

**THE UNIVERSITY OF ZAMBIA
THE SCHOOL OF EDUCATION
DEPARTMENT OF LIBRARY AND INFORMATION SCIENCE**

**PERFORMANCE PREDICTOR: A DATA MINING AND MACHINE LEARNING
SOFTWARE FOR STUDENT PERFORMANCE OUTCOMES**

CHAIBELA MUTUNE	2017012923
CHISHA IVY	2017012966
PUNGWA DAVID	2017012954
SIABBABA DANNY	2017003625
SIMUKOKO BYDON	2017012915

SUPERVISOR: DR. PHIRI LIGHTON

DECEMBER, 2021

ABSTRACT

The intent of this study was to investigate the feasibility of implementing machine learning model for automatically predicting students particularly ICT 1110, who are at risk of failing ICT 1110. Performance predictions as highly important as it is could be very cumbersome as an educator has to analyze large sums of data in order to identify performance especially those at risk of failing so they can administer appropriate correction mechanisms. Over the past 2 years the University of Zambia recorded a high poor performance of students in the ICT 1110 course. There could be many reasons associated with their poor learning outcomes associated with their poor learning outcomes which could be associated to workloads, attendance, lack of resources just name a few. In an attempt to solve this problem, this study presents a performance prediction software that will involve the use of data mining and machine learning in order to train data, associated with student's academic performance that will be able to predict students who are at risk of failing ICT 1110. The discoveries of the study will benefit the educator as it will predict those students who at risk of failing so that the educators can appropriately execute correction mechanisms to help the students at risk. The research carried out was both quantitative and qualitative. The study targeted a sample size of 60 students (total number of participants) which comprised of the ICT 1110 lecturer, tutor and students. Online interviews and questionnaires were employed to collect data via Google Meet and Google Forms concerning the factors that could possibly be related to student academic performance. The quantitative data was analyzed using Microsoft excel. The results showed a number of factors that could be associated with student performance outcomes which include student interest, mode of teaching, prior knowledge, motivation, support structures, time management, workload, guidance attendance, program minors, participation in course activities and engagement with the course activities. Factors used outside the elicitation process included gender, institutional aid and tuition support. After analyzing the factors to determine the data sources associated with for the factors, it was concluded that student interest, workload, engagement, minors, gender, institutional aid and tuition support would be used as machine learning input features for the model in order for the model to make predictions. In conclusion, the factors analyzed were seen to be effective potential features for the model to identify at risk students accurately in order for educators to render corrective measures to these struggling students.

Acknowledgements

To begin with, we would love to thank Almighty God for the overall wisdom, knowledge, skill and strength rendered to us to complete the project. Further we would like to extend our gratitude to our supervisor Dr. Lighton Phiri for the support, commitment, motivation, enthusiasm, generosity. For his knowledge base and guidance, he proved to be an incredible mentor and guide for this project.

Furthermore, we would love to thank the Department of Library and Information Science lecturers, the participants of the study and our classmates for the help and support rendered to us.

Lastly, we would express our gratitude to our family members for their prayers and support towards the completion of this project.

Dedications

After all the hard effort put into this study, we would like to dedicate this study to the following. To Almighty God our source of all knowledge, wisdom, power, health, and overall purpose. We are eternally grateful to God for the strength and endurance to engage in this kind of research. To our dear families, or gratitude goes to you for all the prayers, words of encouragement, motivation and perseverance and for all your support. To our friends, we thank you for your support in every way. To our supervisor, Dr. Lighton Phiri, we thank you for your support, for educating us in doing our work, for your time, your corrections, your tolerance and your guidance. We thank you all from the bottom of our hearts.

Acronyms

API- Application programming interface

CSS- Cascading Style Sheet

HTML- Hypertext Markup Language

ICT- Information Communication Technology

LMS- Learning management system

SIS- Student information system

UNZA- University of Zambia

Keywords: Machine learning, Data mining, Learning outcomes

TABLE OF CONTENTS

1 INTRODUCTION.....	7
1.1 Background of Study.....	8
1.2 Problem Statement.....	8
1.3 Study Objectives.....	9
1.3.1 General Objective	
1.3.2 Specific Objectives	
1.3.3 Research Question	
1.4 Rationale of Study.....	9
1.5 Ethical Considerations.....	10
2 RELATED WORK /LITERATURE REVIEW.....	11
3 METHODOLOGY.....	14
3.1 Introduction	
3.2 Research design and approach	
3.3 Sample population	
3.4 Sampling procedures	
3.5 Data collection instruments	
3.6 Validity and reliability of the study instruments	
3.7 Data analysis procedures	
3.8 Ethical considerations	
4 REQUIREMENT ENGINEERING.....	17
4.1 Requirement elicitation	
4.2 Requirement specification	
4.3 Requirement verification and validation	
4.4 Design and implementation.....	20
4.4.1 Architectural design	
4.4.2 Interface design	
4.4.3 Modeling	
4.4.4 Select modeling technique	

4.4.5	Modeling technique	
4.4.6	Modeling assumptions	
4.4.7	Generate test design	
4.4.8	Test design	
4.4.9	Build model	
5.	SYSTEM EVALUATION.....	26
5.1	Selected machine learning algorithms	
5.2	Testing machine learning algorithms	
6.	RESULTS.....	32
6.1	Results from ICT 1110 students' responses	
6.2	Responses from the ICT 1110 lecturer and tutor	
6.3	Table of interview questions and respective resources	
7.	DISCUSSIONS.....	43
7.1	Features associated with performance prediction	
8.	CONCLUSION.....	49
9.	REFERENCES.....	50
10.	APPENDICES.....	53
	Appendix 1: Interview questions for lecturer	
	Appendix 2: Interview questions for tutor	
	Appendix 3: Questionnaire for students	

CHAPTER ONE

1. INTRODUCTION

Learning outcomes are one of the most important aspects in every level of education as the primary focus is on what the student has achieved than a mere focus on what has been taught. (Kennedy, 2007). As it serves as a grading metric for not only students but for educational institutions as well, learning outcomes, according to Adam (2004), can be referred to as a written statement of what the student is expected to be able to accomplish at the end of a course unit. Learning outcomes are essentially vital in that they, like a GPS, guide learners to desired performance results and they also give a clear idea of what can be achieved by undertaking a particular course. To educators, learning outcomes help in giving precise ideas on how to teach various lessons and also placing strategic measures in how guide both the lesson and the learner in order for students to achieve academic excellence. (Mahajan & Singh, 2017)

In guiding students to academic excellence, it is important for educators to consider the progress towards successful learning outcomes by early predicting the performance outcomes of their students. The prediction of student academic performance draws considerable attention to most learning institutions as it allows educators to identify students at risk of failing as well as those who are attaining academic excellence. Student performance in higher education facilities, like The University of Zambia, is researched extensively to curb academic underachievement, university dropout rates, graduation delays and many other academic challenges. (Namoun and Alshanqiti, 2021)

With the rise of technology in various sectors of society, more especially in the education sector, the prediction of student academic performance has been made digitally possible with the use of various machine learning techniques and applications. These techniques allow educators of learning institutions to tackle issues based around student attrition. This is done by primarily identifying at risk students in a timely manner so as to execute assistive measurements. (Fahd, Miah and Ahmed, 2021)

The intent of this research study was to investigate the feasibility of implementing a machine learning model for automatically predicting students at risk of failing, particularly, ICT 1110 at The University of Zambia. The research further accessed some potential factors that would

affect the ICT 1110 students' performance outcomes and with these factors fostered the making a classification model that would predict struggling students timely enough for educators to execute appropriate and effective correction mechanisms.

1.1 BACKGROUND OF STUDY

Student success resulting from excellent academic performance is imperative to higher learning institutions including The University of Zambia as it serves as a grading metric not only for students but also for the performance of the institution. It is also essential in assessing the quality of the educational institutions. However, at the University of Zambia, recorded a low academic performance among a number of students over the past two years. This drop of academic performance could be due to many factors which would be associated to student demographics, welfare and learning environment. The traditional or rather manual way of educators figuring out at risk students and why they are at risk may tend to be time consuming which in turn may not allow correction mechanisms to be executed on time. Furthermore, it may allow educators to focus on certain factors while leaving out other vital factors that may affect student performance outcomes and thus incorrect predictions may occur as a result. (Alyahyan and Düştegör, 2020)

1.2 PROBLEM STATEMENT

Over the past two years, the School of Education at The University of Zambia has recorded a high poor academic performance of students in the ICT 1110 course. There could be many reasons as to why the performance rate has drastically dropped and reasons could be no less related to certain academic markers that can foretell student learning outcomes such as class attendance, workload, less course content interactions, unfamiliarity with computer software, demographics, educational backgrounds and educator-student interactions.

1.3 STUDY OBJECTIVES

1.3.1 General Objective

To investigate the feasibility of implementing machine learning models for automatically predicting students at risk of failing.

1.3.2 Specific Objectives

1. To identify input features that are correlated with ICT 1110 students.
2. To implement classification models for predicting students at risk of failing ICT 1110.
3. To implement appropriate APIs and user interfaces for interacting with the classification models for predicting students at risk of failing ICT 1110.

1.3.3 Research Questions

1. What features are correlated with student performance in ICT 1110?
2. Is it feasible to implement a classification model for predicting students at risk of failing ICT 1110?
3. How should useful and usable classification models for predicting students at risk of failing ICT 1110 be deployed?

1.4 RATIONALE OF THE STUDY

The purpose of the study was to investigate the feasibility of implementing machine learning models for automatically predicting students at risk of failing, particularly ICT 1110. The results of this study will further allow The University of Zambia educators, especially those instructing the ICT 1110 course to focus on potential factors that could affect student academic performance outcomes in order to executive appropriate and effective correction mechanisms. Furthermore, it will allow for the use of a classification model for predicting students at risk of failing ICT 1110. Also, the project model will provide a user friendly interface in order for the educators to easily identify at risk students with less time and effort.

1.5 ETHICAL CONSIDERATIONS

While conducting the research, measures were undertaken to ensure compliance with ethical issues which included not forcing the respondents to answer the questions. Measures were also taken to ensure that the identities of the respondents are kept confidential. In addition, the respondent's responses were neither interfered nor contested by the researchers. Furthermore, informed consent was obtained from respondents. The researchers also communicated to the participants before involving them in the study. Moreover, the researchers also openly informed the respondents that they had the right to withdraw and the effect of their withdrawal in the study would be explained. Additionally, all respondents were assured of the benefits which would be obtained from the findings of the study. The respondents received equal treatment by the researchers.

CHAPTER TWO

2 RELATED WORK/BACKGROUND

There are a number of available research projects that have taken into account prediction of academic performance outcomes with the usage of datamining and machine learning and their general results have been to identify potentially at risk students and general performance outcomes. This section provides related literature on predicting student performance outcomes and identifies certain gaps and exclusions within the existing literature reviewed.

Prior works in relation prediction of student performance outcomes have used different types of models and machine learning techniques and algorithms for different datasets with various attributes. Dorina et al (2015) proposed a predictive model for student's performance by classifying students into binary class (successful / unsuccessful). The proposed model was constructed under the CRISP-DM (Cross Industry Standard Process for Data Mining) research approach. Different classification algorithms were applied on the given dataset. The results show that the highest accuracy was achieved by the MLP (Multilayer Perceptron) model for identification of successful students while other three models perform better for the identification of unsuccessful students. However, the model was unable to work out for data high dimensionality and class balancing problems.

Hasan et al (2020) aimed to predict student's overall performance with the use of Video Learning Analytics (VLA) and Data Mining Techniques. The article further elucidates that institutions of learning have adopted the used of flipped classrooms or classrooms done through the internet. Also, Learning Management Systems (LMS) and Video Streaming Servers have been essentially useful in disseminating academic content. Through the use of Learning Analytics such as e-Dify, the institutions can keep track of student's video interactions thus providing student information useful to the educators. However, the article points out that less work has been done to combine both Video Learning Analytics and Educational Data Mining Techniques to predict student academic performance and this is the breach the article has undertaken to seal up. The study proposed a supervised data classification model with the aim of predicting the academic performance of students at the end of the semester. The dataset comprised of student academic

data gathered from Student Information System (SIS) and the performance of two modules, selected for the study, using two different learning environment which were Moodle (LMS) and e-Dify (VLA). With the various other tools that were used in this study, the model could predict students at risk of failure and overall student performance.

According to John et al (2016) in order to predict academic failure or drop outs they used a decision tree structure that follows a sequential path and has nodes to help make logical decision. The branches of a decision tree are used to help make logical decision. The branches of a decision tree are used to identify the students who are academically weak and need remedial classes or any other help in order to keep them from failing or dropping a year.

Osmanbegovic and Suliic (2015) proposed a model to predict student academic success in a course by reducing data dimensionality problem. Various machine learning classifiers such as Naïve Bayes, MLP and j48 were evaluated in this study. The result shows that the Naïve Bayes gained the highest accuracy in its prediction. However, the proposed model although did not handle the class imbalance problem.

Bilal et al (2016) presented a student failure prediction model which identified the students that might be at-risk of performing poorly. Four output classes (Average, Risk, below Average and Above Average) were generated by the proposed model based on the CGPA of the students. Six classifiers were applied on the given dataset and The ID3 got the highest accuracy at predicting at-risk students but the model was unable to work out for course imbalance problem.

Shaymaa et al (2014) proposed to predict the student academic performance by giving a student to write their comment on the space provided after each lesson then extracting words and speech frequencies and use the LSA technique to reduce the dimensions of a matrix and obtain the most significant vectors. Finally, the researcher concluded that this study expressed the correlation between self-evaluation descriptive sentence written by students and their academic performance by predicting their grade.

In conclusion of this section, the reviewed works share some commonalities in that their ultimate goal is to predict student academic performance outcomes be it poor or successful performance. However, there are some gaps that the reviewed works did not address. The works focused much on academic features such as marks and left out some other important features such as course

workload, minor courses, course engagement, demographics etc. Also the reviewed works did not show how one would interact with the model after completion.

The reviewed literature also showed how the use of various datasets and modelling techniques that use data mining and machine learning can predict at risk students and overall student performance thus giving our study very useful anticipations and insights on the data preparation stages and the performance of various machine learning techniques that was considered in order to build a predictive model.

CHAPTER THREE

3 METHODOLOGY

3.1 Introduction

This chapter describes the research design, target population, sample population, research instruments and data collection procedures and data analysis.

This chapter outlined and described research methods and techniques that were used in conducting this research. It will start by explaining the area of study, research design and data collection instruments. Population and sample size and technique considered in this study will be explained as well. The methods of data collection, data analysis tools which were used to analyze data are explained, limitations encountered during the study, issue of data validity and reliability as well as ethical consideration will be covered.

3.2 Research design and approach

According to Omari (2011), research design refers to a distinct plan on how a research problem will be attacked. Creswell, (2003) and Kerlinger (1978) defined research design as the plan, structure and strategy of investigation conceived so as to obtain answers to research questions and control variance. In this study the researchers applied a survey research design where the researchers employed cross-sectional survey. Cross-sectional survey is done where a researcher uses different categories of people. However, the study applied both quantitative and qualitative research approaches. Quantitative approach helped to quantify the problem by way of generating numerical data or data from the field and transform them into useable statistics. Qualitative approach helped to study attitudes, opinions, behaviors, and other defined variables of the population.

3.3 Sample population

According to Kerlinger (1978) a sample can be defined as a group or subset of the total populations selected for observation and analysis. The researchers identified and selected respondents that fulfilled or justified the questions the research addressed. The following was the sample which was selected and interviewed; 1 ICT 1110 lecturer and 1 ICT 1110 tutor. In addition, an online questionnaire was distributed to 60 students via their student email and WhatsApp.

3.4 Sampling procedures

The study used two types of sampling procedures which are convenience and purposive sampling methods. Convenience sampling is a sampling method where the sample is taken from a group of people who are easy to contact or to reach while purposive sampling means that respondents are chosen on the basis of their knowledge of the information desired. Moreover, convenience sampling was used in choosing sample units from the entire population of students. Purposive sampling was also used in choosing lecturers and tutors as they were concerned with monitoring the performance of ICT 1110 students. Through convenience sampling process 60 students were selected and through purposive sampling 1 lecturer and 1 tutor were selected (Kothari, 2004).

3.5 Data collection instruments

Research instruments included the following as recommended by Ballantine and Hammek (2009); questionnaires and interviews. Questionnaires were used on ICT 1110 students. Interviews were used on selected course lecturer and course tutor. This was both a quantitative and qualitative research. Online questionnaires were used to collect data through the use of Google Forms concerning the factors that affect students in their performance in ICT 1110. A questionnaire was used to collect primary data among ICT 1110 students. Closed and Likert-scale questions were used because they generated a limited set of responses. A questionnaire was chosen because of the nature of this study so as to get views of the respondents. Respondents replied to them on their own free will without any influence from another person; they were easy to be self-administered within a short time and from the relatively larger groups of people who were scattered geographically. Moreover, its results could easily be tabulated and interpreted. Interviews. This study employed structured interview. Structured interviews can be conducted face to face, online or over the telephone, sometimes with the aid of lap-top computers. But in this study, it was conducted online via Google meet. The researchers provided the respondents with pre-set questions and let them respond on the asked questions by the researchers.

3.6 Validity and reliability of the study instruments

To establish validity of the instruments applied, the researchers conducted a pilot study prior to the actual data collection. The instruments were tested by providing it to group members. The instruments were presented to the supervisor for further comments and improvement hence all necessary adjustments were made for items which were found unsuitable were removed. To ensure reliability of the collected information, some of the items in questionnaire and interviews were

asked more than one time to the respondents to see if there is consistency in responses from the respondents.

3.7 Data analysis procedures

According to Kothari (2004), data analysis is a process of editing, coding, classification and tabulation of collected data. The process involves operations which are performed with the purpose of summarizing and organizing the collected data from the field. Since the study involved both qualitative and quantitative data, the data analysis process was done in two ways.

First the researchers applied Microsoft excel for quantitative data. This is the software which is used to analyze information that is quantitative in nature. In this study, data collected using questionnaire was analyze using Microsoft excel software. The process involved coding of data, sorting and conclusion was drawn.

Secondly, the qualitative data obtained using interviews was analyzed by considering major themes to extract relevant information. This helped the researcher to make description of the data collected from the field basing on research objectives and derived conclusion on what to take regarding its usefulness.

3.8 Ethical consideration

To obtain population of study, data collection and dissemination of the findings, the researchers were sensitive to research ethics and its values. This helped to ensure that good image of research enterprise in the world was maintained (Omari, 2011). The researchers ensured the freedom of participants by adhering to the principal of informed concerned. This principal required the researcher to ensure that participants were aware of the purpose of the study so as to get their concern and participate freely. The statement of the research purpose, description of any potential risks or discomforts, description of potential benefits and the description of confidentiality were assured to the respondents. The researchers assured them not to reveal their identity to anyone other. These findings were stored in such a way that it will be accessible only for the research purpose so as to maintain privacy or confidentiality and anonymity of the respondents in the researchers' personal computers with passwords.

CHAPTER FOUR

4 REQUIREMENT ENGINEERING

A requirement is a singular documented need, what a particular product or service should be or how it should perform. It is a statement that identifies a necessary attribute, capability, characteristic, or quality of a system in order for it to have utility or value to a user. Requirement engineering is the discipline concerned with establishing and managing requirements, (Shaw, M. 1990). It is a process of gathering and defining the services provided by the system. Requirement engineering consists of the following main activities; Elicitation, Specification, Verification and Validation.

4.1 Requirement Elicitation

Requirement elicitation is related to the various ways used to gain knowledge about the project domain and requirements. During elicitation the project team aimed at understanding the project vision and constraints, the context that the product will be deployed into, and the stakeholders that will need to accept the product (Hickey and Davis 2004, Zowghi and Coulin 2005). Such requirements elicitation results in an overview of users, external systems, and other stakeholder viewpoints and a description of their respective background, interests, and expectations.

The various sources of domain knowledge which were used for our project include ICT 1110 students, tutor and lecturer. The Course Register (Workload) datasets, report demographics datasets and Moodle logs (ICT) of the students who happens to be the main stakeholders were fetched from UNZA SIS and UNZA Moodle. The elicitation techniques that were used for requirements elicitation include interviews with the lecturers and tutors and online questionnaire with the ICT 1110 students. Thirty (30) students responded to the questionnaire which were shared online. Elicitation didn't produce the formal models of the requirements understood. Instead, it widened the domain knowledge of our analysis and thus it helped in providing input to the next stage. Prototyping is the type of system analysis that was used. This is a paper or tool-based approximation of the end-systems to increase the tangibility and authenticity of the planned system (Rettig, 1994).

4.2 Requirement Specification

This activity was used to produce the formal software requirement models. All the requirements including the functional as well as the non-functional requirements. Functional requirements refer to the services that the system should provide, how the system reacted to the particular inputs and how the system behaved in particular situations (Zowghi and Coulin, 2005).

The inputs which were used include the following minor course, number of courses, number of times the student interact with Moodle as well as accommodation status of the student. The model enabled the instructor for ICT 1110 to know students who are at risk of failing ICT 1110 and also those not at risk of failing by just entering the student's computer number and the results were brought showing whether he/she is at risk of failing. The context diagram below shows how the system or model responded to the users of the system.

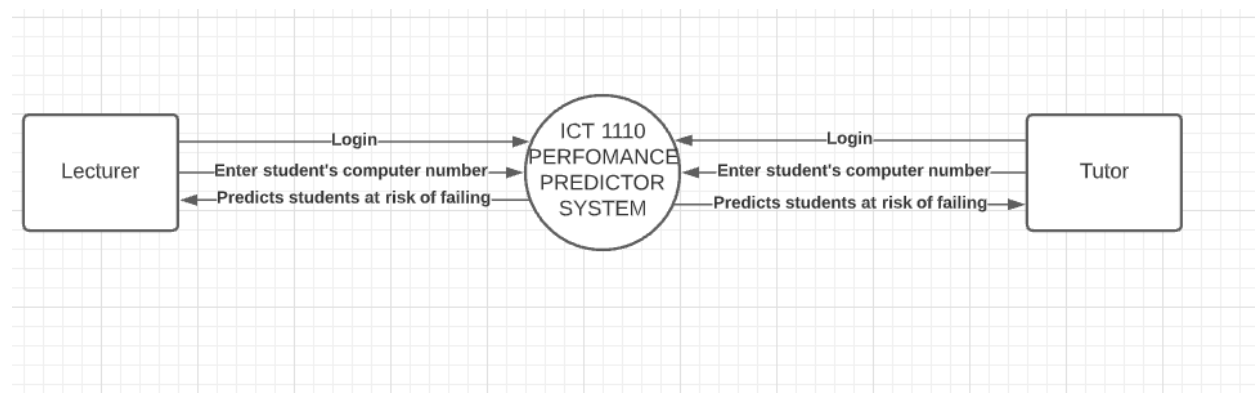


Figure 1 above shows the context diagram

4.3 Requirement Verification and Validation

Verification refers to the set of tasks that ensures that the software correctly implements a specific function. Validation refers to a different set of tasks that ensures that the software that has been developed or built is traceable to the stakeholder's requirements, (Cheng, B. and J. Atlee 2007).

In instances where requirements were not validated with our software, errors in the requirement definitions propagated to the successive stages which resulted in a lot of modification and rework with the software so as to meet the stakeholder's requirements who are the students of ICT 1110, lecturer and tutor. The main steps that were used for this process to be achievable include:

- The requirement should be consistent with all the other requirements.
- The requirement should be complete in every sense,
- The requirement should be practically achievable and for this reviews, buddy checks were some of the methods used for this.

4.4 DESIGN AND IMPLEMENTATION

This section defines the design structure and implementation of the project software. It specifies the architectural design of the software, taking into accounts its structure, components and the relationship and interaction between its components. Also, this section defines the software's interface design for interactions of and with the system.

4.4.1 Architectural design

The software is a web based application that consists of a Machine Learning prediction algorithms which take in various data attributes associated with student performance for its training input to produce a predictive output on student learning outcomes. The model will then take in new input data and produce a prediction on whether or not a student will pass or fail the course of ICT 1110.

4.4.2 Interface Design

The model is hosted on a web with a general user interface that includes elements such as input control that takes into account text fields, data fields and buttons. Navigational components such as tags, icons and other components for easy navigation through the software, Information components like message boxes and Containers. This will enable the user to interact with the model on any operating system. Hypertext transfer protocol is used to enable communication between system interfaces and interfaces with the user.

The ICT 1110 student performance predictor was implemented using the cross industry standard process for data mining (Crisp DM Model). It has 6 phases:

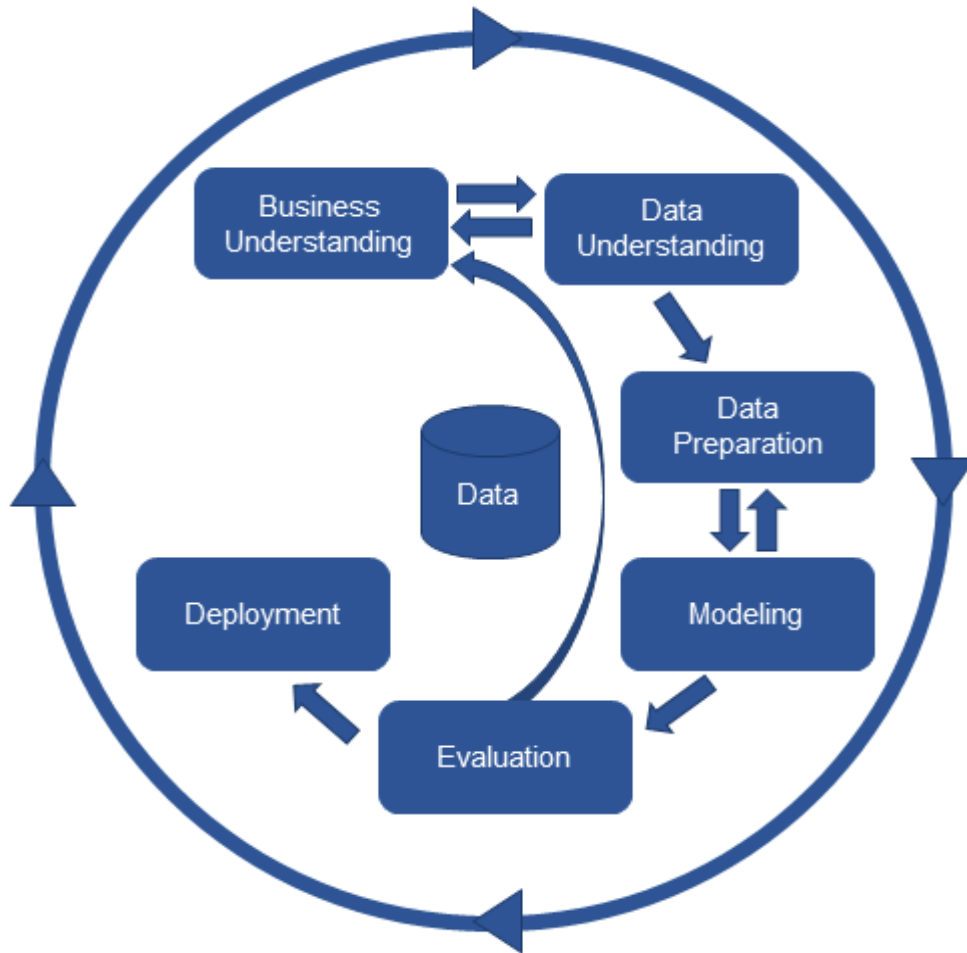


Figure 2 above shows the Crisp DM Model

Business understanding. The first phase was business understanding which involved general and specific objectives. The first objective of the analysts was to thoroughly understand, from a business perspective, what the stakeholders really wanted to accomplish. Often the stakeholders had many competing objectives and constraints that needed to be properly balanced. The analyst's goal was to uncover important factors at the beginning of the project that can influence the final outcome. In this study the main objective was to investigate the feasibility of implementing machine learning models for automatically predicting students at risk of failing

On the other hand, the specific objectives were to identify input features that are correlated with ICT 1110 students. To implement classification models for predicting students at risk of failing ICT1110 and to implement appropriate API's and user interfaces for interacting with the classification models for predicting students at risk of failing ICT1110.

The research questions for this project output background collates the information that is known about the organization's business situation at the start of the project. These details not only serve to more closely identify the business goals that were solved, but also serve to identify resources, both human and material, that may be used or needed during the course of the project. Activities organization, develop organizational charts identifying divisions, departments and project groups.

The target group for this project were the students, the lecturer and tutor. Therefore, the project attempted to answer each of these questions: what features are correlated with student performance in ICT 1110? Is it feasible to implement a classification model for predicting students at risk of failing ICT1110? How should useful and usable classification models for predicting students at risk of failing ICT 1110 be deployed?

The second step was data understanding and its objective was to know what can be expected and achieved from the data collected. It checked the quality of the data such as data completeness, values distributions, data governance compliance. This was one of the crucial parts of the project because it defined how viable and trustworthy the final results are. In this step, team members had to brainstorm on how to extract the best value of the pieces of information collected in relation to the project (O. Grljević, and Z. Bošnjak 1998). This phase or stage, consisted of the following minor stages for it to be successfully done:

Collect Data: This was basically to acquire the data. The data we wanted had to be collected from the stakeholders and the datasets had to be fetched from UNZA Moodle and UNZA SIS.

Describe data: The project team had to examine the data format, number of rows and columns, field identities, and available features.

Explore data: The project team had to describe the relationship between data, visualize the data, and be creative.

Verify data quality: The project team had to verify the data to ensure that data is of good quality by check for the missing values and ensuring that the data collected was appropriate.

Data understanding is where the project team had to show everything they understood about the data and relate it with the business question.

Immediately after the project team understood the data, it was time for the data preparation. This phase or stage rather involved the extract transform and load or extract load and transform processes that turns the pieces of data into something useful by the algorithms and process (Pyle D, 1999).

This phase is what we did to prepare the data for the modeling phase. The phase was made up or included the following stages:

Data Selection: Selecting the dataset, columns, and/or rows we used.

Data Cleaning: Garbage-in, garbage-out normally happens if you did not clean the data properly. This is the task where we made sure the data was right. Cleaning data took a lot of preparation and data understanding.

Feature engineering: Feature engineering proved to be interesting or helpful. This is simply because this is where project team members applied creativity when creating new data from existing data.

Data integration: New data set from combining two or more data sets

Data formatting: Formatting data was performed. For example, when converting the categorical value into numerical value or vice versa.

Data preparation was key to a great modeling process. Likewise, some algorithms perform better under certain parameters, some don't accept no-numerical values and others don't work well with a large variance on values. The project team had to be careful to avoid messing up in this phase this is because the next phase would not have produced any viable result. That is why this is the phase the project team had to focus on the most and describe as much as possible to make the project stand out even more.

4.4.3 Modeling

In this phase, various modeling techniques were selected and applied, and their parameters were calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going

back to the data preparation phase is often necessary. This phase has steps which were undertaken and are explained below:

4.4.4 Select modeling technique

The first step in modeling, selecting the actual modeling technique that had to be used. This task referred to the specific modeling technique logistic regression and decision-tree building.

4.4.5 Modeling technique

The project team used logistic regression and decision tree classifier modeling techniques to build the model.

4.4.6 Modeling assumptions

The project team discovered that many modeling techniques make specific assumptions about the data, for example, all attributes have uniform distributions, no missing values allowed and class attributes must be symbolic.

4.4.7 Generate test design

Before the project team actually built the model, the team generated a procedure or mechanism to test the model's quality and validity. For example, in supervised data mining tasks such as classification, the project team used error rates as quality measures for data mining models. Therefore, the project team typically separated the dataset into train and test sets and built the model on the train set, and estimated its quality on the separate test set.

4.4.8 Test design

The project team had to divide the data into training (80%) and test data (20%).

4.4.9 Build model

Under this step the model tool was able to run on the prepared dataset to create the model.

Evaluation is the fifth step in order to verify that the results are valid and correct. In case the results are wrong, the methodology permitted the review back to the first step, in order to understand why the results were mistaken. Usually, on a data science project, the data scientists divided the data into training and testing. In this step the testing data was used, its objective was to verify that the

model (product of the modeling step) is accurate to reality. Depending on the task and the context, there were diverse techniques. For example, in the context of supervised learning, with the task of classifying items, one way to verify the results was with the confusion matrix (Agrawal, R. 1999).

The sixth and last step is Deployment and it consisted of presenting the results in a useful and understandable manner, and by achieving this, the project had to achieve its goals. A model is not particularly useful unless the customer accesses its results. The model was deployed using flask API (Brodley and Smyth, 1995).

CHAPTER FIVE

5 SYSTEM EVALUATION

System evaluation involves measuring the final system against its initial performance goals.

5.1 SELECTED MACHINE LEARNING ALGORITHMS

The following Machine Learning Algorithms were used to in order to test and train the datasets in the project model:

Logistic Regression

Although similar to Linear Regression, Logistic Regression analyzes the relationship between multiple independent variable and a categorical dependent variable. It estimates the probability of occurrence of an event by fitting data to a logistic curve.

Decision Tree

As one of the widely used techniques in Machine Learning, Decision Tree is a tree based technique which any path starting from the root id described by a data separating sequence until finally a Boolean result at the leaf node is accomplished.

5.2 TESTING MACHINE LEARNING ALGORITHMS

In order to investigate the performance of each of the algorithms being used in the project model, a Hold-Out Method test was used in order to measure the performance of the algorithms. The procedure splits the data into two parts: the first part is the training set where the proposition is trained and the hold-out set and which the performance is measured for validating and testing the model.

To evaluate the performance of the model, a Confusion Matrix, which is used to evaluate the performance of a classification model by comparing the actual target values with those predicted by the machine learning model were used. The performance measure metrics used were the Precision, Recall, F1-Score and Support. These four metrics are defined below as follows:

Precision

This measures the proficiency of the classifier and indicates how well the classifiers label positive predictions. The formula for the precision is as follows:

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Recall

This measures the proficiency of the classifier to predict all the positive instances and indicates how many correct positive labels are assigned by the classifier. The formula for the recall is as follows:

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

F₁Score

This accuracy measure utilizes a combination of Precision and Recall and it is a harmonic average of the two measures where the F₁score is between 0 and 1. The formula for the F₁score is as follows:

$$F_1\text{Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Support

This is the total number of occurrences of each label in the actual values. It is also used to measure imbalances in the data set.

Each feature was measured via the 4 quantities and the results of the accuracy measurements are presented in the figures below.

Table 1 below. Measurement of Model Performance using Logistic Regression

Features	Precision	Recall	F₁Score	Support	Accuracy
Interest	1.33	2.37	1.08	32	0.40
Workload (No. of courses)	1.00	0.84	0.92	32	0.84
Engagement	1.33	2.37	1.08	32	0.40
Minor Courses	1.64	1.57	0.58	25	0.80
Gender	1.90	2.94	1.92	25	0.96
Tuition Support	3.92	4.00	3.96	32	0.96
Institutional Aid	1.7	1.67	1.68	25	0.84
Average	1.83	2.25	1.60	29	0.74

An average accuracy of 0.74%, a precision of 1.83, a recall of 2.25, and an F-measure of 1.60 as well as the support of 29 have been obtained using decision tree classifier.

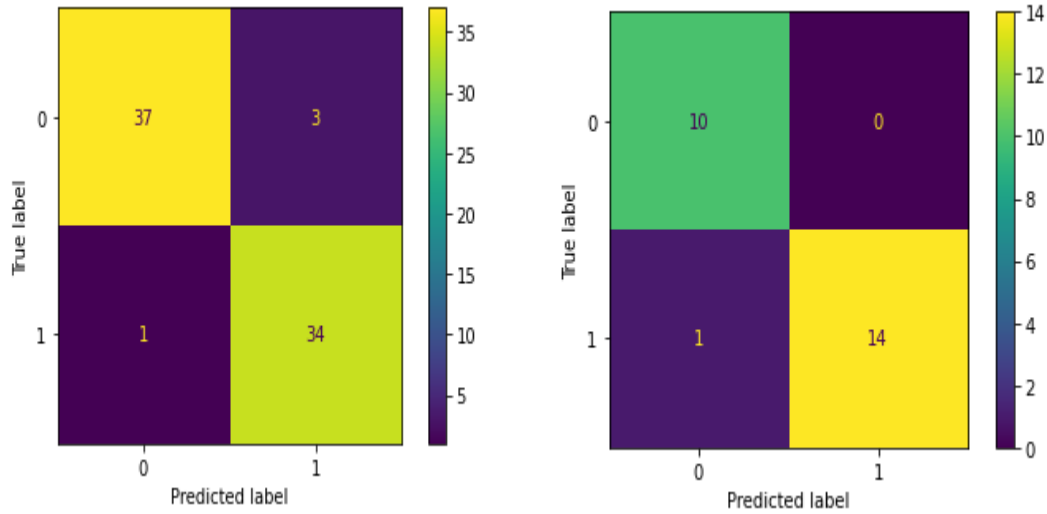
Table 2 below. Measurement of Model Performance using Decision Tree

Features	Precision	Recall	F₁Score	Support	Accuracy
Interest	1.83	1.83	1.83	25	0.92
Workload (No. of courses)	1.66	1.66	1.66	25	0.84
Engagement	1.83	1.83	1.83	25	0.92
Minor Courses	0.76	0.96	0.83	32	0.34
Gender	0.84	1.00	0.92	27	0.84
Tuition Support	1.77	1.77	1.76	25	0.88
Institutional Aid	1.00	0.75	0.86	32	0.75
Average	1.38	1.4	1.38	27.28	0.78

An average accuracy of 0.78, a precision of 1.38, a recall of 1.4, and an F-measure of 1.38 as well as the support of 27.28 have been obtained using decision tree classifier.

The confusion matrices of the performance of the model using Logistic Regression and Decision Tree Algorithms are shown below. (*Note: In the diagrams below, figure 1 seen on the True Label and Predicted Label represents pass, while 0 seen on the True Label and Predicted Label represents fail*)

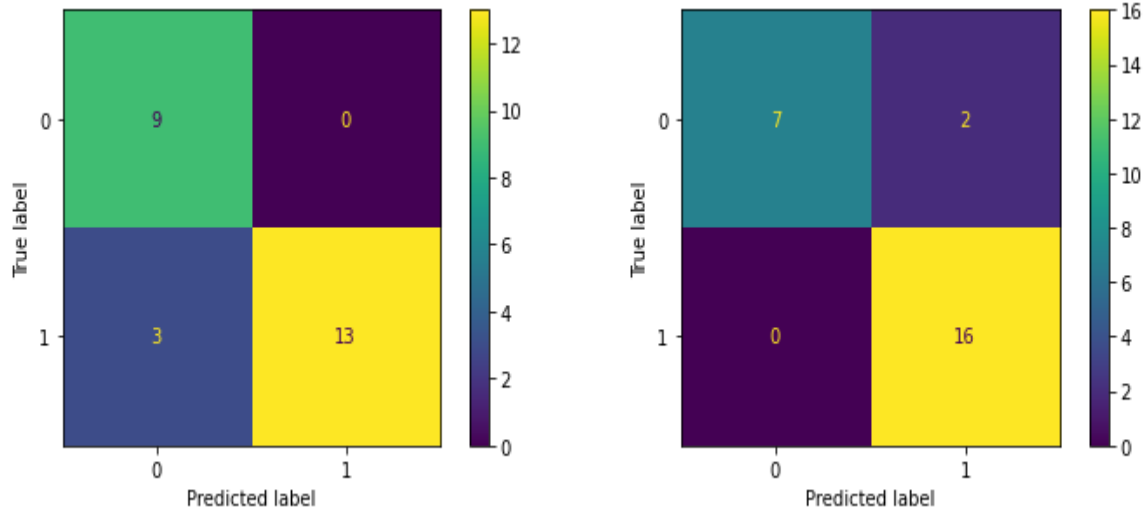
Engagement and Interest based on Moodle Logins



Logistic Regression

Decision Tree

Figures 3. The above figure shows the confusion matrices for Engagement and Interest Workload (No. of courses)

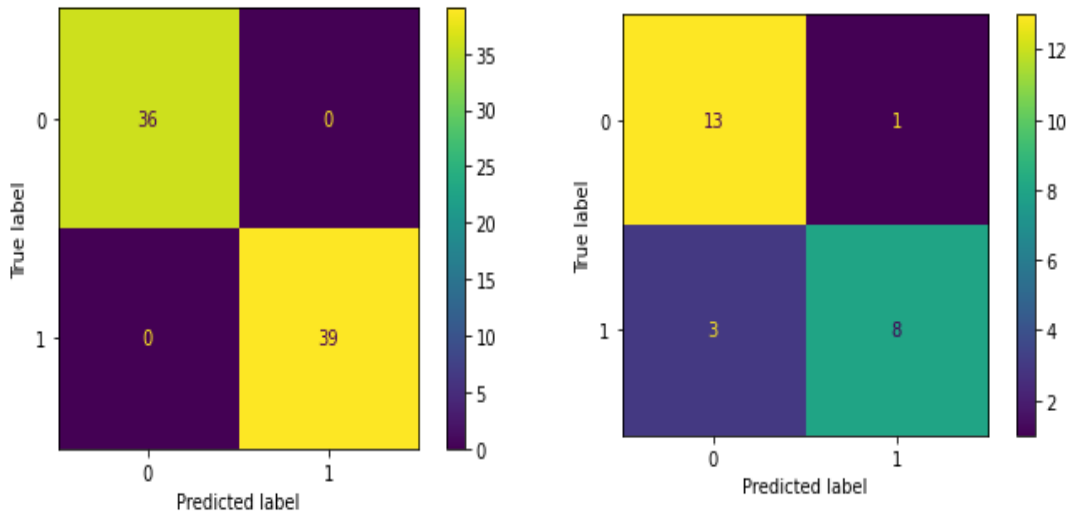


Logistic Regression

Decision Tree

Figures 4. The above figure shows the confusion matrices for workload

Minor Courses

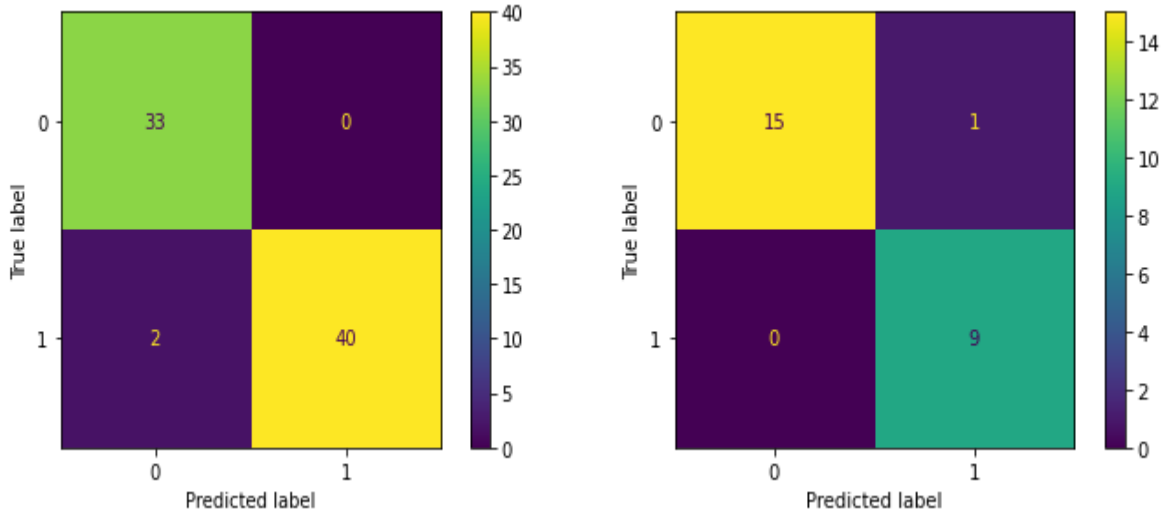


Logistic Regression

Decision Tree

Figures 5. The above figure shows the confusion matrices for Minor Courses

Gender

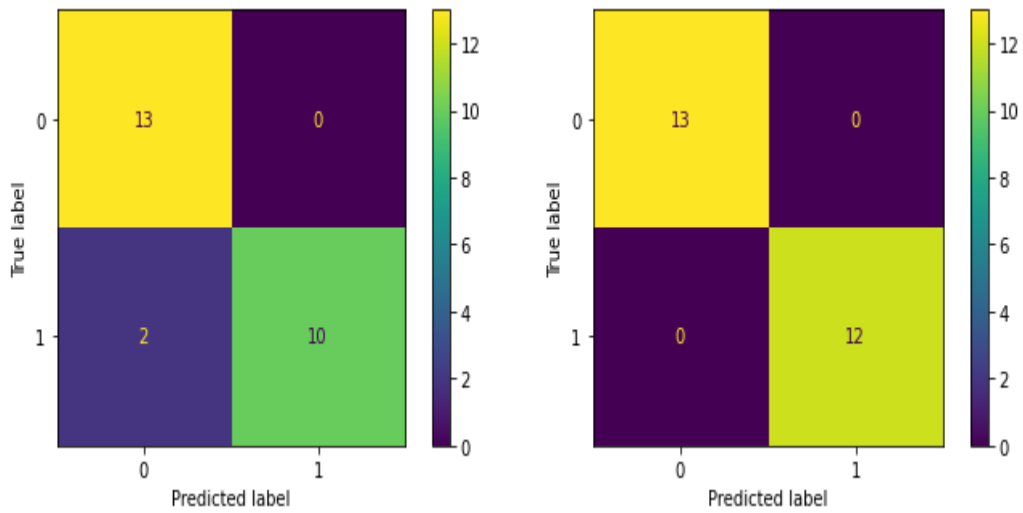


Logistic Regression

Decision Tree

Figures 6. The above figure shows the confusion matrices for Gender

Tuition Support

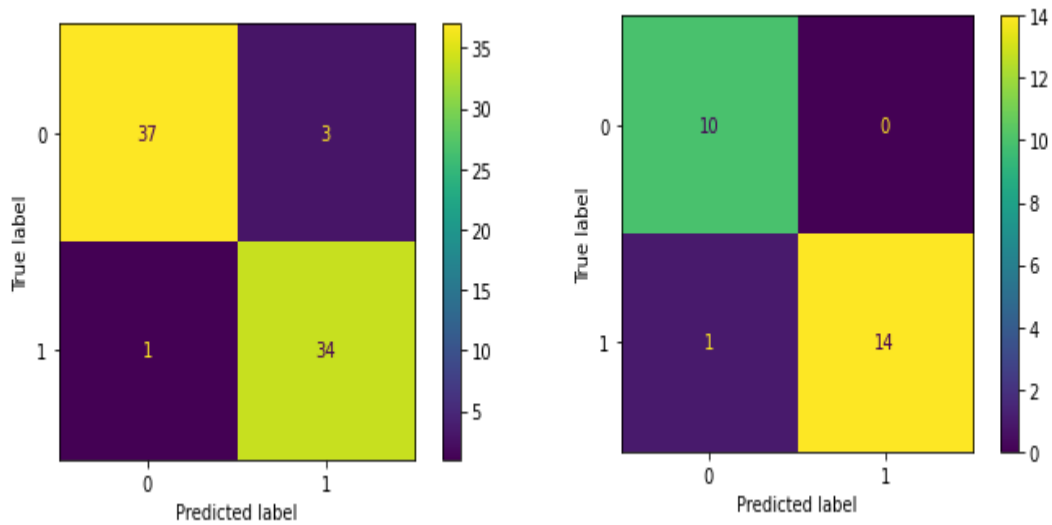


Logistic Regression

Decision Tree

Figures 7. The figure above shows the confusion matrices for Tuition Support

Institutional Aid



Logistic Regression

Decision Tree

Figures 8. The figure above shows the confusion matrices for Institutional Aid

CHAPTER SIX

6 RESULTS

This section presents the results that were obtained from all the research stakeholders using online interviews and questionnaires.

6.1 RESULTS FROM ICT 1110 STUDENTS RESPONSES

The questionnaire sought to identify factors that are more likely to be associated with the academic performance of the ICT 1110 students and enabled the research team to obtain reliable results.

The questionnaire was created and distributed online via Google Forms and was limited to the ICT 1110 students and the results obtained from 30 respondents were used to interpret the following responses.

The responses gathered via the online questionnaire were quantified and presented in Bar Chart form in association with a potential factor.

6.1.1 *Course Interest*

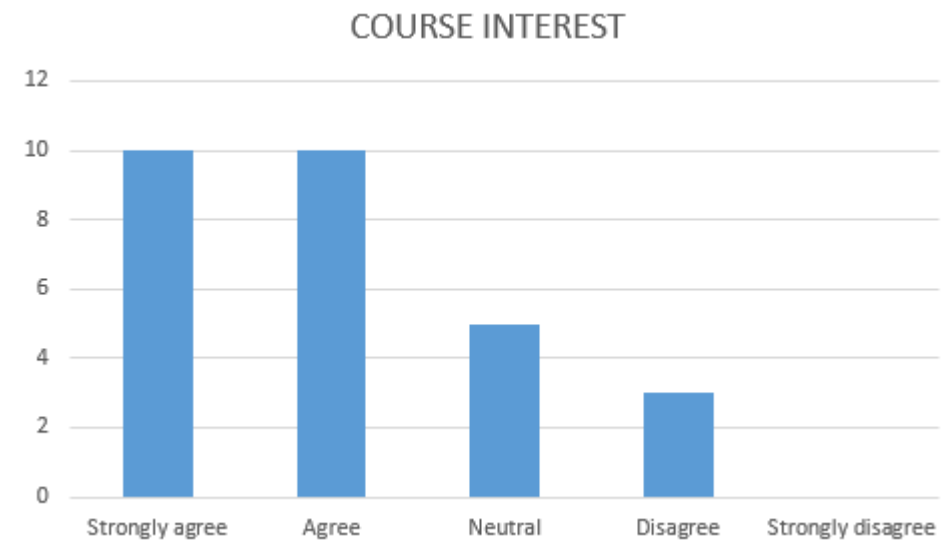


Figure 9: Student interest in course content

Figure 9 shows the number of responses on students' interest in the course content of ICT 1110. The results show that 10 students strongly agreed that interest in the course was associated to

academic performance, 10 students agreed, and 5 students were neutral while 3 students disagreed that course interest wasn't associated to the performance in ICT 1110.

6.1.2 Mode of Teaching

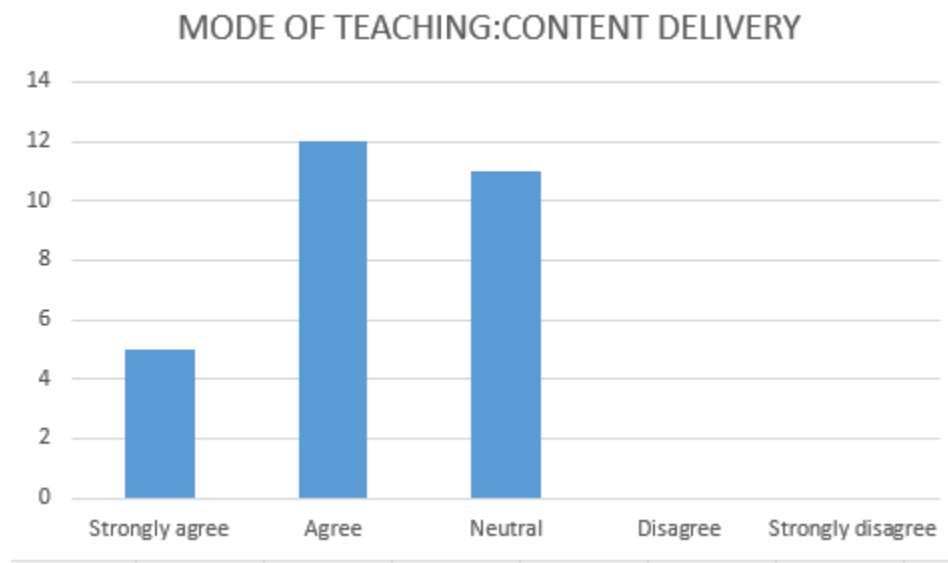


Figure 10: Mode of teaching in ICT 1110

Figure 10 indicates the results of student responses in regards to the mode of content delivery being associated to academic performance. The results indicate that 5 students strongly agreed and 12 students agreed that the mode of teaching contributed to their academic performance while 11 students were neutral.

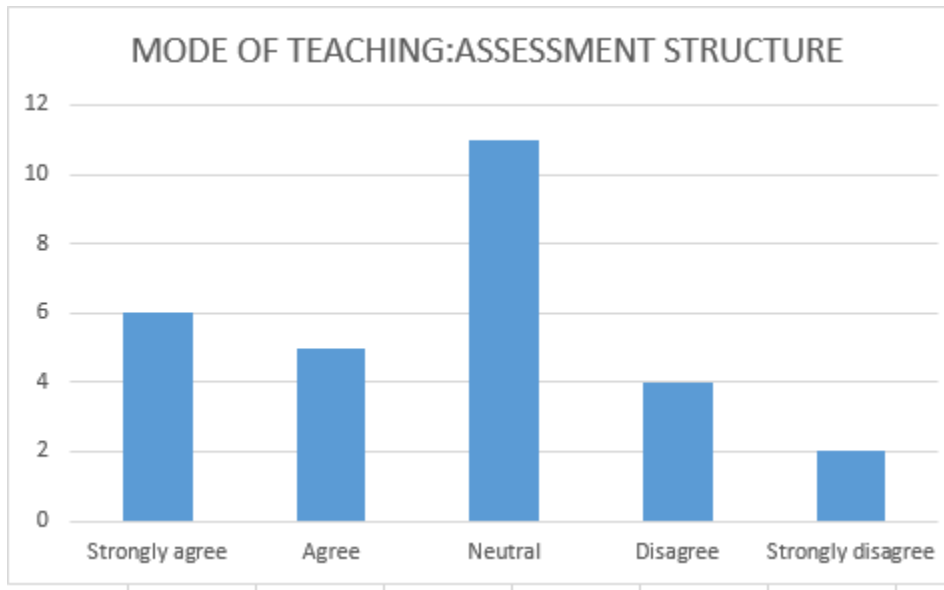


Figure 11: Assessment structure associated with academic performance

The results shown in figure 11 above indicates that 6 strongly agreed that the assessment structure in ICT 1110 affected their academic performance and 5 only agreed while 11 remain neutral and 4 disagreed while 2 strongly disagreed.

6.1.3 Prior Knowledge

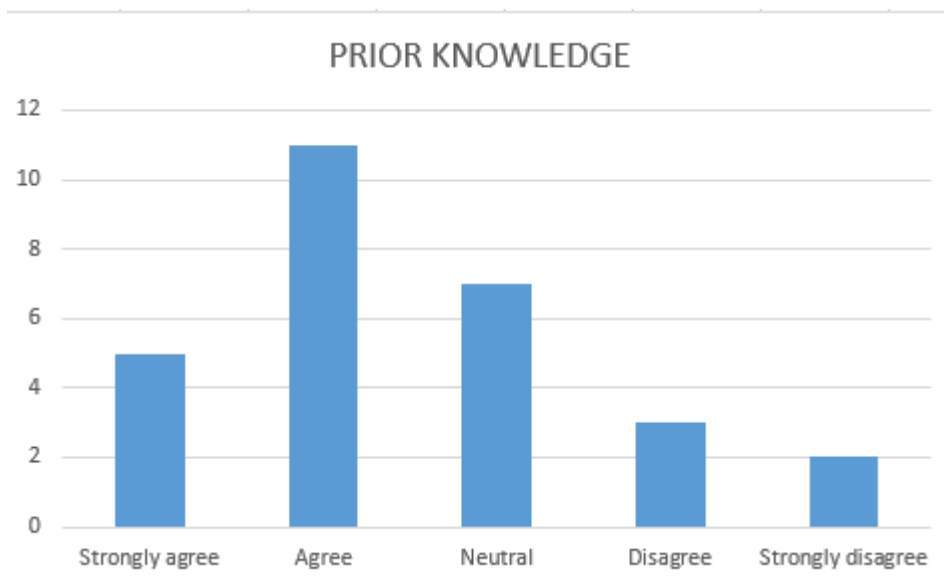


Figure 3: Student prior knowledge in Computer Studies

The figure above shows the results of the responses associated to students' prior knowledge in Computer Studies affecting their academic performance. The results indicate that 5 of the students strongly agreed and 11 students agreed while 7 were neutral and 3 disagreed and 2 of the students strongly disagreed that prior knowledge in computer studies affects their performance in ICT 1110.

6.1.4 Support Structures

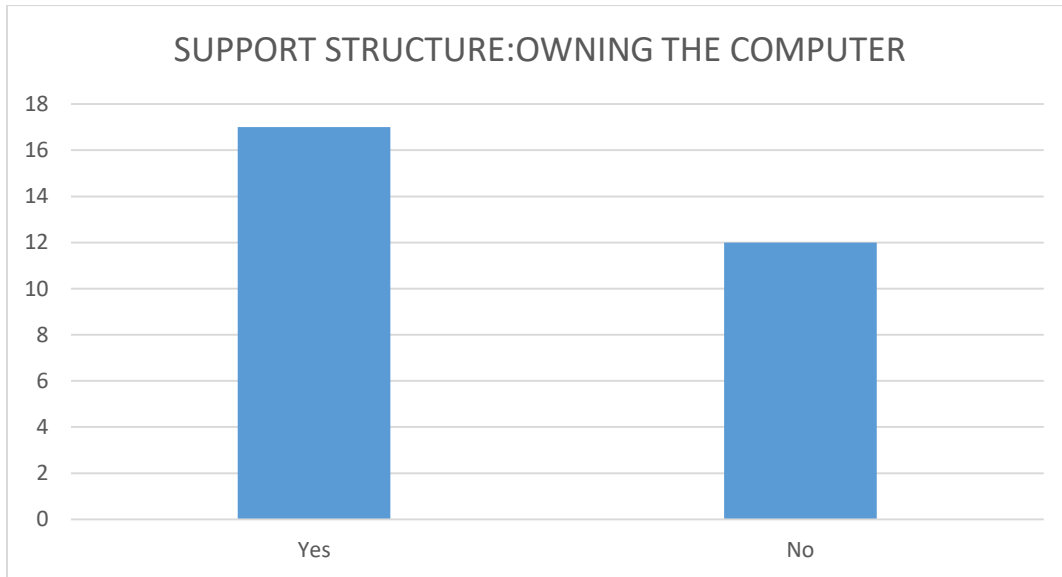


Figure 12: Owning of a Computer

The figure 12 above shows the number of students who own a computer. The results show that 17 of the students owned a computer while only 12 students did not possess computers.

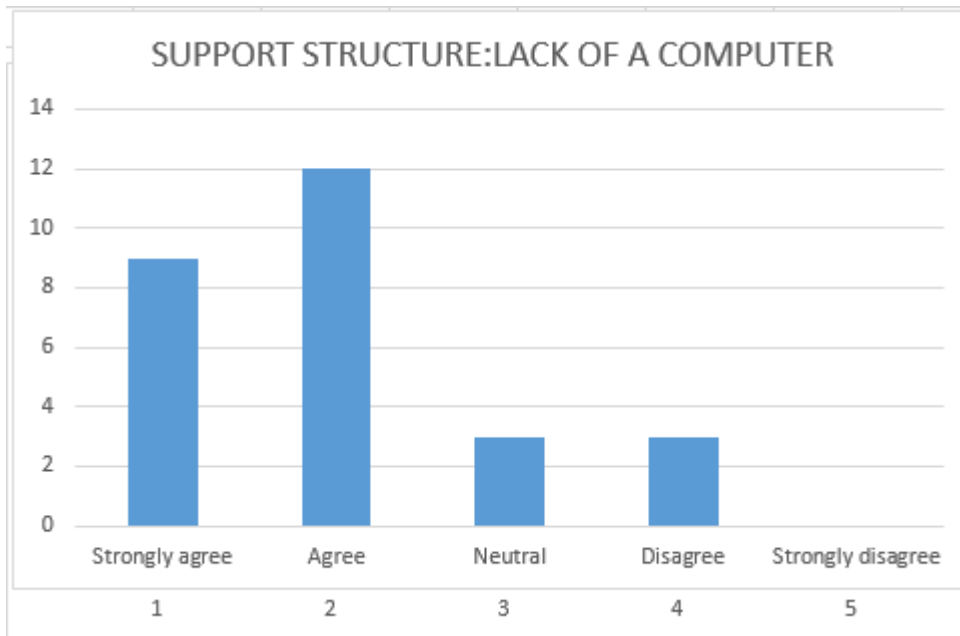


Figure 13: Lack of a computer affecting academic performance

Figure 13 shows the results with regards to academic performance in association with owning a computer. The results show that 9 of the students strongly agreed that lack of computers affected students' academic performance, another 12 agreed while 3 of the students were neutral and 3 disagreed that support structure does not affect student's performance.

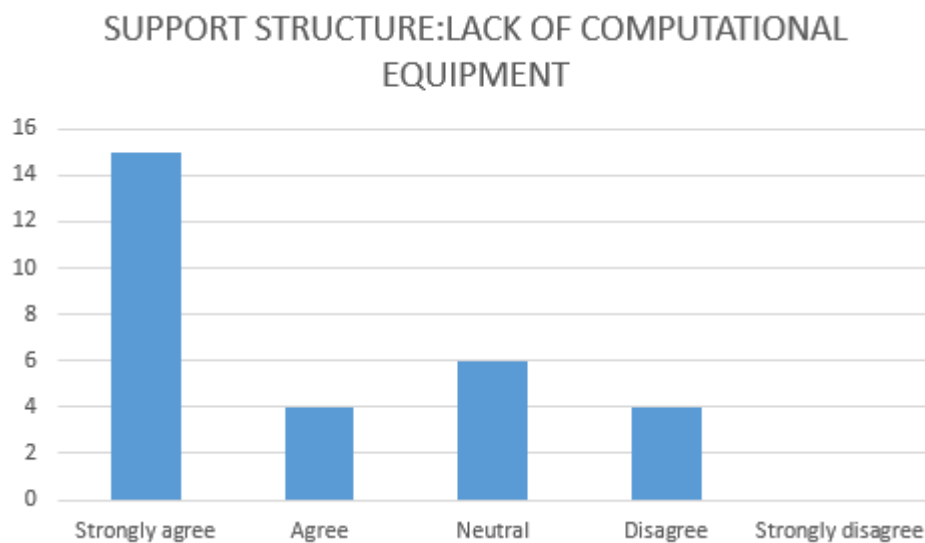


Figure 14: Lack of computational equipment

Figure 14 above shows the results of responses of students with regards to lack of computational equipment and resources affecting academic performance. The results thus indicated that 15 of the students strongly agreed that lack of computational material affected their academic performance and 4 merely agreed while 6 of the students remained neutral and 4 students disagreed.

6.1.5 Student Attendance

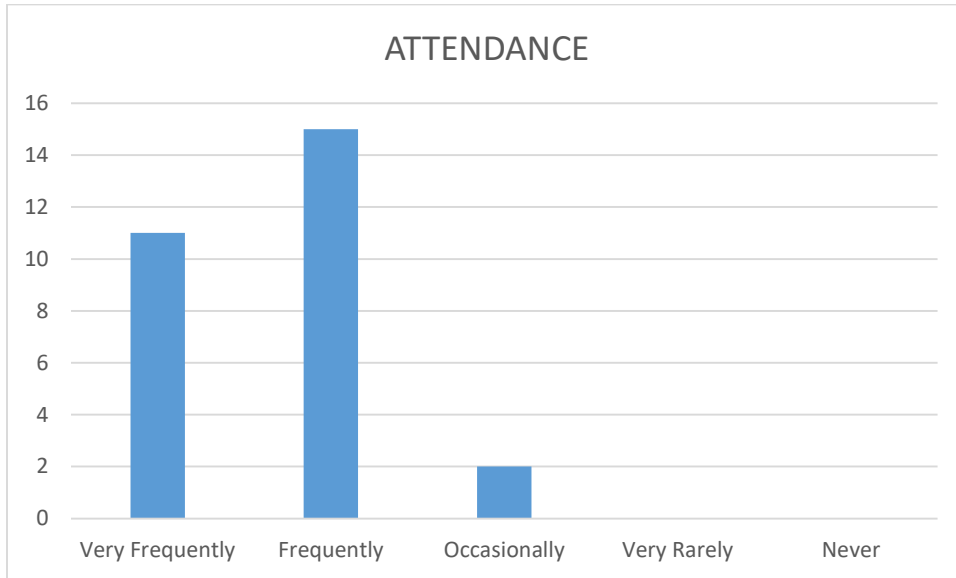


Figure 15: Student lecture and tutorial attendance

The figure 15 above shows the results of students who attended lectures and tutorials in ICT 1110. The results further indicate that 11 of the students attended lectures and tutorials very frequently and 15 attended frequently while 2 attended lectures and tutorials occasionally.

6.1.6 Student Motivation

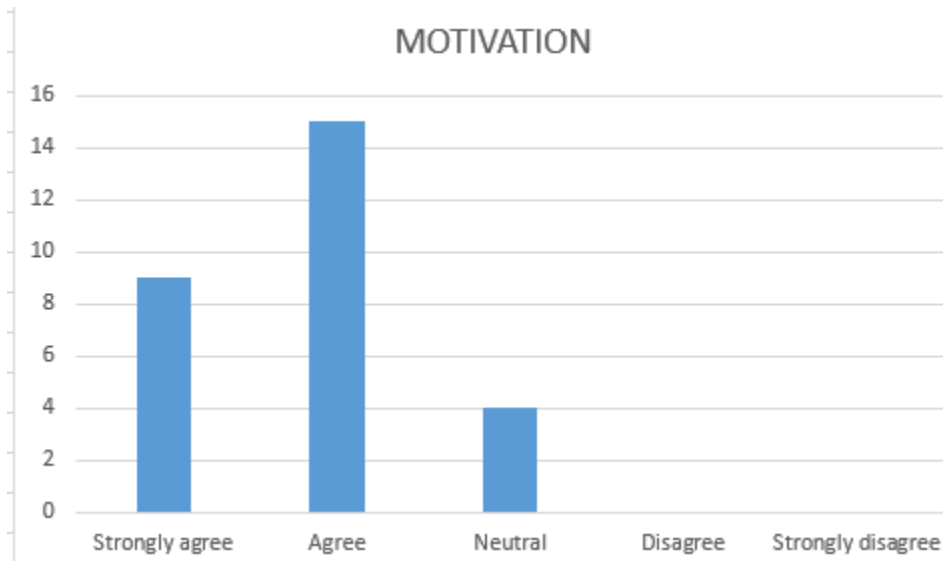


Figure 16: Student motivation in ICT 1110

The figure 16 above reveals the results of students' lack of motivation in association to academic performance in the ICT 1110 course. The information from the results show that 9 of the students strongly agreed with the suggestion that student motivation played a role in academic performance and 15 only agreed while 4 of the students remained neutral.

6.1.7 Time Management

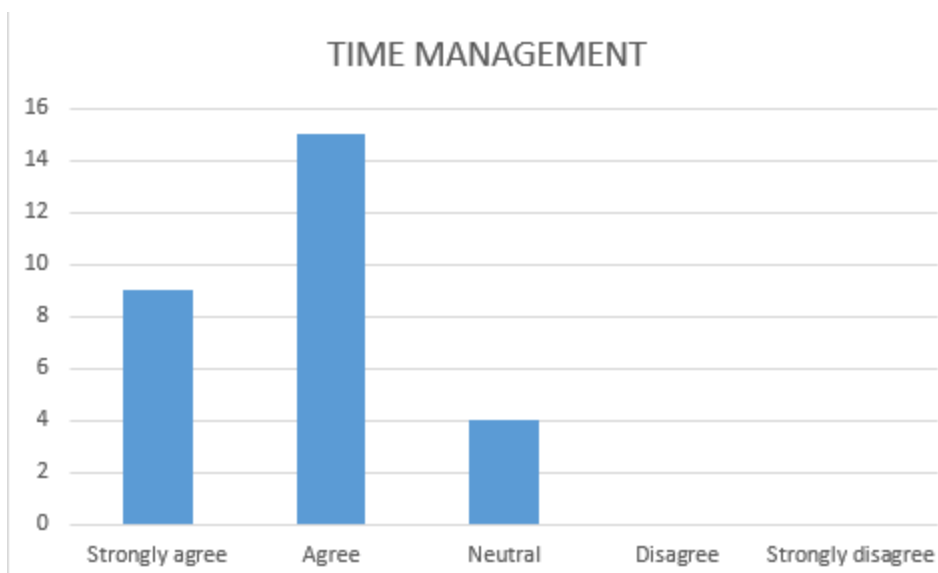


Figure 17: Lack of time management associated with academic performance

Figure 17 shows the results of students' responses associated to whether time management affected their academic performance or not. The results indicate that 9 of the students strongly agreed, 15 of the students agreed while 4 of the students remained neutral.

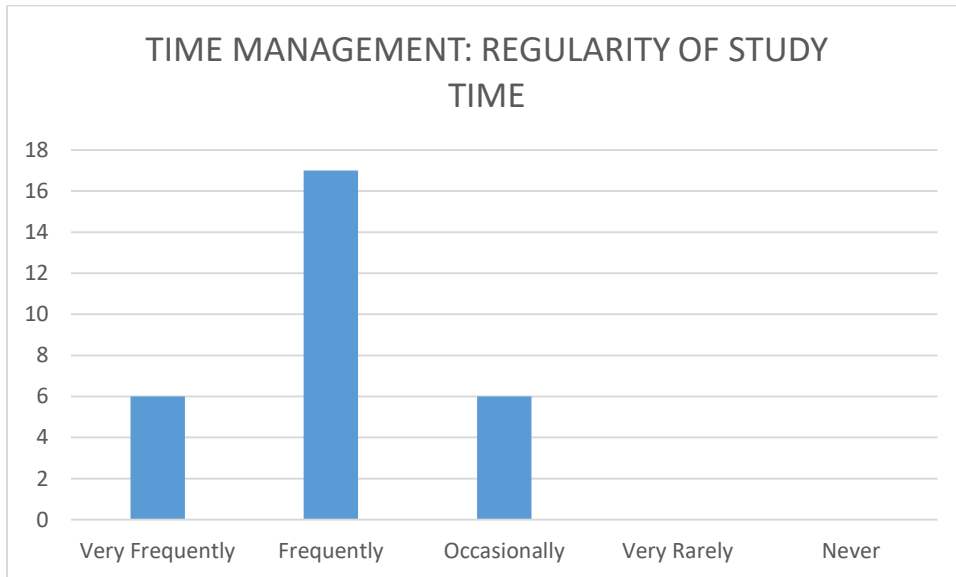


Figure 18: Regularity of study time

The above figure 18 shows the results of how frequently students study the ICT 1110 course. The evidence indicates that only 6 of the students studied the course very frequently while 17 of the students studied the course frequently and 6 of the students studied the course occasionally.

6.1.8 Minor Courses

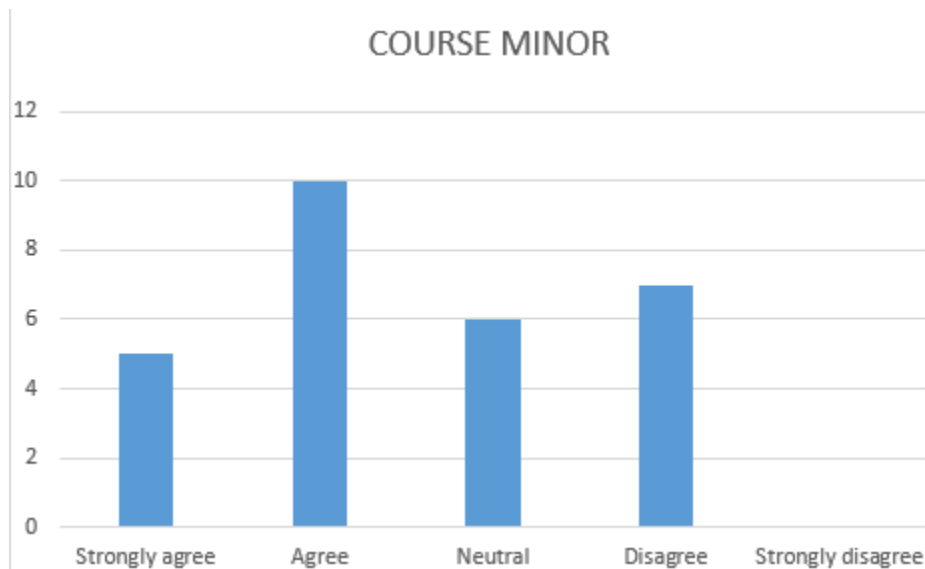


Figure 19: Students minor courses associated with academic performance

Figure 19 above shows results with regards to students' minor courses in association with their academic performance and the results reveal that 5 of the students strongly agreed that minor course workload affected their academic performance and 10 agreed while 6 were neutral and 7 disagreed.

6.1.9 Student Guidance

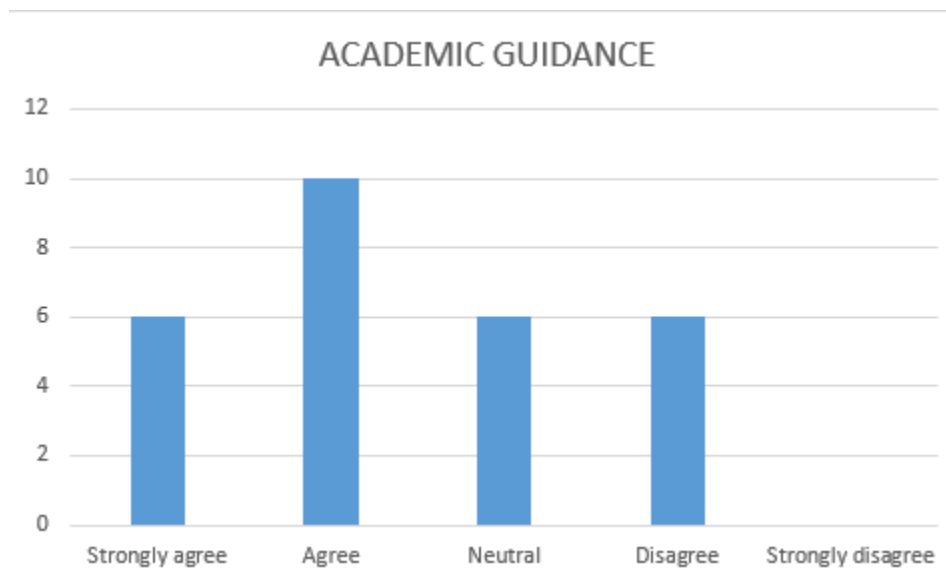


Figure 20: Student orientation in ICT 1110 in association to academic performance

Figure 20 shows the results of student responses with regards to whether lack of orientation in ICT 1110 affects their academic performance or not. The evidence of the results indicates that 6 of the students strongly agreed and 10 of the students only agreed while 6 were neutral and the other 6 disagreed that lack of orientation in ICT 1110 affects their academic performance.

3.2 RESPONSES FROM THE ICT 1110 LECTURERS AND TUTORS

The interview sought to acquire the possible factors that could be associated with the academic performance of the ICT 1110 students and enabled the research team to obtain reliable results from the educators themselves.

The interview was conducted online via Google Meets and two educators were interviewed being 1 lecturer and 1 tutor respectively.

The responses gathered via the online interview were quantified, summarized and presented in tabular form.

3.3 TABLE OF INTERVIEW QUESTIONS AND RESPECTIVE RESPONSES

Questions	Lecturer's response	Tutor's response
In Association with Attendance In the lecture sessions, be it online or face to face interactions, how would you describe the class attendance?	'Overall participation is really bad.'	'Mostly almost everyone tried by all means to make it.'
In Association with Participation How active are your students when it comes to participation and giving back feedback when questions are asked in class?	'Overall participation is really bad.'	'The overall participation was average.'
In Association with Performance Outcomes Relative to the lecture session you conduct and have conducted in ICT 1110, what could be some of the factors that influence the poor academic performance of the students?	Workload 'There is a correlation between the minor program that the student is pursuing and their overall performance. It's a factor that is correlated to an outcome.' Engagement 'The students that regularly access those	'Absenteeism, lack of interest'

	resources tend to perform better.'	
Support Structures How would you describe the efficiency of the computer labs at UNZA?		'In terms of the number of computers, the number of computers are enough but not all of them were working at the same time.'

Table 1 above shows the responses from the lecturer ad tutor of ICT 1110.

CHAPTER SEVEN

7 DISCUSSIONS

7.1 FEATURES ASSOCIATED WITH PERFORMANCE PREDICTION

The findings from the online interviews conducted among the ICT 1110 educators and online questionnaires distributed among the ICT 1110 students revealed the following factors that would make possible features for the project model:

- ***Student Interest***

The results showed that students' interest in the course content is likely to affect their academic performance. Students with less interest in the course content are more likely to fail the course than those with more interest.

- ***Mode of Teaching***

The results from the questionnaires revealed that the mode of course delivery and assessment structure affects the performance outcomes of the students. Familiarity and adaptation to lesson delivery and assessment structures was crucial to the students' academic performance.

- ***Prior Knowledge***

The results revealed that students' prior knowledge to any Computer based subject or program has an effect on students' academic performance.

- ***Motivation***

The evidence from the results revealed that students' motivation in the course content affects performance outcomes.

- ***Support Structures***

The results indicated that inadequate support structures such as computer labs, equipment, resources etc. affect the academic performance outcomes of students.

- ***Time Management***

The findings disclosed that lack of time management on the part of the students affects their academic performance for instance, the inability to plan ahead and stick to goals results in poor efficiency and performance.

- ***Course minor***

The findings showed that minor course among the students has effects on their academic performance in that attention is usually skewed to certain courses than others.

- ***Guidance***

The results showed that lack of orientation to the overall ICT 1110 course content affects the students' academic performance.

- ***Attendance***

The findings revealed that students' lecture and tutorial attendance affects the students' performance outcomes as that is the place where course content is delivered.

- ***Participation***

From the results obtained, student participation in the ICT 1110 course affects students' academic performance as participation in the course content indicates students learning progress and understanding.

- ***Engagement***

The results revealed that students' engagement with the course content e.g. accessing necessary resources from learning management system platforms like Moodle, looking up prescribed readings, accessing repositories for certain information, etc., affects the overall performance of students' academic outcomes.

The above stated factors are the potential factors that could contribute to the performance outcomes of the students. The factors gathered were closely analyzed in order to source data associated with the factors and there after come up with features whose data will be used for the prediction model. After analyzing the data and searching for data sources for each factor, it was concluded that the feasible factors that would serve as potential features for the project model would be as follows:

- ***Student Interest***

For this feature, data was drawn from how many times a student would log onto the UNZA Moodle learning management system platform. This would indicate whether or not the student is interested with the course content and its activities as majority of the resources of the course are found on the Moodle.

- ***Minor Courses***

Though elicited under workload, this feature was obtained from sources provided by the ICT 1110 course instructor that takes in to account the minor courses of the students. As

some minors may contain more academic work than others, this feature could potentially affect the performance outcomes of the students.

- ***Engagement***

In order to use this as a potential feature, the data that was sourced from the UNZA Moodle logins would be utilized in order to indicate how much students engaged with the course with regards to how many times they accessed resources and performed course activities.

Other notable features that were used in the project model outside the elicitation process that could also potentially affect academic outcomes are as follows:

- ***Gender***

This demographic feature served as potential factor as prior research reveals a correlation between student gender and performance outcomes. The data was sourced from student demographic details obtained from UNZA Student Information System (UNZA SIS) by the course lecturer.

- ***Workload (based on total number of courses)***

With this feature, the data was obtained from the course register that the course lecturer provided that indicated each students' total number of courses in an academic year. As the total number of courses differ, the course weight could potentially affect the academic performance of the students.

- ***Institutional Aid***

Whether or not a student is accommodated or not could serve as a factor affecting performance outcomes as distance and resource availability could be notable factors to consider when it comes to accommodation and performance outcomes. The data from this feature was sourced from student demographic details obtained by the lecturer.

- ***Tuition Support***

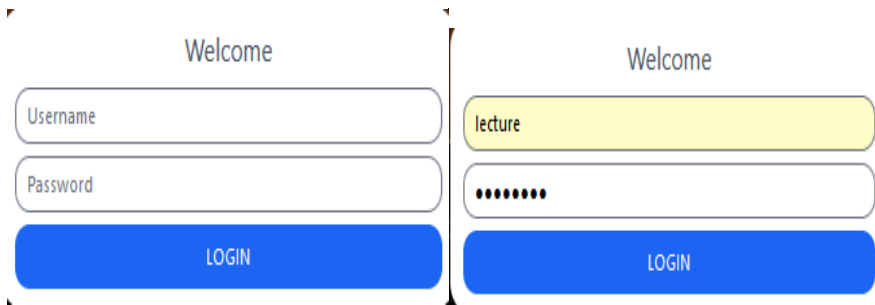
Whether or not a student is sponsored either by the government or any other sponsorship institution or organization could potentially affect their academic performance as lack of monetary resources could affect how a student accesses course information. The data obtained for this feature was sourced from student demographic details obtained from the course lecturer.

7.2 MODEL IMPLEMENTATION

The model was implemented as a web based application built using HTML, CSS, JavaScript and Python programming languages. The machine learning algorithm used was logistic regression and decision tree classifier.

The following screenshots show exactly how the model operates in order to predict students who are at risk of failing ICT 1110.

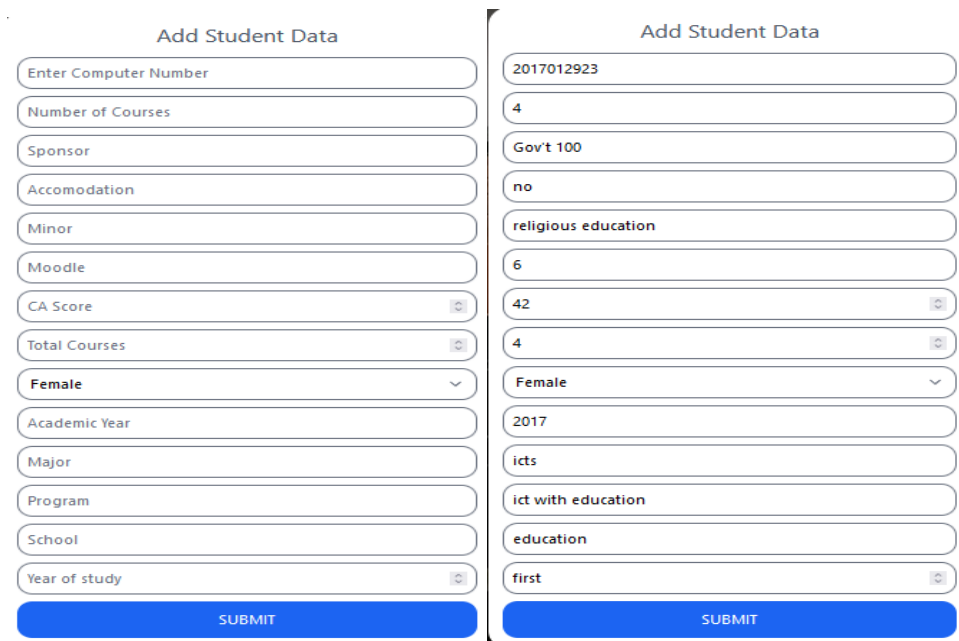
User Login



The image displays two side-by-side screenshots of a user login interface. Both screenshots feature a 'Welcome' heading at the top. The left screenshot shows the login form with empty input fields for 'Username' and 'Password', and a blue 'LOGIN' button at the bottom. The right screenshot shows the same form with the 'Username' field filled with the text 'lecture' and the 'Password' field filled with a series of dots, representing a masked password. The 'LOGIN' button is also present in this screenshot.

The screenshots above show the user login interface

Input field for student details



The image displays two side-by-side screenshots of an 'Add Student Data' form. The left screenshot shows the form with empty input fields for: Enter Computer Number, Number of Courses, Sponsor, Accomodation, Minor, Moodle, CA Score, Total Courses, a dropdown menu set to 'Female', Academic Year, Major, Program, School, and Year of study. A blue 'SUBMIT' button is at the bottom. The right screenshot shows the same form filled with the following data: 2017012923, 4, Gov't 100, no, religious education, 6, 42, 4, Female, 2017, icts, ict with education, education, and first. The 'SUBMIT' button is also present in this screenshot.

The screenshots above show the input field for entering student details

Prediction Interface

Student Results Prediction Model

2017012923

PREDICT RESULT

Predicted Results
Examination status:At risk

Logout

The screenshots above show the prediction interface that displays the prediction of a student's performance once user clicks the 'PREDICT RESULTS' button on the screen.

Input field for student details

Add Student Data

Add Student Data

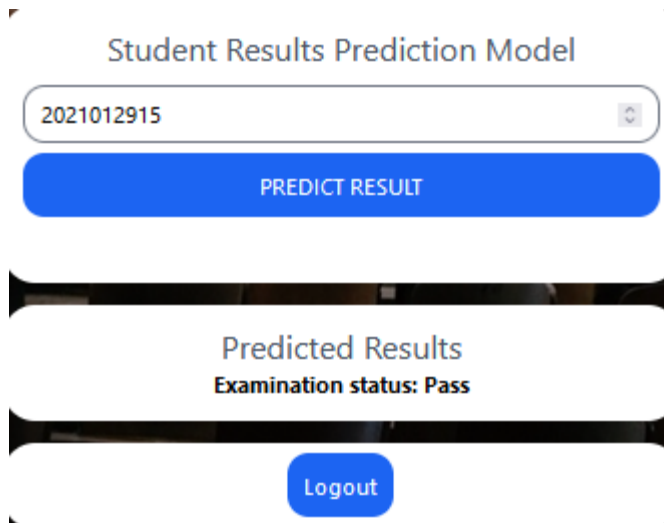
Field	Value
Enter Computer Number	2021012915
Number of Courses	4
Sponsor	govt 75
Accomodation	yes
Minor	history
Moodle	5
CA Score	50
Total Courses	4
Gender	Male
Academic Year	2021
Major	ict
Program	ict with education
School	education
Year of study	first

SUBMIT

SUBMIT

The screenshot above shows the input field for entering student details

Prediction Interface



The screenshot displays a web interface for a 'Student Results Prediction Model'. At the top, the title 'Student Results Prediction Model' is centered. Below it is a text input field containing the student ID '2021012915'. A blue button labeled 'PREDICT RESULT' is positioned below the input field. The interface then shows the predicted results: 'Examination status: Pass'. At the bottom, there is a blue 'Logout' button.

The screenshot above shows the prediction interface that displays the prediction of a student's performance once user clicks the 'PREDICT RESULT' button on the screen.

8 CONCLUSION

This study was based on identifying factors that contribute to students' low performance and building a model that was able to predict students who are at risk of failing ICT 1110 and require high intervention. The study was carried out using data collected from the ICT 1110 stakeholders which included lecturer, tutor and ICT 1110 first year students. Using a dataset of input features that include student interest, engagement, workload, minor, gender, tuition support and institutional aid, the study constructed a prediction model that was able to identify students who need high intervention with a reasonable degree of accuracy. The key observations from the experiments were that predictions for performance may be made with a reasonably good accuracy and that it is possible to determine a greatly reduced feature subset which can achieve predictions similar to using all the features. Overall, this study has shown that data collected from the stakeholders can be used to build data-driven intervention prediction models for academic performance. The aim of this model was to identify students who at risk of failing the ICT 1110 course so that the course instructors and educators can ascertain and execute appropriate and effective correction mechanisms to help at risk students achieve good performance and learning outcomes.

9 REFERENCES

Adam, S. (2004). **Using Learning Outcomes: A Consideration of the Nature, Role, Application and Implications for European Education of Employing Learning Outcomes at the Local, National and International Levels. Report on United Kingdom Bologna Seminar, July 2004.** Herriot-Watt University.

Agrawal, R. (1999). **Data Mining: Crossing the Chasm.** Presentation at International Conference on Knowledge Discovery and Data Mining, San Diego.

Alyahyan, E. and Düşteğör, D. (2020). Predicting academic Success in Higher Education: Literature Review and Best Practices. **International Journal of Educational Technology in Higher Education.**

Ballantine, J.H. and Hammck, F.M (2009). **The Sociology of Education: A systematic Analysis. 6th ed.** London: Pearson

Bilal Mehboob, Rao Muzamal Liaqat and Nazar Abbas Saqib. (2016). Predicting Student Performance and Risk Analysis by Using Data Mining Approach. **International Journal of Computer Science and Information Security (IJCSIS)**, vol. 14, No 7.

Brodley, C. E.; Smyth, P. (1995): **The Process of Applying Machine Learning Algorithms.** Presented at the Workshop on Applying Machine Learning in Practice, 12th International Machine Learning Conference (IMLC 95), Tahoe City, CA.

Cheng, B. and J. Atlee (2007). **Research Directions in Requirements Engineering. Future of Software Engineering (FOSE'07).** Washington, DC, USA.

Creswell, J. W. (2003). **Research Design: Qualitative, quantitative and mixed method approaches** (2nded.). California: Sage.

Dawson, C. (2002). **Practical Research Methods, A user-friendly guide to mastering research techniques and projects.** How to Books Ltd, 3 Newtec Place: United Kingdom.

Dorina Kababchieva. 2015. Student performance prediction using data mining classification algorithms. **International Journal of computer science and management research**, vol.1.

Edin Osmanbegovic and Mirza Suljic. 2015. Data mining approach for predicting student performance. **Journal of Economics and Business**, Issue 1.

Fahd, K., Miah, S.J. & Ahmed, K. (2021). **Predicting Student Performance in a Blended Learning Environment using Learning Management System Interaction Data. Applied Computing Informatics.** Emerald Publishing Limited.

Grljević, O and Bošnjak, Z. (1998). **Data understanding** (“CRISP-DM methodology Utilization in Preprocessing Small and Medium Sized Enterprises Data”), Book of proceedings of XXXV Symposium on OR, SYM-OP-IS 2008, ISBN: 978-86-7395-248-2, pp. 275-279, 2008.

Hasan, Palaniappan, Mahmood, Abbas, Sarker and Sattar. 2020. **Predicting student performance in higher educational institutions using video learning analytics and data mining techniques.** Applied sciences 10 (11), 3894.

Hickey, A. and A. Davis (2003). **Elicitation Technique Selection: How Do Experts Do It?** 11th IEEE International Requirements Engineering Conference. Monterey Beach, CA, USA.

John. Jacob, K. Jha. P Kotak and S. Puthran. 2016. Data mining techniques and their application. **International Journal of a computer science and information technologies**, vol. 5 (4).

Kennedy, D. (2007). **Writing and Using Learning Outcomes. A Practical Guide.** Quality Promotion Unit, UCC.

Kerlinger, M. (2007). **Research Methods Education and Social Sciences.** London: Edward Arnold.

Kothari, C. K. (2004). **Research Methodology; Methods and Techniques (2nd Edition).** New Age International Publishers: India.

Mahajan, M. & Singh M.K.S. (2017). Importance and Benefits of Learning Outcomes. **Journal of Humanities and Social Science**, Vol.22.

Namoun, A. & Alshantiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. **Applied science.**

Omari, I. M. (2011). **Concept and Methods in Educational Research: “A Practical Guide Based on Experience”.** Dar es Salaam: Oxford University Press.

Pyle D, (1999). **Data Preparation for Data Mining.** Morgan Kaufman Publisher Inc.

Rettig, M. (1994). "Prototyping for Tiny Fingers." Communications of the ACM 37(4): 21-27.

Shaw, M. (1990). **Prospects for an Engineering Discipline of Software**. IEEE Software, 7(6): 15-24

Shaymaa E, Tsunenori M, Kazumasa G, and S. H, .2014. Efficiency of LSA and K-means in predicting Students' Academic Performance Based on Their Comments Data. **6th International Conference on Computer Supported Education**, pp. 63-74, 2014.

Zowghi, D. and C. Coulin (2005). **Requirements Elicitation: A Survey of Techniques, Approaches, and Tools**. **Engineering and Managing Software Requirements**. A. Aurum and C. Wohlin. Berlin, Germany, Springer

10 APPENDICES

APPENDIX 1: INTERVIEW QUESTIONS FOR THE LECTURER

ICT 4014 PROJECT TEAM 1: REQUIREMENTS ELICITATION INTERVIEW GUIDE FOR ICT 1110 LECTURER

1. Introduction of the project team members and the project
2. Briefing of the interview
3. Welcoming the interviewee
4. Outline of what is intended to be obtained from the interviewee
5. Ethical and consent discussions
6. Questions to ask:
 - a. How many students do you lecture in this ICT 1110 cohort?
 - b. In the lecture sessions, be it online or face to face interactions, how would you describe the class attendance?
 - c. How active are your students when it comes to participation and giving back feedback when questions are asked in class?
 - d. Relative to the lecture session you conduct and have conducted in ICT 1110, what could be some of the factors that influence the poor academic performance of the students?
 - e. When you assign to them an assessment be it in the form of quizzes or tests, have you experienced instances where students fail to submit assessments?
 - f. If your answer above was a yes, what have been the main reasons why?
 - g. Have you been able to predict students who are at risk of failing?
 - h. If your answer was yes, how have you been able to do this?
 - i. How would a performance prediction model be of benefit to you?
 - j. What type of input would you prefer to use to find out students who are at risk of failing?
 - k. What way would you prefer to notify students who are at risk of failing so that appropriate correction mechanisms can be executed?
7. Conclusion

APPENDIX 2: INTERVIEW QUESTIONS FOR THE TUTOR

ICT 4014 PROJECT TEAM 1: REQUIREMENTS ELICITATION INTERVIEW GUIDE FOR ICT 1110 TUTORS

1. Introduction of the project team members and the project
2. Briefing of the interview
3. Welcoming the interviewee
4. Outline of what is intended to be obtained from the interviewee
5. Ethical and consent discussions
6. Questions to ask:
 - a. How many students do you tutor in this ICT 1110 cohort?
 - b. Relative to the tutorial sessions you conduct in ICT 1110, what do you think could be the factors that influence poor performance outcomes?
 - c. How would you describe the class attendance?
 - d. How active are your students when it comes to participating in computer lab session activities?
 - e. When you assign to them activities, have you experienced instances where students fail to attempt the activities?
 - f. If your answer above was a yes, what have been the main reasons why?
 - g. How would you describe the efficiency of the computer labs at UNZA?
 - h. How would you describe the students' familiarity with the computer software in the lab sessions?
 - i. Have you been able to predict students who are at risk of failing?
 - j. If your answer was yes, how have you been able to do this?
 - k. How would a performance prediction model be of benefit to you?
 - l. What type of input would you prefer to use to find out students who are at risk of failing?
 - m. What way would you prefer to notify students who are at risk of failing so that appropriate correction mechanisms can be executed?
7. Conclusion

APPENDIX 3: QUESTIONNAIRE FOR THE STUDENTS

QUESTIONNAIRE

University of Zambia

School of Education

Department of Library and Information Science

Questionnaire: Performance predictor

Dear respondent,

We are fourth year students in the Department of Library and Information Science and we are developing a software which will be able to predict students who are at risk of failing ICT 1110. You have been conveniently selected to participate in this study and please take time to go through it before responding to the question on this questionnaire, and be assured that any information given to us will be treated with confidentiality and your response will be used specifically for academic purposes.

If you have any queries regarding this questionnaire, please contact

Cell No.: 0979025025

Your cooperation will be greatly appreciated.

Yours faithfully,

Project Manager:

Mutune Chaibela

Instructions

- Please give your answer either by ticking in the box or by writing on the blank space where applicable.
- Please note that try as much as possible to make your answers accurate and specific, as they will make our research easy.

1. What is your gender?

Male []

Female []

2. Does your interest in ICT 1110 play a major role in relation to your performance?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

3. Does the mode of teaching in ICT 1110 contribute to your academic performance?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

4. Do you think student's prior knowledge in computer studies does have an impact on their performances in ICT 1110?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

5. Do you own a computer?

Yes []

No []

6. Does lack of owning a computer affect student performance in ICT 1110?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

7. How often do you attend lectures and tutorials?

Very frequently []

Frequently []

Occasionally []

Very rarely []

Never []

8. How often do you study ICT 1110?

Very frequently []

Frequently []

Occasionally []

Very rarely []

Never []

9. Do you think lack of motivation in ICT 1110 does play a role in relation to their performances?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

10. Does lack of computational equipment and resources affect your academic performance?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

11. Does lack of time management affect your performance in ICT 1110?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

12. Does the assessment structure in ICT 1110 affect your academic performance?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

13. Does having a lot of courses (minor) affect your academic performance in ICT 1110?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

14. Does lack of orientations in ICT 1110 course affect your academic performance?

Strongly agree []

Agree []

Neutral []

Disagree []

Strongly disagree []

Thank you for participating in our survey