

R-Shiny packages and codes are applied to create a user interface that allows ease of navigation through the tabs of the web-app. The user can easily input the required fields and prompt the application to predict their results on the predict button. The app ensures that all inputs are available as the predict button is enabled upon filling of all fields. New data (user input) is sent to the backend. The back-end server runs through functions that will perform data predictive modeling and provide timely insights to the users via email.

The figure below depicts the architectural design of the system and the interaction between components of the application under this study.

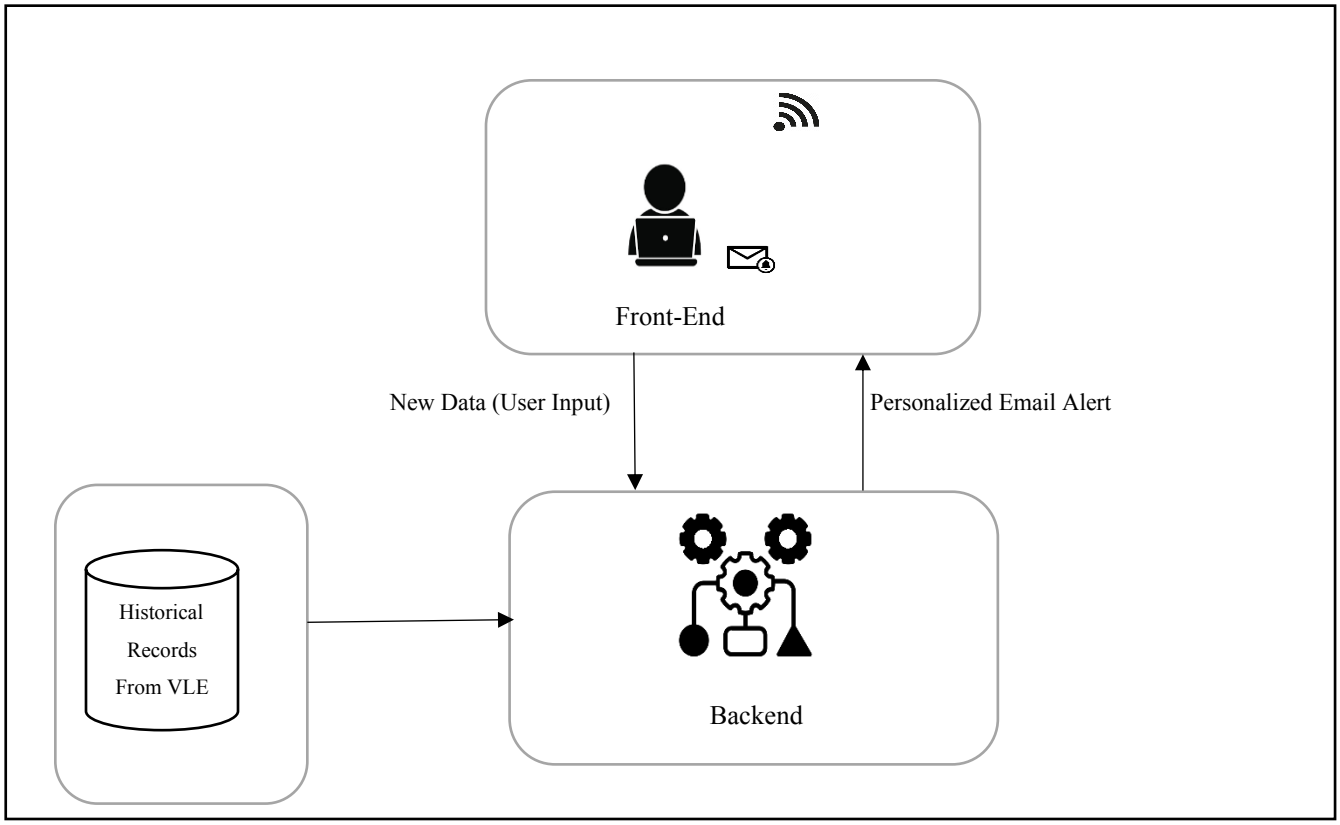


Figure 4.1 System Architecture

4.3.4 Data Flow

With the system design discussed in the previous section, the data flow between the components begins with student input from the R-Shiny interface on the front end of the system. The students' input is then securely transmitted to the backend where it is analyzed. In real-time the data is analyzed using comprehensive algorithms and feedback sent via the email provided by the student in the initial stage on the R-Shiny interface.

The data flow between various components in the system under this study is shown in the figure 4.2

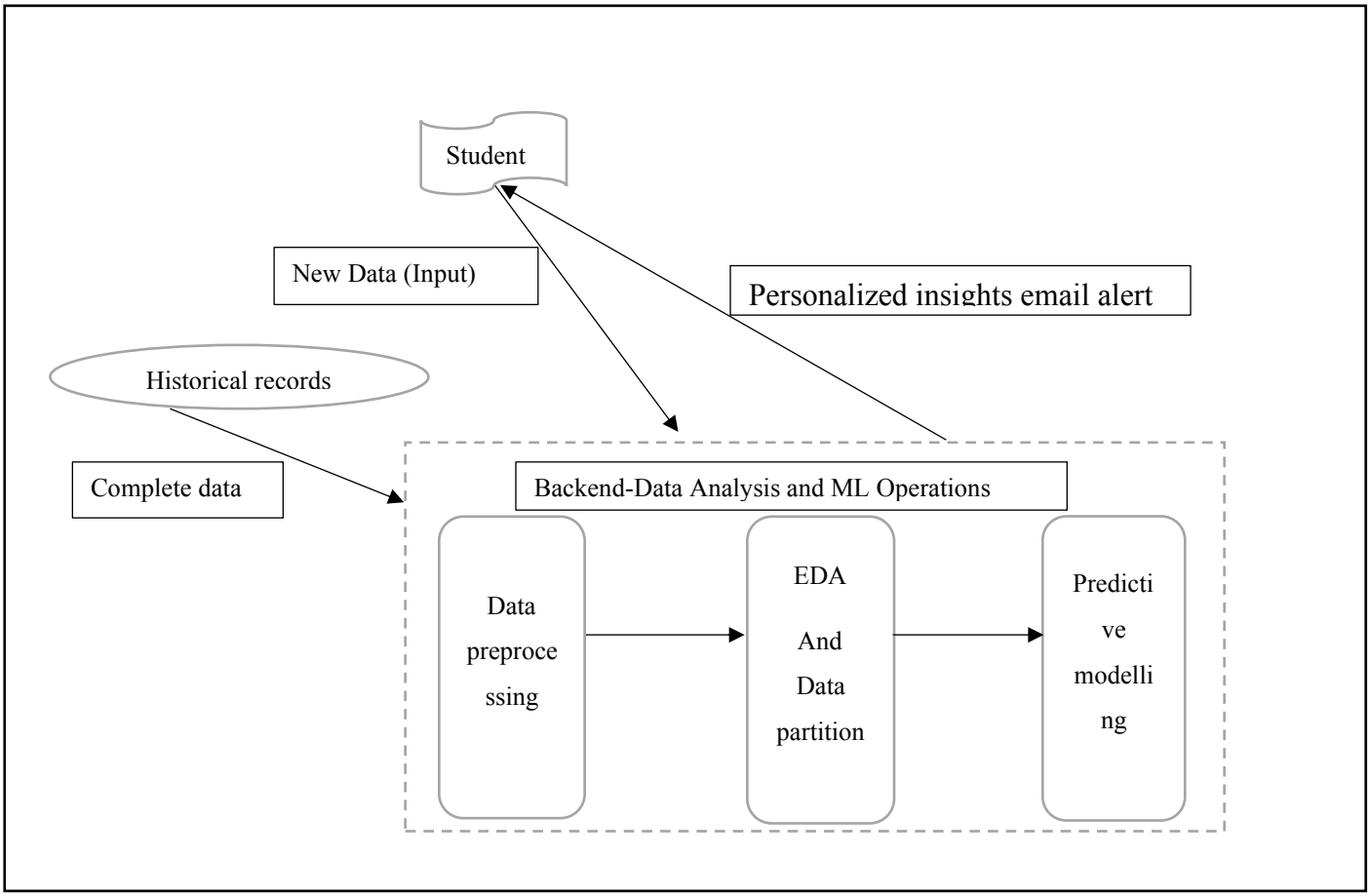


Figure 4.2 Data Flow

4.4 User Interface Design

The UI design is key for the implementation of the solution under this study since the new users' inputs play a fundamental role to the realization of the solution. The tabs in the UI serve a distinct purpose to improve user experience and give required information. Clear instructions on the usage of the app are provided on the 'About' section and further assistance can be provided since the 'Contact' section provides contact for the service providers of the solution. The UI is also designed to obtain all the required fields from the users since this is a mandatory stage that allow the execution of the predict prompt for timely feedback. With shinyjs, the app utilizes Java Script operations to enable predict button when all fields are filled and disable when the user has not provided any entry. This is to ensure we capture all the necessary details for seamless user experience of this solution.

CHAPTER FIVE: RESULTS

5.1 Introduction

This chapter presents the results obtained from the learning analytic practices applied towards the realization of the solution in this study. This chapter focuses on the results attained in light of the first two objectives in section 1.4.2 that were used in the implementation of the tangible solution in for this study. The objectives are, to identify key indicators for predicting students' performance and to examine which algorithm performs best in predicting students' performance. Emphasis is placed on exploratory data analysis, model performance evaluation and interpretation of the results.

5.2 Exploratory Data Analysis

In this section, the characteristics of the dataset utilized in this study will be concisely discussed. Focus on the univariate analysis and the multivariate analysis will be given to have a comprehensive view of the dataset.

5.2.1 Univariate Analysis

The data consists of 21216 rows, with the distribution of the target variable ('final_result') as 15381 of 'Pass' and 5835 of 'Fail'. From the distribution this is clearly a case of an imbalanced class dataset. Figure 5.1 gives a visual representation the distribution of the dataset.

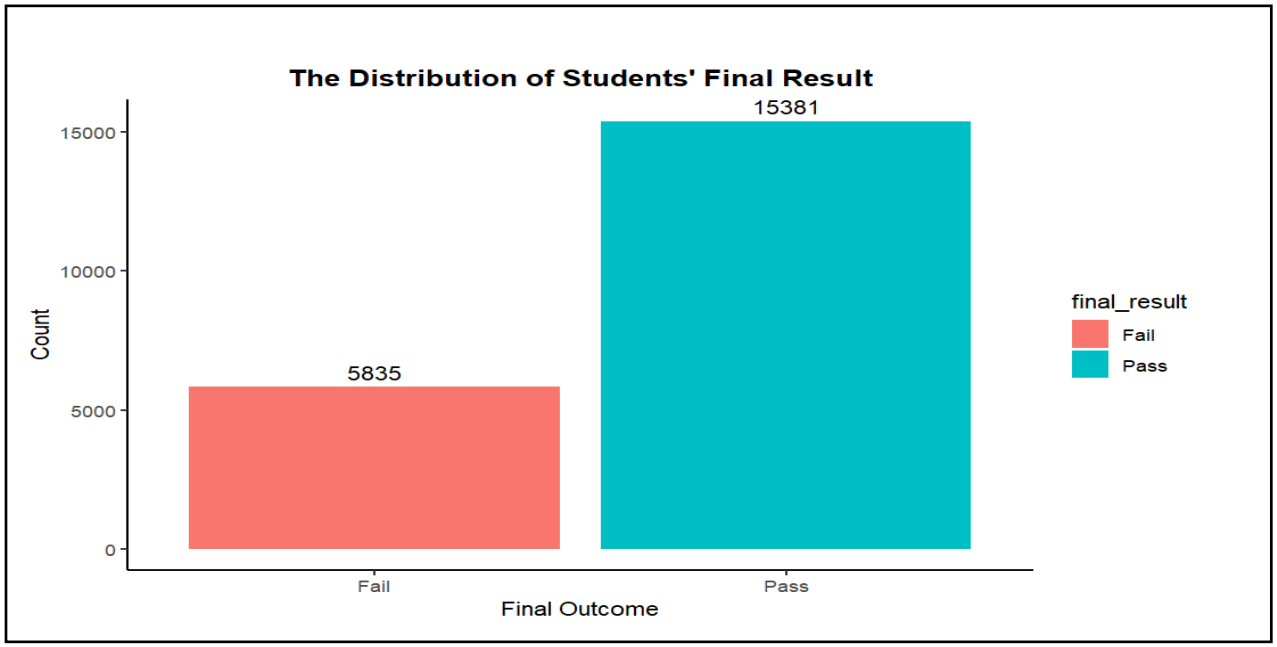


Figure 5.1 The distribution of students' final result

5.2.2 Multivariate Analysis

This study applied correlational analysis for the realization of the first objective of this study that is, to identify key indicators for predicting students' performance. The exploration in this section is aimed at examining relationship between various variables and the target variable (students' performance). The heatmap presented in figure 6.2 illustrates the relationship between variables and in focus to that of the target variable

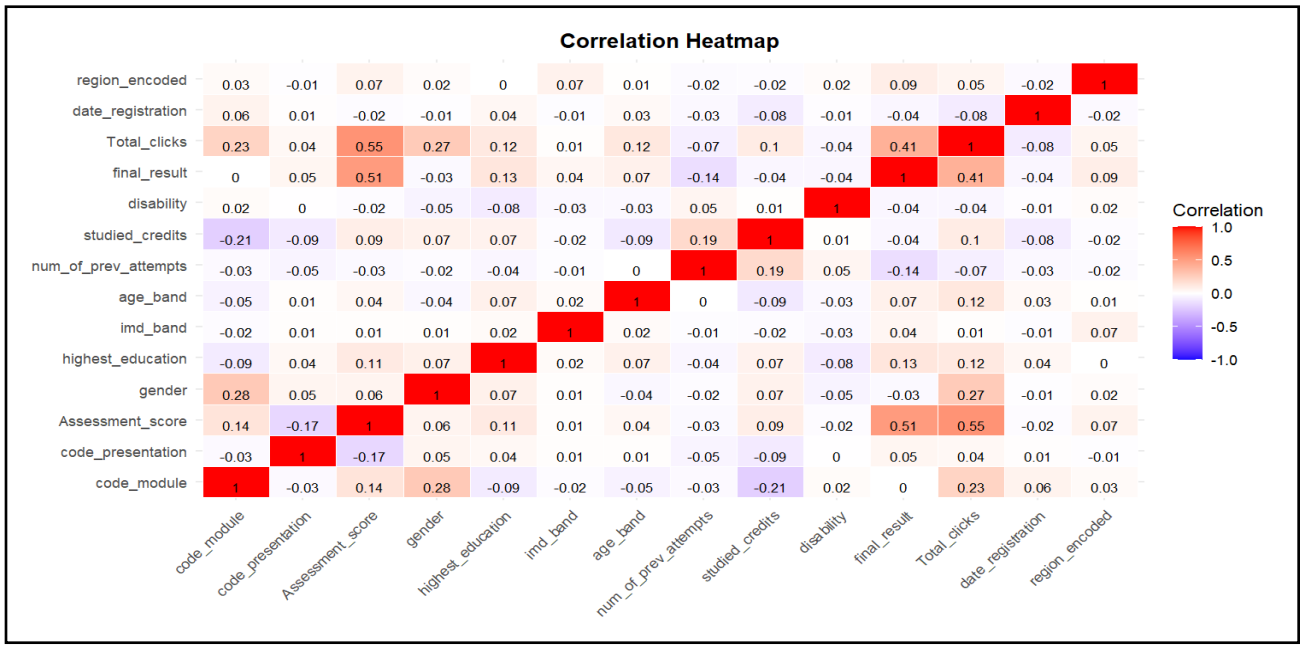


Figure 5.2 Correlation of variables within the dataset

In line with the operationalization of first objective of this study, a moderate positive correlation is recorded between the students' assessment score and the final_result (0.51) as well as total clicks within the VLE and the student performance (final_result) (0.41). This study therefore considers the assessment score as well as the behavioral aspect of accessing VLE as indicators that provide an input towards student performance. These features are also correlated and this meant that in our model one feature was utilized to avoid redundancy as they both strongly capture the impact towards learners' performance. Including both would be unnecessary.

5.3 Class Imbalance

Prior to modeling, the training data was balanced using ROSE technique to enhance performance of the classification models applied in this study. Figure 5.3 shows the illustration of the class distribution before and after handling class imbalance.

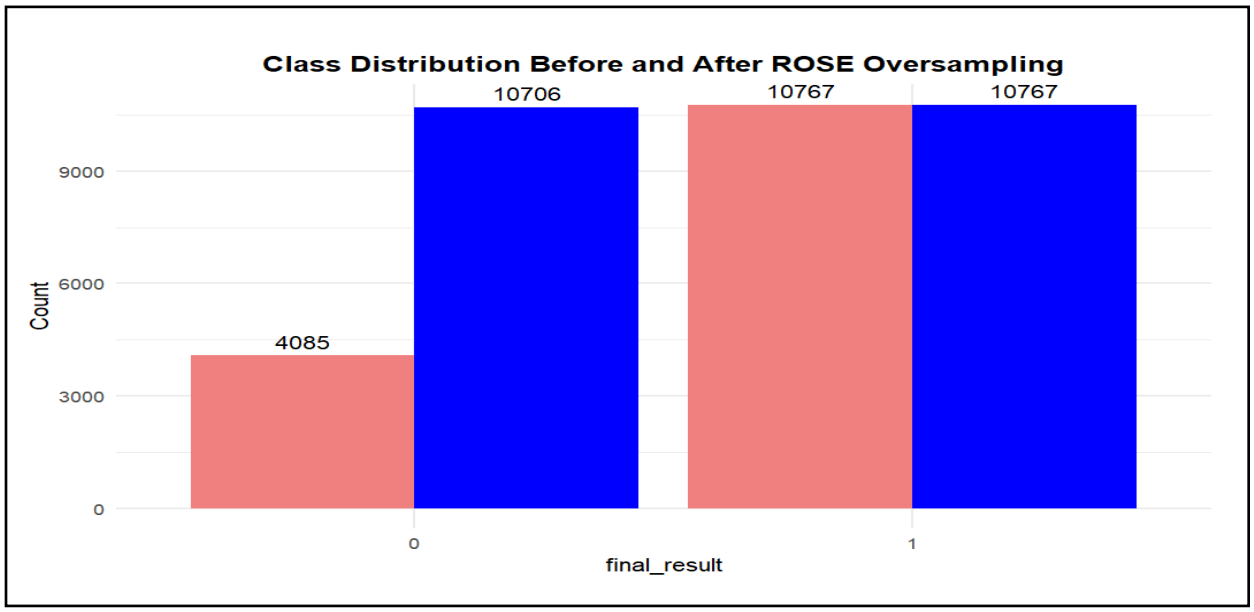


Figure 5.3 Illustration of class distribution before and after ROSE Oversampling

5.4 Model Performance Evaluation

The performance of the models used under this study were evaluated and the results are presented in this section. Considering the nature of the dataset utilized in this study (imbalanced), to examine which algorithm performed best in predicting students' final outcome the metrics used were Precision, Recall and F1 Score. The metrics applied are less sensitive to class imbalances. A graphical representation comparing the performance of the classification models used in this study is shown in figure 5.4.

Figure 5.4 Comparison of the model performance

In line with the operationalization of the second objective of this study, i.e to examine which algorithm performs best in predicting students' performance, logistic regression performs best. It records the highest precision, F1 Score and recall compared to the other models. A precision of 90% implies that the model correctly identifies students at risk of failing significant times. In this study, missing to identify students at risk of failing is costly and therefore with high recall, the model is reliable. With an F1 score of 95%, logistic regression is significantly reliable with reduced cost of misclassification to identify students at risk of failing which is important since the solution developed under this study is meant to optimize learners' performance. Therefore, there will be proper intervention and resource allocation towards learners who are actually at risk of failing. The SVM model has relative precision, recall and F1score. Naïve Bayes and KNN have recall of 73% and 71% respectively, they are able to catch positive instances i.e correctly identifying learners at risk of failing a good number of times however, the low precision, makes it moderately fit in as far as optimization purposes are concerned.

CHAPTER SIX: SYSTEM IMPLEMENTATION AND TESTING

6.1 Introduction

In this chapter, we cover the implementation and testing of the R-Shiny powered solution developed under this study. It focuses on applying the results achieved based on the first two objectives in section 1.4.2 discussed in Chapter 5 of this study for the realization of the solution. This chapter is embedded on the last objective of this study which is, to develop and deploy a Performance Prediction and Early Alert Tool using R-Shiny. The R-Shiny implementation and functionality testing of the system will be discussed.

6.2 R-Shiny Implementation

In this section, we provide a key focus to the concepts behind the functionality of the various components discussed in the previous chapter. An emphasis on the user interface implementation and the server function implementation is placed under this section. The figures associated with R-shiny implementation are detailed in the Appendix of this study.

6.2.1 User Interface Implementation

The user interface in this study which was developed using R-Shiny is accessible on web. The UI is designed to provide users with clear instructions (considering the non-tech users) and access to additional information and support. It displays various tabs that allow ease of navigation and enhance user experience. The UI has elements such as the tabs and input fields that enhance user interaction through the application. A code snippet of the Shiny interface implementation that provides a glimpse on the front-end operation is shown in figure A.1 in the Appendix of this study.

6.2.2 Server Function Implementation

The server function implementation entails R function codes for backend operations which includes real-time data extraction, data analysis and machine learning. The server function presents the logic that observes user input, enables or disables the predict button, makes predictions using a pre-trained model and sends prediction results to the provided email. The app ensures that all required fields are filled before enabling the predict button. A code snippet that illustrates the backend operation's implementation to complement the front-end operations is shown in the figure A.2 in the Appendix of this study.

6.2.3 Web Interface Overview

To highlight the key features and aspects that support the functionality of the solution developed under this study, this sub-section provides visual representation of the web interface. The demonstration clearly elucidates the realization of implementing an interactive user interface discussed in section 6.2.1. Figure 6.1 shows an overview of the web interface of the alert tool developed under this study.

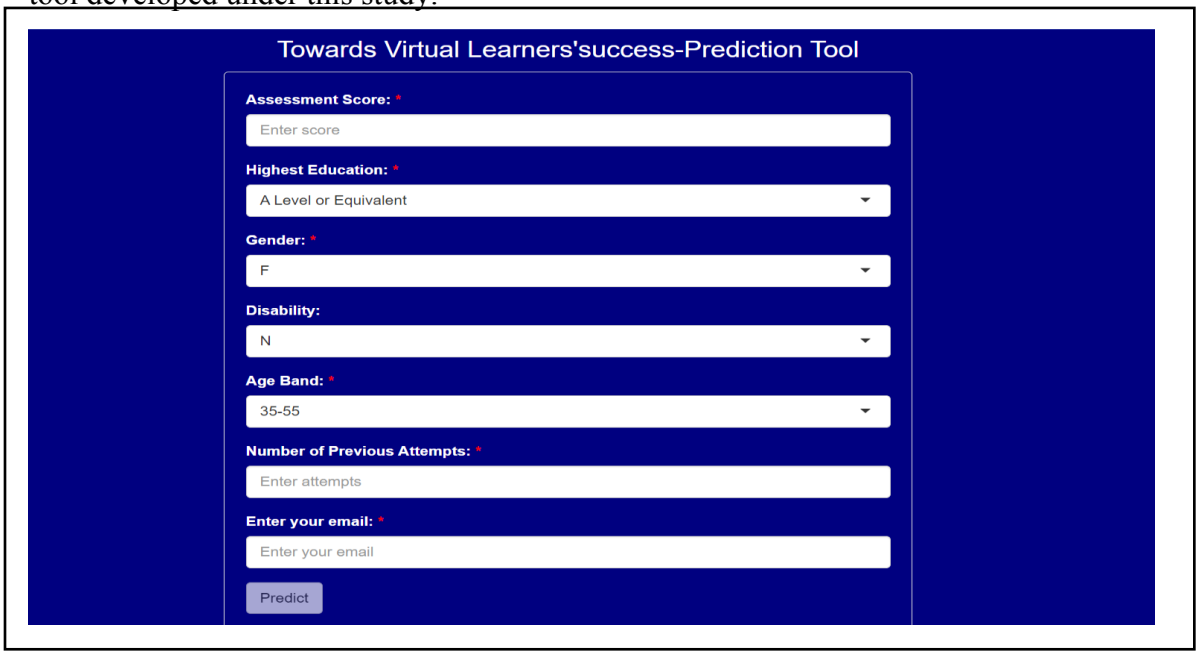


Figure 6.1 Web Interface overview

6.2.4 Alert Generation Mechanism

In context of this study, the system was designed to give automated response to the users (learners). The UI has a mandatory field input for the user's email through which the insights are sent. Through the seamless integration between the front-end and back-end components the alert generation mechanism ensures that the main objective of this study is met through timely and relevant alerts. A code snippet to the actualization of alert generation is provide in Figure A.3 in the appendix of this study.

The actualization of the backend processes of the system applied in this study and the general objective of this study under section 1.4.1 is seen in the email alert on the figure 6.2. This marks a pivotal point as we uncover the realization of the R shiny alert tool that predicts learners' academic

outcome. We utilized a record (id_student = 101279), for information added on the input in the user interface to give the prediction alert below which actually reflects the final outcome of this particular student from the dataset used in this study.

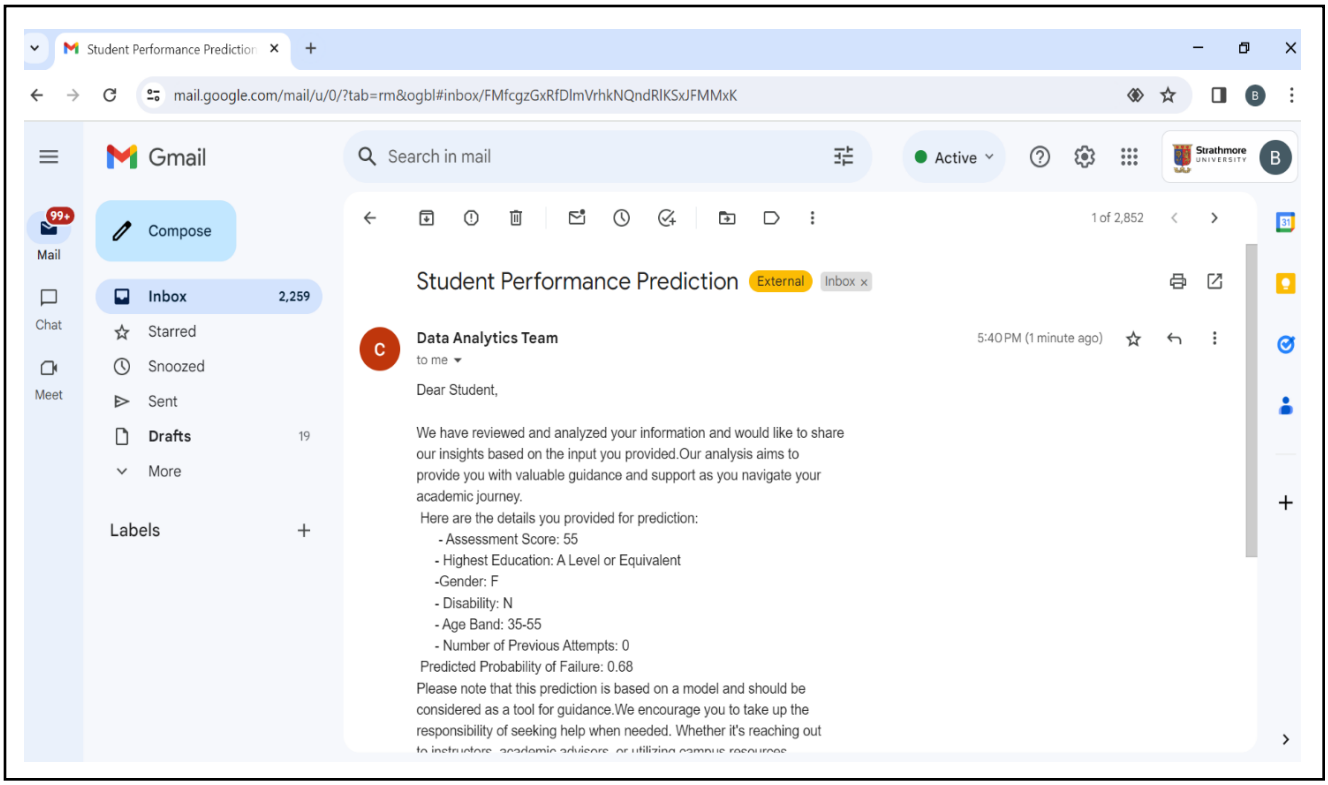


Figure 6.2 Generated Email Alert

6.3 System Functionality Testing

The functionality of the various components of the system under this study was assessed to ensure its effectiveness and reliability. Testing various elements including, system responsiveness and the integration between the front-end and back-end components is crucial. The system is fast and is able to generate feedback instantly once a user prompts the predict button. The server execution time is 0.00657, it took less than a hundredth of a second for the server to complete the extraction of user input and analyze the data for actionable insights. Essentially, this output tells how long the server spent processing the input and generating predictions. The email alerts are almost instant once the user presses the predict button. Figure A.4 shows the system responsive in as far as user interaction and reliability is concerned.

CHAPTER SEVEN: SUMMARY

7.1 Introduction

This chapter provides focus into the summary of the findings drawn from the learning analytics practices employed into the development of an R-Shiny tool designed to predict and alert learners about their final course outcome. The summary will entail the findings in as far as the objectives of this study is concerned and their relation to the related works discussed earlier in chapter 2. This chapter also emphasizes on the implication of these findings to various stakeholders and provide avenues for future work.

7.2 Summary

The goal of this study was to leverage learning analytics practices to develop an R-Shiny powered solution that predicts and provides personalized alert on virtual learners' academic outcome. The study was steered by three major objectives namely; to identify key indicators for predicting students' performance, to examine which algorithm performs best in predicting students' performance, to develop and deploy a Performance Prediction and Early Alert Tool using R-Shiny. The Open University UK's dataset was used for realization of the solution developed under this study. R software was used to perform data manipulation, data preprocessing, exploratory data analysis, modeling, development and deployment of the solution under this study. In this section, the summary of findings on each objective in this study will be comprehensively discussed.

7.2.1 Key factors influencing students' performance.

To identify key factors influencing students' performance this study took a correlational and feature importance analysis concurrently. This study revealed that age, assessment score, disability, gender, total clicks and number of previous attempts as factors contributing to the final outcome of a learner. Through the correlational analysis in section 6.2.2 we were able to rule out redundancy in some features and this study prioritized certain features in the models to curb multicollinearity. These findings align with the research conducted by Akçapınar, Altun and Askar (2019), that employed the use of some of these indicators (demographic data, assessments and interaction with VLE) to develop one of the most profound systems at Purdue University known as Course Signals used to predict learners' outcome. These findings are also supported by the research conducted by

Nghe, Janecek and Haddawy (2007), who observed that gender, age as well as entry mark were beneficial at predicting learners' academic outcome.

7.2.2 Model Selection

To examine which algorithm performs best in predicting students' performance, this study employed the use of performance supervised classification algorithms namely, Logistic regression, Support Vector Machine, K-Nearest Neighbors and Naïve Bayes. Each of these algorithms were evaluated using performance metrics namely, Recall, Precision and F1 Score. This study found that Logistic regression outperformed the other algorithms in predicting learners' performance. This finding is validated by the study conducted by Hashim and fellow researchers comparing the best performing models in predicting learners' performance in which the Logistic regression outperformed the other models.

7.2.3 Development and deployment R-Shiny tool

To develop and deploy a Performance Prediction and Early Alert Tool using R-Shiny, this study relied on versatile R framework to the realization of the solution. The packages used for the actualization of the tool include, shiny (for shiny framework), shinyjs (for integrating JavaScript with shiny), shinythemes (for themes on the UI), gmailr (for backend operation of sending mails) , dplyr (for data manipulation within the server) and glmnet (for backend operation of the Logistic regression- predictive supervised machine learning model). R function codes were used to develop the front-end and back-end and deployed on web hosted by shiny.

7.3 Implications of the findings

In this section, we provide a focus on actionable insights to all the stakeholders within the realms of the learning analytics space in support of the significance of the study discussed earlier in Section 1.6 of this paper.

7.3.1 School Administrators and Policy Makers

The study revealed valuable insights on the key factors influencing virtual learners' performance. Policy makers and school administrators can use the information to design policies and targeted interventions that address the specific needs identified such the academic history of a student (number of previous attempts). The solution under this study can be integrated into existing support

structures to provide personalized alert to enhance early intervention strategies and overall learners' success.

7.3.2 Learners

This study uncovered the factors shaping the overall outcome of the students, this empowers them to put in place strategies that boost areas of interest such as the assessment. The findings of this study enable the learners to comprehend on the aspects that influence their academic journey. The solution developed under this study, puts actionable insights directly into the hands of learners through the personalized alerts through their email. The R-Shiny solution serves as helpful resource for the learners to stay informed and take proactive steps towards their academic achievement. The tool developed under this study fosters a sense of responsibility to students allowing them to seek support when needed and proactively address the areas of concern.

7.3.3 Tutors

This study provides the teachers with invaluable insights that can be tailored to support learners in their academic outcome. They are able to focus on specific areas of need and provide interventions that will assist students in their academic journey to enhance their performance. The alert tool allows teachers provide personalized interventions to students who seek support thus enhancing learners' academic success.

7.3.4 Data Solution Providers

This study not only provides a tangible solution of students' performance optimization but also contributes to the growing body of knowledge on learning analytic practices providing solutions in the education sector. Data solution experts can rely on every input given into the operationalization of the R-Shiny powered solution designed to optimize the performance of virtual learners. The findings of this study contribute to the ongoing studies on the learning analytic practices applied in enhancing and optimizing learners' academic outcome. This study showcases the immense contribution of leveraging learning analytic practices to provide data-driven interventions in the education sector.

7.4 Institutional Adoption

The institutional adoption of the solution under this study represents a transformative step to foster personalized learning and enhance academic success of its learners. For adoption, alignment with institutional goals, robust security measures (to ensure students' data privacy) and seamless integration with existing systems are some of the key factors to be put into consideration. The benefits of the tool developed under this study include data-driven interventions on individual student needs to enhance their performance and ultimately enhance institutional reputation.

7.5 Further Studies.

This study focused on developing an R shiny solution designed to predict and alert virtual learners on their academic outcome. An area for research is exploring integration strategies for this solution at an institutional level with existing educational frameworks.

The variables applied for in this study are limited to the data provided by The Open University, in order to contribute to the generalizability and robustness of the machine learning algorithms used, a promising avenue for future research could be to expand the scope by incorporating additional variables from diverse educational institutions.

REFERENCES

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*(1), 1-20

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270).

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. *Office of Educational Technology, US Department of Education*.

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013, December). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *2013 IEEE 5th conference on engineering education (ICEED)* (pp. 126-130). IEEE.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683695.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International journal of Technology Enhanced learning*, *4*(5-6), 318-331.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304-317

Fiaidhi, J. (2014). The next step for learning analytics. *It Professional*, *16*(5), 4-8.

Gull, H., Saqib, M., Iqbal, S. Z., & Saeed, S. (2020, November). Improving learning experience of students by early prediction of student performance using machine learning. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-4). IEEE.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

https://shiny.rstudio.com/

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing.

Igual, L., & Seguí, S. (2017). Introduction to data science. In *Introduction to data science* (pp. 14). Springer, Cham.

Kirkwood, Adrian, and Linda Price. "Technology-enhanced learning and teaching in higher education: what is 'enhanced'and how do we know? A critical literature review." *Learning, media and technology* 39.1 (2014): 6-36.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). A learning management system-based early warning system for academic advising in undergraduate engineering. In *Learning analytics* (pp. 103-119). Springer, New York, NY.

Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).

Looi, C. K., Seow, P., Zhang, B., So, H. J., Chen, W., & Wong, L. H. (2010). Leveraging mobile technology for sustainable seamless learning: A research agenda. *British journal of educational technology*, *41*(2), 154-169.

Lourens, A., & Bleazard, D. (2016). Applying predictive analytics in identifying students at risk:

A case study. *South African Journal of Higher Education*, *30*(2), 129-142.

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in Moodle learning management system: A case of Mbeya University of Science and Technology. *The Electronic Journal of Information Systems in Developing Countries*, *79*(1), 1-13.

Marczyk, G. R., DeMatteo, D., & Festinger, D. (2010). *Essentials of research design and methodology* (Vol. 2). John Wiley & Sons.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In *2007 37th annual frontiers in education conferenceglobal engineering: knowledge without borders, opportunities without passports* (pp. T2G-7). IEEE.

Ogwoka, T. M., Cheruiyot, W., & Okeyo, G. (2015). A model for predicting students' academic performance using a hybrid of K-means and decision tree algorithms. *International Journal of Computer Applications Technology and Research*, *4*(9), 693-697.

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of asynchronous learning networks*, *16*(3), 9-20.

Prinsloo, P. (2018). Context matters: An African perspective on institutionalizing learning analytics. *INCludE uS All! dIRECtIONS fOR AdOPtION Of lEARNINg ANAlYtICS IN thE glObAl SOuth*, 24.

Prinsloo, P., & Kaliisa, R. (2022). Learning Analytics on the African Continent: An Emerging Research Focus and Practice. *Journal of Learning Analytics*, 1-18.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education. *London: Jisc. Accessed February*, *8*(2017), 176.

Siemens, G., Dawson, S., & Lynch, G. (2013). Improving the quality and productivity of the higher education sector. *Policy and Strategy for Systems-Level Deployment of Learning Analytics. Canberra, Australia: Society for Learning Analytics Research for the Australian Office for Learning and Teaching*, *31*.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18.

Society for Learning Analytics Research. (2011). Open Learning Analytics.retrieved from; https://www.solaresearch.org/about/what-is-learning-analytics/

Walji, S., Deacon, A., Small, J., & Czerniewicz, L. (2016). Learning through engagement: MOOCs as an emergent form of provision. *Distance Education*, *37*(2), 208-223.

Zulfikri, M. F., Shaharudin, S. M., Rajak, N. A. A., & Ibrahim, M. S. B. (2021). Predictive Analytics on Academic Performance in Higher Education Institution during COVID-19 using Regression Model. *International Journal of Biology and Biomedical Engineering*.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11).

APPENDICES

Appendix A: R-Shiny Implementation

A.1: User- Interface Implementation

```
      )
    ),
    tabPanel("Predictive Analytics",
      fluidRow(
        column(
          width = 6, offset = 3,
          tags$div(
            h3(class = "predictive-title", "Towards Virtual Learners'success-Prediction Tool"),
            sidebarLayout(
              sidebarPanel(style = "background-color: navy; color: white;",
                width = 18,
                textInput("Assessment_score_input", create_label("Assessment Score:", TRUE), value = "", placeholder = "Enter score"),
                selectInput("highest_education_input", create_label("Highest Education:", TRUE),
unique(training_data$highest_education), multiple = FALSE),
                selectInput("gender_input", create_label("Gender:", TRUE), unique(training_data$gender), multiple = FALSE),
                selectInput("disability_input", create_label("Disability:", FALSE), unique(training_data$disability), multiple =
FALSE),
                selectInput("age_band_input", create_label("Age Band:", TRUE), unique(training_data$age_band), multiple = FALSE),
                textInput("num_of_prev_attempts_input", create_label("Number of Previous Attempts:", TRUE), value = "", placeholder =
"Enter attempts"),
                textInput("email_input", create_label("Enter your email:", TRUE), placeholder = "Enter your email"),
                actionButton("predict_button", "Predict", disabled = TRUE),



                shinyjs::hidden(
                  div(
                    textOutput("prediction_output"),
                    textOutput("email_output")
                  )
```

Figure A.1: Code snippet for realization of the front-end operation.

A.2: Server Function Implementation

```
server <- function(input, output, session){
  # Reactive values for prediction and email
  rv <- reactiveValues(prediction = NULL, email = NULL)
  # Enable or disable the "Predict" button based on whether all required fields are filled
  observe({
    required_fields <- c("email_input", "Assessment_score_input", "highest_education_input", "age_band_input",  "gender_input",
"disability_input","num_of_prev_attempts_input")
    all_fields_filled <- all(sapply(required_fields, function(field) {
      !is.null(input[[field]]) && input[[field]] != ""
    }))
    if (all_fields_filled) {
      shinyjs::enable("predict_button")
    } else {
      shinyjs::disable("predict_button")
    }
  })
 reference_ranges <- list(
  Assessment_score = range(training_data$Assessment_score),
  highest_education = c(0, length(mapping))  # Assuming the range for highest_education remains intact
)
# Define a function to perform scaling for a single variable
scale_variable <- function(input_value, reference_range) {
  scaled_value <- (input_value - reference_range[1]) / (reference_range[2] - reference_range[1])
  return(scaled_value)
}
  # Event handler for the predict button
  observeEvent(input$predict_button, {
    # Create a data frame with user input
    new_student <- data.frame(
      Assessment_score =scale_variable(as.integer(input$Assessment_score_input),reference_ranges$Assessment_score),
```

Figure A.2: Code snippet for Back-end operation

A.3: Alert Generation Mechanism

```
1766      email_body <- paste(
1767        "Dear Student,\n",
1768    "\nWe have reviewed and analyzed your information and would like to share our insights based on the input you provided.Our
        analysis aims to provide you with valuable guidance and support as you navigate your academic journey.\n",
1769        "Here are the details you provided for prediction:\n",
1770        paste0(
1771          "     - Assessment Score: ", input$Assessment_score_input,"\n",
1772          "     - Highest Education: ", input$highest_education_input,"\n",
1773          "     -Gender: ", input$gender_input,"\n",
1774          "     - Disability: ", input$disability_input,"\n",
1775          "     - Age Band: ", input$age_band_input,"\n",
1776          "     - Number of Previous Attempts: ", input$num_of_prev_attempts_input,"\n"
1777        ),
1778        if (!is.na(glm_predictions)) {
1779          if (glm_predictions=="0") {
1780            paste("Predicted Probability of Failure:", round(glm_prob, 2))
1781          } else {
1782            paste("Predicted Probability of Passing:", round(glm_prob, 2))
1783          }
1784        } else {
1785          "Prediction is not available.\n"
1786        },
1787        "\nPlease note that this prediction is based on a model and should be considered as a tool for guidance.We encourage you to
        take up the responsibility of seeking help when needed. Whether it's reaching out to instructors, academic advisors, or utilizing
        campus resources, seeking support is a sign of strength and determination to succeed.If you have any concerns or questions, feel
        free to reach out to us.\n",
1788        "\n",
1789        "Best regards,\n",
1790        "Analytics Team.\n"
1791      )
```

Figure A.3: Code Snippet for Back-end Email Alert Generation

A.4: System Responsiveness

```
e any concerns or questions, feel free to reach out to us.\n",
+       "Best regards,\n",
+       "Analytics Team.\n"
+     )
+
+
+     # Send email
+     gmailr::gm_send_message(
+       gm_mime(
+         to = input$email_input,
+         subject = "Student Performance Prediction",
+         body = email_body
+       )
+     )
+     # Show the hidden div
+     shinyjs::enable("hidden")
+   })
+
+   # Output for displaying prediction
+   output$prediction_output <- renderText({
+     if (!is.null(rv$prediction)) {
+       paste("Prediction:", rv$prediction)
+     }
+   })
+ }
>
> shinyApp(ui,server)

Listening on http://127.0.0.1:7027
[1] "Server execution time: 0.0065760612487793"
```

Figure A.4: System Responsiveness

A.5: Data Product User Guide

All fields are required for the "**Predict button**" to be enabled.

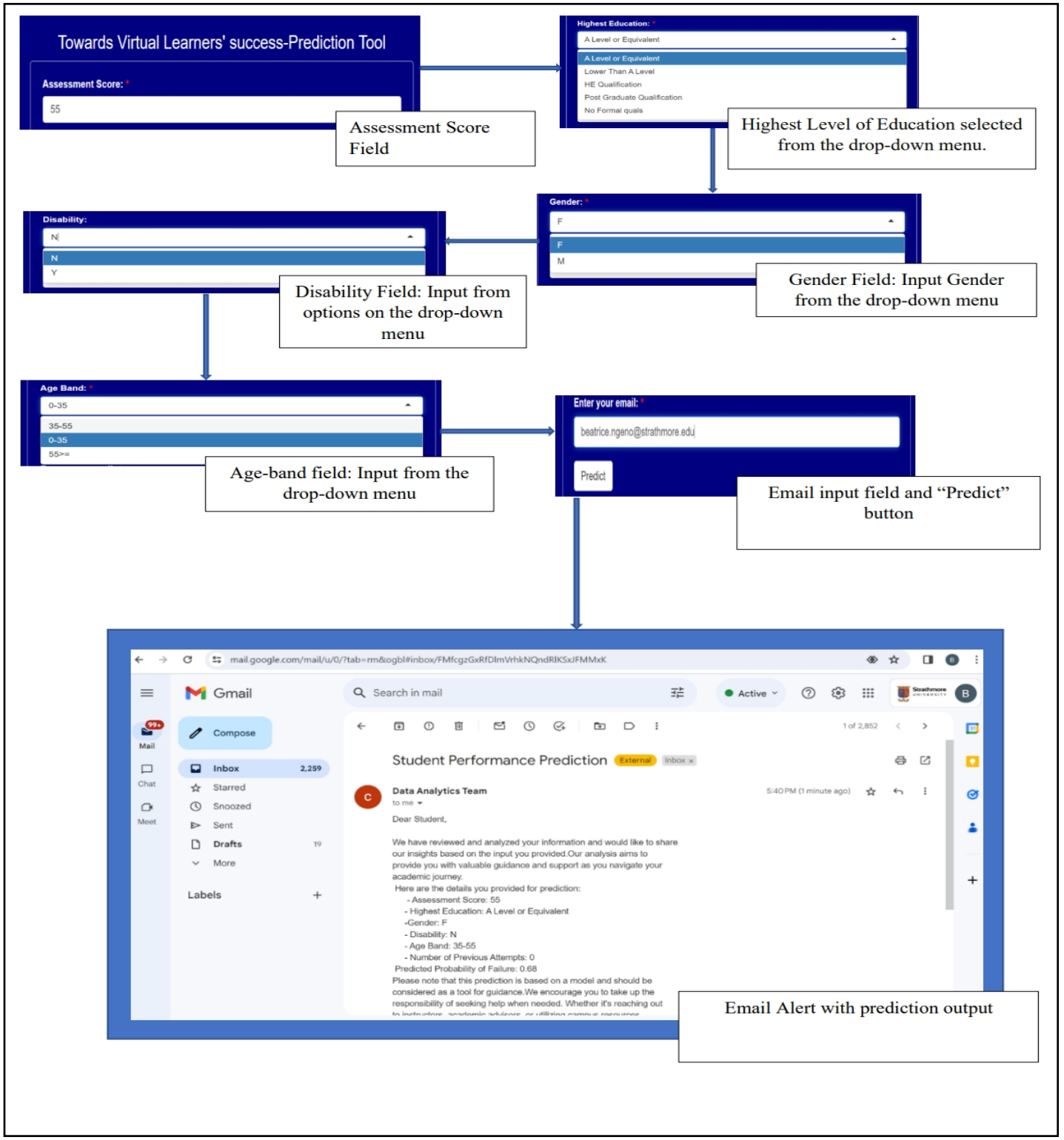


Figure A.5: Data Product User Guide

Appendix B

B.1 Ethical Review Committee Approval

Strathmore UNIVERSITY

21st November 2023

Ms Ngeno Beatrice,
beatrice.ngeno@strathmore.edu

Dear Ms Ngeno,

RE: Developing an R-Shiny Powered Early Alert Tool to Optimize Virtual Students' Performance: Case Study of The Open University, UK

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1914/23.** The approval period is from **21st November 2023 to 20th November 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
21-Nov-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

B.2 Similarity Report