ADVANCED SCHOOL OF SYSTEMS AND DATA STUDIES (ASSDAS)

DEPARTMENT OF COMPUTER SCIENCE

TOPIC

ACADEMIC PERFORMANCE PREDICTION SYSTEM

BY

ADA GLORY

(ADS21B00235Y)

THESIS SUBMITTED TO THE ACCRA INSTITUTE OF TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT OF THE DEGREE IN COMPUTER SCIENCE

DATE

OCTOBER, 2022

ABSTRACT

This project is written to develop a prediction system that can help predict students' future performance using their past educational records. The research was carried out despite several challenges such time constraints, financial constraints, and the inability to obtain a lot of information that may have been useful. Even yet, the researcher managed to produce something that is understandable and helpful to anybody who reads this thesis.

SIGN COPYRIGHT/DECLARATION OF AUTHORSHIP

This is to declare, that the under-mentioned student under the supervision of the under-mentioned supervisor has carried out the research work underlying this thesis. Both the students and supervisor certify that the work documented in this thesis is the output of the research conducted by the student as part of his final year project work in partial fulfillment of the requirements of Bachelor of Science in Computer Science Degree.

(Student)
ADA GLORY

_____
Signature

Date: _____

(Supervisor)
MR. DAVID BOTWE

_____
Signature

Date: _____

DEDICATION

This thesis is dedicated to the almighty, our lord, provider of knowledge and wisdom and protector of mankind, for his unrepeated love towards humanity, and for his continuous governance in the affairs of men.

ACKNOWLEDGEMENTS

In a special way I give glory to God Almighty, the creator of all things visible and invisible, for his guidance and protection all through the duration of this program. I owe a great debt of gratitude to all who have combined efforts with us in order to complete this project firstly my supervisor, MR. DAVID BOTWE. He was the guiding force and prime inspiration to lift us from the initialization stage to the successful completion of the project. His friendly guidance and discussions made completing this project possible, May God bless you and your family. I specially want to thank our parents for their support and encouragement throughout the years. It is with immense gratitude I appreciate all who made this program a reality morally and financially. God bless you all.

TABLE OF CONTENT

ABSTRACT ……..…….……i

SIGN COPYRIGHT/ DECLARATION OF AUTHORSHIP ….…….……ii

DEDICATION …….……...…….iii

ACKNOWNLEDGEMENT ……..…….iv

TABLE OF CONTENTS…….…..…….……v

LIST OF TABLES ……..…….……x

LIST OF FIGURES ……..…...……xi

CHAPTER ONE: ……..……1

GENERAL INTRODUCTION AND SUMMARY …….…….....…….1

1.0 **Introduction…….…….……...1**

1.1 **Field and Subject Area of Study.……...……...…….1**

1.2 **Purpose of Research.…….…….……1**

1.3 **Objectives of the study.…….……...…….…..2**

1.4 **Problem Statement....…….…….……....2**

1.4.1 Prior Works and Study.…….…….……...3

1.4.2 Focus of the Study.…….…...…….…….5

1.5 Background and Justification of Study.…….…..5

1.6 Scope of the study.…….…...…...6

1.7 Research Methodology.……...……..…….7

1.8 Expected Outcome of the Research.……...…….....7

1.9 Research Implementation Schedule.……...……....7

CHAPTER TWO: ……8

LITERATURE REVIEW.……...…….…8

2.0 Introduction.…….....8

2.1 Definition and Description of Terms.……8

2.1.1 Data Collection.……9

2.1.2 Data Preprocessing and cleaning.……...9

2.1.3 Features Encoding.……10

2.1.4 Features Scaling.……10

2.1.5 Dataset Splitting.……10

2.1.6 Regression.……...10

2.1.7 Support Vector Machine.……...10

2.1.8 Random Forest.……...11

2.1.9 K-Nearest Neighbors (KNN).……11

2.1.10 Artificial Neural Network (ANN).……...11

2.1.11 Linear Regression (LR).…….……..11

2.2 The Web Based Prediction System…….…….12

2.3 Empirical Review.……12

2.3.1 The Impact of Using Different Features on the Performance of the Regressions Models.……14

2.3.2 Discussion.……...14

2.3.3 Features Importance.……15

2.3.4 Experimental Settings.……..…….15

2.4 Conclusion.…….……15

CHAPTER THREE: ……..……17

RESEARCH METHODOLOGY……..17

3.0 Introduction.……..……....17

3.1 Design and Development Process of Academic Prediction System……...18

3.1.1 Defining and Understanding the Problem.……18

3.1.2 Requirement Analysis of the entire System.……..18

3.1.3 Design and Development of the System.…….18

3.1.4 Testing and Evaluation.…….……...19

3.1.5 Implementation and Integration.……19

3.1.6 Maintenances.……..…..19

3.2 System Analysis Materials to Be Used.……19

3.2.1 Survey as method of data collection.……...…20

3.2.2 Interview as a method of data Collection.……..21

3.3 System Design and Implementation Methodology.……22

3.4 System Security….…….....23

3.5 Conclusion.……...…..23

CHAPTER FOUR: ……...…….…25

SYSTEM ANALYSIS AND DESIGN.……...…..……...25

4.0 Introduction.....25

4.1 Overview of the Proposed System.....25

4.2 General Requirements.....26

4.2.1 Functional Requirements for Computerized Product Distribution Monitory System….……....26

4.2.1.1 Input Design.....26

4.2.1.2 Output Design.....26

4.2.1.3 Logical Design.....26

4.2.2 Non-Functional Requirements for a Computerized Product Distribution Monitory

System.....27

4.2.3 Software Requirement Specification.....28

4.2.4 Hardware Requirement Specification.....28

4.3 System Design.....29

4.3.1 Feasibility Study.....29

4.3.1.1 Technical Feasibility.....30

4.3.1.2 Economic Feasibility.....30

4.3.1.3 Operational Feasibility.....31

4.3.1.4 Schedule Feasibility.....31

4.3.1.5 Resource Feasibility.....31

4.4 Machine Learning.……...……..32

4.5 Data Flow Diagram (DFD).....33

4.5.1 Context Dataflow Diagram.....34

4.6 Use Case Design.....34

4.6.1 Use Case Diagram for a Computerized Product Distribution Monitory System...34

4.6.2 Use Case Diagram.....34

4.7 Entity Relationship Diagram for a Computerized Product Distribution Monitory System.……..35

4.7.1 Entity Relationship Diagram.....36

4.8 Flowchart Design of Computerized Product Distribution Monitory System.....37

4.8.1 Flowchart Symbols.....38

4.8.2 Flowchart Diagram.....38

4.8.3 Sequence Diagram.....40

4.8.4 Sequence Diagram of the Academic Performance System....…….…….....41

4.9 UML Activity Diagram.…….....42

4.10 System Implementation.....43

4.10.1 System Testing.....44

4.10.2 Software Testing Strategies.....44

4.10.3 System Security.....44

4.11 Conclusion.....45

CHAPTER FIVE.……46

FINDINGS, CONCLUSION AND RECOMMENDATIONS….……46

5.0 Introduction.....46

5.1 User Interface.....46

5.2 Programming Languages.....47

5.2.1 Why Using Python 47

5.3 Implementation 48

5.4 Conclusion of the Study.....48

5.5 Maintenance.....49

5.6 Implications.....49

5.7 Future Works and Recommendation.....49

5.8 Conclusion.....50

REFERENCES.....51

APPENDICES.....54

List of Tables

Table 4.1 Software Requirement Specification.....25

Table 4.2 Hardware Requirement Specification.....26

Table 4.3 Feasibility Study.....29

Table 4.5 Data Flow Diagram Symbols.……33

Table 4.8.1 Flowchart Symbols.....38

List of Figures

Figure 2.3 Empirical Review.....15

Figure 4.4 Machine Learning Phase.....32

Figure 4.5.1 Context Dataflow Diagram.....34

Figure 4.6.2 Use Case Diagram.....34

Figure 4.7.1 Entity Relational Diagram1.....36

Figure 4.7.1 Entity Relational Diagram2.....36

Figure 4.8.2A Flowchart Model Diagram.....37

Figure 4.8.2B Flowchart Prediction System Diagram.……38

Figure 4.8.4 Sequence Diagram.....39

Figure 4.9 UML Activity Diagram.…….……...…41

CHAPTER ONE

INTRODUCTION

1.0 Background of Study

Education facilitates the learning or acquisition of knowledge, skills, values, morals, beliefs, habits, and self-development. Educational methods include education, training, storytelling, discussions, and directed questions. Education often takes place under the direction of educators; However, students can also train themselves. Education can take place in formal or informal settings, and any experience that has a formative effect on the way one thinks, feels, or acts can be considered educational. The teaching methodology is called pedagogy.

Formal education is generally formally divided into levels such as preschool or kindergarten, elementary school, high school, and then college, university, or education. In most regions, schooling is compulsory up to a certain age.

Education is a very important aspect of Ghana's Government and it takes an expenditure of 3.99% of its GDP over 80perecent of its children population are enrolled in kindergarten and primary schools as cited by its ministry of education. The Ghana Education system is broken down into several levels kindergarten, Primary Secondary (Junior and High), and tertiary. Although there is a higher percentage of students enrolled in kindergarten to secondary than that of tertiary the percentage of students enrolled for tertiary education in Ghana is about 18.88% and just about 10% out of this 18.88% end up graduating. The purpose of this research is to predict student performance using their past records thereby giving insights on how to retain students and also increase the number of enrollments and graduation in tertiary education.

1.1 FIELD AND SUBJECT AREA OF STUDY

The Field of study for this research is Data Mining to be more Precise Educational Data Mining which is interdisciplinary of Computer Science and Statistics.

1.2 PURPOSE OF RESEARCH

The purpose of this research is to develop a prediction system that can help predict students' future performance using their past educational records

1.3 OBJECTIVE OF STUDY

Main Objective

The main objective of this research is to design and implement an academic performance prediction system

Specific Objective

- Create a web-based prediction system using educational data mining
- To allow this system to give future insight into students' academic performance
- Allow schools to make informed decisions from this prediction in order to retain and also improve the learning process of students

1.4 PROBLEM STATEMENT

As written in the introductory statement the percentage of students who graduate from secondary school and then enrolls in tertiary schools in Ghana is about 18.88% but not this percentage follow through and graduate. This is as a result of different factors ranging from financial problems, social problems, and most importantly to this research difficulty to pull through and understand and as a result dropping out of school.

The Main demographic targeted for this research is the tertiary level because of its low percentage enrollment ratio and its even lower graduation ratio.

This problem is prevalent in different countries in the world not just in Ghana, but advanced countries like the USA and China have started tackling this problem by using educational data mining to predict students' performance at the point of enrollment, during studies, and so on in order to make informed decisions and also assist the students to make an informed decision. The tertiary education system is a highly competitive system where average and less average student can get overwhelmed.

Predicting student performance can help predict weak students and help tertiary institutions management make strategies and decisions related to improving student performance. The institution management can take steps at an early stage to prevent failure or dropouts. Knowing

the prediction, the weak student is also expected to improve his performance and get a better score.

1.4.1 PRIOR WORK AND STUDY

There are prior works in regards to predicting student academic performance but we will be taking an in-depth look at the work of Predicting Academic Performance in the School of Computing & Information Technology (SCIT) at the University of Technology Jamaica. There are different criteria used in screening students before giving them admissions and the UTJ uses in the School of Computing & Information Technology, the main admission criteria are appropriate scores in an aptitude test and pass in at least five Caribbean Examination Council subjects including Mathematics and English. The research carried out by the school of computing and information technology examined relationships between students' demographics, aptitude test scores, performance in first-year courses, their overall performance in the program, and so on.

The School of Computing and Information Technology (SCIT) at Jamaica University of Technology (UTJ) has had more than seven hundred (700) applicants per year for its programs over the past few years. Of that number, about two hundred and twenty-five are accepted. The faculty offers two courses, a Bachelor of Computing and Information Technology (BSc IT) and a Bachelor of Computing with Management Studies (BCMS). The BSc IT title is the more popular of the two and attracts around two-thirds of applicants. The selection of students for both programs includes the evaluation of the results of the tests of the Caribbean Examination Council (CXC) and/or the General Certificate Ordinary Level (GCE O'Level) of the applicants, as well as the score obtained in the test of SCIT suitability. The current enrollment requirements for the programs are five (5) CXC or GCE-O`Level subjects, including mathematics and English, and preferably one natural science subject. With its application to admission ratio approximately 3:1 question was raised about the admission criteria for the program, are they selecting crème de la crème and to what extent does the admission decision variable affect students' academic performance, in addition, the failure rate in their Introduction to Programming (ITP) course has risen from 30% to just over 40%. This is a cause concern as ITP is a foundation course and a prerequisite for several other programming courses including Programming using C, Data Structure using C, and Advance Programming using Java. Although specific actions have been

taken to address the course failure rates problems the question of whether or not it is possible to predict which student will perform well on the programs has persisted. And they asked questions like: If we can identify the factors that indicate which students are more likely to succeed in the programs then we will be able to optimize the selection process? They sampled 96 students, 68 males, and 28 females. To matriculate to the SCIT programs students are required to take an aptitude test and to have passed at least five subjects at CXC and/or GCE O'Level as earlier stated in their admission criteria. The subjects should include English Language, Mathematics and preferably a science subject. Fifty-five (57%) of the matriculated candidates passed between 7 and 10 O'Levels and 35 (approximately 37%) had between 4 and 6 subjects, just about six (6%) had between 1 and 3 subjects. the majority of the matriculated students did not have A 'Levels. Sixty-four (66.7%) of the students had no passes at A 'Levels, while twelve (12.5%) had one subject pass and another twelve (12.5%) had two subject's passes. Regarding the GPA only four students (4.2%) received a first-class honor, 18 students or (approximately 19%) received upper second class, 48 students or (50%) received lower second-class honors and three students or (3%) received a pass. Twenty-three students (24%) discontinued. There were no failures. We examined a number of predictor variables; gender, scores in the aptitude test, passes in Chemistry, Physics, Math's, Additional Math's, GCE Advanced Level (A 'Level) Math, Accounting, Information Technology at both the O'Level and A' Level, and the number of CXC, O'Level and A' Level passes. We also included passes in first-year first-semester computing courses; Introduction to Programming (ITP), and Computer Logics and Digital Design (CLDD). The dependent variable in this study was the student's final year GPA. GPA was divided into five categories, first-class (3.45 – 4), upper second class (3.05 – 3.44), lower second class (2.40 – 3.04), pass (1.70 – 2.39), and fail (0 –1.69). All the data that was gathered was analyzed using stepwise multiple regression analysis. The stepwise approach eliminates variables already in the model that are no longer significant predictors. Three models were used to determine the prediction and each model was used with different predictors. The first model determined to what extent the aptitude test, and science subjects (Chemistry and Physics at O'Level and Physics at A 'Level related to academic performance. The results indicated that neither of these factors had any effects on GPA, as all were removed from the stepwise analysis. The second model examined the predictive value of Mathematics (both O'Level and A 'Level), Additional Mathematics, Principles of Accounting (both O and A-level), and Information Technology (both

O and A levels) the result indicated that they were weak predictors for academic performance The third model examined the number of A-Level subjects passed and the number of O-Level subjected passed. The expectation is that people who have passed more O and A-level subjects would perform better than those who did not. This made them make informed decisions in the admission criteria and better improve how to classify the students in the two courses the school had to offer.

1.4.2 FOCUS OF THE STUDY

In an article written by the Univar newsroom on March 9, 2021, the president of Ghana Akufo Addo stated that the government's target was to increase enrollment for tertiary institutions from 18.88% to 40% by 2030 and this research aims at helping tertiary institutions make informed decisions about the admission process and several other things to improve students' performance thereby increasing its retention rate also.

This research tries to answer the following questions: "Can we predict the performance of students throughout their stay in tertiary institutions, using only their grades from secondary school without any socioeconomic factors. "and" can we identify freshman and second-year courses that are effective predictors for students' performance at the end of the degree?". From an administrative point of view, it is easier to gather the marks of students than their socio-economic data. Therefore, if a reasonable prediction can be reached without socioeconomic data, it makes the implementation of the performance support system in tertiary institutions easier."

1.5 BACKGROUND AND JUSTIFICATION OF STUDY

The enormous growth of electronic educational data provides the opportunity to extract information that can be used to predict overall student success, predict student dropout rate, evaluate teacher and instructor performance, evaluate learning material according to the needs of students to improve and much more. This research aims to examine the latest trends in predicting student achievement in higher education and then apply and improve the methods used in mining these data. Higher education plays an integral role in the development of society (Pinheiro et al. 2015), so increasing academic success is a long-term goal of academic institutions.

Understanding and defining academic success is important for increasing student success rates. The definition of academic success is very complex and meaningful. Therefore, it is often

misused in education and research. The study by York et al. (2015). which is often measured either through course evaluations or institutional surveys, persistence, which is measured by staying between certain academic years and graduation rates, acquisition of skills and competencies that can be measured by assignments, and course evaluations, Achievement of learning goals, which can also be measured by course assignments and evaluations, and finally Professional success, which can be determined by success rates, track record promotion, job satisfaction and occupations - goal achievement. A second key requirement for maximizing academic success is identifying the factors that impact academic achievement. Awareness of student success factors could help achieve the highest level of education (Yassin et al. 2017). Potentially, it can help provide a clear and meaningful description of the types of knowledge and behavior associated with proper performance. This awareness can be obtained by using data mining (DM) methods through educational materials. The practice of DM methods applied to educational data is known as Educational Data Mining (EDM) (Baker and Yosef 2009). It comes from a variety of fields including DM and machine learning, psycho-metrics, and other areas of statistics, information visualization, and computational modeling (Romero and Ventura 2007). In general, EDM refers to techniques and tools designed to automatically extract useful information and patterns from huge stores of data related to human learning activities in an educational setting (Nithya et al. 2016). These tools use machine learning algorithms, database systems, statistical analysis, and artificial intelligence. DM techniques include regression, clustering, class identification, association, and prediction.

The use of DM in education systems is presented as an iterative cycle of training, testing, and refinement of hypotheses, whereby systems can be adapted to support the specific needs of each participant in the educational process.

1.6 SCOPE OF THE STUDY

This project focuses on designing and developing an academic performance prediction system using Educational Data Mining. It will be limited to articles, documents, research, and journal papers only.

1.7 RESEARCH METHODOLOGY

For the purpose of this research, we will be analyzing two or more data mining techniques that are suitable for predicting academic performance and then choosing the best possible options. And will be mentioning the two most probable approach.

These techniques include

- The first approach is based on classification and regression algorithms that search for patterns in study-related data and also data about students' social behavior.
- The second approach is based on collaborative filtering techniques. We predict the final grades based on previous achievements of similar students

1.8 EXPECTED OUTCOME OF RESULT

This research aims at helping schools (tertiary institutions) predict student performance by using data they already have thereby giving them the power to make informed decisions that will better enhance student performance and increase the retention rate on students.

1.9 THE RESEARCH IMPLEMENTATION SCHEDULE

The research work would follow the timeline as provided below: -

- The first chapter would be developed in 2 weeks
- The second chapter will be developed in 2 weeks
- The third chapter will be developed in 3 weeks
- The fourth chapter would be developed in 3 weeks
- The fifth chapter will be developed in 2 weeks

CHAPTER TWO

LITERATURE REVIEW

2.0 INTRODUCTION

This chapter will be based on a review and discussion of relevant literature in connection with the research on Data Mining to be more Precise Educational Data Mining. I will carry out a censorious analysis of several components of the published body of Knowledge out through a summary, comparison of prior research studies, articles, findings, and observation of their limitation, advantages, and recommendation.

For the purpose of this study, the literature review is focused on functionalities, processes, and several components of the reviewed systems. The technology used and the functionalities provided by the system are the main focus when selecting the system for systematic review and comparison among the selected systems. The purpose of this literature review is to study the requirement specification, weaknesses, and strengths of selected systems. After considering these systems, ideas can be generated and implemented in the proposed system to overcome the weaknesses of this type of technology.

Besides this, functions that are commonly used among the selected systems could be selected as functions of the proposed system as well. The technology used by the selected systems is also studied to further enhance the proposed system.

2.1 DEFINITION AND DESCRIPTION OF TERMS

2.1.1 Data Collection

The dataset will be collected from the Accra Institute of Technology (AIT) and other systems. The data contains numeric data, which are the students' course scores, and categorical data which are course names and the city. Each course has different columns depending on the course's nature. Some columns were shared in all courses, such as the "mid-term" column. Other columns had been exceptional among courses, such as "participation", "workshop", and "discussion". These columns had been blended in a single column called "mid activities". Moreover, the dataset includes demographic facts approximately the scholars accumulated thru a questionnaire, which are:

Marital Status Data: Includes whether the student is married and has children. Health data: Includes whether the student is suffering from any medical condition. Residence data: Includes the city in which the student lives.

We have ten features; ten features are inputs and the last one is the output. The feature descriptions include the feature name, the description, and the domain. The domain means the numeric values of each feature. In features marital status, children, and health, “1” means “Yes” and “0” means “No”, while in level and class the domain of the feature is the range between specific numbers.

2.1.2 Data Pre-processing and Data Cleaning

Pre-processing is an important step in data mining. The purpose of data preprocessing is to convert the data into a suitable form that can be used by algorithms. Three main pre-processing steps have been applied to the dataset which are data cleaning, features encoding, and features scaling. The pre-processing was implemented using Python language and Microsoft Excel.

Data in the real world is often noisy and unstructured. The data cleaning step aims to fix anomalies in the data by filling in missing values and smoothing out noise when finding outliers. Handling Missing Values: In our dataset, several students did not fill out the questionnaire which led to missing values in the demographic data columns.

2.1.3 Features Encoding

In Machine Learning models, all inputs and outputs are required to be numerical variables. Therefore, when there is categorical data, it must be encoded before using it in the model, and this is called features encoding. So, encoding is the pre-processing of the categorical data when working on a model of ML algorithms. There are several techniques to encode categorical.

features, such as Label Encoding, One-Hot Encoding, Frequency Encoding, Ordinal Encoding, Binary Encoding, Hash Encoding, and Mean/Target Encoding.

In our dataset, we have two categorical features, namely “course” and “city”. The course feature contains the abbreviation for the course names, and the city feature is the name of the city where the student currently lives. To encode these two features into numeric features, we started by applying Label Encoder and One-Hot Encoding techniques to determine which technique suits our dataset. Label Encoder gave us the lower MAPE, so we used it in our dataset.

2.1.4 Features Scaling

In our dataset, we have two categorical features, namely “course” and “city”. The course feature contains the abbreviation for the course names, and the city feature is the name of the city where the student currently lives. To encode these two features into numeric features, we started by applying Label Encoder and One-Hot Encoding techniques to determine which technique suits our dataset. Label Encoder gave us the lower MAPE, so we used it in our dataset.

2.1.5 Dataset Splitting

The dataset is split into training datasets and testing datasets. The training dataset is used to build the model and the testing dataset is used to evaluate the model. The model has been validated by using different validation techniques which are: percentage split and cross-validation. The validation technique that gave better results in terms of prediction’s MAPE has been chosen.

2.1.6 Regression

Regression techniques are used to predict continuous results rather than predicting individual class labels. The purpose of this study is to predict the student's overall score. Therefore, the prediction problem is considered a regression problem. Compare different algorithms (that is, SVM, RF, KNN, ANN, and LR) to determine the appropriate algorithm for predicting the total student score for each course. Regression model-based predictions take features as input, run processes on them, and predict total scores as output. We then used metrics to compare the performance of all regression models. The algorithm with the lowest predicted MAPE is used in the prediction system.

2.1.7 Support Vector Machines (SVM)

SVMs are one of the most popular supervised ML algorithms used primarily for classification tasks (Bithari, T.B 2020). It was developed by Vapnick in 1995 to deal with the problem of prediction and pattern recognition and for analyzing and mapping both linear and nonlinear functions.

The basic idea of how the SVM algorithm works is to find the hyperplane that can separate data belonging to two classes with a maximum margin, and builds a hyperplane, or a group of hyperplanes (classes) in a high dimensional space. It classifies the object into categories, above or

lower plane depending on the features of the object and by using kernel techniques, it can convert nonlinear to linear before partitioning (Adejo O.W 2018).

2.1.8 Random Forest (RF)

RF is a supervised learning algorithm that was developed by Leo Breiman in 2001. RF is a form of ML method based on the aggregation of many decision trees used for regression and classification. In general, the more trees in the forest, the more robust the prediction.

RF has become a standard tool for data analysis, making it an effective tool for prediction (Akar, Ö 2012). Decision tree regression predicts a number as the output of specific data. RF regression receives the average of these predictions as the "final" output.

2.1.9 K-Nearest Neighbors (KNN)

The KNN method was first explained in the early 1950s. KNN is a supervised ML algorithm that can be used to solve problems such as classification and regression. KNN is purely lazy and maintains the entire training set. Ferring all inductive generalization attempts before regression time (Han, J. 2011).

2.1.10 Artificial Neural Network (ANN)

McCulloch's work on simulating the biological nervous system culminated in the development of the ANN method in the 1940s. An NN is a set of interconnected units whose properties are determined by the topology of the network and the properties of the neurons. NN is one of the most widely used and efficient learning systems today. ANN learns, trains, and changes in the same way that humans do in the brain.

2.1.11 Linear Regression (LR)

Regression methods are used to describe the relationship between response and explanatory variables. LR is one of the simplest ML algorithms for supervised learning techniques and is used to solve regression problems. LR is used to predict continuous dependent variables using independent variables. When predicting using one independent variable, it is called a single LR, and when there are three or more independent variables, it is called multiple LRs.

2.2 THE WEB BASED PREDICTION SYSTEM

The main goal of developing a student academic performance prediction system is to enable decision makers and university advisors to easily and quickly predict a student's overall score. After training the model using various algorithms: SVM, RF, KNN, ANN, and LR. The RF algorithm has the lowest percentage of MAPE and was chosen as the basic algorithm for website prediction systems.

This system allows users to enter student scores and demographics and use predictive models to predict the student's overall score. The table will then display a report containing all the information about the student (that is, general and demographic information, expected total score, expected grades). You can also print this report.

At the start of the system flow, we need a condition to call two predictive models and handle missing value errors.

The system operates under the following conditions:

1. If the values of the mid activities and lab scores are null, then the first prediction model is called, which is based on the midterm exam score only from academic data. Then it is transferred to the report template.
2. If the value of the mid activities score is null and the value of the lab score is not null, the user is warned that they must be entered together.
3. If the value of the mid activities score is not null and the value of the lab is null, the user is warned that they must be entered together.
4. If all the data are entered, then the second prediction model is called based on all three-academic data. Then it is transferred to the report template.

2.3 EMPIRICAL REVIEW

Machine Learning models were developed using five algorithms, namely SVM, RF, KNN, ANN, and LR. In the experiments, Adjusting the hyper-parameters of each algorithm is out of the scope.

. Parameter settings of the algorithms.

Algorithm	Parameters	Value
SVM	kernel	linear
	epsilon	0.1
	cost	1.0
RF	random_state	0
	n_estimators	500
KNN	n_neighbors	5
	weights	uniform
	algorithm	auto
	leaf_size	1
	Power	4
ANN	random_state	0
	hidden_layer_sizes	100
LR	intercept	True

Figure 2.3 empirical review

Performance Metrics

The following performance metrics were used to determine whether the model is accurate or not

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is one of the most common performance metrics. It is used to calculate the prediction error of the model. The MAE measures the average magnitude of the errors in a set of predictions.

MAE is given by the following Equation (1):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$$

where Yi represents the actual value, Yiˆ represents the predicted value of Yi, and N represents the number of instances.

Mean Absolute Percentage Error (MAPE)

The MAPE is calculated as follow.

$$MAPE = 100 * (MAE/y_i)$$

where MAE represents the value of MAE and Yi represents the actual value.

2.3.1 The Impact of Using Different Features on the Performance of the Regressions Models

The objective of this research is to know the best features that affect students' academic performance, through which it is possible to predict the total score of students early, and take necessary measures when failures are discovered. Hence, two scenarios in which to investigate the effect of the features set on model performance were set. The first scenario when using the first dataset (DS1), which contains only 8 features, the midterm exam score and all demographic features (i.e., course, level, mid-term, marital status, children, health situation, living place, and class which is the total score for students). The second scenario when using the second dataset (DS2), which contains all the academic and demographic features (i.e., course, level, mid-term, mid activities, lab, marital status, children, health situation, living place, and class which is the total score for students).

Scenario 1: Midterm Exam Score Only and All Demographic Features

In the first scenario, predict the total score by only one of the academic features, which is the midterm exam score, and all demographic features.

Scenario 2: All-Academic and Demographic Features

The second scenario predicts the total score by using three academic features: midterm exam, mid activities, lab scores, and all demographic features.

2.3.2 Discussion

Two scenarios were conducted to find out the effect of a set of features on the performance of the regression's models. It can be noticed from Scenario 1 and Scenario 2 that the performance of the regression models that are using the DS2 achieve lower MAPE scores. The reason for this is the increase in the number of important features used for prediction.

2.3.3 Features Importance.

The dataset may contain many features, but some may not affect students' academic performance. Hence, it is recommended to know the most important features. The purpose of this is to identify the features that have a significant impact on students' academic performance and through which it is possible to discover whether the student will fail. After training the prediction model, the RF calculated the feature's importance by ranking the importance of the features. Features ranking aims to sort the features regarding the index's quality that reflects the individual relevance of a feature [**31**]. Ranking techniques were used to rank features, and it assigned rank values to each feature according to its influence.

2.3.4 Experimental Settings

Machine Learning models were developed using five algorithms, namely SVM, RF, KNN, ANN, and LR. In the experiments, Adjusting the hyperparameters of each algorithm is out of the scope.

2.4 CONCLUSION

This project aims to develop a Machine Learning model to predict the students' academic performance at the early stages based on the course level. The students' data were collected from the Accra Institute of Technology (AIT) containing academic and demographic factors. Several algorithms (SVM, RF, KNN, ANN, and LR) were applied to predict students' total scores. The RF algorithm obtained the lowest MAPE and then it was adopted in the Web-Based prediction system. Academic factors had a higher impact on students' academic performance. Other features such as marital status, children, health, and the living place had little effect on students' academic performance.

The model can be used as an early warning mechanism to identify and improve the performance of at-risk students. With the help of Machine Learning methods, tutors are in a position to know which of their students will complete a course with the lowest error rate and which student should be admitted into the school. In addition, creating a web page helps tutors to enter data easily, and print a report with academic students' performance. Moreover, this paper contributed to considering a new factor that has not been studied in previous studies, which is living place

(city). In future work, the system can be developed to add more features such as the academic history of the students, the effect of the English Language on the students' academic performance, and frequent absence.

CHAPTER THREE

RESEARCH METHODOLOGY

3.0 INTRODUCTION

This chapter takes into consideration the various findings made in the review of the existing system. Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises a collection of procedures, techniques, tools and documentation aid, which will help system development in an effort to implement a new system. System analysis and design will be performed to determine if the proposed system is feasible to design and implement and will meet user requirements in ways that eliminate the weakness of the existing system.

This project aims to build a Machine Learning model to research Data Mining to be more Precise Educational Data Mining (EDM) which is interdisciplinary of Computer Science and Statistics with specified requirements using python libraries. This involved researching the Data Collection, Data Pre-processing and cleaning, and the use of different Algorithm models such as the SVM, RF, KNN, ANN, and LR. The chosen approach is to integrate

all the necessary modules to achieve a result of proper prediction of Student performances, (written in python) and uses of various machine learning libraries.

To address the research questions in the current study, we used a mixed-method approach. First, a systematic review of the literature on current EDM techniques used to profile SRL was carried out. The review followed five steps of the systematic review methodology (Khan et al., 2003). The review stages included (a) framing the research questions, (b) identifying relevant literature, (c) setting the articles' assessment criteria, (d) presenting review results, and (e) discussing the results. This review formed the foundation for the second study which involved an experimental evaluation of EDM algorithms to establish the optimal algorithm to identify SRL profiles from a dataset obtained from the Open University in the UK.

Finally, correlation analysis was used to identify the association between the SRL profiles and students' academic performance.

The articles reviewed in this study were repeatedly searched in international journals and databases, including Google Scholar, SCOPUS, Science Direct, Elsevier, ERIC, IEEE Xplore,

and ACM Digital Libraries. We searched for articles using keywords: "educational data mining techniques" and "learner analysis" and "measurement of self-adjusting learning" and "evaluation of self-adjusting learning" and "clickstream data" and "student behavior". And "online learning" and "self-regulation". A total of 742 papers have been identified. Read the full text of each article, Khan et al., 48 articles have been deleted after applying the inclusion criteria described in 2003. Twenty-four articles, twelve journal articles, and twelve conference articles met the selection criteria.

3.1 DESIGN AND DEVELOPMENT PROCESS OF ACADEMIC PREDICTION SYSTEM

To design the prediction system, we looked into current thoughts and research on the best possible development process approach, and these approaches are listed below.

3.1.1 Defining and Understanding the Problem.

A problem statement is a brief description of the problem this project is trying to address. A problem statement identifies the current state, the desired future state, and the gap between the two. A problem statement is an important communication tool that helps ensure that everyone involved in a project understands the issues to be addressed and why this project is important. And the problem definition has been stated clearly in chapter 1 of this research paper.

3.1.2 Requirement Analysis of the entire System.

Once the problem definition and understanding are complete, the next step is to define and document the requirements for the way the system will work. This is done through an SRS (Software Requirements Specification) document that contains all the product requirements that need to be designed and developed during the life cycle of the project.

3.1.3 Design and Development of the System.

In this Phase, the SRS document created is transformed into a more logical structure so that it can later be implemented in a programming language. Operations, training, and maintenance plans are all created so that when developing we know what to do at each subsequent stage of the cycle. The development phase is where the actual codes are written and the application is built according to previous design documents and outlined specifications. Different programming

language are usually used to develop a system for this system we will be using python to achieve the machine learning part and HTML, CSS and JavaScript to achieve the frontend.

3.1.4 Testing and Evaluation

Evaluation includes performing functional testing such as unit testing, code quality testing, integration testing, system testing, security testing, performance testing, acceptance testing, and non-functional testing. If an error is detected, the developer will be notified. A verified (actual) error is fixed and a new version of the software is created. The best way to ensure that all tests run regularly is to implement automated tests. Continuous integration tools support this need.

3.1.5 Implementation and Integration.

After testing, the overall design of the software is put together. Various modules or designs are integrated into the main source code through the efforts of developers. A training environment is typically used to detect further errors and defects. Information systems are integrated into the environment and eventually installed. After this phase, the software is theoretically ready for the market and available to all end users.

3.1.6 Maintenance.

In the maintenance mode we practice the necessary activities to resolve issues reported by end users. Additionally, the developer is responsible for implementing any necessary changes to the software after deployment. This may include addressing remaining bugs that could not be patched prior to release or resolving new issues that surfaced based on user reports (Synops 2015). Larger systems may require longer maintenance windows than smaller systems.

3.2 SYSTEM ANALYSIS MATERIALS TO BE USED

(Spence and Grout, 1978) proposed that systems analysis and design (SAD) should be part of the computer science curriculum. Validate requirements using a system modeling and prototyping approach. Define requirements as a basis for why we should develop an academic prediction system and use survey and interview techniques as information gathering methods in understanding the scope of our investigation.

3.2.1 Survey as method of data collection.

A survey is a simple tool for gathering information. A poll usually consists of a series of questions designed to assess a participant's preferences, attitudes, traits, and opinions about a particular topic. As a research method, questionnaires allow concepts to be counted or quantified. That is, use a wider sample or subset of the audience, which allows the findings to be applied to a wider population. For example, in a given year he has 100,000 unique users of his website. If he collects information from 2,000 of these users, he can confidently apply that information to all 100,000(Chris Gray 2016). There are different types of questions in surveys, depending on the circumstances of the survey and the type of information you want to access. Many surveys combine open-ended and closed-ended questions, such as rating scales and semantic scales. This means it can be used for qualitative and quantitative research. Surveys come in two main forms. Paper form or online form. Paper surveys are a more traditional method of data collection and can easily lose data. Paper forms are also cumbersome to organize and process.

When it comes to the digital space, surveys can be used for a variety of purposes, including:

- Collect feedback during a live product or pilot.
- Research because people visit her website and evaluate their experience of that visit (e.g. True Intent surveys).
- Quantify the results of qualitative research activities such as contextual research and interviews. When Evaluating usability such as System usability scale.

Surveys can be effective for identifying:

- Who your users are;
- What your users want;
- What they purchase;
- Where they shop;
- What they own

Benefits of Surveys are:

- It provides information to better understand end-users in order to develop better products.
- Reduce the risk of designing the wrong or inappropriate solution for your users.

- Provides confidence to stakeholders that the design is or will be effective. Collecting a larger sample size compared to qualitative research often speaks the language of business people. Like it or not, research often gives the impression that more is better.

Disadvantages of Surveys are:

- The validity of the research data can be affected by survey response bias.
- High survey dropout rates can also affect the number of responses received in your survey.

3.2.2 Interview as a method of data Collection

Interviews are used to collect data from small groups on a wide range of topics. You can use structured or unstructured interviews. Structured interviews are similar to surveys, with the same questions asked in the same order for each topic and multiple-choice answers. In unstructured interviews, questions vary by topic, may depend on answers to previous questions, and do not have fixed answer options. Where face-to-face interviews cannot be conducted, study participants/respondents may be interviewed using video conferencing tools in addition to regular telephone calls. You can safely use the different video conferencing tools such as: Microsoft Teams, Google meet etc.

The main purpose of interviews as a data collection tool is to collect data comprehensively and centrally. As Pauline Young pointed out, the purpose of the interviews could be the exchange of ideas and experiences, the collection of information related to a very wide range of data in which respondents repeat the past and define their present and future desires. there is. discuss options

Importance of Interview:

- It is the most suitable method for assessing personal qualities.
- It has clear value for diagnosing and treating emotional problems.
- This is one of the essential foundations for implementing the recommended procedures.
- It provides information that complements other data collection methods.
- It can be used not only for observation but also for verifying information obtained by the response method.

Types of Interviews

There are different types of interviews used to collect survey data. Interviews are either structured or unstructured, depending on whether a formal questionnaire is formulated, and questions are asked in a pre-determined order. Interviews can be direct or indirect, depending on whether the purpose of the question is explicitly stated or deliberately hidden. Classifying these two traits against each other yields four different types of interviews. That is, interviews can be (1) structured direct, (2) unstructured direct, (3) structured indirect, or (4) structured It may be an indirect one. Types (1) and (2) are basically objective types. (3) and (4) are subjective types.

3.3 SYSTEM DESIGN AND IMPLEMENTATION METHODOLOGY

The Adopted methodology for the proposed system is Agile. Agile software development refers to a set of iterative software development methodologies in which requirements and solutions evolve through collaboration between self-organizing cross-functional teams (Jack Mchill 2005). Agile methods and processes generally promote a disciplined project management process that encourages frequent inspection and adaptation, a leadership philosophy that encourages teamwork, self-organization, and accountability, a set of engineering best practices designed to allow for the rapid delivery of high-quality software, and a business approach that aligns development with customer needs and company goals. Agile Methodology refers to a practice that encourages continuous iteration of development and testing throughout the project's software development lifecycle. Unlike the Waterfall model, both development and testing activities are carried out concurrently in the Agile model of software testing. The Agile software development methodology is one of the most straightforward and efficient methods for translating a vision for a business need into software solutions. Agile software development approaches use continuous planning, learning, improvement, team collaboration, evolutionary development, and early delivery. It promotes adaptable responses to change.

Four core values are emphasized in agile software development.

- Interactions between individuals and teams regarding processes and tools
- Working software trumps extensive documentation.
- Customer involvement in contract negotiations

- Responding to change by sticking to a plan

Advantages of Agile Methodology

- It is flexible and adaptable
- Lower Cost
- Improved Quality
- End-user satisfaction

Disadvantages of Agile Methodology

- Scalability
- Skill Required

3.4 SYSTEM SECURITY

The problem of system security is divided into four interconnected issues: security, integrity, privacy, and confidentiality. As stated in the data rationale, they determine the file structure, data structure, and access procedures.

1. System security refers to the technical enhancements and procedures implemented in hardware and operating systems to protect against intentional or unintentional damage from a defined threat. Data security, on the other hand, is the protection of data from loss, disclosure, modification, and destruction. The application selects the device number (IMEI number) and sends it to the API database.
2. System integrity refers to the proper operation of programs, as well as appropriate physical security and protection from external threats such as eavesdropping and wiretapping. Data integrity, on the other hand, ensures that do not differ from others, and how the organization can be protected from unwanted, unfair, or excessive dissemination of information about it.
3. The term confidentiality refers to the special status given to sensitive information in a database in order to minimize potential invasions of privacy. It is a characteristic of information that defines its need for protection. The technical means of providing such protection is system security. In contrast, privacy is primarily a matter of how information is used.

3.5 CONCLUSION

In conclusion, methodology is the systematic, theoretical analysis of the methods applied in developing a software. Agile methodology has been chosen as the adopted methodology for this project because of the nature of the project and the advantages attached to Agile methodology with respect to the project. Various information gathering and system analysis and design techniques are considered so as to come up with an accurate system.

CHAPTER FOUR

SYSTEM ANALYSIS AND DESIGN

4.0 INTRODUCTION

The process of determining user expectations for a new or modified product is known as requirements analysis. These characteristics, known as requirements, must be quantifiable, relevant, and detailed. Such requirements are frequently referred to as functional specifications in software engineering. The analysis of requirements is an important part of any project.

Requirements analysis entails frequent communication with system users to determine specific feature expectations, resolving conflict or ambiguity in requirements as demanded by various users or groups of users, avoiding feature creep, and documenting all aspects of the project development process from beginning to end.

Requirements analysis is a collaborative effort that necessitates knowledge of hardware, software, and human factors engineering, as well as interpersonal skills.

4.1 OVERVIEW OF THE PROPOSED SYSTEM

Due to the large volume of data in educational databases, predicting student performance is a critical job. This task is being handled by educational data mining (EDM). EDM creates methods for discovering data derived from the educational environment. These techniques are used to comprehend students and their learning environments. Educational institutions are frequently interested in how many students will pass/fail for necessary arrangements. Previous research has found that many researchers are intent on selecting an appropriate algorithm for just classification while ignoring solutions to problems encountered during data mining phases such as data high dimensionality, class imbalance, and classification error, among others. These types of issues reduced the model's accuracy. In this domain, several well-known classification algorithms are used, but this paper proposed a student performance prediction model based on supervised learning decision tree classifiers. In addition, an ensemble method is used to improve the classifier's performance. The ensemble methods approach is intended to address classification and prediction problems.

4.2 GENERAL REQUIREMENTS

In order to implement an efficient and effective system one needs to understand the business and technical requirements of the implementation process. The main requirement of the proposed system is categorized into:

1. Functional requirement
2. Non-functional requirement
3. Software Requirement Specification
4. Hardware Requirement Specification

4.2.1 Functional Requirement for Academic Prediction System

Functional requirements are system features that developers must include in order for users of the system to achieve their objectives. They define the fundamental system behavior under certain conditions.

4.2.1.1 Input Design

User-oriented inputs are transformed into a computer-based system format in the input design. The design of menus and prompts is a key component of input design. The user's options are predefined for each possibility. Logical flow, data storage, source, and destination are shown in the data flow diagram. The gathering and grouping of input data into groups of related data.

4.2.1.2 Output Design

The goal of the output design is to show output on a screen in a specific format. The capabilities of the output device, the necessary response times, etc. are taken into account. In order to provide users with information that is clear, precise, and quick, the form design elaborates on how output is displayed and the layout accessible for information capture.

4.2.1.3 Logical Design

The logically suggested data is what logical data design is all about. It is possible to construct a form so that the meaning may be understood by each and every piece of data. A clear knowledge

and notion of the associated data required to build a form and process information and events should be provided by logical data design.

Few functional requirements of the system shall perform;

1. The system shall allow students to view their current academic performance
2. The system shall generate reports for each student’s performance.
3. The system shall allow students to input necessary details for prediction

4.2.2 Non-Functional Requirement

Non-functional requirements are descriptions of how the system should operate; they list all current requirements that the functional requirement does not address. These nonfunctional criteria include things like;

1. Usability: The simplicity of this system's operation is its most crucial criterion. The system is designed in such a manner that users who are unfamiliar with the technical details of computer systems would find it easy to use and comfortable.
2. Maintainability: The suggested solution is highly adaptable to new trends and platform development.
3. Interoperability: This is the application's capacity to work with many versions of Windows, including Windows 7, Linux, and MAC, which are the most recent trends in computer system devices.

Few non-functional requirements of the system shall perform;

1. The system should present the required feature at any given time.
2. The system should present text using readable font style i.e., no demographic font.
3. The system should generate the report in less than 5 second with an internet speed greater or equal to 200kbps.
4. The system should be accessible to end user at any location in the world.
5. The system should be available at every hour of the day.
6. The system should be responsive to any user device.
7. The system should sort post by time and date.

8. The system should display the date and time a report is created.
9. The system's functionalities should work on every recent browser.

4.2.3 Software Requirement

The developed system is not platform dependent in order to effectively solve the problem that led to the development of this project. As a result, it can run on any platform. The reviews of the developed system's specifications are listed below.

Operating System	Windows OS, Linux, and Mac OS (Python must be installed on the selected Operating System used.)
Front-End Technologies	HTML5, CSS3, JavaScript
Server-side Technologies	Python3, Flask, WebSocket, Event Emitter
Server	Python3
Back-End Tools	CSV, SQLite, Jupiter Notebook

Table 4.1 Software Requirement Specification

4.2.4 Hardware Requirement

Devices	Desktop Devices, Laptop
Processor	x64 Processor: AMD Opteron, AMD Athlon 64, Intel Xeon with Intel EM64T support, Intel Pentium IV with EM64T support or higher
Memory	Minimum of 1GB RAM
Storage	Minimum 1GB

Table 4.2 Hardware Requirement Specification

4.3 SYSTEM DESIGN

The process of defining the elements of a system such as the architecture, modules, and components, the various interfaces of those components, and the data that flows through that system is known as system design (Gates 2007). It is intended to meet specific needs and requirements of a business or organization by engineering a coherent and well-functioning system.

A systematic approach to system design is implied by the term systems design. It may take a bottom-up or top-down approach, but in either case, the process is systematic in that it considers all related variables of the system that needs to be created—from the architecture to the necessary hardware and software, all the way down to the data and how it travels and transforms throughout its journey through the system. Then there's systems analysis, systems engineering, and systems architecture to consider.

When engineers were attempting to solve complex control and communications problems just before World War II, the systems design approach first appeared. They needed to be able to formalize their work into a formal discipline with proper methods, particularly in new fields such as information theory, operations research and computer science generally

4.3.1 Feasibility Study

A feasibility study is simply an evaluation of the viability of a proposed project plan or method. This is accomplished through an examination of technical, economic, legal, operational, and time feasibility factors. As the name implies, you're wondering, "Is this possible?" Do you, for example, have or can you develop the technology to do what you propose? Do you have the necessary people, tools, and resources? And, will the project generate the expected ROI?

Feasibility studies have several advantages, including assisting project owners in determining the pros and cons of undertaking a project before investing significant time and capital into it.

Feasibility studies can also provide critical information to a project owner preventing them from entering into a risky and unattainable venture. They will have a better understanding of how they will operate, what potential obstacles exist, who their competitors are, and what the market is. Feasibility studies can also persuade investors and bankers that investing in a specific project or business is a wise decision.

Feasibility analysis Facts;

1. Technical feasibility
2. Economic feasibility
3. Operational feasibility
4. Schedule feasibility
5. Resource feasibility

4.3.1.1 Technical Feasibility

By evaluating the new system's technological viability, one may determine whether it will function as intended and whether the proposed system can be built. The technical evaluation provides information on issues including whether the system's required technology is available, how challenging it will be to construct, and if the knowledge and expertise now held are sufficient to operate it.

The following are some of the technical concerns with feasibility that were brought up during the feasibility stage of the investigation:

1. The software (Visual Studio Code) is running on a windows operating system
2. The minimum hardware required is 2.4 GHz Intel Core 2.
3. The system can be easily expanded.

4.3.1.2 Economic Feasibility

Finding out what positive economic advantages the proposed system would provide to the company is the goal of the economic feasibility evaluation. All of the anticipated advantages are identified and quantified. A cost-benefit analysis often forms part of this evaluation. Economic evaluation, which deals with elements that can be measured, quantified, and compared in monetary terms, is an essential component of investment appraisal (Chen 1996). This feasibility analysis compares the development and operating costs to show the prefect's tangible and intangible advantages.

4.3.1.3 Operational Feasibility

Operational feasibility is a metric used to assess how successfully a proposed system addresses issues, seizes opportunities, and complies with requirements found during the requirements

analysis stage of system development. In terms of development timetable, delivery date, corporate culture, and current business procedures, the operational feasibility study examines how well the planned development projects fit into the existing business environment and objectives.

4.3.1.4 Schedule Feasibility

The system is socially feasibly because it can be delivered within the required deadline given our technical expertise.

4.3.1.5 Resource Feasibility

This raises issues like how much time is available to develop the new system, when it can be developed, if it obstructs regular company activities, the kind and quantity of resources needed, dependencies, etc. These plans also include mitigation and contingency measures, ensuring that the business is prepared in case the project does go over budget.

Feasibility Questions	**Answers**	**Comment**
With the short time available, can this task be completed successfully?	YES	Depending on the accessibility of the necessary technology and the usability of the selected platforms and technologies, the project can be completed.
Are the resources needed for the project on hand?	YES	The project's success can be improved with the help of the available resources.
Would the project interfere with ongoing corporate operations?	NO	Because the project may be created outside the firm, it won't in any way interrupt or halt ongoing commercial operations.

Are there any backup plans in place in case the system malfunctions to carry on with routine company operations?	NO	When the system malfunctions, users can utilize the manual method until the system is restored since it is not a crucial system (one whose functionality cannot be done manually).

Table 4.3 Feasibility Study

4.4 MACHINE LEARNING

The brain of machine learning is where all learning occurs. The way a computer learns is comparable to how a person learns. Experience is how people learn. The easier it is to forecast, the more we know. By analogy, our chances of success are lower than they would be in a known circumstance when we encounter one. Machines receive the same training. The computer observes an example in order to create a precise forecast. The machine is capable of predicting the result when we provide a comparable case. However, much like a person, the machine has trouble predicting if it is given a new example.

Learning and inference are at the heart of machine learning. The first way the machine learns is by identifying patterns.

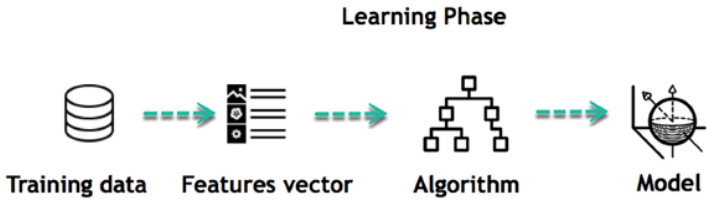


Figure 4.4 Machine learning phase

4.5 DATA FLOW DIAGRAM (DFD)

Based on Martin and Estrin's "data-flow graph" model of computing, Larry Constantine, the original creator of structured design, created data-flow diagrams.

A data flow diagram (DFD) is a diagram that shows how data moves through a system and how different data is changed. It is a technique for structured analysis and design that may be applied to flowcharting in place of or along with data. System flowcharts that are process-oriented and orientated. Data-flow diagrams (DFDs) were established and made popular for structured analysis and design in the late 1970s (Gane and Sarson 1979). DFDs display the flow of data from external sources into the system as well as how the data was transferred between processes and logically stored.

A data-flow diagram (DFD) shows how data "flows" through an information system graphically. DFDs may also be used for structured design, which visualizes data processing. (2005) Dennis J. et al. Data flow diagrams are created using just four fundamental symbols. Data sources, data flows, data transformations, and data storage are all represented by these symbols. Nodes, which are often enclosed figures like circles, show the locations where data is altered.

Data Flow Diagram Symbols:

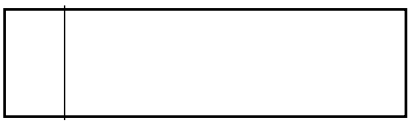
	Source or destination of data. Also called external entity
	Arrow / Data flow
	Process
	Data store / storage

Table 4.5 data flow diagram symbols

4.5.1 Context Dataflow Diagram

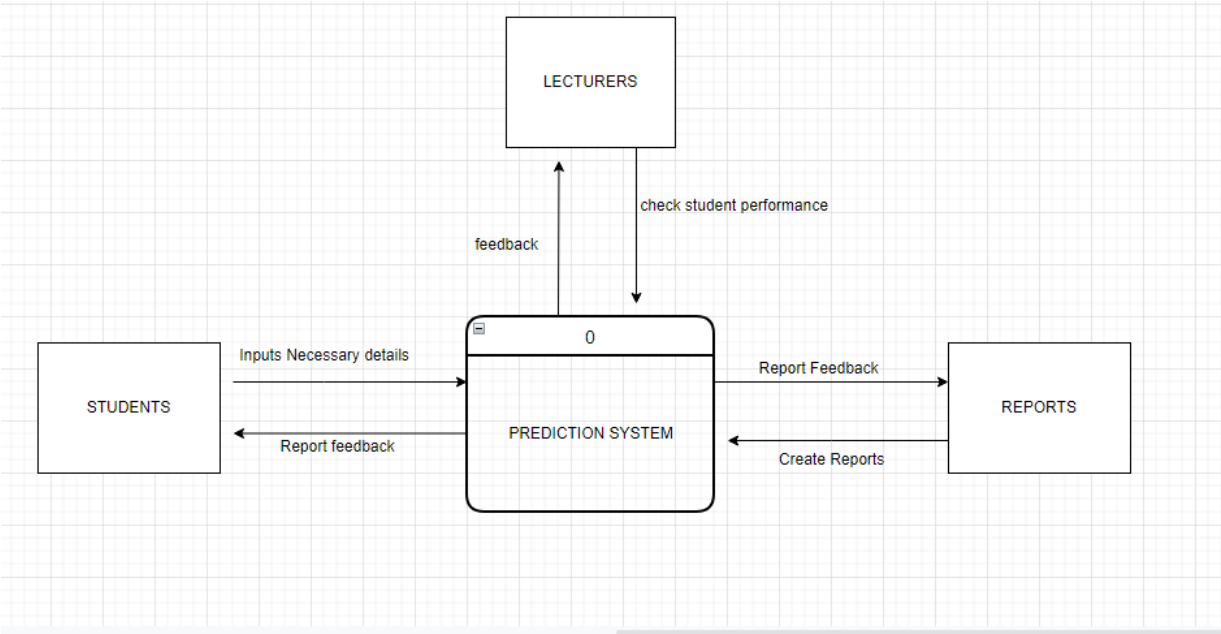


Figure 4.5.1 Context Dataflow Diagram

4.6 USE CASE DESIGN

4.6.1 Use case diagram

A graphical depiction of the high-level system scope is a use case diagram. It contains actors, who are the system's users, and use cases, which are functional components the system will offer. Use cases depict how a system often interacts with its users (end users and other systems). These interactions show how the system appears to users from an external, or functional, standpoint. A scenario is the name given to each route taken through the use case.

4.6.2 Use Case Diagram for the Academic performance system

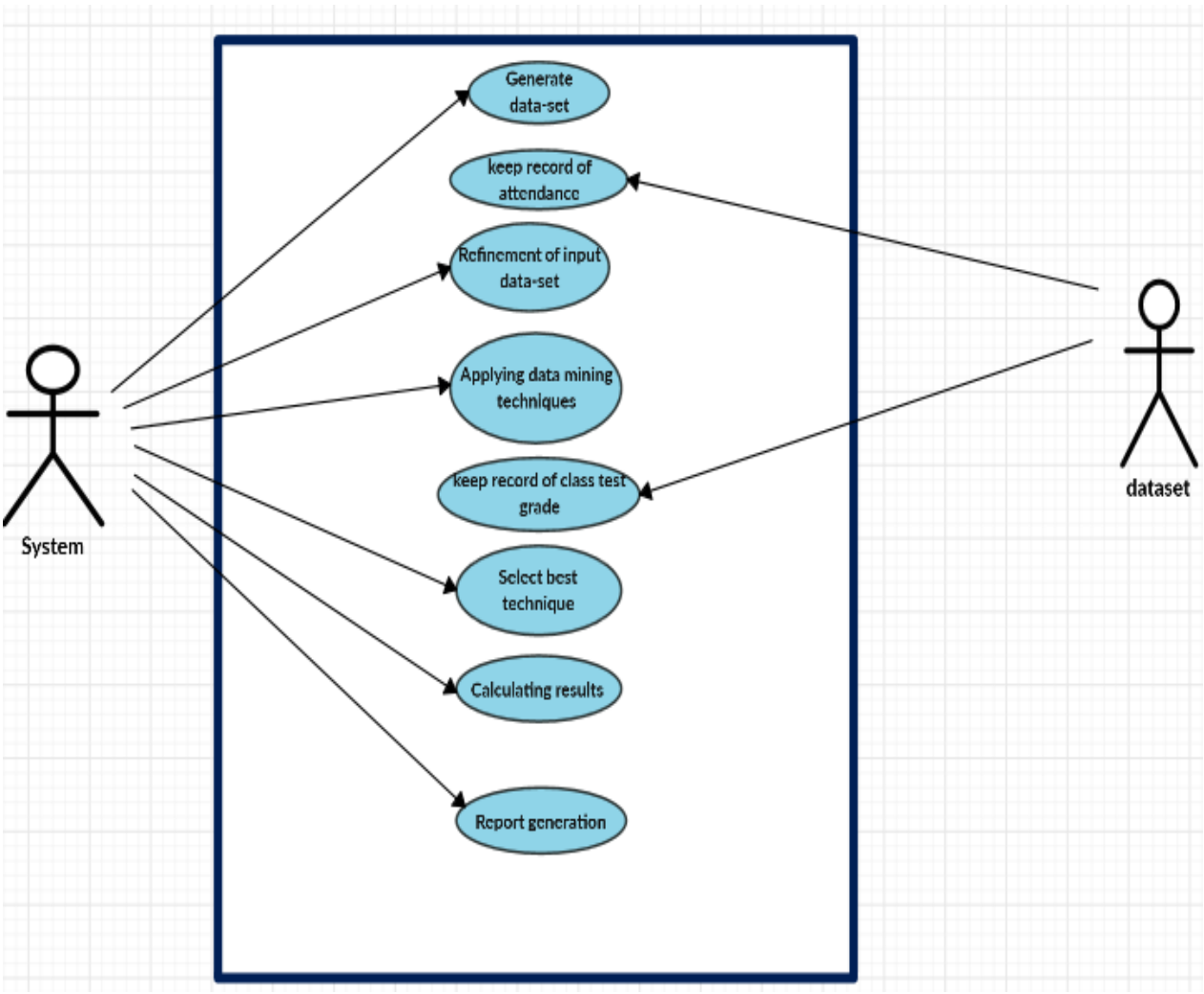


Figure 4.6.2 use case diagram of the Academic predicting system

4.7 ENTITY RELATIONSHIP DIAGRAM FOR THE ACADEMIC PERFORMANCE SYSTEM

An abstraction and conceptual representation of data is called an entity-relationship model (ERM). Entity Relationship Modeling is a database modelling technique that creates a certain kind of conceptual system schema. An information system's entities and their connections are graphically represented by an entity-relationship diagram, which is a data modelling approach.

Its components consist of three, they are;

1. The entity is a thing, person, place, or event about which data is being gathered. Consider the information system for a business as an example. Entities would comprise not just customers but also orders and the address of each client.
2. The interaction between the entities constitutes the relationship. The consumer puts an order in the aforementioned example; hence the word "places" refers to the interaction between that specific customer and the order or orders that they place.
3. The link between the things is described mathematically by their cardinality. An entity may not be necessary. Cardinality notation comes in a variety of forms, with the crow's foot notation being one of the most used. One-to-one, also known as 1:1, one-to-many, or 1:M, and many-to-many, also known as M: N, are the three primary cardinal relationships.

4.7.1 Entity Relationship Diagram

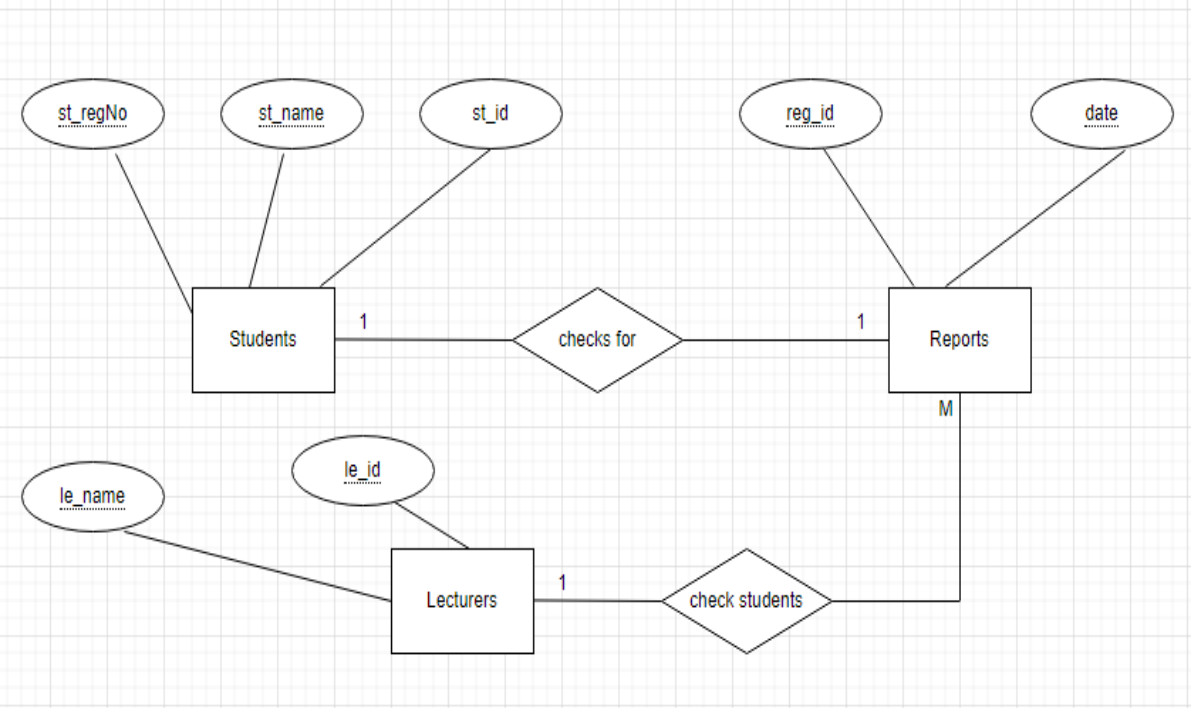


Figure 4.7.1 entity relationship diagram1

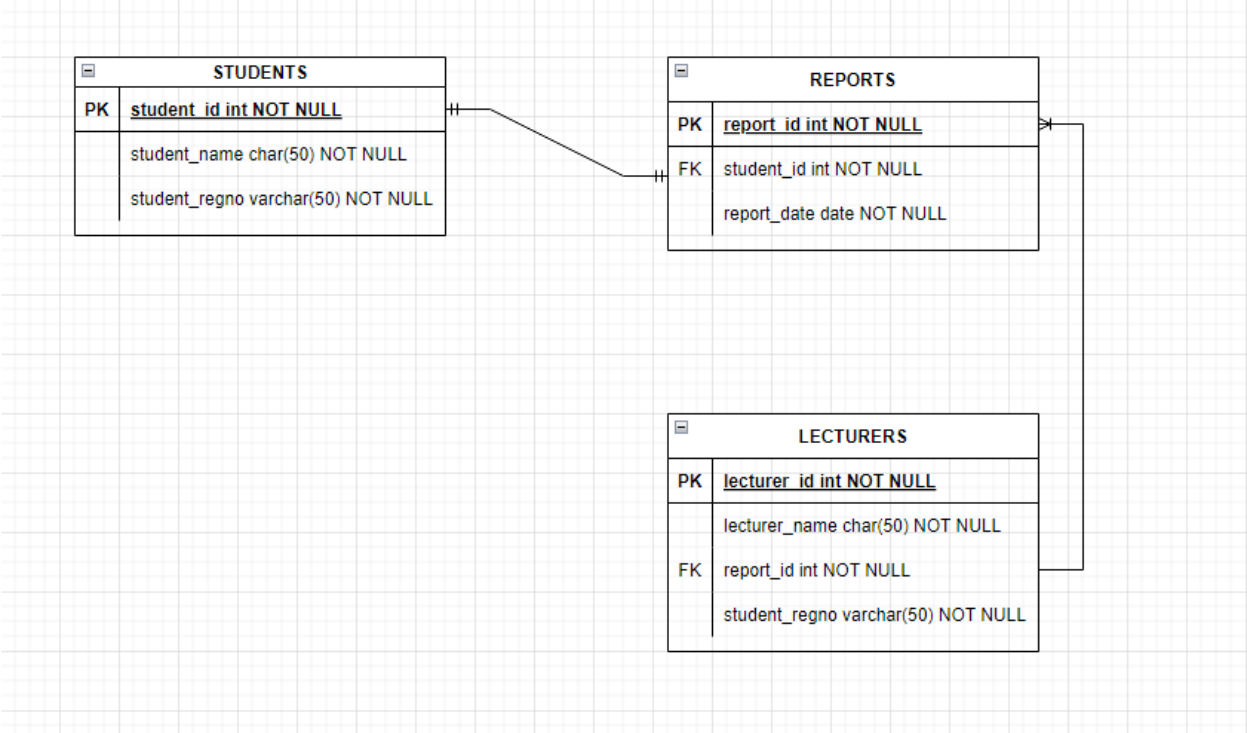


Figure 4.7.1 entity relationship diagram2

4.8 FLOWCHART DESIGN OF THE ACADEMIC PERFORMANCE SYSTEM

A flowchart is a diagram that shows how a system, computer algorithm, or process works. They are frequently used in many different disciplines to examine, organize, enhance, and convey frequently complicated processes in simple, understandable diagrams. Rectangles, ovals, diamonds, and perhaps many more shapes are used in flowcharts, also known as flow charts, to indicate the kind of step and connecting arrows to indicate flow and sequence. They might be anything from straightforward hand-drawn charts to detailed computer-drawn diagrams showing several processes and pathways. Considering all the different variations, flowcharts are among the most widely used diagrams on the world, being utilized by both technical and non-technical persons in a wide range of industries. The terms Process Flowchart, Process Map, Functional Flowchart, Business Process Mapping, Business Process Modeling and Notation (BPMN), or Process Flow Diagram are sometimes used to refer to flowcharts (PFD). They have a connection with other well-known diagrams like Data Flow Diagrams (DFDs) and UML Activity Diagrams.

4.8.1 Flowchart Symbols

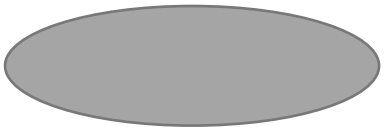
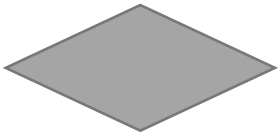
Symbol	Name	Function
	Start / end	An oval represents a start or end point
	Arrows	A line is a connector that shows relationships between the representative shapes
	Input / Output	A parallelogram represents input or output
	Process	A rectangle represents a process
	Decision	A diamond indicates a decision

Table 4.8.1 Flowchart Symbols

4.8.2 Flowchart Diagram

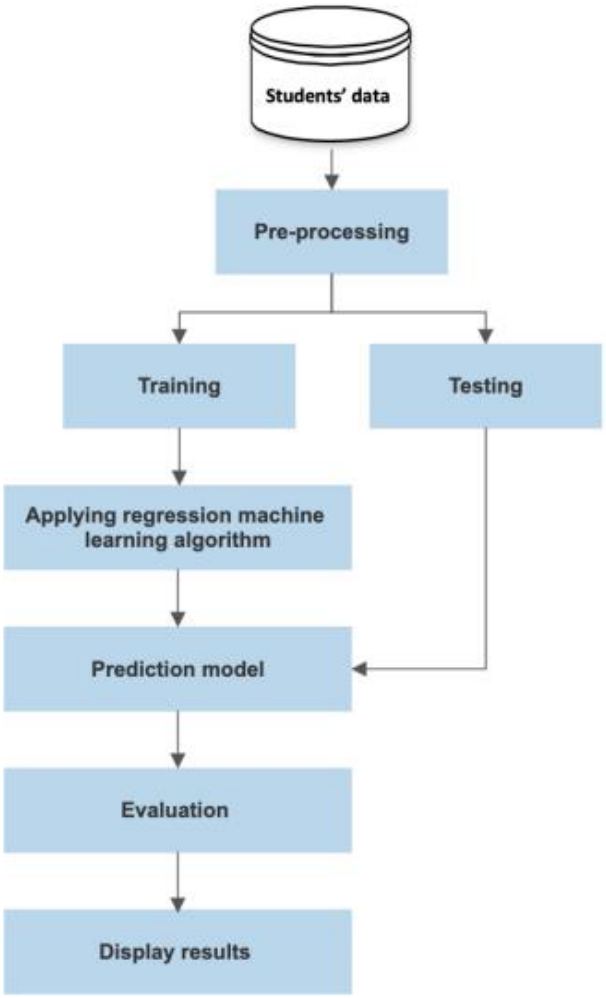


Figure 4.8.2A Flowchart model of the Academic Performance System

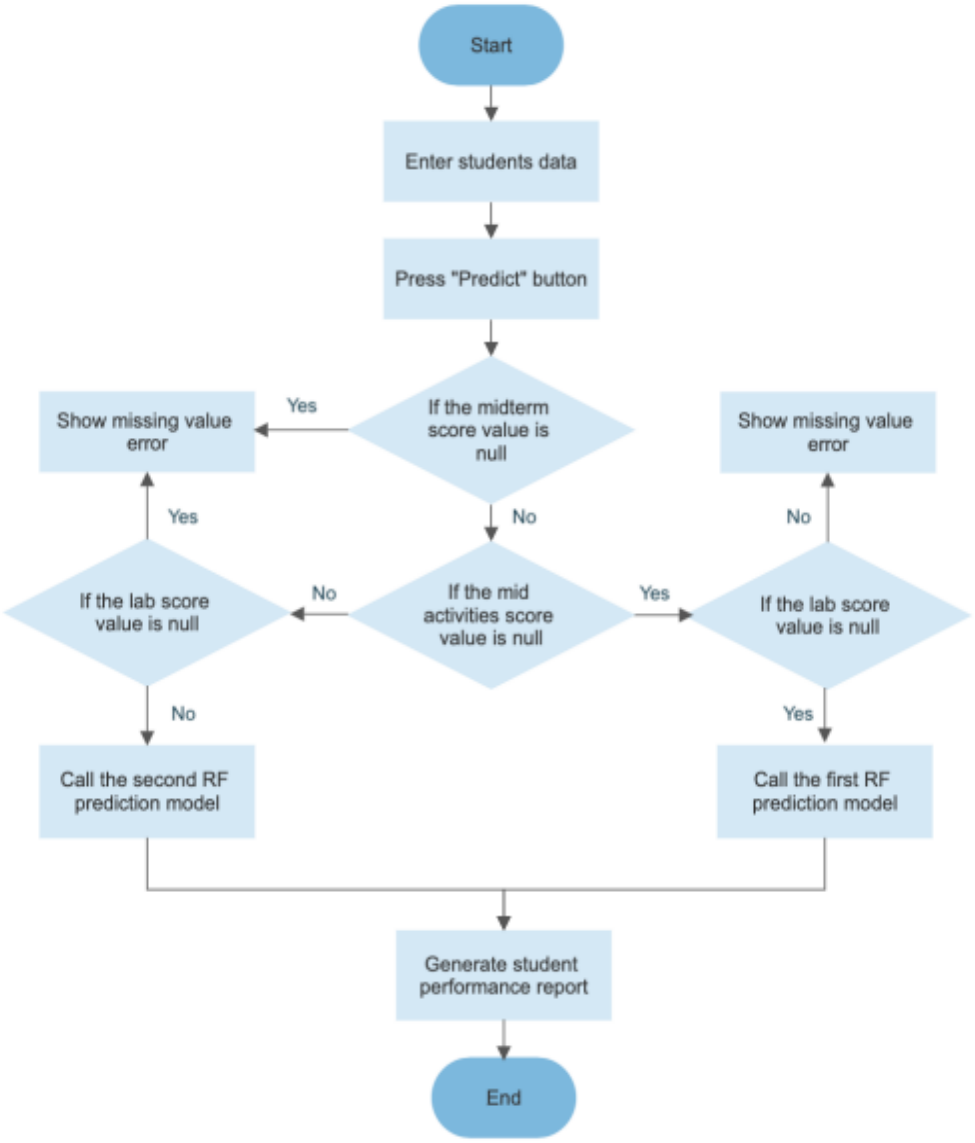


Figure 4.8.2B. Flowchart of the prediction System

4.8.3 Sequence Diagram

A sequence diagram is an interaction diagram that demonstrates the order and interactions of several things. It is an implementation of a message sequence diagram.

Object interactions are arranged in temporal sequence in a sequence diagram. It shows the classes and objects involved in the scenario as well as the flow of messages that must be

exchanged for the objects to work as intended. In the Logical View of the system being developed, sequence diagrams are often connected to use case realizations. Event diagrams and event scenarios are other names for sequence diagrams.

4.8.4 Sequence Diagram of the Academic Performance System

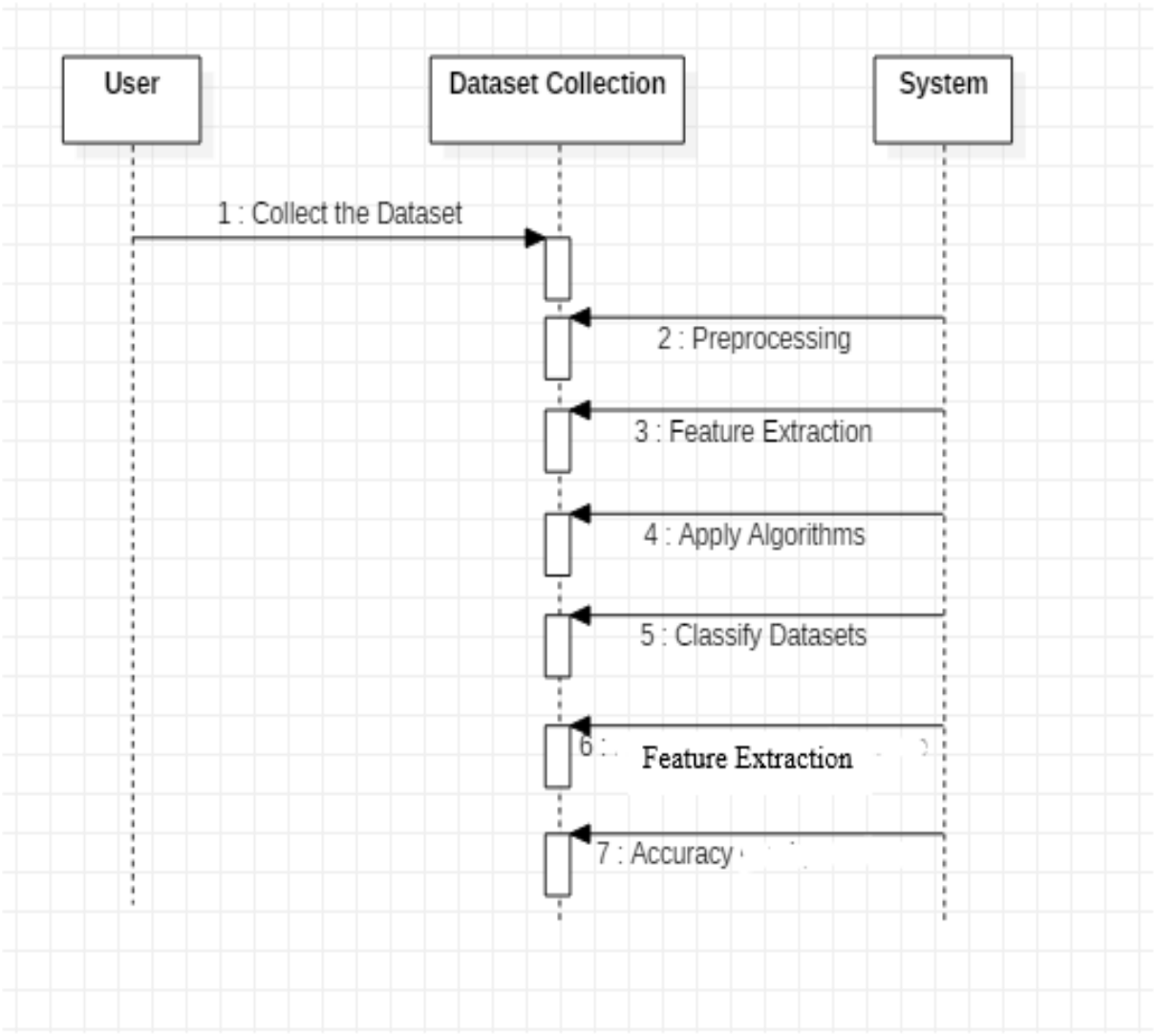


Figure 4.8.4 sequence diagram of the system

4.9 UML ACTIVITY DIAGRAM

Another crucial behavioral diagram in the UML used to depict the system's dynamic elements is the activity diagram. An activity diagram is simply a more complex flow chart that models the transition between activities.

UML Activity Diagram;

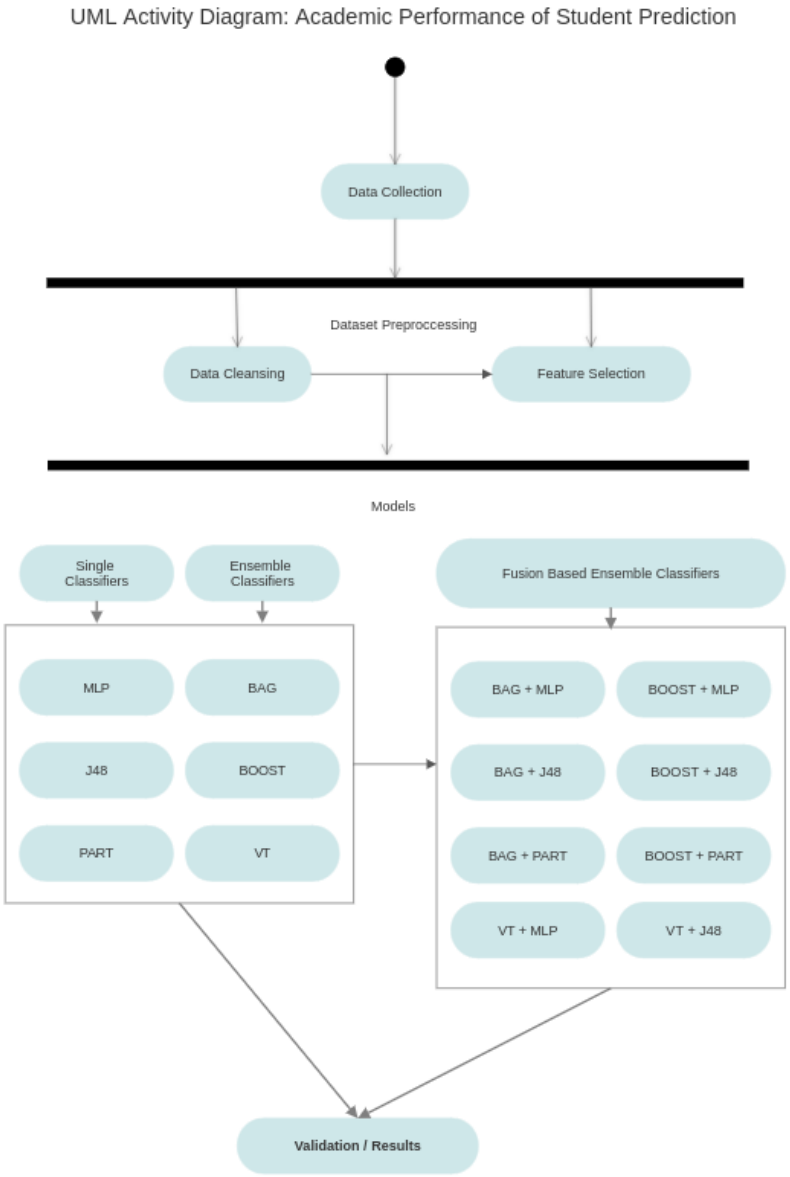


Figure 4.9 UML Activity diagram

4.10 SYSTEM IMPLEMENTATION

The system implementation approach that was used was the pilot implementation. This is because the Academic Performance System is a new system for most institutes especially for this project focus college (AIT). A company can validate its strategy for full application deployment through a pilot implementation. Running an application pilot can reveal operability problems related to production-like circumstances and give a chance to fix these problems before a complete application roll out. An organization should create a thorough pilot strategy, choose the right participants, organize the pilot environment, and decide how to monitor the pilot in order to properly be ready for a pilot deployment. This might entail either exploring a brand-new idea or method or applying a tried-and-true method that has been suggested by outsiders but has never been used by the organization.

4.10.1 System Testing

The process of testing demonstrates that a program complies with its specification and operates as intended. Testing is used to find mistakes or circumstances where the program's behavior is unacceptable. Testing aims to demonstrate thoroughness, enhance software quality, and offer maintenance assistance. Several guidelines that can work well as testing objectives are stated by Glen Myers [MYE79] in his book "The Art of Software Testing" from 1979:

1. Testing is the process of executing a program with intent of finding errors.
2. A good test case is one that has a high probability of finding an as-yet undiscovered error.
3. A successful test is one that uncovers an as-yet undiscovered error.

The suggested system was tested using the following methodologies:

- Careful scrutiny of procedural detail is the cornerstone of white-box software testing. By providing test cases that put certain sets of circumstances and/or loops to the test, the program is tested along its logical routes. It is possible to check the "state of the program" at various times to see if the asserted or expected status matches the actual situation. (Pressman, R.S., & Herron, S.R., 1991).
- The focus of black-box testing, also known as behavioral testing, is on the functional specifications of software. The software engineer may determine the input circumstances

that will completely exercise all of a program's requirements using this testing methodology. Black box testing attempts to find the errors like;

1. Incorrect functions
2. Interface errors
3. Mistakes while accessing other databases or data structures
4. Mistakes in behavior or performance
5. Initiation and termination mistakes

In Black box testing software is exercised over a full range of inputs and outputs are observed for correctness.

4.10.2 Software Testing Strategies

The strategies for testing of the system involve;

1. Unit Testing: Unit testing is the basic level of testing. Unit testing is used to make sure that every software has been thoroughly tested.
2. Integration testing: Integration testing is the next phase. To make sure the software requirements are satisfied, separate program components or programs are combined and tested as a whole system.
3. Acceptance testing: Planning and carrying out various sorts of tests as part of acceptance testing is necessary to show that the developed software system fits the requirements. Finally, after passing all of the testing phases, our product satisfies the criteria.

4.10.3 System Security

Security, integrity, privacy, and confidentiality are the four connected problems that make up the system security challenge. They choose the access protocols and file and data structures.

1. When referring to system security, we mean the technical advancements and practices used on the hardware and operating systems to guard against intentional or unintentional damage from a specified danger. Data security, on the other hand, is the safeguarding of data against loss, disclosure, alteration, and erasure.
2. System integrity refers to the programs' proper operation, suitable physical security, and protection from external dangers like wiretapping and eavesdropping. Data integrity, in

contrast, ensures that data is not altered or damaged in a way other than what authorized users intended.

3. Confidentiality is a unique designation given to sensitive data in a database to reduce potential privacy breach. Information's requirement for protection is characterized by this quality. The technological method of offering such protection is system security. Contrarily, privacy is mostly a procedural issue related to the way information is utilized.

4.11 CONCLUSION

The creation and execution of a computerized product distribution monitoring system were discussed in this chapter. This system was created using the Agile technique. The UML methodology was used to evaluate the design and functions of the system. The system functionalities were designed using a use case diagram. System databases that store and retrieve data were designed using entity relational modelling (ERM). The breakdown of the various entities and the interaction between the system and the user were done using data flow diagrams. This chapter also covered the system testing technique.

CHAPTER FIVE

FINDINGS, CONCLUSION AND RECOMMENDATIONS

5.0 INTRODUCTION

The construction phase of the project began once the thorough design for individual modules was finished. During this phase, integration of the numerous source codes into the overall system was accomplished. for system users to quickly become accustomed to the electronic wallet system. This chapter's main goal is to describe the transition from the system design phase to the operational system and overall, this chapter also includes an overview.

Academic Performance System implements web-based technologies such as;

- Python 3
- Visual Studio Code
- CSV: A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format.
- SQLite: SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine
- Jupyter Notebook: The Jupyter Notebook is the original web application for creating and sharing computational documents.

The system's implementation tools must take the project's requirements and scope into account. The developer's experience must also be taken into account. However, the developer's insufficient experience might have an impact on the creation of a solution.

5.1 USER INTERFACE

Assuring that users can complete their tasks effectively, efficiently, safely, and with enjoyment is a key objective of interface design. To achieve this, a designer must be aware of the available technology as well as the needs, tasks, and characteristics of the system's potential users (Yeates and Wakefield) (2004, p.318).

The Graphical User Interface (GUI) is used by the information system in its design process. The user points at and clicks on buttons with a pointing device to carry out particular actions. Users are more likely to adopt and routinely utilize the system if they have GUI access.

According to Nielsen (2006), there are ten guidelines for a well-built user interface, which include:

1. System state visibility
2. System compatibility with the outside world
3. User control and freedom
4. Consistency and standards
5. Error avoidance
6. Recognition as opposed to memory recall
7. Flexibility and effectiveness of usage
8. Beautiful and minimalist design
9. Guidance and instructions

5.2 PROGRAMMING LANGUAGES

The language that was used to develop the Academic Performance System is a machine language called PYTHON. PYTHON is a widely used, flexible, and all-purpose programming language. It works well as a first language since it is clear and simple to understand, and it is a useful language to have in any programmer's toolkit because it can be used for everything from creating websites to creating software for scientific purposes. It is easy-to-understand grammar makes it the ideal language for someone who is just starting to learn computer programming.

5.2.1 Why Using Python for the Academic Prediction system?

In order to show data (information) visually and often to draw conclusions and make decisions based on the information, data analysis and visualization is utilized. Python is the best language for making graphs like bar charts and line charts, which may be used to express information in an image. The information or data offered is frequently numerical and may be based on a certain time frame. Charts can be as basic or as sophisticated as necessary and there may be several aspects at play.

Benefits;

1. A great library ecosystem

2. A low entry barrier
3. Flexibility
4. Platform independence
5. Readability
6. Good visualization options

5.3 IMPLEMENTATION

A strategy, a technique, or any other design for carrying out anything is implemented when it is carried out, executed, or put into effect. Implementation is therefore the activity that must come after any initial planning in order for anything to truly occur. Implementation in the context of information technology includes all the steps necessary to get new hardware or software functioning correctly in its environment, including installation, setup, running, testing, and making the necessary adjustments. Sometimes, the terms "deployment" and "deployment" are interchangeable.

5.4 CONCLUSION OF THE STUDY

We are in an era of technological advancement and rapid progress in almost all aspects of living. It is therefore of great importance if our schools are designed to assist student in their educational performances as well as lecturers to better the future of education.

This research aims at helping schools (tertiary institutions) predict student performance by using data they already have thereby giving them the power to make informed decisions that will better enhance student performance and increase the retention rate on students. The main objective of this project is to design and implement an academic performance prediction system for schools in such a way student can be able to track their performances in their academic and also for lecturers to be able to track students' performance so as to help increase teachings.

In achieving the main objective mentioned, this project will specifically concentrate on the following;

- The system shall maintain better user relationship.
- The system shall be able to predict students' performance

- The system shall display accurate performance
- The system shall help lectures and students improve learning process

5.5 MAINTENANCE

Making updates to a system after it has been deployed is known as system maintenance. The system needs to be modified in a variety of ways, including small adjustments to fix code problems, larger adjustments to fix design flaws, big additions to fix specification errors, and adding new features that couldn't be included in a timely manner. Three different kinds of system maintenance exist:

1. Repairing fault: is expensive since it may need a significant system redesign in the case of a requirements mistake. Coding mistakes may be easily fixed in this situation.
2. Environmental Adaptation: This sort of maintenance is necessary when the hardware, the operating system platform, or other supporting software, among other environmental factors, change. After then, the application system is changed to accommodate these environmental changes.
3. Functionality Addition: This sort of maintenance is required when organizational or business changes affect the system needs. When compared to other forms of maintenance, the size of the modifications that must be made to the program is frequently substantially larger.

5.6 IMPLICATIONS

The main implication of this study for this project is that the system enhances effectiveness and efficiency in support activities for the semester. The system would be of immense importance for schools and any other learning institution. Also, the system opens up opportunities for schools worldwide and also for students who finds it difficult in learning. It will help the student know the area he /she needs to improve in their respective academic career.

5.7 FUTURE WORKS AND RECOMMENDATION

The researcher advises moving on with the system's implementation while evaluating it for the addition of new modules or components that have been left out due to time and financial constraints.

Schools using this system should also;

1. As new issues occur, increase the number of solutions.
2. Administrators who will use or administer the system should have proper training.
3. Regularly backup database in case of a calamity or system breakdown.
4. Make sure that hardware components are constantly functioning properly. If necessary, replacement should be made.

The developer should be engaged for a system or program upgrade as the management's needs grow.

5.8 CONCLUSION

In the chapter one, the problem statement of current system was explained and also the project objectives which are to design an Academic Performance Prediction System. In chapter two the researcher conducted a literature review of existing systems. In Chapter three, the researcher underscored the various methodologies that will be used to implement the system. Chapter four, the researcher used the methodology stated in chapter three to build the system thus, System Analysis, System Design, System implementation and testing the developed system.

REFERENCES

1. Bithari, T.B.; Thapa, S.; Hari, K. (**2020**, *2*, 89–98), *Predicting Academic Performance of Engineering Students Using Ensemble Method. Tech. J.* [**Google Scholar**] [**CrossRef**]

2. Adejo, O.W.; Connolly, T. (**2018**, *10*, 61–75*), Predicting student academic performance using multi-model heterogeneous ensemble approach. J. Appl. Res. High. Educ.* [**Google Scholar**] [**CrossRef**]

3. Jakkula V. (2006), *Tutorial on Support Vector Machine (svm)*; School of EECS, Washington State University: Washington, DC, USA; Volume 37. [**Google Scholar**]

4. Jia, Y.S.; Jia, C.Y.; Qi, H.W. (August 2005), *A new nu-support vector machine for training sets with duplicate samples*. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21; Volume 7, pp. 4370–4373. [**Google Scholar**]

5. Breiman, L. Random forests. (**2001**), *Mach. Learn*, *45*, 5–32. [**Google Scholar**] [**CrossRef**]

6. Akar, Ö.; Güngör, O. (**2012*),*** *Classification of multispectral images using Random Forest algorithm. J.Geod. Geoinf, 1*, 105–112. [**Google Scholar**] [**CrossRef**]

7. Han, J.; Pei, J.; Kamber, M. (2011), *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Burlington, MA, USA. [**Google Scholar**]

8. Russell, S.J.; Norvig, P. (2016.), *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson: London, UK [**Google Scholar**]

9. Seber, G.A.; Lee, A.J. (2012), *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, Volume 329. [**Google Scholar**]

10. Kavitha, S.; Varuna, S. (19 November 2016); Ramya, R. *A comparative analysis on linear regression and support vector regression*. In Proceedings of the 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India; pp. 1–5. [**Google Scholar**

11. Wojtas, M.; Chen, K. **(2020),** *Feature Importance Ranking for Deep Learning*. *arXiv*, arXiv:2010.08973. [**Google Scholar**]

12. Ly-Huong T. Pham, Teja Desai-Naik, Laurie Hammond, & Wael Abdel Jabbar (Apr 8, 2021), *Implementation Methodologies*

13. (Hevner et al, m. e. (1992, 1995, 2004). *design science research. design science research*.

14. Adams et al, n. e. (1991,1992,2004). *design science methodology. theoretical frameworks to justify design research studies.*

15. alechina, N. (2014). *Entity-relationship*. Entity-relationship model.

16. Archer. (1984). *data collection*. data collection.

17. beal, V. (1982). application software. *application software*

18. Cooper. (1990, and 2000). design science methodologies. *design or development methodologies.*

19. Eekels et al, h. e. (1991, 1992, 2001). design science methodologies. *design or development methodologies.*

20. Simplilearn (2022), *Feasibility Study and Its Importance in Project Management*

21. Costa, E.B.; Fonseca, B.; Santana, M.A.; de Araújo, F.F.; Rego, J. *Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses*. Comput. Hum. Behav. 2017, 73, 247–256.

22. Liao, S.N.; Zingaro, D.; Thai, K.; Alvarado, C.; Griswold, W.G.; Porter, L. *A robust machine learning technique to predict low-performing students.* ACM Trans. Comput. Educ. (TOCE) 2019, 19, 1–19.

23. Hu, Y.H.; Lo, C.L.; Shih, S.P. *Developing early warning systems to predict students' online learning performance*. Comput. Hum. Behav. 2014, 36, 469–478.

24. Jayaprakash, S.; Krishnan, S.; Jaiganesh, V. *Predicting Students Academic Performance using an Improved Random Forest Classifier*. In Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 12–14 March 2020; pp. 238–243.

25. Francis, B.K.; Babu, S.S. *Predicting academic performance of students using a hybrid data mining approach*. J. Med Syst. 2019, 43, 162.

26. Nandeshwar, T. Menzies and A. Nelson (2011), *Learning patterns of university student retention.*

27. Alexander Ryabtsev (2019), *Why Python is Good for Artificial Intelligence and Machine Learning*. March 11.

APPENDICES

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M↓Data visualisation\M↓2)features visualisation\M↓b) Distribution plot\M↓b.1) Student status By goout\ # going out

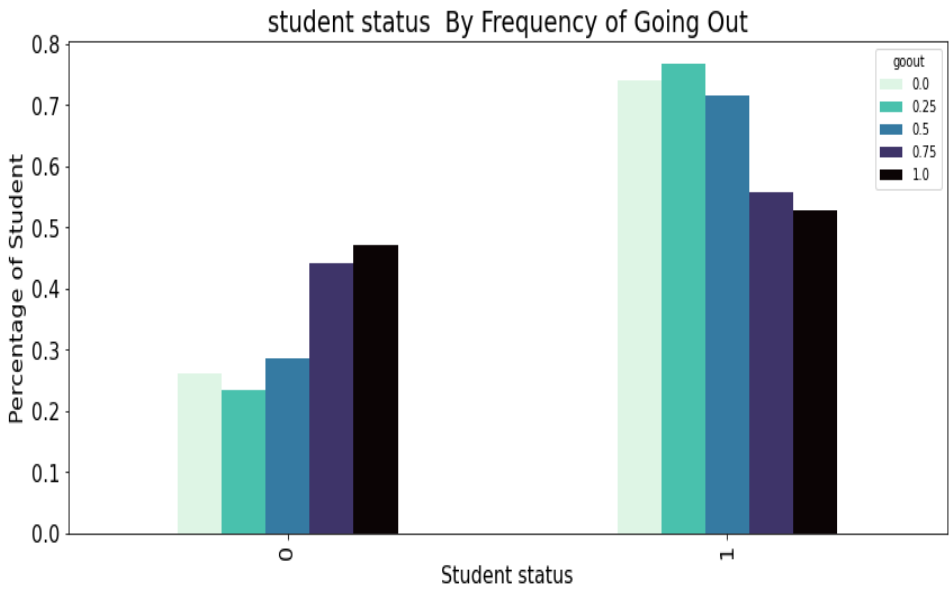
+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

```
array([0.75, 0.5 , 0.25, 0.  , 1.  ])
```

```
# going out
perc = (lambda col: col/col.sum())
index = [0,1]
out_tab = pd.crosstab(index=df.passed, columns=df.goout)
out_perc = out_tab.apply(perc).reindex(index)
out_perc.plot.bar(colormap="mako_r", fontsize=16, figsize=(14,6))
plt.title('student status  By Frequency of Going Out', fontsize=20)
plt.ylabel('Percentage of Student', fontsize=16)
plt.xlabel('Student status', fontsize=16)
```

[15] Python

```
Text(0.5, 0, 'Student status')
```



code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\Data visualisation\2)features visualisation\b) Distribution plot\b.2)Student status by romantic relation:

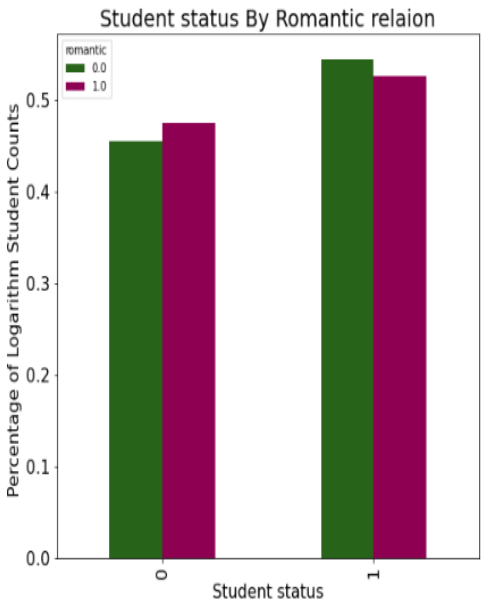
+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

b.2)Student status by romantic relation:

```
# romantic status
romance_tab1 = pd.crosstab(index=df.passed, columns=df.romantic)
romance_tab = np.log(romance_tab1)
romance_perc = romance_tab.apply(perc).reindex(index)
plt.figure()
romance_perc.plot.bar(colormap="PiYG_r", fontsize=16, figsize=(8,8))
plt.title('Student status By Romantic relaion', fontsize=20)
plt.ylabel('Percentage of Logarithm Student Counts ', fontsize=16)
plt.xlabel('Student status', fontsize=16)
plt.show()
# 0 in romantic mean no romantic relation
```

[16] Python

```
<Figure size 432x288 with 0 Axes>
```

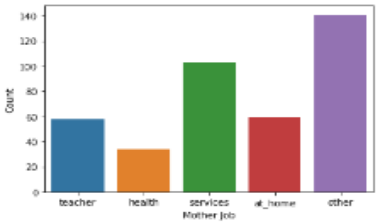


code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M4 Data visualisation\M4 2)features visualisation\M4 b) Distribution plot\M4 b.3)Student status by mother job and mother education\ # 1) mother j

+ Code + Markdown | Run All Clear Outputs of All Cells | Outline ··· Select Kernel

```
# 1) mother job
# Mjob distribution
f, fx = plt.subplots()
figure = sns.countplot(x = 'Mjob', data=dfv, order=['teacher','health','services','at_home','other'])
fx = fx.set(ylabel="Count", xlabel="Mother Job")
figure.grid(False)
```

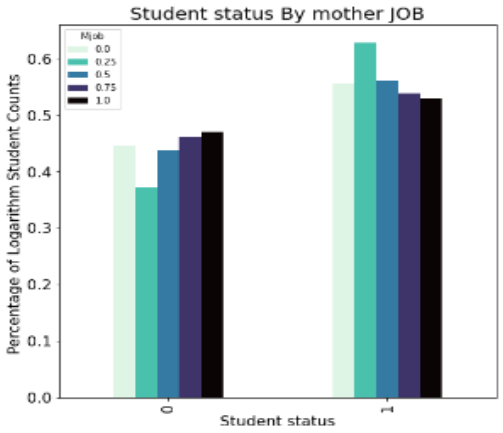
[17] Python




```
mjob_tab1 = pd.crosstab(index=df.passed, columns=df.Mjob)
mjob_tab = np.log(mjob_tab1)
mjob_perc = mjob_tab.apply(perc).reindex(index)
plt.figure()
mjob_perc.plot.bar(colormap="mako_r", fontsize=16, figsize=(8,8))
plt.title('Student status By mother JOB', fontsize=20)
plt.ylabel('Percentage of Logarithm Student Counts ', fontsize=16)
plt.xlabel('Student status', fontsize=16)
plt.show()
#'teacher': 0, 'health': 1, 'services': 2, 'at_home': 3, 'other': 4
```

[18] Python

```
<Figure size 432x288 with 0 Axes>
```



code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M↓Data visualisation\M↓2)features visualisation\M

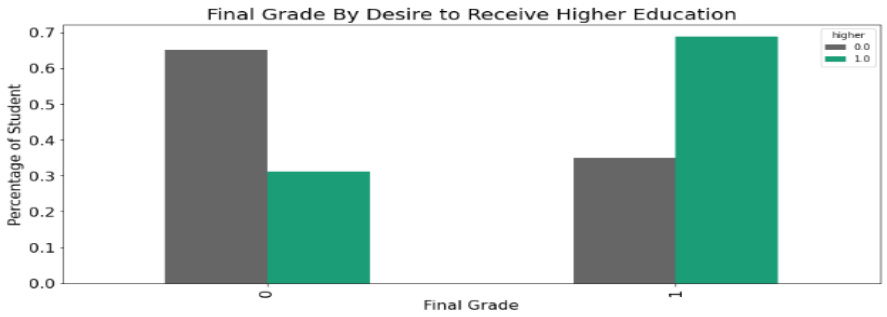
+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

b.4)Student status by deseire to take heigher education:

```
higher_tab = pd.crosstab(index=df.passed, columns=df.higher)
higher_perc = higher_tab.apply(perc).reindex(index)
higher_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('Final Grade By Desire to Receive Higher Education', fontsize=20)
plt.xlabel('Final Grade', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[20] Python

```
Text(0, 0.5, 'Percentage of Student')
```



C4:Most of people who passed the exam want to take heigher education sow it could be a good idea to encourage your kids or students to take heigher education.

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M↓Data visualisation\M↓2)features visualisation\M

+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

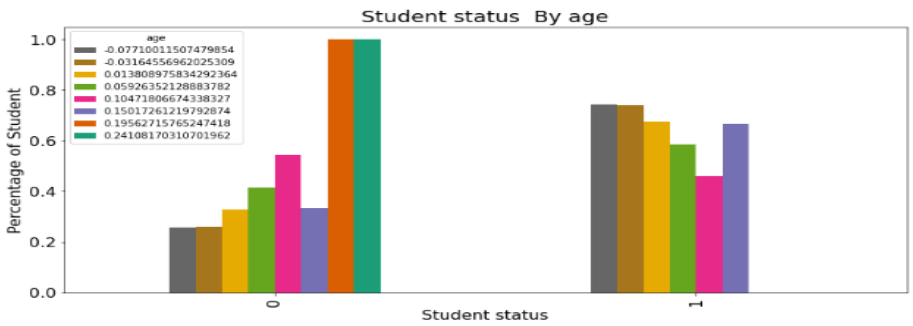
be a good idea to encourage your kids or students to take heigher education.

b.5)Student status by age

```
#impact of age
higher_tab = pd.crosstab(index=df.passed, columns=df.age)
higher_perc = higher_tab.apply(perc).reindex(index)
higher_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('Student status  By age', fontsize=20)
plt.xlabel('Student status', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[21] Python

```
Text(0, 0.5, 'Percentage of Student')
```



C5:Age also play an important role in student sucess, most of people who passed the

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\Data visualisation\2)features visualisation\

+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

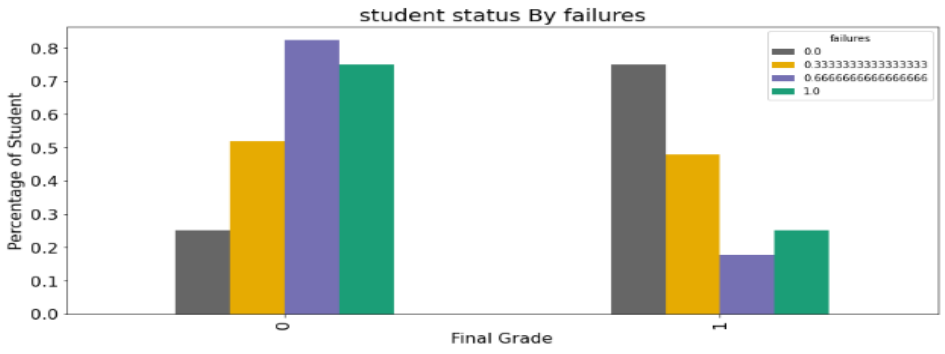
```
#impact of failures
```

[22] Python

```
fail_tab = pd.crosstab(index=df.passed, columns=df.failures)
fail_perc = fail_tab.apply(perc).reindex(index)
fail_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('student status By failures', fontsize=20)
plt.xlabel('Final Grade', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[23] Python

```
Text(0, 0.5, 'Percentage of Student')
```



C6:most of people who passed the exam had 0failures sow it could be a good choice to study hard and pass all the grades

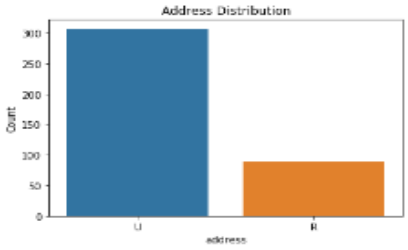
code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M4 Data visualisation\M4 2)features visualisation\M4 b) Distribution plot\M4 b.3)Student status by mother job and mother education\ # 1) mother j

+ Code + Markdown | Run All Clear Outputs of All Cells | Outline ··· Select Kernel

```
#first let's see the destribution of students who live in urban or rural area
f, fx = plt.subplots()
figure = sns.countplot(x = 'address', data=dfv, order=['U','R'])
fx = fx.set(ylabel="Count", xlabel="address")
figure.grid(False)
plt.title('Address Distribution')
```

[24] Python

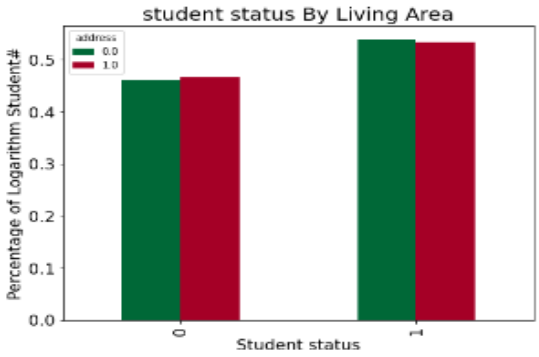
```
Text(0.5, 1.0, 'Address Distribution')
```




```
ad_tab1 = pd.crosstab(index=df.passed, columns=df.address)
ad_tab = np.log(ad_tab1)
ad_perc = ad_tab.apply(perc).reindex(index)
ad_perc.plot.bar(colormap="RdYlGn_r", fontsize=16, figsize=(8,6))
plt.title('student status By Living Area', fontsize=20)
plt.ylabel('Percentage of Logarithm Student#', fontsize=16)
plt.xlabel('Student status', fontsize=16)
```

[25] Python

```
Text(0.5, 0, 'Student status')
```



C7:Area doesn't had an impact on student performance even people with good results live in contry side

b.8) Student status by alchool consumption :

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M4 Data visualisation\M4 2)features visualisation\M4 b) Distribution plot\M4 b.3)Student status by mother job a

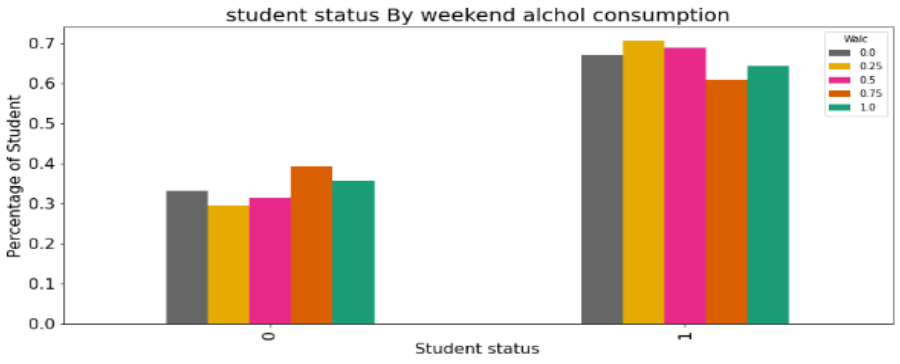
+ Code + Markdown | Run All Clear Outputs of All Cells | Outline ··· Select Kernel

b.8) Student status by alchool consumption :

```
#impact of weekend alcohol consumption in student performance
alc_tab = pd.crosstab(index=df.passed, columns=df.Walc)
alc_perc = alc_tab.apply(perc).reindex(index)
alc_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('student status By weekend alchol consumption', fontsize=20)
plt.xlabel('Student status', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[26] Python

```
Text(0, 0.5, 'Percentage of Student')
```




```
# weekend alcohol consumption
# create good student dataframe
good = df.loc[df.passed == 1]
good['good_alcohol_usage']=good.Walc
# create poor student dataframe
poor = df.loc[df.passed == 0]
poor['poor_alcohol_usage']=poor.Walc
plt.figure(figsize=(10,6))
p1=sns.kdeplot(good['good_alcohol_usage'], shade=True, color="r")
p1=sns.kdeplot(poor['poor_alcohol_usage'], shade=True, color="b")
plt.title('Good Performance vs. Poor Performance Student Weekend Alcohol Consumption', fontsize=20)
plt.ylabel('Density', fontsize=16)
plt.xlabel('Level of Alcohol Consumption', fontsize=16)
```

[27] Python

```
<ipython-input-27-3df7c3a0a2b6>:4: SettingWithCopyWarning:
```

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M↓Data visualisation\M↓2)features visualisation\M↓b) Distribution plot\M↓b.3)Student status by mother j

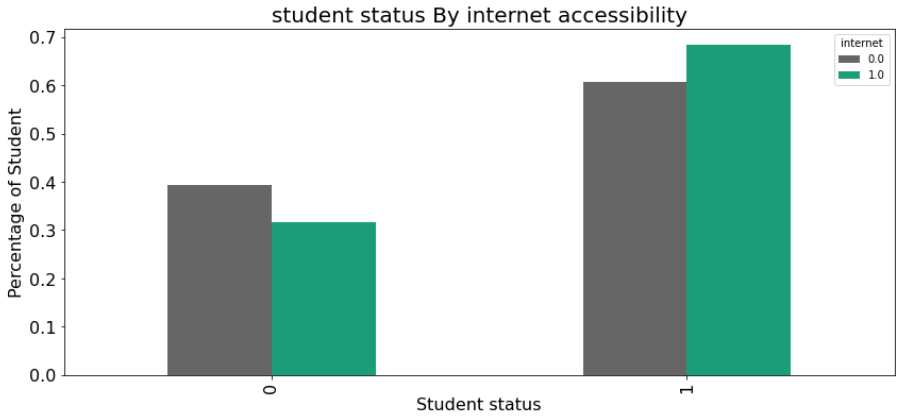
+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

b.9) Student status by internet accessibility:

```
alc_tab = pd.crosstab(index=df.passed, columns=df.internet)
alc_perc = alc_tab.apply(perc).reindex(index)
alc_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('student status By internet accessibility', fontsize=20)
plt.xlabel('Student status', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[28] Python

```
Text(0, 0.5, 'Percentage of Student')
```



code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M↓Data visualisation\M↓2)features visualisation\M↓b) Distribution plot\M↓b.3)Student status by mother job and mother education:\ # 1) mot

+ Code + Markdown | Run All | Clear Outputs of All Cells | Outline ··· Select Kernel

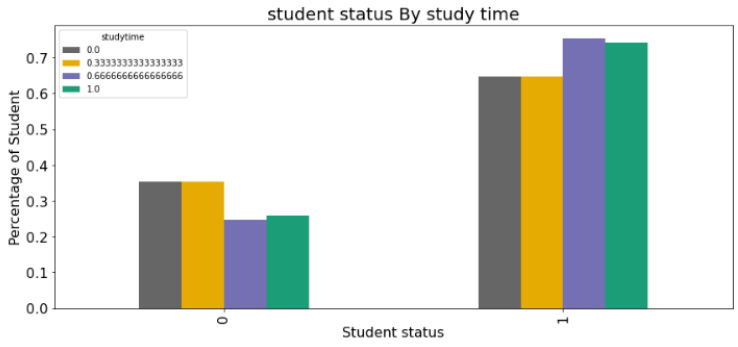
C9:Most of people who passed the exam had the accessibility to internet ,sow we should provide a fair materials's education

b.10) Student status by wekelly Study time :

```
stu_tab = pd.crosstab(index=df.passed, columns=df.studytime)
stu_perc = stu_tab.apply(perc).reindex(index)
stu_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('student status By study time', fontsize=20)
plt.xlabel('Student status', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

[29] Python

```
Text(0, 0.5, 'Percentage of Student')
```



C10:Most of people who passed the exam study 5-10 hours weekely

code.ipynb C:\Users\pc\Desktop\ada\spp\code.ipynb\M4 Data visualisation\M4 2)features visualisation\M4 b) Distribution plot\M4 b.3)Student status by mother job and mother education:\ # 1) mot

Code | Markdown | Run All | Clear Outputs of All Cells | Outline | Select Kernel

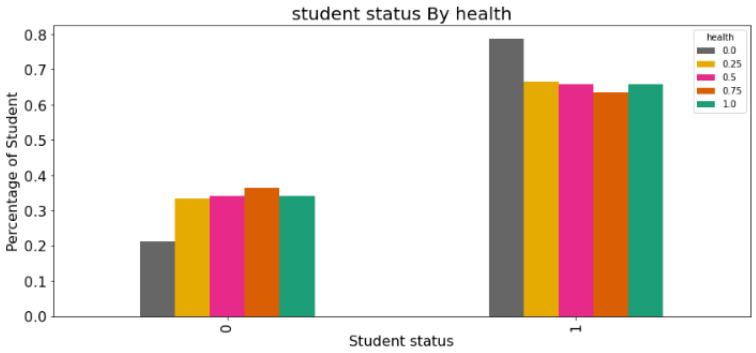
C10:Most of people who passed the exam study 5-10 hours weekely

b.11)Student status by health:

```
he_tab = pd.crosstab(index=df.passed, columns=df.health)
he_perc = he_tab.apply(perc).reindex(index)
he_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
plt.title('student status By health', fontsize=20)
plt.xlabel('Student status', fontsize=16)
plt.ylabel('Percentage of Student', fontsize=16)
```

Python

```
Text(0, 0.5, 'Percentage of Student')
```



C11:most of student who fails the exam don't have a good health