Sudan University of Science and Technology

College of Graduate Studies

A Dynamic Data Mining Model to Predict Student Graduation Level

(Case Study: College of Computer Science - Sudan University of Science and Technology)

نموذج تنقيب بيانات ديناميكي للتنبؤ بمستوي تخرج الطالب.

(دراسة حالة كلية علوم الحاسوب بجامعة السودان للعلوم والتكنلوجيا)

Dissertation

Submitted in Partial Fulfilment of the requirements

For the degree of Doctor of Philosophy in Computer Science and Information Technology

By:

Haitham Alagib Alsuddig Hamza

Supervisor:

Professor Piet Kommers

NOV-2020

DECLARATION

I hereby declare that this thesis is the result of my own investigation, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at Sudan University of Science and Technology or other institutions.

Haitham Alagib Alsuddig Hamza

Signature Date **30/11/2020**

ACKNOWLEDGEMENTS

I am grateful to the Almighty Allah for giving me the opportunity to complete my Ph.D. thesis. May peace and blessing of Allah be upon His beloved Prophet Muhammad (SAW), his family, and his companions?

Firstly, I would like to express my sincere gratitude to my advisor Professor Piet Kommers for the continuous support of my Ph.D. study, and for patience, motivation, and immense knowledge. His guidance helped me all the time regarding the research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Also, I would like to express my sincere gratitude to all the family of the college of computer science and information technology.

I would like to thank my all friends for encouragement and support, valuable comments, and encouragement during the Ph.D. journey.

I deeply thank my parents, Alagib Alsuddig and Zainab Taha for their unconditional trust, timely encouragement, and endless patience. It was their love that raised me up again when I got weary.

I would like to thank my all brothers and sisters for supporting me spiritually throughout writing this thesis and my life in general.

Last but not least, I thank with love to Zainb and Ahamed and my wife and my daughter and my son for her great companion, love, support, encouragement, interest, and help throughout this agonizing period in the most positive way.

ABSTRACT

This study sought to develop an intelligent and dynamic model to predict early on a student's graduate level after graduation. For the purpose of early evaluation and to avoid graduating by a critical level, through a number of independent variables: some academic achievement scores in the high school diploma, the result of the first year and the score of the achievement test for some subjects for the first year of study, the student community for the scores of 2012, 2013 and 2014 at the Sudan University of Science and Technology, which numbered 326 Student. This study is based on three research questions Q1: It possible build a model to predict the student's graduation level in a dynamic way that makes it easy to deploy and generalize? Q2: Is it possible to reprocess the model data set and choose the inputs automatically? Q3: can we identify courses that most influence the student's graduation level through available academic data? In this research, we used the concept of data mining and machine learning (ML) algorithms and techniques to build a dynamic predictive model that can be used to perform the prediction process, where a feature selection technique was used to let the model select the best attributes automatically Which have high relationship and effect of the dependent variable, and then a linear regression algorithm was used to predict the student's rate upon graduation. We also used coefficient and correlation analysis and other statistical methods that were implemented in order to obtain the expected initial results and to know the possibility of building the model as a preliminary analytical study for the research. And we used R-squared method to evaluate the model. And we have reached through research that a model has been built and we can predict a student's graduation level in a dynamic way that facilitates to deploy it and generalization. We also reached the ability to pre-process data and choose appropriate inputs for the model automatically, which made the model flexible, reliable and usable. The model was applied and tested on students who graduated for the years 2016, 2017 and 2018. The research concluded with a number of recommendations and future work that could become a continuation of this study. An application has also been designed that allows using the model and obtaining the results of student's rate expecting at graduation. The application also allows uploading any data that is required to be analyzed and trained on the model through easy and simple user interfaces.

مستخلص البحث

سعت هذه الدراسة إلى تطوير نموذج ذكي وديناميكي للتنبؤ مبكرًا بمستوى تخرج الطالب. لغرض التقويم المبكر وتفادي تخرجه بمستوي حرج ، من خلال عدد من المتغيرات المستقلة مثل درجات التحصيل الأكاديمي في شهادة الثانوية العامة ، ونتيجة السنة الأولى ودرجة الاختبار التحصيلي لبعض المواد للسنة الأولى من الدراسة . مجتمع الدراسة شمل طلاب الدفعات 2012 و 2013 و 2014 كلية علوم الحاسوب بجامعة السودان للعلوم والتكنولوجيا والبالغ عددهم 326 طالب. تستند هذه الدراسة إلى ثلاثة أسئلة بحثية :Q1 هل من الممكن بناء نموذج للتنبؤ بمستوى تخرج الطالب بطريقة ديناميكية تجعل من السهل نشره وتعميمه؟ س 2: هل من الممكن إعادة معالجة مجموعة بيانات النموذج واختيار المدخلات بطريقة تلقائية؟ س 3: هل يمكننا تحديد الكورسات التي تؤثر بشكل أكبر على مستوى تخرج الطالب من خلال البيانات الأكاديمية المتاحة؟.استخدمنا في هذا البحث مفهوم التنقيب عن البيانات وخوارزميات وتقنيات التعلم الآلي (ML) لبناء نموذج تنبؤي ديناميكي يمكن استخدامه لأداء عملية التنبؤ ، حيث تم استخدام تقنية اختيار الميزة للسماح للنموذج بتحديد أفضل السمات تلقائيًا والتي لها علاقة عالية وتأثير علي المتغير التابع ، ومن ثم تم استخدام خوارزمية الانحدار الخطي للتنبؤ بمعدل الطالب عند التخرج. كما استخدمنا تحليل المعامل والارتباط وطرق إحصائية أخرى تم تنفيذها من أجل الحصول على النتائج الأولية المتوقعة ومعرفة امكانية بناء النموذج كدراسة تحليلية أولية للبحث. واستخدمنا طريقة R-squared لتقييم النموذج. وقد توصلنا من خلال البحث إلى أنه تم بناء نموذج الذي يمّكننا التنبؤ بمستوى تخرج الطالب بطريقة ديناميكية تسهل نشره وتعميمه وكذلك توصلنا أيضًا إلى القدرة على المعالجة المسبقة للبيانات واختيار المدخلات المناسبة للنموذج تلقائيًا ، مما جعل النموذج مرنًا وموثوقًا به وقابل للإستخدام. تم تطبيق النموذج واختباره على الطلاب الذين تخرجوا للأعوام 2016 و 2017 و 2018. واختتم البحث بعدد من التوصيات والعمل المستقبلي التي يمكن ان تصبح مواصلة لهذه الدراسة. كما تم تصميم تطبيق يسمح باستخدام النموذج والحصول على نتائج معدل الطالب فيما يتعلق بالتخرج .كما يسمح التطبيق بتحميل أي بيانات مطلوب تحليلها وتدريبها على النموذج من خلال واجهات مستخدم سهلة وبسيطة.

Table of Contents

CONTENTS

LIST OF PUBLICATIONS

1- Hamza, Haitham Alagib Alsuddig, and Piet Kommers. "A review of educational data mining tools & techniques." International Journal of Educational Technology and Learning 3.1 (2018): 17-23.

2- Haitham Alagib Alsuddig, Piet Kommers. "A Prediction Model of Students level at Graduation Using Educational Data Mining". International Journal of Computer Science Trends and Technology (IJCST) V7 (5): Page (47-51) Sep-Oct 2019. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DM	Data mining.
EDM	Education Data Mining.
ICT	Information communication technology.
AIT	Asian Institute of Technology.
BN	Bayesian Net.
NN	Neural Net.
DT	Decision Tree.
NB	Naïve Bayes.
ML	Machine Learning.
AI	Artificial Intelligence.
KNN	K-Nearest Neighbors.
MLR	Multiple Linear Regression.
MANFIS-S	Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets.
ANOVA	Analysis of variance.
MANOVA	Multivariate analysis of variance.
GCR	Greatest Characteristic Root.
GLM	Generalized Linear Model.
MDA	Multiple Discriminant Analysis.

LIST OF APPENDICEX

CHAPTER ONE INTRODUCTION

1.1 Overview

Data mining (DM) is the process of analyzing massive volumes of data to discover business intelligence that helps companies solve problems, mitigate risks, and seize new opportunities. (talend.com, 2020) This branch of data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for ore. Both processes require sifting through tremendous amounts of material to find hidden value. (talend.com, 2020) DM can answer business questions that traditionally were too time consuming to resolve manually. Using a range of statistical techniques to analyze data in different ways, users can identify patterns, trends and relationships they might otherwise miss. They can apply these findings to predict what is likely to happen in the future and take action to influence business outcomes. DM is used in many areas of business and research, including sales and marketing, product development, healthcare, and education. When used correctly, data mining can provide a profound advantage over competitors by enabling you to learn more about customers, develop effective marketing strategies, increase revenue, and decrease costs. (talend.com, 2020).

Machine learning is a computer programming technique that uses statistical probabilities to give computers the ability to "learn" without being explicitly programmed. In essence, machine learning is getting computers to learn the way humans do, improving their learning and knowledge over time autonomously. The idea is to get computers to act without being explicitly programmed. Machine learning utilizes development programs that can adjust when exposed to different external inputs. (Kotsiantis, 2017)

Regression techniques are useful for identifying the nature of the relationship between variables in a dataset. Those relationships could be causal in some instances, or just simply correlate in others. Regression is a straightforward white box technique that clearly reveals how variables are related. Regression techniques are used in aspects of forecasting and data modeling. (Fox, 2019)

Prediction is a very powerful aspect of data mining that represents one of four branches of analytics. Predictive analytics use patterns found in current or historical data to extend them into the future. Thus, it gives organizations insight into what trends will happen next in their data. There are several different approaches to using predictive analytics. Some of the more advanced involve aspects of machine learning and artificial intelligence. However, predictive analytics doesn't necessarily depend on these techniques —it can also be facilitated with more straightforward algorithms (talend.com, 2020).

There are five Key Steps of Data mining techniques:

Data Integration

First, you need to collect data from various sources and integrate them into one portal (database). This data could be anything from useful to not-so-useful, qualitative to quantitative and continuous to discrete.

Data Selection

As in the first step, we collected the data for our database. Now in this step, we have to mark and select the most relevant data which we need to carry forward.

Data Cleaning

As we had collected the data from various sources. So, there are chances that the data may contain some missing figures, errors, or inconsistency. So, to get rid of that, we need to apply different techniques.

Modeling

For proper modeling of data first, we need to create data sets. Each data set contains information about a particular subject. And, the immediate next step should be the testing of data to confirm its quality.

Evaluation

In this phase, data is evaluated in such a manner that it will meet the business objectives. Moreover, in this phase, some new business r

Owing to digitization of academic processes, universities are generating a huge amount of data pertaining to students in electronic form. It is crucial for them to effectively transform this massive collection of data into knowledge which will help teachers,

administrators and policy makers to analyze it to enhance decision making. Furthermore, it may also advance the quality of the educational processes by providing timely information to different stakeholders (Han, 2011).

In knowledge management process, data mining technique can be used to extract and discover the valuable and meaningful knowledge from a large amount of data. Nowadays, data mining has given a great deal of concern and attention in the information industry and in society as a whole. This technique is an approach that is currently receiving great attention in data analysis and it has been recognized as a newly emerging analysis tool.

Education is the foundation of nation building and development. Accordingly, the last decade has seen increasing interest in finding ways to improve the quality of teaching and to improve teaching and learning. The development of educational information systems and the proliferation of learning techniques led to the availability of many data on the educational process, which led researchers to think the need to apply methods of data mining to extract useful information from the data provided by the educational systems, and led to the emergence of independent research areas such as exploration in educational data. Education Data Mining (EDM) (Learning Analytic) Learning data mining techniques seek to access educational data repositories and extract useful information that helps to better understand the educational process and improve the teaching and learning process. In the educational data the same approach used in traditional methods of data mining of the need to understand the environment to be dealt with and then collect the data and then cleaned and arranged and select the techniques that can be applied and finally interpret the results and verify the validity of the techniques applied.

Education Data Mining is growing at a very fast pace. The main aim of EDM is to develop methods to explore the unique type of data that comes from educational institutes and to use those methods to understand the students and their learning environments. EDM deals with mining of large data sets of educational data to answer educational research questions. These data sets may come from learning management systems, interactive learning environments, intelligent tutoring systems, or any system used in a learning context. (Hamza, 2018)

The educational institutions management process is one of the difficulties faced by the supervisors because of the large size and complexity of its structure and the multiple sources of data. Therefore, the educational institution faces several problems during the management of the educational process, including academic, financial and administrative (brief, 2012).

. With the abundance of existing data and stored in so-called databases, many researchers have become the subject of question, and with the increase in the spread of huge storage warehouses (data warehouses) it has become necessary to find techniques, methods and means to extract information and knowledge from such data accumulated and exploited in solving problems and making decisions , Using modern computer applications Which is considered a modern smart technology based on making the computer "think as human thought and do man" and what is known as artificial intelligence, the idea of disclosure and exploration of these data came in smart ways to help in solving problems and making decisions. Artificial statistics, machine generalization, and databases are considered a step-in exploring knowledge of databases.

One of the biggest problems affecting in the performance of the educational process in universities is Student cannot know which the main factors that have big impact of his academic performance in specific field of study. Higher education institutions are beginning to use analysis to improve services they provide and for increasing student upgrade and retention. The U.S. National Department of Education and Technology Plan, as one part of its model for 21 century learning powered by technology, envisions ways of using data from online learning systems to improve instruction (Donlevy, 2005).

The purpose of this research to know some of the methodologies and tools & Techniques of EDM which used for build predictive model for expected student level at the final year of his study and evaluate the performance of the algorithms for classification and exploration of data to obtain the highest accuracy and the lowest proportion of the line when used in mining and build the models.

Data mining performs two basic operations as follows:

1) Forecasting: Data mining aims to generate forecasts with attribution for the attribute Generic or object attributes of unknown classification data, and uses the learning model Available for forecasting, classification and regression are the two basic types of prediction

model. The first is used to predict the discrete or symbolic value, while regression is used for prediction Continuous values, that is, the answer to a question about purchasing goods over the internet either X or Y and this applies to the first case, which is the classification, as for the prediction of prices Stocks and trends do so through regression functions. Forecasting models can determine Market benefits and risks, as can predict the rates of consumption of land resources.

2) Description: A potential data model available that summarizes relationships

Documentary and interpretative role, relationship analysis is often used to describe a model with characteristics Strong relational to derive important models for finding relationship between data. Expresses the derivation Formula properties for general properties of a data set from a data warehouse, or Finding other features to distinguish between the characteristics of the same method, such as: the derivation of features and to distinguish it from other cases. Despite its logic, it is possible to dig a role Aggregation creates many important interactions, and this is called market basket analysis the famous, who was a secret weapon in the supermarkets, where he could analyze the market basket helping to find sales of stores and their goods (Padhy, 2012).

This study belongs to Data mining field, which is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

Predictive causal analytics used to build a model which can predict the possibilities of a particular event in the future, you need to apply predictive causal analytics.

Prediction is the process of making estimates for the future based on past and present data and the most trend analysis. Prediction helps to make decisions with a temporal and spatial dimension because of its large role in tactical and strategic decision making until it is said that the decision maker is only a consumer of information produced by the forecasting device. (www. Coursehero.com)

This research using only students' academic records to build a dynamic prediction model, ignoring by that other variables that could affect students 'learning outcomes, like: attendance, instructor course delivery, and many others. That was because the main focus in this research was to generate the dynamic predictive model to investigate feasibility of

utilizing from available dataset size in predicting student performance and to shed the light on the possibility of benefiting from the academic students 'records data.

1.2 Motivation and problem statement

Building dynamic Prediction model can use for student's graduation level at an average starting from the early stages of his or her academic career which helps the educational administration to develop appropriate methodologies to improve the level of students expected to perform poorly or to support students who are expected to perform well.

Experiments have shown that graduates with low rates find it difficult to recruit and find good university admissions to complete their studies. Thus, there is a need to limit students who are expected to graduate at a critical level to help them correct early. By securing an appropriate academic plan after discovering flaws and weaknesses, the early detection was the greater the chance of correcting the path.

Using information technology in building a dynamic model can help for Prediction of student's graduation level at an average starting from the early stages of his or her academic career helps the educational administration to develop appropriate methodologies to improve the level of students expected to perform poorly or to support students who are expected to perform well.

1.3 Research Questions

To build Dynamic Model that can be used to predict the level of graduating students from universities and colleges as general, this thesis concentrates on three main research questions:

Question 1: It possible build a model to predict the student's graduation level in a dynamic way that makes it easy to deploy and generalize?

Question 2: Is it possible to reprocess the model data set and choose the inputs automatically?

Question 3: Can we identify courses that most influence the student's graduation level through available academic data?

1.4 The potential Benefits

The importance of this research lies in employing information communication technology (ICT) and data mining methods and the concept of machine learning to provide a dynamic predictive model that can be applied in various educational environments and all disciplines to predict the student's level upon graduation.

ICT can potentially provide many benefits such as:

• Contribute to the study of how to improve the quality of the learning process outputs as measured by the graduate level, starting from ensuring the quality of some elements of the educational process.

• A generalizable model proposal to study the effectiveness of study plans.

•Prediction helps to make decisions with a temporal and spatial dimension because of its large role in tactical and strategic decision making until it is said that the decision maker is only a consumer of information produced by the forecasting device.

1.5 Aim Objectives

To this end with the conjunction of the problem statement and research questions sections, this research is concerned with finding a general dynamic predictive model that can be used as an early warning performance to know the expected level of students 'graduation in the final year.

The major objectives here are:

- To propose a generalizable a dynamic model to predict the level of the student graduation from early stages of the academic career to intervene for the purpose of evaluation and reorientation.
- To suggest dynamic predictive model for providing assistance in discovering the most important causes and factors that may affect the student's graduation at a critical level.
- To clarify the possibility of predicting students 'graduation level from an early date depending on the student's available academic record.

- To Study the effect of the student's academic achievement results on the graduation rate.

1.6 Methods

Methods from Data Mining and Machine Learning Algorithms had been used in this thesis for building the dynamic predictive model.

Different methods and techniques of data mining were using during the building the model.

To utilize from the provisioned dataset, Feature selection have been used to prepare the dataset and prediction variable and Microsoft Excel and Python jupyter Environment were used for that.

For building the model and training the dataset we have been used linear regression algorithm and the accuracy_score method was used to evaluate the model.

For applying the model, we used the data which collected from Ministry of Higher Education Admission Office and Sudan University of Science and Technology Faculty of computer of Science, academic year 2011-2012, 2012-2013, and 2013-2014. To evaluate the performance of the model and then obtain the expected result at the level of the student upon graduation.

The researcher used the following list of technologies and tools:

- Python
- Excel
- SPSS

1.7 Contribution

The main contributions of this thesis are an attempt to use the data mining techniques and machine learning algorithm to build a dynamic model can be generalized to predict the level of student graduation in the early stages of their academic path.

1.8 Structure of the Thesis

In the present chapter, the background about the theme, problem and motivations, objectives and research questions, and potential benefits are illustrated. While the next paragraphs will describe the organization of the remaining chapters as follows:

Chapter 2 Literature Review

Chapter will present an overview of related research in regard Educational Data Mining Tools & Technique for Building Predictive Model of student performance.

Chapter 3 Methodology.

In this chapter, we will explain the method used in this thesis in more details including concepts, model, phases and tools.

Chapter 4 Analysis and Findings.

In this chapter, we will explain how to build predictive model method used in this thesis in more details including concepts, analysis, model, phases and tools.

Chapter 5 Design and implementation.

In this chapter we will explain how to construct, test, evaluate and apply the proposed model.

Chapter 6 Results and Discussion.

This chapter will present the results after implementation and test the model and evaluated the results by comparing the actual results and predictive results and then discusses them.

Chapter 7 Conclusion.

This chapter presents the final takeaways and main findings from this thesis. In addition, it gives recommendations for future works and limitation need to be covered.

References

Appendixes

CHAPTER TWO LITERATURE REVIEW

2.1 Literature Review

Our study Centered of how can use the data mining techniques, machine learning technology's and prediction algorithms for building dynamic predictive model. This review identifies strengths and shortcomings in the existing literature and highlights the unique contribution that the study makes to the field.

Many studies have been conducted regarding Education Data Mining and prediction of student performance, in relation to different topics and from various perspectives, in order to highlight the most important factors for developing and implementing effective techniques for building predictive model. There are many studies can be presented. This review identifies strengths and shortcomings in the existing literature and highlights the unique contribution that the study makes to the field.

One of studies applied data mining techniques in predicting students' academic performance by considering the data of two different academic institutes; Asian Institute of Technology (AIT), Thailand, and Can the University (CTU), Vietnam. The AIT datasets included the Master programs. The students' GPA at the end of first year of their Master program is predicted from their admission information, including academic institute, entry GPA, English proficiency, marital status, Gross National Income, age, gender, and TOEFL score. In the case of the CTU dataset, the students' GPA at the end of the third year is predicted using attributes such as English skill, entry marks range, field of study, faculty, gender, age, family, job, religion, and also second-year GPA. For both case studies, the authors have done predictions for 4 classes (Fail, Fair, Good, and Very Good), 3 classes (Fail, Good, and Very Good) and 2 classes (Fail and Pass). Two data mining algorithms were applied, namely decision trees and Bayesian network. Decision trees produced better accuracies. For 2 classes the ac-curacy was: CTU 92.86% and AIT 91.98%; for 3 classes: CTU 84.18% and AIT 67.74%; and for 4 classes CTU 66.69% and AIT 63.25%. The accuracy of predictions was measured using a 10-fold cross-validation: 9/10 of the data was used to build the model that was tested on 1/10 of the data, and this process was repeated 10 times. Thus, a single cohort was used to build the prediction model and to evaluate it (Nghe, 2007).

Another research studied the academic performance of students in high school and bachelor degree studies in Iran, and compared their results with the results of a similar study done in India. They considered the data of 500 students having a high school level and 600 students having a Bachelor degree level. They applied various classifiers such as naïve Bayesian networks, C4.5 decision tree, Random Forest and Neural Networks, and meta-classifiers such as Bagging, Boosting or AdaBoost to classify students into 2 classes: Pass, Fail. The results revealed that features such as parent educational level, past examination results and gender impact the prediction. Best accuracy of 96% was obtained with C4.5 decision tree. The results were comparable with similar studies conducted in India (Oskouei, 2014).

Another study applied the decision trees and Naïve Bayes algorithms to predict the likelihood of success/failure at university. The dataset consisted of 11,873 undergraduate students from the Debre Markos University, Ethiopia. His findings indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, gender, number of students in a class, number of courses given in a semester, and field of study were the major factors affecting the student performance. The highest prediction accuracy was 92.34% obtained with the decision tree algorithm using 10-fold cross validation (Yehuala, 2015).

One of the studies analyzed how well undergraduate achievements can predict graduate-level performance. They used the data of 171 student records in the Bachelor and Master programs in Computer Science at ETH Zurich, Switzerland. Employing linear regression models in combination with different variable-selection techniques, their findings showed that undergraduate level performance can explain as much as 54% of the variance in graduate-level performance. They identified the third-year grade point average as the most significant explanatory variable, whose influence exceeds the one of grades earned in challenging first-year courses (Zimmermann, 2015).

In one research the relationship between students' demographic attributes, qualification on entry, aptitude test scores, performance in first year courses and their overall performance in their program using regression technique was investigated. In their study, based on the data of a single cohort comprising 85 students of the School of Computing and Information Technology at the University of Technology, Jamaica (UTECH), they found a strong correlation between

performance in a first-year computer science courses and the students' overall performance in the program, with a correlation of 0.499 that explains 70.6% of the students' overall performance (Golding, 2006).

Many studies around the world were interested in applying data mining algorithms to discover knowledge in universities. One of the most important of these studies is a study concerned with the applications of data mining in the field of higher education, Focused on the input of the educational process and its outputs and how they affect each other, The study used the method of neural networks to explore data, The results showed diverse relationships between curricula , multiple hours , the nature of students and between the graduates and the jobs they occupy, As well as other useful conclusions for decision-makers at universities.

Another study presents an applied study in data mining and knowledge discovery. It aims at discovering patterns within historical students' academic and financial data at UST (University of Science and Technology) from the year 1993 to 2005 in order to contribute improving academic performance at UST. Results show that these rules concentrate on three main issues, students' academic achievements (successes and failures), students' drop out, and students' financial behavior. Clustering (by K-means algorithm), association rules (by Apriority algorithm) and decision trees by (J48 and Id3 algorithms) techniques have been used to build the data model. Results have been discussed and analyses comprehensively and then well evaluated by experts in terms of some criteria such as validity, reality, utility, and originality. In addition, practical evaluation using SQL queries have been applied to test the accuracy of produced model (rules), the shortcomings of this study are as follows:

a) Not included of Associate student data, educational staff and other university branches.

b) Not included in scholarship data and lots of personal data for students.

c) Not included attendance and absence data for students (Al-Shargabi, 2010).

Another study was conducted in Ethiopia In Debre_Markos University study has shown that data mining techniques can be applied by higher education institutions or universities in determining student failure/success rate so that managing students' enrolment at the beginning of the year, assist students before they reached risk of failure, effective resource utilization and cost

minimization, helping and guiding administrative officers to be successful in management and decision making. The study applied data mining technology to the data of university students for the purpose of forecasting the success or failure of students, the study used CRISP methodology the analysis was carried out by the WEKA program and the forecast model was built the study found the main class, number of courses given in a semester, and field of study are the major factors affecting the student performances (Asif, 2017).

One of the studies discussed that Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. This study is an attempt to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses (Silva, 2017).

Ramaswamy has written research paper which focused on predicting students' characteristics or academic performances in various educational institutions. This paper focus on students' performance as a slow learner or fast Lerner. For that they applied various data mining techniques and compare the accuracy based on student's attributes. For assessing the goodness of a predictor, an extensive study on the student data set was conducted by applying five individual classifiers J48 (J48), Bayesian Net (BN), Neural Net (NN), Decision Tree (DT), and Naïve Bayes (NB) (Ramaswami, 2014).

Other research use data mining methodologies to study and analyses the school students' performance based on classification techniques which is useful to gauge students' performance and deals with the accuracy, confusion matrices and the execution time taken by the various classification data mining algorithms. The decision tree classifier C4.5 (J48), Random Forest, Neural Network (Multilayer Perception) and Lazy based classifier (IB1) Rule based classifier (Decision Table) were enforced in weak (Mythili, 2014).

One of researches have conducted comparison of data mining algorithms for clustering published. These algorithms are among the most influential data mining algorithms in the research community. A Knn algorithm is more sophisticated approach, k-nearest neighbor (KNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood.

KNN classification is an easy to understand and easy to implement classification technique. Despite its simplicity, it has done perform well in many situations (Kushwah, 2012).

Another study explored the opportunities of the Education data mining for improving students' performance. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. The study used Classification methods like decision trees, Bayesian network etc. Which can be applied on the educational data for predicting the student's performance in examination. This prediction will help to identify the weak students and help them to score better marks. The C4.5, ID3 and CART decision tree algorithms are applied on engineering student's data to predict their performance in the final exam. The results of this study provided predicted the number of students who are likely to pass, fail or promoted to next year, steps to improve the performance of the students who were predicted to fail or promoted and the comparative analysis of the results states that the prediction has helped the weaker students to improve and brought out betterment in the result (Yadav, 2012).

One of the studies showed how useful data mining can be in higher education in particularly to improve student performance through educational data mining to analyses learning behavior. This study collected students' data from Database Course After pre-processing and applied data mining techniques to discover association, classification, clustering and outlier detection rules, in each of these four tasks, the study extracted knowledge that describes students' behavior, Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest neighbors, Naive Bayes, support vector machines and others. The study also recommended for used pre-process and data mining algorithms could be embedded into eLearning system so that anyone using the system can benefited from the data mining techniques.

One of the studies discussed suggests that legal access to alcohol does affect student performance. The study concluded to prediction of teenager's alcohol addiction by using demographic, family and other data related to student, different classifiers are studied and the experiments are conducted to find the best classifier for predicting the performance of the students who consume alcohol. The study propose an approach to predict the performance using data mining techniques, the study also shows that the most important attributes which most affected the performance of students who consume the alcohol during their study are the previous grades which

is gained by students and other attributes are absence in the class, father's job, mother's job, extra educational support, extra paid classes within the course subject, wants to take higher education, reason to choose this institution and also some other attributes (Pal, 2017).

Another study looks at and compare well performing algorithms such as Naïve Bayes, decision tree (J48), Random Forest, Naïve Bayes Multiple Nominal, K-star and IBk. And it mentions Educational Data mining is a relatively new field and has a lot of potential to help society if used in the proper manner. The study compared six algorithms J48 (Decision Tree), Random Forest, Naive Bayes, Naive Bayes Multinomial, K-star, IBk. In the comparative study of all these algorithms can see that the closest we got in terms of getting an accurate prediction was the Random Forest Technique which narrowly edged the J48 to claim the top spot. This was that was done on a relatively larger dataset hence random forest becomes more accurate with the number of entries but all algorithms need modification if they can ever be used because the current amount of accuracy is low for this to be implemented on a large scale in the present state (Kapur).

Another research considered that one of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this study, classification algorithm of Naïve Bayes and C5.0 are used to build classifier models. Their performances are compared over a dataset composed of answer of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics. Although all the classifier models show comparably high classification performances, Naïve Bayes classifier is the best with respect to accuracy, precision, and specificity. In addition, an analysis of the variable importance for each classifier model is done. This research describes the performances of classification algorithms used in building a model does not necessarily indicate that the one that used the least time is the best model to use. Some Algorithms can take the least time but may not produce the best result in term of accuracy. This research used classification algorithms and data mining techniques such as, Naïve Bayes classifier, C5.0 as well as data from universities. Naïve Bayes classifier is the best with respect to accuracy, precision, and specificity (Patil, 2017).

Another research used data mining techniques for predicting the students' graduation performance in final year at university using only pre-university marks and examination marks of early years at university, no socio-economic or demographic features are use. The result of the study shows that

can predict the graduation performance in a four-years university program using only pre-university marks and marks of first- and second-year courses, no socio-economic or demographic features, with a reasonable accuracy, and that the model established for one cohort generalizes to the following cohort. It makes the implementation of a performance support system in a university simpler because from an administrative point of view, it is easier to gather marks of students than their socio-economic data. The result also shows that decision trees can be used to identify the courses that act as indicator of low performance. By identifying these courses can give warning to students earlier in the degree program (Asif, 2017).

This research presented a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fail in introductory programming courses. Although several works have analyzed these techniques to identify students' academic failures. The study evaluated the effectiveness of four prediction techniques on two different and independent data sources on introductory programming courses available from a Brazilian Public University: one comes from distance education and the other from on campus. The results showed that the techniques analyses in this study are able to early identify students likely to fail, the effectiveness of some of these techniques is improved after applying the data pre-processing and/or algorithms fine-tuning, and the support vector machine technique outperforms the other ones in a statistically significant way (Costa, 2017).

In a one of the studies, the researchers used ID3 algorithm to classify students and predict what causes student failure so that teacher can help them to avoid failure (Baradwaj, 2012).

In another study, the researcher used a wide range of tree resolution algorithms rule-learners, J48, Bayes, Naiva, Bayes and IBK. The researcher concluded that the algorithm of J48 is the most accurate algorithm in terms of the accuracy of the prediction, noting that the accuracy of the prediction of the previous algorithms in general was not satisfactory and it cannot be relied on it, where the accuracy of the prediction was promoted between (67% -52%). And use more data to improve the data collection process (Kabakchieva, 2013).

In further study, the researchers predicted the performance of students based on their academic levels in several areas; using the detection of the rules of correlation to ensure that the factors affecting the final outcome of the student linked to each other, having calculate the correlation

coefficient while the student attributes were shown and the result was different from the resulting relationship using the detection of the rules of correlation (Borkar, 2013).

In this cited study below, the researchers used three techniques for Data Mining: classification, clustering and detection of the rules of association on the data collected by students from an e-learning system. This study is a theoretical and practical guide to how to apply data mining techniques to the educational system (Romero).

In another study, researchers apply a classification ID3, J48 and Apriori algorithm to reveal the correlation rules in the system data; Moodle to compare the accuracy of the algorithms in terms of their ability to predict the outcome of the student, whereas they conclude the algorithm ID3 is more accurate than the rest of the algorithms with a probability of 83.916% (Kularbphettong, 2012).

In one case study authors try to extract useful knowledge from graduate student's data collected from the college of Science and Technology – Khanyounis. This study discovered association rules and sorted the rules using lift metric. Then other used two classification methods which are Rule Induction and Naïve Bayesian classifier to predict the Grade of the graduate student. Finally, outlier detection to detect all outliers in the data, two outlier methods are used which are Distance-based Approach and Density-Based Approach. Each one of these tasks (Abu Tair, 2012).The other study used to apply selected data mining algorithms for classification on the university sample data reveal that the prediction rates are not remarkable (vary between 52-67 %). Several different algorithms are applied for building the classification model, each of them using different classification techniques. The WEKA Explorer application is used, each classifier is applied for two testing options – cross validation (using 10 folds and applying the algorithm 10 times – each time 9 of the folds are used for training and 1 fold is used for testing) and percentage split (2/3 of the dataset used for training and 1/3 – for testing) (Kabakchieva, 2013).

There is need to identify factors that lead to a student's success or failure. This will allow the teacher to provide appropriate counselling and focus more on such factors. Hence, a model for forecasting student's performance academically is of a pronounced significance, therefore, data mining techniques in classifying and forecasting the academic performance of students was put into application in this research study. K-means clustering and Multiple Linear Regression (MLR) were used for assessing student's performance. The results showed that student's test scores, quiz

and assignment were the major factors that could be used in predicting academic performance of students. Also, two clusters were derived with the use of elbow method to group all the students into clusters (Omolewa, 2019).

In this research, a new method for handling the Multi-Input Multi-Output Student Academic Performance Prediction (MIMOSAPP) problem is proposed. The MIMO SAPP aims to predict the future performance of a student after being enrolled into a university. The existing methods have limitations of using a parameter set and an unsuitable training strategy. Thus, the new method called MANFIS-S (Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets) uses multiple parameters sets and a special learning strategy to resolve those weaknesses. Specifically, the idea of multiple parameter sets is to approximate the MANFIS-S model with many meaningful parameters to ensure the performance of system (Fujita, 2019).Another research showed that the data mining techniques become very popular among the data analyst. It became an effective tool for finding the uncovered information from a big database. Due to this feature data mining are adopted by many areas like education, telecommunication, retail management etc. to resolve their business problems. In this paper, for building classification models for 'student performance' dataset consisting of 649 different instances with 33 different attributes implement algorithms like NaiveBayes, Decision Tree (J48), RandomForest, RandomTree, REPTree, JRip, OneR, SimpleLogistic and ZeroR. After implementing these algorithms on student performance dataset, we evaluate and compare the implementation result for better accuracy of prediction. The result of this study is extremely significant and hence provides a greater insight for evaluating the student performance and underlines the significance of data mining in education. It also shows that how students attributes affect the student performance (Salal, 2019).

This study discovered that Prediction of student's performance became an urgent desire in most of educational entities and institutes. That is essential in order to help at-risk students and assure their retention, providing the excellent learning resources and experience, and improving the university's ranking and reputation. However, that might be difficult to be achieved for start-up to mid-sized universities, especially hose which are specialized in graduate and post graduate programs, and have small students 'records for analysis. So, the main aim of this project is to prove the possibility of training and modelling a small dataset size and the feasibility of creating a prediction model with credible accuracy rate. This research explores as well the possibility of

identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using visualization and clustering algorithms. Best indicators were fed into multiple machine learning algorithms to evaluate them for the most accurate model. Among the selected algorithms, the results proved the ability of clustering algorithm in identifying key indicators in small datasets. The main outcomes of this study have proved the efficiency of support vector machine and learning discriminant analysis algorithms in training small dataset size and in producing an acceptable classification's accuracy and reliability test rates (Zohair, 2019).

One of the studies show the review which methods and algorithms of DM can be used in the analysis of educational data to improve decision-making. Furthermore, it evaluates these algorithms using a dataset composed of student data in the computer science school of a private university. The core of the analysis is to discover trends and patterns of study in the graduation rate indicator. Finally, it compares these methods and algorithms and suggests which has the best precision in certain scenarios. Our analyses suggest that random trees had better precision but had limitations due to the difficulty of interpretation while the J48 algorithm had better possibilities of interpretation of results in the visualization of the classification of data and only had slightly inferior performance (Moscoso-Zea, 2019).

Another research explained the application of data mining is widely prevalent in the education system. The ability of data mining to obtain meaningful information from meaningless data makes it very useful to predict students' achievement, university's performance, and many more. According to the Department of Statistics Malaysia, the numbers of student who do not manage to graduate on time rise dramatically every year. This challenging scenario worries many parties, especially university management teams. They have to timely devise strategies in order to enhance the students' academic achievement and discover the main factors contributing to the timely graduation of undergraduate students. This paper discussed the factors utilized by other researchers from previous studies to predict students' graduation time and to study the impact of different types of factors with different prediction methods. Taken together, findings of this research confirmed the usefulness of Neural Network and Support Vector Machine as the most competitive classifiers compared with Naïve Bayes and Decision Tree. Furthermore, our findings also indicate that the academic assessment was a prominent factor when predicting students' graduation time (Nurafifah, 2019).

In the same field there is a study referring to enrollments and class sizes in postsecondary institutions have increased, instructors have sought automated and lightweight means to identify students who are at risk of performing poorly in a course. This identification must be performed early enough in the term to allow instructors to assist those students before they fall irreparably behind. This study describes a modeling methodology that predicts student final exam scores in the third week of the term by using the clicker data that is automatically collected for instructors when they employ the Peer Instruction pedagogy. The modeling technique uses a support vector machine binary classifier, trained on one term of a course, to predict outcomes in the subsequent term. We applied this modeling technique to five different courses across the computer science curriculum, taught by three different instructors at two different institutions. Our modeling approach includes a set of strengths not seen wholesale in prior work, while maintaining competitive levels of accuracy with that work. These strengths include using a lightweight source of student data, affording early detection of struggling students, and predicting outcomes across terms in a natural setting (different final exams, minor changes to course content) across multiple courses in a curriculum and across multiple institutions (Liao, 2019).

The impression from review of the aforementioned works is that it is possible to predict performance of students with reasonable accuracy; the more aggregated the performance, e.g. pass/fail, the higher the accuracy. The studies mentioned differ in the features they select from students' personal information like age, gender, religion, place of living, family, job, total score from previous education etc., to predict students' performance, but recognize that earlier marks are essential for good prediction.

A review of the literature reveals that predicting performance at the tertiary level has involved significant interest in the recent past and continued to focus on research and discussion. A number of studies have investigated student performance at the top level and a review of the literature on the performance prediction mentioned above shows that it is possible to predict student performance with reasonable accuracy.

However, there is no previous literature that has created clear practical and applied solutions that can be generalized in making use of historical data in various fields to predict the future. As most of the previous literature was limited to a comparison between the best methods, tools and methods that can be followed, and the trade-offs between them.

CHAPTER THREE METHODOLOGY

3.1 The Concept of dynamic Predictive Model

The main goal of this research is to use the data mining technique and machine learning algorithms to suggest a dynamic model for predicting the students expected to graduate to a critical level by relying on data available on students after the end of the first school year so that the educational administration can monitor and correct from early.

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised. Predictive modeling is often associated with meteorology and weather forecasting, but it has many applications in business and education and healthcare (Pujari, 2001).

One of the most common uses of predictive modeling is in online advertising and marketing. Modelers use web surfers' historical data, running it through algorithms to determine what kinds of products users might be interested in and what they are likely to click on (Rouse, 2018).

This research uses Machine Learning (ML) techniques in order to build dynamic prediction model that can be used as general to compute the students' performance. Once data scientists gather this sample data, they must select the right model. Linear regressions are among the simplest types of predictive models. Linear models essentially take two variables that are correlated -- one independent and the other dependent -- and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points. Data scientists can use this to predict future occurrences of the dependent variable.

In this research, there are two types of techniques that we can use, data mining techniques to get the dataset used for training and testing the model, and machine learning techniques to build the predictive model.

3.2 Data Mining Techniques

Data mining is the process of looking at large banks of information to generate new information. Intuitively, you might think that data "mining" refers to the extraction of new data, but this isn't the case; instead, data mining is about extrapolating patterns and new knowledge from the data you've already collected. Relying on techniques and technologies from the intersection of database

management, statistics, and machine learning, specialists in data mining have dedicated their careers to better understanding how to process and draw conclusions from vast amounts of information. But what are the techniques they use to make this happen? (SCH{\"A}FFER, 1994).

3.2.1 Classification. Classification is a more complex data mining technique that forces you to collect various attributes together into discernible categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

3.2.2 Association. Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

3.2.3. Clustering. Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

3.2.4. Regression. Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

3.2.5. Prediction. Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future. (NURAFIFAH, REVIEW ON PREDICTING STUDENTS' GRADUATION TIME USING MACHINE LEARNING ALGORITHMS, 2019)

In this research, we can use regression technique to obtain data for Sudan University of Science and Technology students for three batches 2012, 2013 and 2014 to apply this study.

3.3 Machine Learning Techniques

Machine learning is a type of Artificial Intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data.

Machine learning involves a computer to be trained using a given data set, and use this training to predict the properties of a given new data. For example, we can train a computer by feeding it 1000 images of cats and 1000 more images which are not of a cat, and tell each time to the computer whether a picture is cat or not. Then if we show the computer a new image, then from the above training, the computer should be able to tell whether this new image is a cat or not (introduction-machine-learning-using-python/, n.d.).

The process of training and prediction involves the use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. These algorithms are Linear Regression Logistic Regression**.**

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it.

Many machine-learning algorithms have been found that are effective for some types of learning tasks. They are especially useful in poorly understood domains where humans might not have the knowledge needed to develop effective knowledge-engineering algorithms. Generally, Machine Learning (ML) explores algorithms that reason from externally supplied instances (input set) to produce general hypotheses, which will make predictions about future instances. The externally supplied instances are usually referred to as training set. To induce a hypothesis from a given training set, a learning system needs to make assumptions (biases) about the hypothesis to be learned. A learning system without any assumption cannot generate a useful hypothesis since the number of hypotheses that are consistent with the training set is usually huge. Since every inductive

learning algorithm uses some biases, it behaves well income domains where its biases are appropriate, while it performs poorly in other domains (Schaffer 1994).

3.4 Types of Machine Learning Algorithms

There are three types of machine learning (ML) algorithms:

3.4.1 Supervised Learning Algorithms:

Supervised learning uses labeled training data to learn the mapping function that turns input variables (X) into the output variable (Y). In other words, it solves for f in the following equation:

$$Y = f(X)$$

This allows us to accurately generate outputs when given new inputs.

We'll talk about two types of supervised learning: classification and regression.

Classification is used to predict the outcome of a given sample when the output variable is in the form of categories. A classification model might look at the input data and try to predict labels like "sick" or "healthy."

Regression is used to predict the outcome of a given sample when the output variable is in the form of real values. For example, a regression model might process input data to predict the amount of rainfall, the height of a person, etc.

Linear Regression, Logistic Regression, CART, Naïve-Bayes, and K-Nearest Neighbors (KNN) — are examples of supervised learning.

Assembling is another type of supervised learning. It means combining the predictions of multiple machine learning models that are individually weak to produce a more accurate prediction on a new sample.

3.4.2 Unsupervised Learning Algorithms:

Unsupervised learning models are used when we only have the input variables (X) and no corresponding output variables. They use unlabeled training data to model the underlying structure of the data. There are three types of unsupervised learning:

Association is used to discover the probability of the co-occurrence of items in a collection. It is extensively used in market-basket analysis. For example, an association model might be used to discover that if a customer purchases bread, s/he is 80% likely to also purchase eggs.

Clustering is used to group samples such that objects within the same cluster are more similar to each other than to the objects from another cluster.

Dimensionality Reduction is used to reduce the number of variables of a data set while ensuring that important information is still conveyed. Dimensionality Reduction can be done using Feature Extraction methods and Feature Selection methods. Feature Selection selects a subset of the original variables. Feature Extraction performs data transformation from a high-dimensional space to a low-dimensional space. Example: PCA algorithm is a Feature Extraction approach.

3.4.3 Reinforcement learning:

Reinforcement learning is a type of machine learning algorithm that allows an agent to decide the best next action based on its current state by learning behaviors that will maximize a reward.

Reinforcement algorithms usually learn optimal actions through trial and error. Imagine, for example, a video game in which the player needs to move to certain places at certain times to earn points. A reinforcement algorithm playing that game would start by moving randomly but, over time through trial and error, it would learn where and when it needed to move the in-game character to maximize its point total.

We apply different algorithms on the train dataset and evaluate the performance on the test data to make sure the model is stable. The framework includes codes for Random Forest, Logistic Regression, Linear Regression, Naive Bayes, Neural Network and Gradient Boosting. We can add other models based on our needs. In this research we can use the Logistic Regression and Linear Regression as provided below.

3.5 Regression Algorithms

Regression is an important and broadly used statistical and machine learning tool. The key objective of regression-based tasks is to predict output labels or responses which are continues numeric values, for the given input data. The output will be based on what the model has learned in training phase. Basically, regression models use the input data features (independent variables) and their corresponding continuous numeric output values (dependent or outcome variables) to learn specific association between inputs and corresponding outputs.

3.5.1 Linear Regression

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$$Y=mX+b$$

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b. The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above.

In this equation:

- Y – Dependent Variable
- a – Slope

- X – Independent variable
- b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

Look at the below example. Here we have identified the best fit line having linear equation **y = 0.2811x+13.9**. Now using this equation, we can find the weight, knowing the height of a

Person.

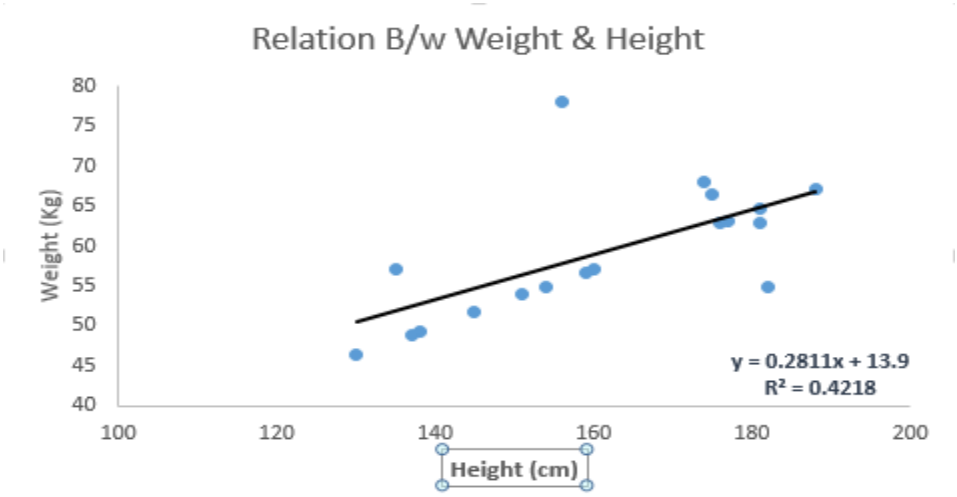


Figure 3. 1 Linear Regression Diagram.

Logic code

```
#Load Train and Test datasets
#Identify feature and response variable(s) and values must be numeric and
numpy arrays
x_train <- input_variables_values_training_datasets
y_train <- target_variables_values_training_datasets
x_test <- input_variables_values_test_datasets
x <- cbind(x_train, y_train)
# Train the model using the training sets and check score
```

```
linear <- lm(y_train ~ ., data = x)
summary(linear)
#Predict Output
predicted= predict (linear, x_test)
```

3.5.2 Logistic Regression

It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variables(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logic function. Hence, it is also known as logic regression. Since, it predicts the probability, its output values lie between 0 and 1 (as expected).

Let's say your friend gives you a puzzle to solve. There are only 2 outcome scenarios – either you solve it or you don't. Now imagine, that you are being given wide range of puzzles / quizzes in an attempt to understand which subjects you are good at. The outcome to this study would be something like this – if you are given a trigonometry based tenth grade problem, you are 70% likely to solve it. On the other hand, if it is grade fifth history question, the probability of getting an answer is only 30%. This is what Logistic Regression provides you. Coming to the math, the log odds of the outcome is modelled as a linear combination of the predictor variables. (Jain, 2019)

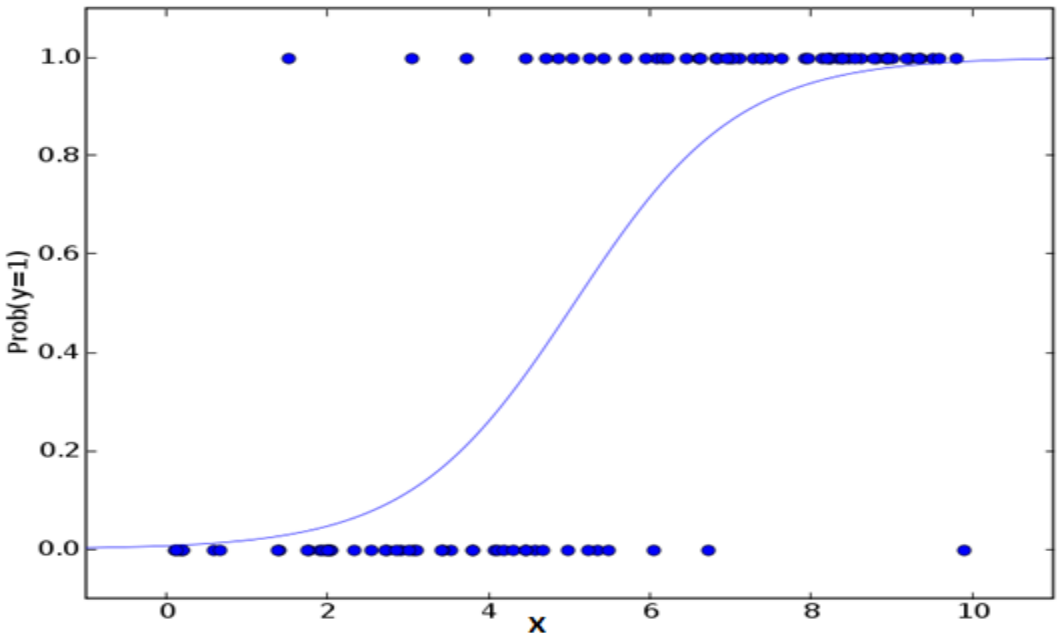


Figure 3. 2 Logistic Regression Diagram.

Logic code

```
x <- cbind(x_train,y_train)
# Train the model using the training sets and check score
logistic <- glm(y_train ~ ., data = x,family='binomial')
summary(logistic)
#Predict Output
predicted= predict (logistic, x_test)
```

Through the techniques that have been explained and according to the type of variables and the excepted outputs, we find that the appropriate technique to be used in research is the linear regression algorithm for developing the predictive model and we used the python tool for applying the model.

3.6 Predictive Model Development Process

Designing Predictive model methodology has many processes that can be used in the designing process but there are however some similarities between them. According to ways developed a model that consists three phases preparatory stage, building the model stage and apply and evaluation stage, all these stages can show in Diagram 3.1 below:

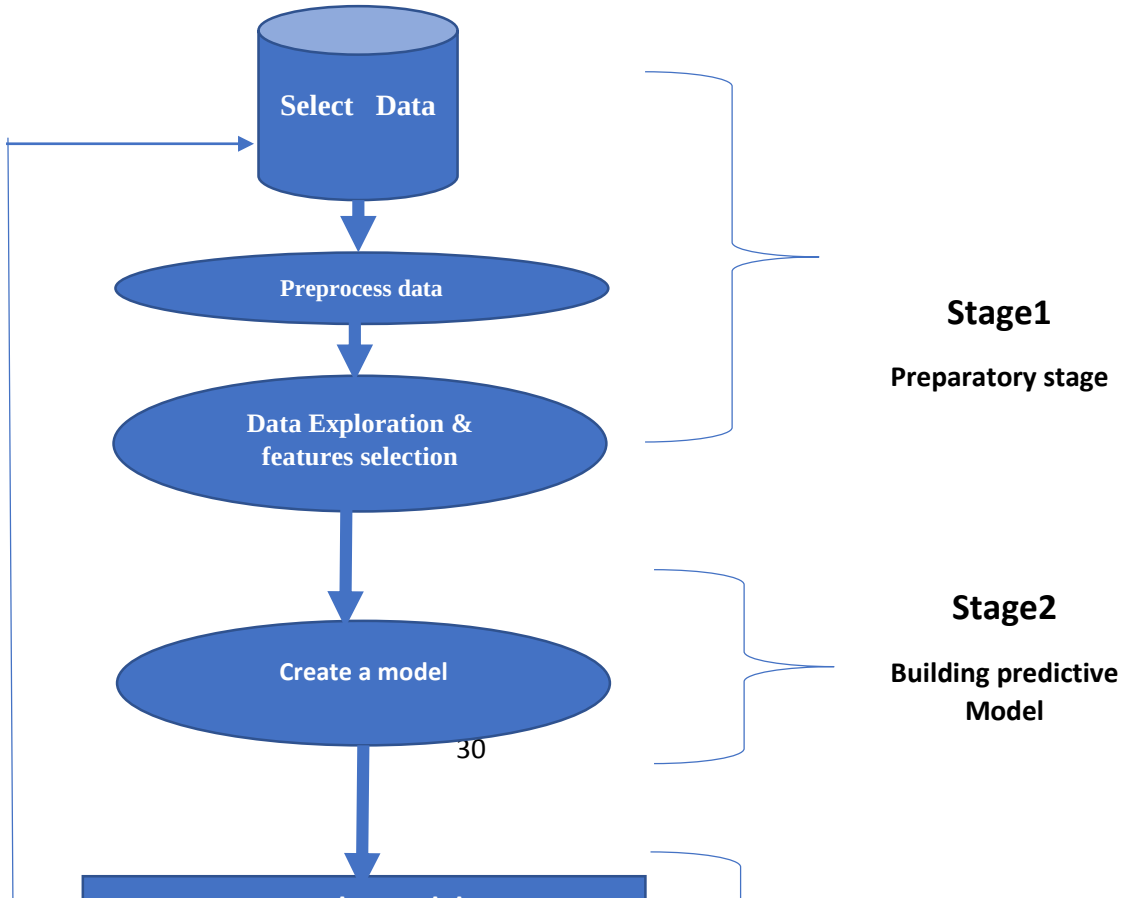


Figure 3. 4 methodology of building the model.

Every phase of the process has its own purpose and gives input for the next phase. Each phase is briefly described below:

To verify the objectives of this study and to get answer of the main questions of this research, there are several stages that had followed to build the proposed dynamic predictive model and obtain the expected results, and then Then the questions are answered. All the questions are built to achieve the main goal of this research, which is building a dynamic prediction model to know the student's graduation level.

To approve or deny these questions, we follow the following these stages in building the predictive model:

3.6.1 Step 1: Data collection

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

It is the stage of selecting the candidate data for the study, from the database, according to the purpose of the study. We use historical data to train our model. The data is usually scattered across multiple sources and may require cleansing and preparation. Data may contain duplicate records and outliers; depending on the analysis and the business objective, we decide whether to keep or remove them. Also, the data could have missing values, may need to undergo some transformation, and may be used to generate derived attributes that have more predictive power for our objective. Overall, the quality of the data indicates the quality of the model.

We have collected data from the Directorate General of Admission - Ministry of Higher Education and Scientific Research and Faculty of Computer Science, Sudan University of Science and Technology for three batches: 2012, 2013 and 2014. The total sample 347 students.

We obtained the targeted data from the student admission database, admission department at the Ministry of Higher Education, and the student results database at the Faculty of Computer Science, Sudan University of Science and Technology.

Data mining techniques and some software such as Oracle, MS -access and MS- excel to build our datasets as shown in the table below.

We created the database for training and building the predictive model, where we determined the size of the training database for all three students in 2012, 2012 and 2014, and the 347 students.

For each student, the values of the approved inputs and the graduation rate were extracted, as the record took the following form:

Table No: 3. 1 sample of database

NO	FRMNO	Home	SMATH	CHEMISTRY	PHYSICS	H-S-Result	F-Y-Result	F_GRADE
1	**157020**	**0**	**87**	**75**	**77**	**83**	**2.75**	**2.71**
2	**151189**	**0**	**90**	**77**	**78**	**83.1**	**2.8**	**2.48**
3	**25761**	**1**	**63**	**66**	**67**	**71.9**	**0**	**2.22**
4	**187203**	**1**	**82**	**78**	**84**	**84.9**	**2.6**	**2.51**
5	**188795**	**1**	**79**	**78**	**80**	**82.7**	**2.55**	**2.42**
6	**2073**	**1**	**70**	**70**	**66**	**71.7**	**2.48**	**2.42**

7	**163678**	**1**	**76**	**85**	**78**	**83.1**	**2.9**	**2.67**
8	**151911**	**1**	**76**	**79**	**79**	**83**	**2.82**	**2.81**
9	**58980**	**1**	**76**	**71**	**76**	**83.6**	**2.66**	**2.71**
10	**170584**	**1**	**77**	**84**	**78**	**84**	**2.61**	**2.57**
11	**181596**	**1**	**84**	**78**	**84**	**83**	**2.55**	**2.51**
12	**19270**	**1**	**61**	**61**	**68**	**70.9**	**2.51**	**2.4**
13	**147585**	**1**	**77**	**83**	**85**	**84**	**3.02**	**3.08**
14	**97038**	**0**	**88**	**80**	**82**	**82.9**	**2.41**	**2.41**
15	**116521**	**1**	**75**	**84**	**83**	**83.4**	**2.3**	**2.47**
16	**23411**	**1**	**63**	**68**	**75**	**76.9**	**2.8**	**2.67**
17	**23588**	**1**	**70**	**67**	**71**	**73.4**	**2.26**	**2.22**
18	**182555**	**1**	**86**	**77**	**82**	**82.7**	**2.93**	**2.66**
19	**96884**	**1**	**90**	**79**	**81**	**84.4**	**2.8**	**2.73**
20	**187298**	**1**	**79**	**85**	**84**	**84.7**	**2.58**	**2.69**
21	**159287**	**1**	**79**	**78**	**89**	**83.6**	**2.73**	**2.56**
22	**3778**	**1**	**67**	**67**	**73**	**74.1**	**2.36**	**2.35**
23	**17023**	**1**	**80**	**81**	**78**	**82.1**	**2.69**	**2.71**
24	**105787**	**0**	**74**	**82**	**78**	**84.1**	**2.85**	**2.64**

3.6.2 Step 2: Preprocessing the Data

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately (Smolinska, 2014). In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis. It is the stage of removing data that contains interference or noise from the dataset to get that clean database (Gon{\c{c}}alves, 2019).

Before we start designing the experiment, it is important to preprocess the dataset. In most cases, the raw data needs to be preprocessed before it can be used as input to train a predictive analytic model. From the earlier exploration, we have noticed that there are missing values in the data, as a precursor to analyzing the data, these missing values we cleaned.

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

Most of the real-world data are messy, some of these types of data are:

1. Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application.

2. Noisy data: This type of data is also called outliners; this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.

3. Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

These are some of the basic preprocessing techniques that can be used to convert raw data.

1. Conversion of data: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2. Ignoring the missing values: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.

3. Filling the missing values: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

For this experiment, we will substitute the missing values with a designated value. In addition, the normalized-losses column will be removed as this column contains too many missing values.

3.6.3 Step 3: Data Exploration & features Selection

In this stage we used the features selection techniques are used to help in visualizing the relations between variables and in identifying the main indicators that could help in predicting dissertation and courses' grades.to determine the best attributes of candidate input variables that have big effect the student's graduation rate. Variables according to the highest degree of correlation and influence on the expected end result of the student were chosen, we repeat this step until it gets improve the accuracy of the model, we apply different feature selection techniques available.

Need for Feature Selection

- Helps train the model faster: We have reduced numbers of relevant features, so training is much faster.
- Increase model interpret-ability and simplifies the model — It reduces the complexity of the model by including only the most relevant features and hence easy to interpret. This is very helpful in explaining the predictive model
- Improves accuracy of the model: We include only features that are relevant for our prediction and that increases the accuracy of the model. Irrelevant features introduce noise and reduce the accuracy of the model
- Reduces Over-fitting: Over-fitting is when the predictive model does not generalize well on test data or unseen data based on the training. To reduce over fitting, we need to remove the noise in the data set and include the features that most influence the prediction. Noise comes from irrelevant features in the data set. When a predictive model has learned the noise as part of training then it will not generalize well on unseen data.

Different methods for Feature Selection

- Filter
- Wrapper
- Embedded methods

Filter method for feature selection

The filter method ranks each feature based on some uni-variate metric and then selects the highest-ranking features. Some of the uni-variate metrics are

- variance: removing constant and quasi constant features
- Chi-square: used for classification. It is a statistical test of independence to determine the dependency of two variables.
- correlation coefficients: removes duplicate features
- Information gain or mutual information: assess the dependency of the independent variable in predicting the target variable. In other words, it determines the ability of the independent features to predict the target variable.

Advantages of Filter methods

- Filter methods are model agnostic
- Rely entirely on features in the data set
- Computationally very fast
- Based on different statistical methods

The disadvantage of Filter methods

- The filter method looks at individual features for identifying its relative importance. A feature may not be useful on its own but maybe an important influencer when combined with other features. Filter methods may miss such features.

Filter criteria for selecting the best feature

Select independent features with

- High correlation with the target variable
- Low correlation with other independent variables
- Higher information gain or mutual information of the independent variable

Wrapper method for feature selection

The wrapper method searches for the best subset of input features to predict the target variable. It selects the features that provide the best accuracy of the model. Wrapper methods use inferences based on the previous model to decide if a new feature needs to be added or removed.

Wrapper methods are

- Exhaustive search: evaluates all possible combinations of input features to find the input feature subset that would give the best accuracy for a selected model. Computationally very expensive when the number of input features gets larger;

- Forward selection: start with a null feature set and keeping adding one input feature at a time and evaluate the accuracy of the model. This process is continued till we reach a certain accuracy with a predefined number of features;
- Backward selection: start with all the features and then keep removing one feature at a time to evaluate the accuracy of the model. Feature set that yields the best accuracy is retained.

Advantages

- Models feature dependencies between each of the input features
- Dependent on the model selected
- selects the model with the highest accuracy based on feature subset

Disadvantages:

- Computationally very expensive as training happens on each of the input feature set combination
- Not model agnostic

Embedded method for feature selection

Embedded methods use the qualities of both filter and wrapper feature selection methods. Feature selection is embedded in the machine learning algorithm.

Filter methods do not incorporate learning and are only about feature selection. Wrapper methods use a machine-learning algorithm to evaluate the subsets of features without incorporating knowledge about the specific structure of the classification or regression function and can, therefore, be combined with any learning machine

Embedded feature selection algorithms include

- Decision Tree
- Regularization — L1(Lasso)and L2(Ridge) Regularization

3.6.4 Step 4: Create the Model

For building dynamic predicting model that can use to expect the students' graduation performance at the end of the degree and answer Research Question 1, there are several machine learning algorithms can used. Also, we need to split our data into two sets: training and test datasets. We build the model using the training dataset. We use the test data set to verify the accuracy of the model's output. Doing so is absolutely crucial. Otherwise we run the risk of over fitting our model - training the model with a limited dataset, to the point that it picks all the characteristics that are only true for that particular dataset. A model that's over fitted for a specific data set will perform miserably when you run it on other datasets. A test dataset ensures a valid way to accurately measure your model's performance. Here we can use the linear regression algorithm because Regression techniques are useful for identifying the nature of the relationship between variables in a dataset. Those relationships could be causal in some instances, or just simply correlate in others. Regression is a straightforward white box technique that clearly reveals how variables are related. Regression techniques are used in aspects of forecasting and data modeling.

When constructing a predictive model, we first needed to train the model, and then validate that the model is effective. In this stage, we will train a regression model and use it to predict the e expecting the student's level upon graduation. Specifically, we will train a simple linear regression model. After the model has been trained, you will use some of the modules available in Machine Learning to validate the model.in this step, we will use 80% of the data to train the model and hold back 20% for testing. We will have chosen the python environment to be the application environment, because it provides a comprehensive environment that enables us to apply the selected techniques in building the model. It also provides graphs resulting from the application of these technologies.

3.6.5 5 Step 5: Testing the model

At this stage we will test the model that was built by applying it to predict the level of students who actually graduated and then compare the results with the actual results.

We first prepared the inputs and then implemented the system that was built to predict the rates of students who actually graduated.

3.6.6 Step 6: Evaluation the model

Model Evaluation is an integral part of the model development process. It helps to find the best attributes that represents our data and how well the chosen model will work in the future.

In order to improve the model, we tuned the hyper-parameters of the model and try to improve the accuracy to increase the number of true positives and true negatives.

For calculated the accuracy of the model we use R-squared method, R-squared is a statistical measure of how close the data are to the fitted regression line. ... 0.0 indicates that the model explains none of the variability of the response data around its mean. 1.0 indicates that the model explains all the variability of the response data around its mean. R2. Score is made for continuous variables, such as for regression problems, the calculate the R2 value as:

R2 = 1- sum (residual squared)/ (N*variance of data).

CHAPTER FOUR: ANALYSIS AND FINDING

4.1 Overview

This chapter presents the analysis and findings of the data that have been collected to suggestion a dynamic predictive model for students who are expected to graduate at a critical level, based on data available to students before entering the university and after the end of the first academic year. In this chapter, we conduct an analytical study of the eighth candidate variables, to become inputs for the proposed model, using appropriate statistical methods like statistical significance and correlation coefficient to choose the appropriate inputs for the proposed predictive model.

Also, to confirm/reject the hypotheses of the study and choose the most influential variables on the student's result at the graduation, we conducted several analytical statistical processes as shown in the analysis of the following:

4.2 Correlation coefficient

A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. For example, height and weight are correlated—as height increases, weight also tends to increase. Consequently, if we observe an individual who is unusually tall, we can predict that his weight is also above the average (Zou, 2003).

In statistics, a correlation coefficient is a quantitative assessment that measures both the direction and the strength of this tendency to vary together. There are different types of correlation that we can use for different kinds of data. In this research, we use the most common type of correlation—Pearson's correlation coefficient.

According to the assumptions below, we used Pearson correlation coefficient analysis to determine the extent of correlation between the selected variables and the student's graduation rate in the final year and then choose the inputs of the proposed predictive model.

The table below shows the results obtained after Pearson correlation coefficient analysis:

Table 4.1 (Pearson correlation coefficient analysis between Candidate variables and final rate)

Correlations		
Variables		**Results**
Home	Pearson Correlation	**0.012**
	Sig. (2-tailed)	**0.832**

	N	**326**
Math	Pearson Correlation	**0.465**
	Sig. (2-tailed)	**0.000**
	N	**326**
PHYSICS	Pearson Correlation	**.318**
	Sig. (2-tailed)	**0.000**
	N	**326**
CHEMISTRY	Pearson Correlation	**0.318**
	Sig. (2-tailed)	**0.000**
	N	**326**
H_school	Pearson Correlation	**0.453**
	Sig. (2-tailed)	**0.000**
	N	**326**
Programming	Pearson Correlation	**0.561**
	Sig. (2-tailed)	**0.000**
	N	**326**
F_year	Pearson Correlation	**0.625**
	Sig. (2-tailed)	**0.000**
	N	**326**
Enter computer	Pearson Correlation	**0.403**
	Sig. (2-tailed)	**0.000**
	N	**326**

4.3 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

4.3.1 The student's grades in Mathematics at the secondary level is a significant predictor for the level of the student at the graduation.

In order to answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student’s grade in mathematics and the student’s graduation level.

The tables below show test results and statistical significance and Correlation coefficient.

Table No:4. 2 (test results and the statistical significance between Math and the student's final result)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.223	36	.090	1.639	.015
Within Groups	15.784	289	.055		
Total	19.007	325			

Table No:4. 3 (Correlation between Math and final rate)

Correlations			
		Final	MATH
Pearson Correlation	Final	1.000	0.465
	MATH	0.465	1.000
Sig. (1-tailed)	Final	.	.000
	MATH	.000	.
N	Final	326	326
	MATH	326	326

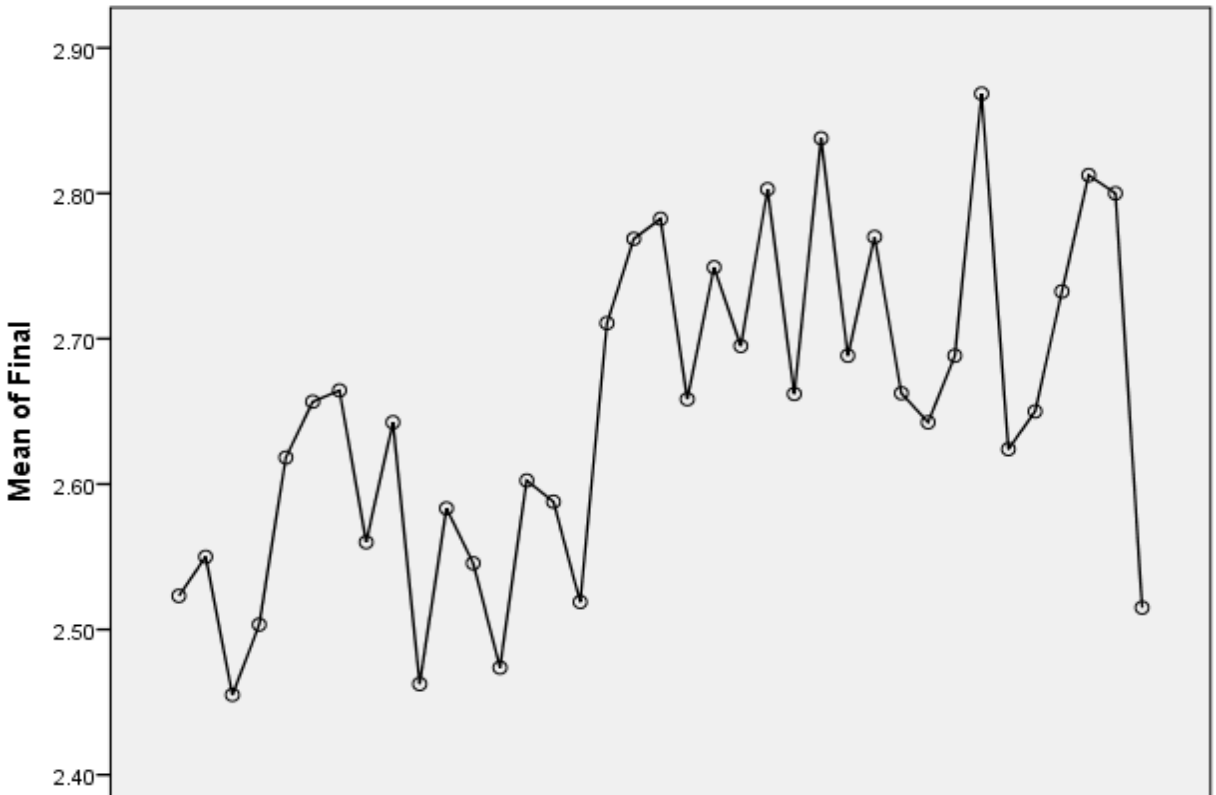


Figure: 4.1 (Diagram of Math and Final Rate)

The result and above figure, showed that there is significant effect on student's degree in Mathematics at the secondary level on a result of the student's graduation.

4.3.2 The student's degree in Chemistry at the secondary level has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's grade in Chemistry and the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Table No:4. 4 (test results and the statistical significance between Chemistry and the students' final results)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.742	32	.117	2.244	.000
Within Groups	15.266	293	.052		
Total	19.007	325			

Table No:4. 5 (Correlation between Chemistry and final grades)

Correlations			
		Final	CHEMISTRY
Pearson Correlation	Final	1.000	.318
	CHEMISTRY	.318	1.000
Sig. (1-tailed)	Final	.	.000
	CHEMISTRY	.000	.
N	Final	326	326
	CHEMISTRY	326	326

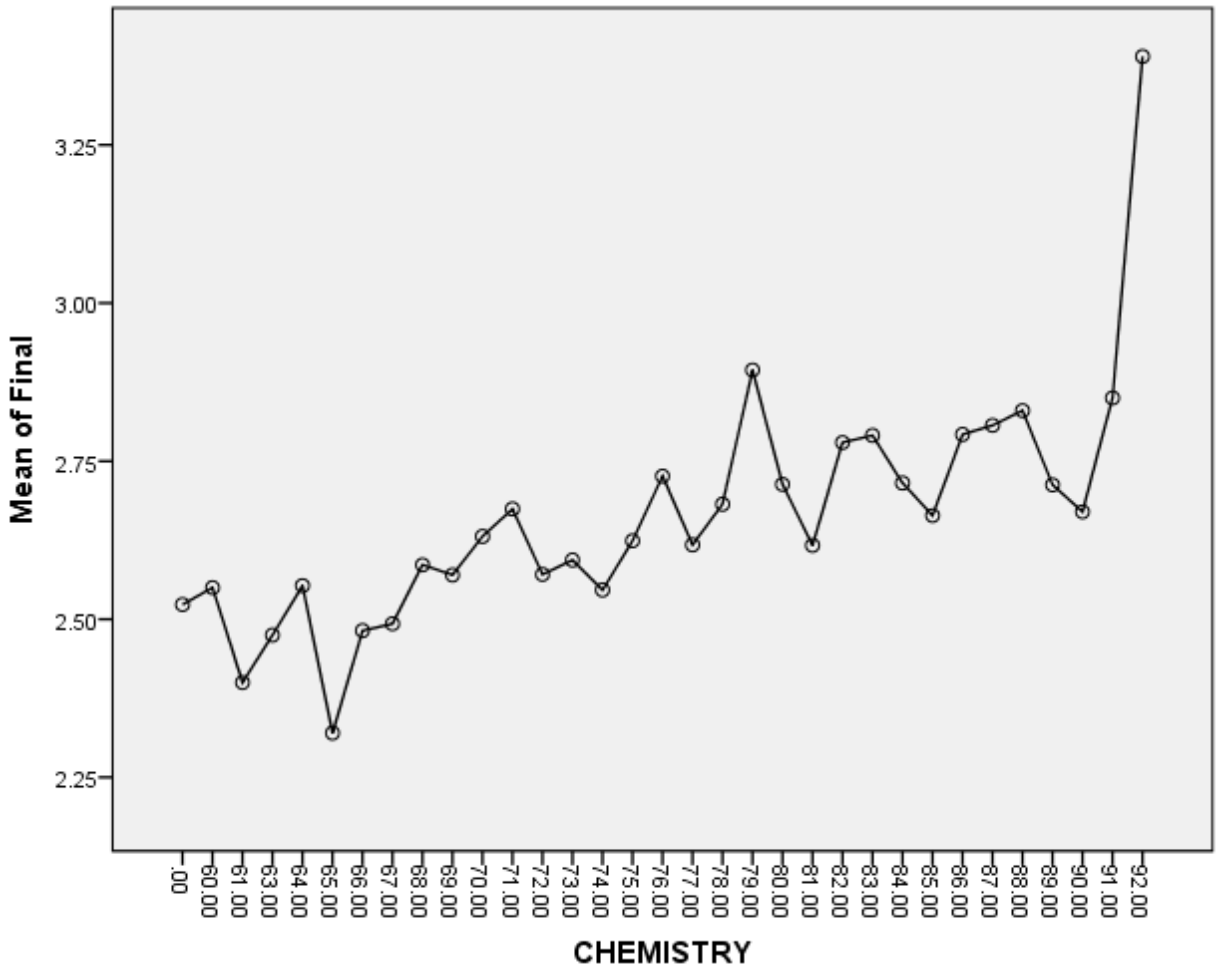


Figure No 4.2 :(Diagram of Chemistry and Final Rate)

The result and above figure, showed that there is no significant effect on the student's degree in Chemistry on a result of the student's graduation.

4.3.3 The students' grades in Physics at the secondary level has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way - variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's grade in Physics and the student's graduation level.

The tables below show test results and a statistical significance and Correlation coefficient.

Table No:4.6 (test results and the statistical significance between Physics and the student's final result)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.447	26	.133	2.548	.000
Within Groups	15.560	299	.052		
Total	19.007	325			

Table No:4.7 (Correlation between Physics and final rate)

Correlations			
		Final	PHYSICS
Pearson Correlation	Final	1.000	. 318
	PHYSICS	. 318	1.000
Sig. (1-tailed)	Final	.	.000
	PHYSICS	.000	.
N	Final	326	326
	PHYSICS	326	326

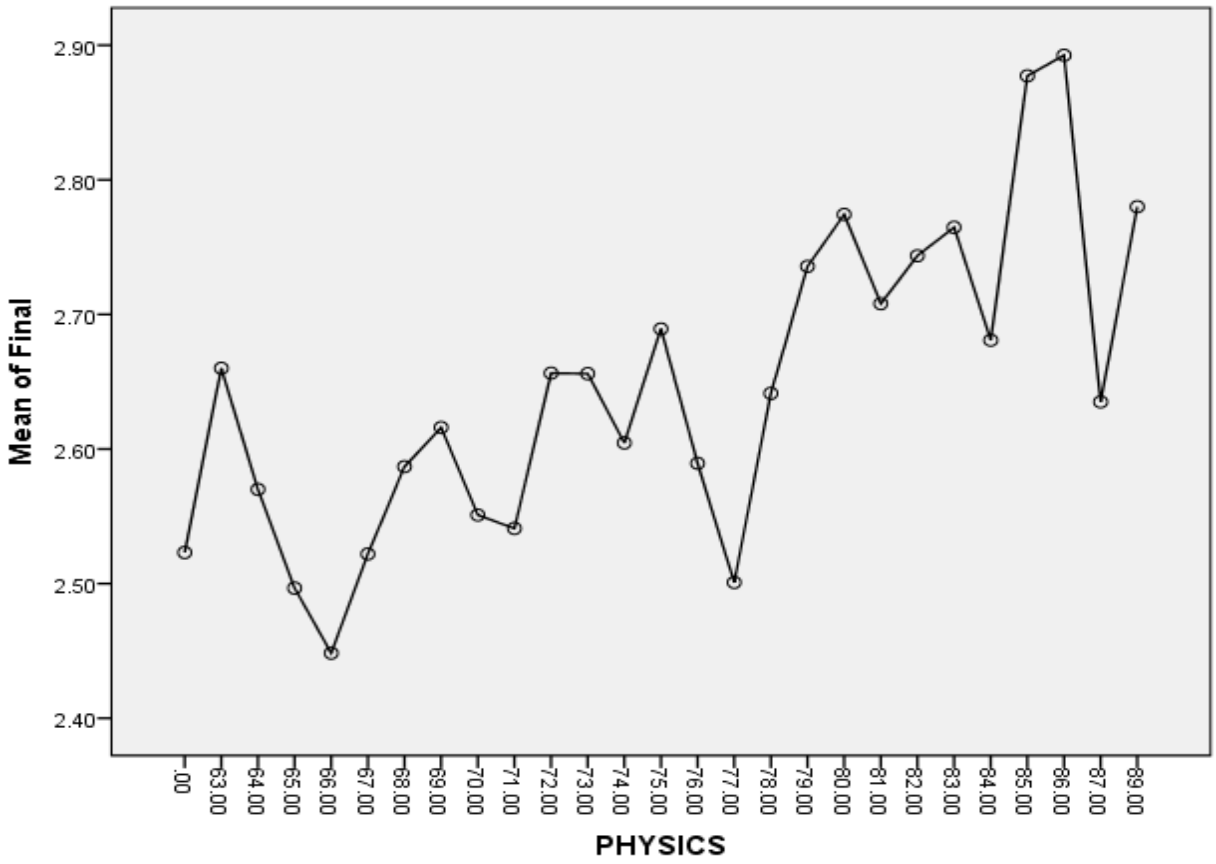


Figure No: 4.3 Diagram of Physics and Final Rate)

The result and above figure, showed that there is no significant effect on the student's degree in Physics on a result of the student's graduation.

4.3.4 The student's degree in High school graduation level has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's grade in High school graduation and the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Table No:4. 8 (test results and the statistical significance between High school graduation degree and the student's final result)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.437	98	.086	1.848	.000
Within Groups	10.437	224	.047		
Total	18.874	322			

Table No:4. 9 (Correlation between High school graduation degree and final rate)

Correlations			
		Final	H.school
Pearson Correlation	Final	1.000	.463
	H.school	.463	1.000
Sig. (1-tailed)	Final	.	.000
	H.school	.000	.
N	Final	323	323
	H.school	323	323

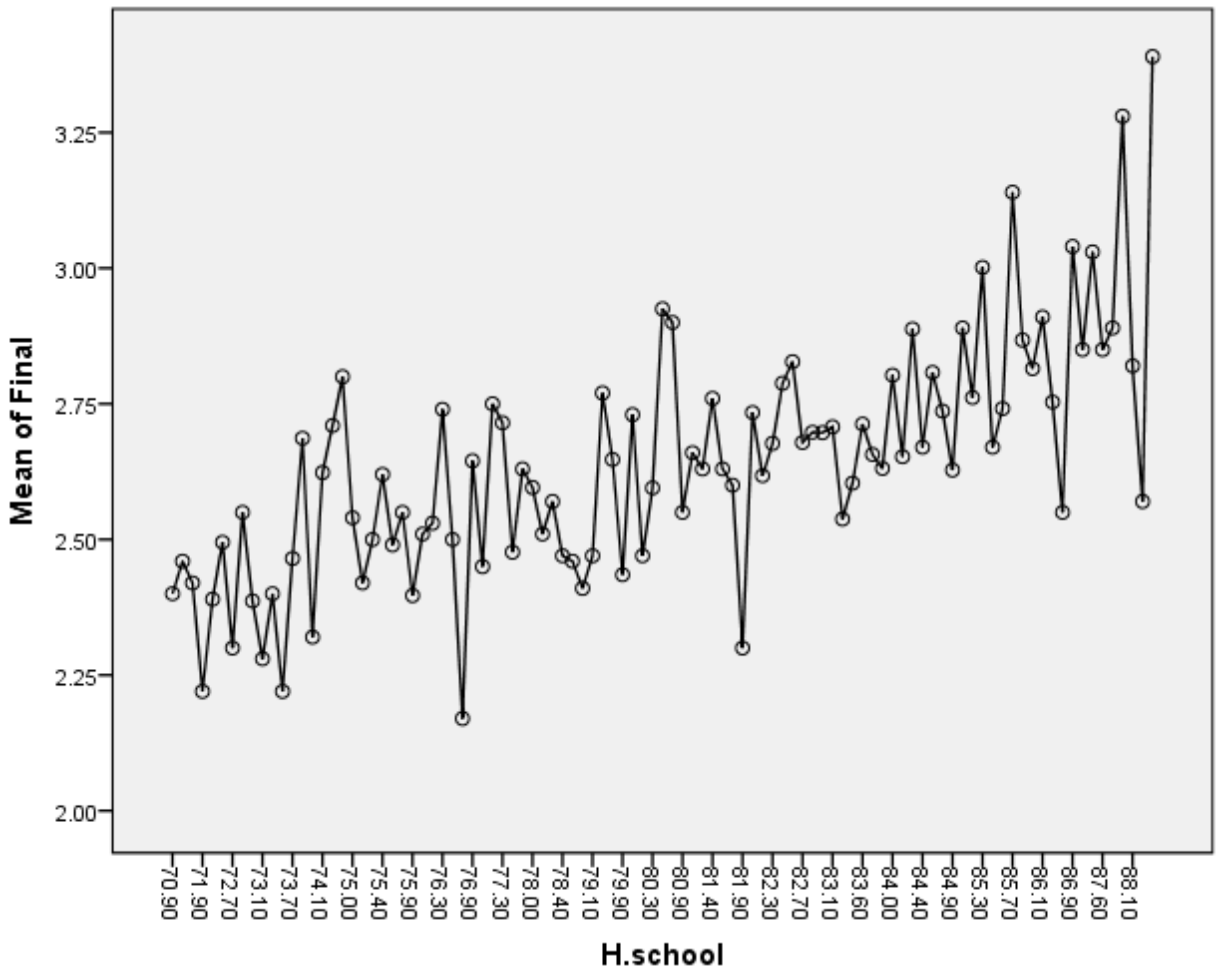


Figure No4.4: (Diagram of High school graduation degree and Final Rate)

The result and above figure, showed that there is significant effect on the student's degree in High school graduation degree on a result of the student's graduation.

4.3.5 The student's degree in Frist year has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's grade in first year **and** the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Table No: 4.10 (test results and the statistical significance between First year degree and the student's final result)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13.618	117	.116	4.492	.000
Within Groups	5.390	208	.026		
Total	19.007	325			

Table No: 4.11 (Correlation between Frist year degree and final rate)

Correlations			
		Final	F.grade
Pearson Correlation	Final	1.000	.625
	F.year	.625	1.000
Sig. (1-tailed)	Final	.	.000
	F.year	.000	.
N	Final	326	326
	F.year	326	326

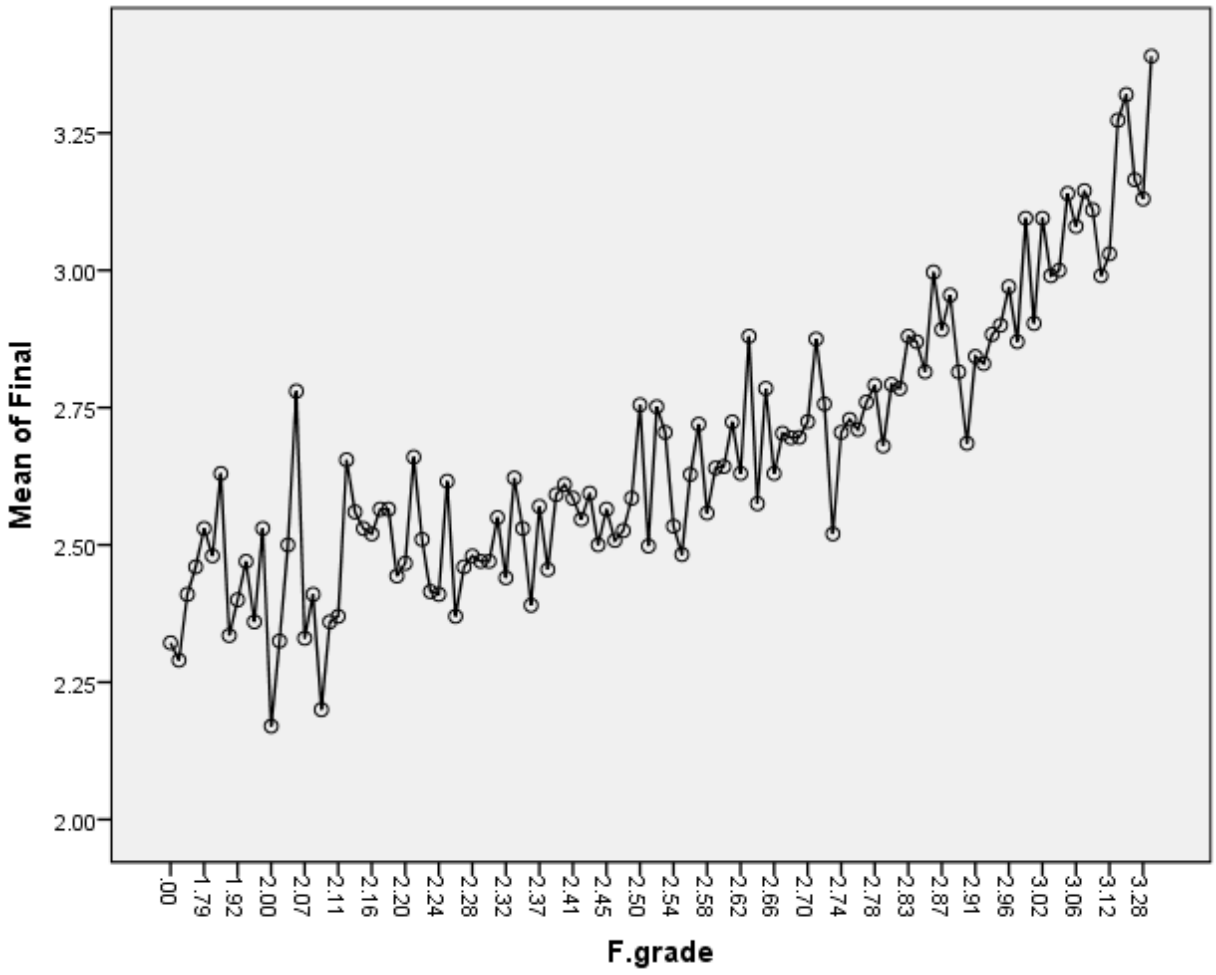


Figure No4.5 :(Diagram of Frist year degree and Final Rate)

The result and above figure, showed that there is significant effect on the student's degree in Frist year degree on a result of the student's graduation.

4.3.6 The student's degree in Programming method is a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's grade in programming method the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Table No: 4.12 (test results and the statistical significance between programming method and the student's final result)

ANOVA[a]					
Final					
	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	.657	1	.657	12.492	.001[b]
Within Groups	4.158	79	.053		
Total	4.815	80			

Table No: 4.13 (Correlation between programming method degree and final rate)

Correlations			
		Final	programming
Pearson Correlation	Final	1.000	.561
	programming	.561	1.000
Sig. (1-tailed)	Final	.	.000
	programming	.000	.
N	Final	78	78
	programming	78	78

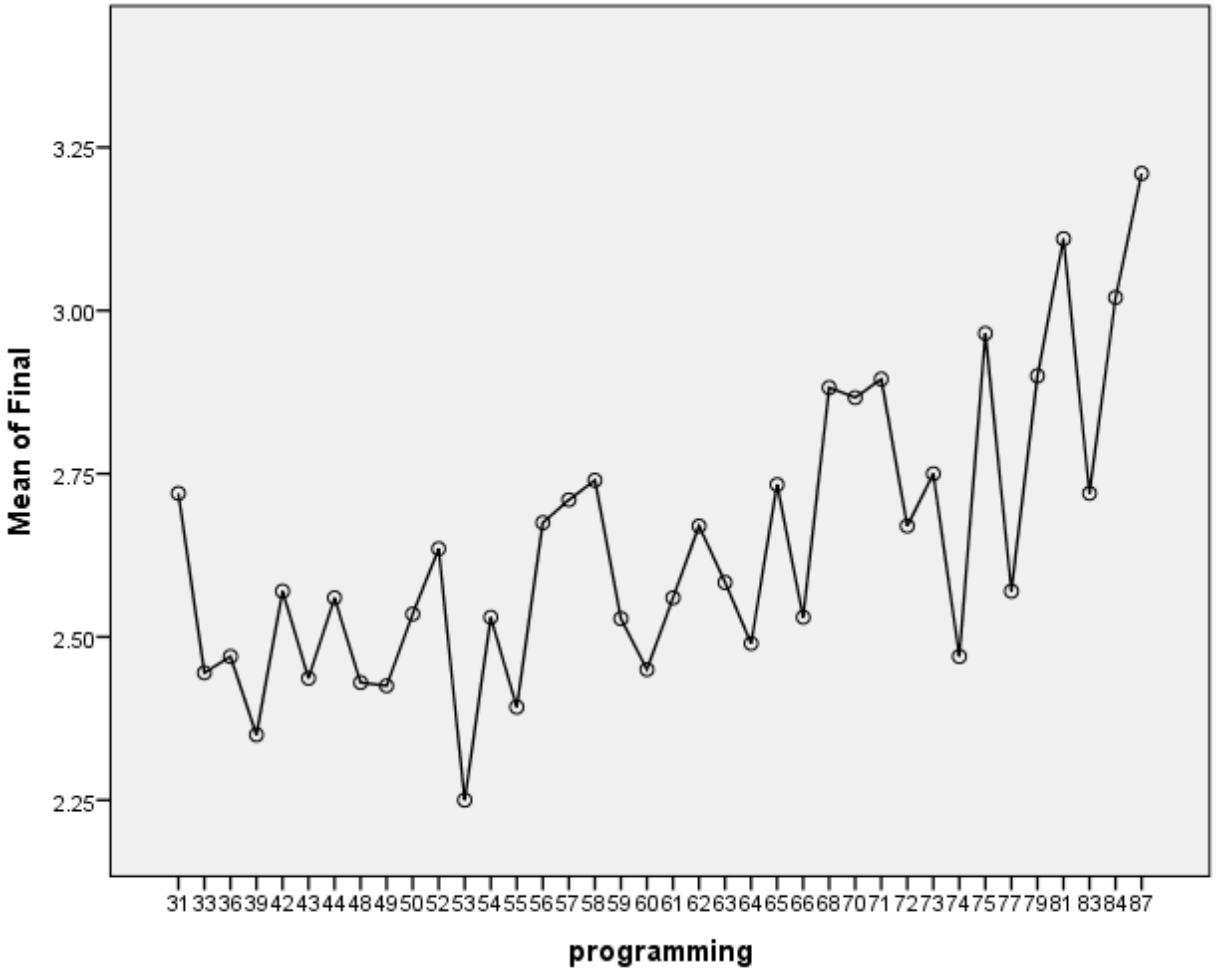


Figure No4.6 :(Diagram of programming degree and Final Rate)

The result and above figure, showed that there is significant effect on the student's degree in programming degree on a result of the student's graduation.

4.3.7 The student's place of home has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's place of home and the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Table No: 4. 14 (test results and the statistical significance between student's place of home and the student's final result)

ANOVA					
grade					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.000	1	.000	.000	.998
Within Groups	20.242	345	.059		
Total	20.242	346			

Table No: 4.15 (Correlation between student's place of home and final rate)

Correlations			
		Final	home
Pearson Correlation	Final	1.000	.000
	home	.000	1.000
Sig. (1-tailed)	Final	.	.499
	home	.499	.
N	Final	347	347
	home	347	347

4.3.8 The student's E_computer degree has a significant effect on the level of the student at graduation.

To answer this hypothesis, one way -variance test (ANOVA) and Correlation coefficient were conducted to discover whether there were statistically significant differences between the student's E_computer degree the student's graduation level.

The tables below show test results and statistical significance and Correlation coefficient

Figure No: 4.23 (test results and the statistical significance between student's E_computer degree and the student's final result)

ANOVA					
Final					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.862	37	.077	1.940	.021
Within Groups	1.595	40	.040		
Total	4.458	77			

Table No: 4.16 (Correlation between student's E_computer degree and final rate)

Correlations			
		Final	E_computer
Pearson Correlation	Final	1.000	.403
	E_computer	.403	1.000
Sig. (1-tailed)	Final	.	.000
	E_computer	.000	.
N	Final	73	73
	E_computer	73	73

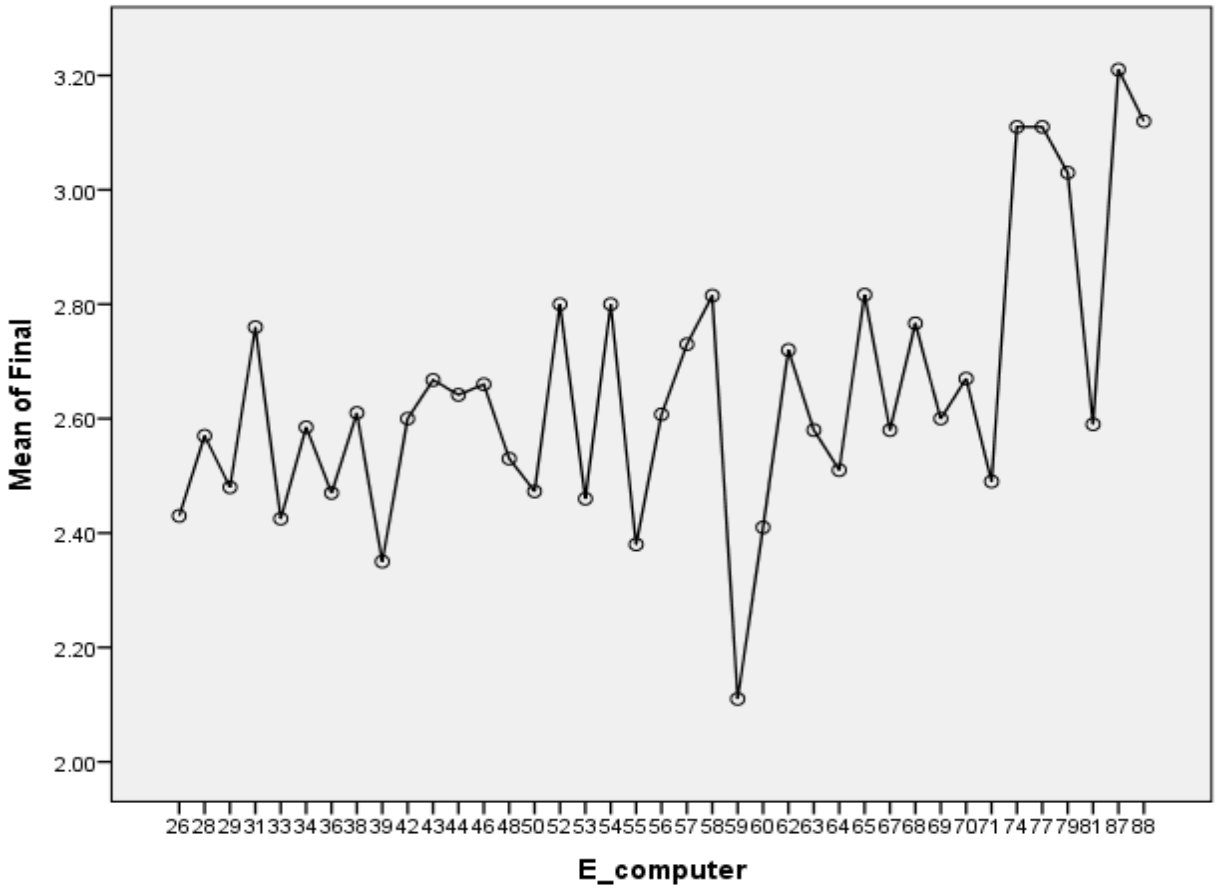


Figure No4.8 :(Diagram of student's E_computer degree and Final Rate)

The result and above figure, showed that there is no significant effect on the student's degree in E-computer degree on a result of the student's graduation.

By applying the Pearson correlation coefficient and one way -variance test (ANOVA) we can get the candidate variables for the inputs of the model, we found that there are a correlation and statistical significance to all the eight selected income elements. Especially in the student's mark in high school, First year and math. That mean can take the student's mark in high school, First Year and the student's mark in mathematics, programming methods as Inputs for the prediction model.

The table below shows the results obtained from conducting the ANOVA testing and Pearson correlation coefficient:

Table No: 4.17 (The results of the analysis candidate input)

NO	Candidate variables	Coefficient of correlation	ANOVA Testing	Evaluation
1	The student's First Year Grade	0.625	.000	Good correlation and a strong impact
2	The student's mark in programming	0.561	.001[b]	Good correlation and a strong impact
3	The student's mark in high school	0.463	.000	Good correlation and a strong impact
4	The student's mark in mathematics	0.465	.015	Good correlation and a strong impact
5	The student's mark in chemistry	0.318	.000	Weak correlation and a strong impact
6	The student's mark in Physic	0. 318	.000	Weak correlation and a strong impact
7	The student's mark in Enter__computer	0.403	.021	Weak correlation and a Weak impact
8	The student's home place	0.000	.998	Weak correlation and a Weak impact

Through the hypotheses of this research, which focused on the effect of student achievement in some academic subjects on his final result upon graduation, and after using various tools of statistical analysis tools, we came to the most influential academic subjects to become the inputs of the predictive model

No	Model inputs based on hypotheses	Expected output
1	First year grade	Student graduation rate
	Second year grade	
2	The student's mark in mathematics	
3	The student's mark in high school	
4	The student's mark in programming	

Table No: 4.18 (The candidate input of the model)

To obtain more accurate results and access to independent variables that have a greater impact on the level of student graduation from the university, we use more statistical analysis tools like MANOVA.

4.4 Multivariate analysis of variance (MANOVA)

Multivariate analysis of variance (MANOVA) is an extension of the uni-variate analysis of variance (ANOVA). In an ANOVA, we examine for statistical differences on one continuous dependent variable by an independent grouping variable. The MANOVA extends this analysis by taking into account multiple continuous dependent variables, and bundles them together into a weighted linear combination or composite variable. The MANOVA will compare whether or not the newly created combination differs by the different groups, or levels, of the independent variable. In this way, the MANOVA essentially tests whether or not the independent grouping variable simultaneously explains a statistically significant amount of variance in the dependent variable.

4.5 Explanation of MANOVA analysis results

Multiple analysis of variance (MANOVA) is used to see the main and interaction effects of categorical variables on multiple dependent interval variables. Dependent variables typically are treated as a s et because they are correlated (if they were not correlated, uni-variate GLM would be appropriate).

MANOVA uses one or more categorical independent Variable s as predictors, MANOVA tests the differences in the centroid (vector) of means of the multiple interval dependents, for various categories of the independent variables.

The researcher may also perform planned comparisons or post -hoc comparisons to see which values of a factor contribute most to the explanation of the dependent variables.

4.6 There are multiple potential purposes for MANOVA

. • Compare group differences. To compare groups formed by categorical independent variables on group differences in a set of interval dependent variables.

• Improve model parsimony. To use lack of difference for a set of dependent variables as a criterion for reducing a set of dependent variables to a smaller, more easily modeled number of variables.

• Rank predictor variables by discriminant effect. To identify the independent variables which differentiate values in a set of dependent variables the most.

Multivariate tests answer the question, "Which predictor effects are significant?" or more specifically, “Is each effect significant for at least one of the dependent variables?" That is, where the between -subjects F test focuses on the significance of the relationship of each predictor variable to each dependent variable, the multivariate tests focus on relationship of each predictor variables and any interaction effects to the set of dependent variables. These tests appear in the “Multivariate Tests" table of SPSS output. The multivariate formula for F is based not only on the sum of squares between and within groups but also on the sum of cross products - that is, it takes covariance into account as well as group means.

SPSS gives us four different approaches to calculate the F value for MANOVA. All of them are used to test whether the vector of means of the groups are from the same sampling distribution or not. We can choose any of them for interpretation.

Hotelling's T -Square is the most common, traditional test where there are two groups formed by the independent variables. SPSS prints the related statistic, Hotelling's Trace (a.k.a. Lawley - Hotelling or Hotelling -Lawley Trace). To convert from the Trace coefficient to the T-Square coefficient, multiply the Trace coefficient by (N-g), where N is the sample size across all groups and g is the number of groups. The T-Square result will still have the same F value, degrees of freedom, and significance level as the Trace coefficient. The larger the Hotelling's trace, the more the given effect contributes to the model.

Wilks' lambda, U. This is the most common, traditional test where there are more than two groups formed by the independent variables. Wilks' lambda is a multivariate F test, akin to the F test in univariate ANOVA. It is a measure of the difference between groups of the centroid (vector) of means on the independent variables. The smaller the lambda, the greater the differences. The Bartlett's V transformation of lambda is then used to compute the significance of lambda. Wilks's

lambda is used, in conjunction with Bartlett's V, as a multivariate significance test of mean differences in MANOVA, for the case of multiple interval dependents and multiple (>2) groups formed by the independent(s). The t-test, Hotelling's T, and the F test are special cases of Wilks's lambda. Wilks' lambda ranges from 0 to 1, and the lower the Wilks' lambda, the more the given effect contributes to the model.

Pillai's trace, also called Pillai -Bartlett trace, V. Multiple Discriminant Analysis (MDA) is the part of MANOVA where canonical roots are calculated. Each significant root is a dimension on which the vector of group means is differentiated. The Pillai -Bartlett trace is the sum of explained variances on the discriminant variates, which are the variables which are computed based on the canonical coefficients for a given root. Found V to be the most robust of the four tests and is sometimes preferred for this reason, especially if the homogeneity of variance assumption or other assumptions have not been met. Specifically, if Box's M is significant, then Pillai's trace is preferred over the usual Wilks' lambda. The larger the Pillai's trace, the more the given effect contributes to the model. Pillai's trace is always smaller than Hoteling's trace.

Roy's Greatest Characteristic Root (GCR), called "Roy's largest root" in SPSS, is similar to the Pillai-Bartlett trace but is based only on the first (and hence most important) root. Specifically, let lambda be the largest Eigenvalue, then GCR = lambda/ (1 + lambda). Note that Roy's largest root is sometimes also equated with the largest eigenvalue, as in SPSS's GLM procedure (however, SPSS reports GCR for MANOVA(. GCR is less robust than the other tests in the face of violations of the assumption of multivariate normality. The larger the root, the more that effect contributes to the model. Note, however, that Roy's largest root sets a lower bound for the significance value, tending to make this test more prone to Type 1 error (false positives) than the nominal significance level might suggest -that is, it is a more liberal test.

6.7 Significance tests of between-subjects' effects (F tests)

In SPSS output, the "Tests of between Subjects Effects" table provides an F test of the significance of the model overall (the "omnibus test", in the "Corrected Model" row). This test is also called the omnibus F-test and it answers the question, "Is the model significant for at least one of the predictors?" This table also reports the significance of the intercept and of each predictor variable

in the model. There is an F significance test for each dependent variable (here, income and educ), both for the model and for each predictor variable.

The tables below show the results of the analysis of the candidate independent variables as inputs to the proposed model:

Table No: 4.18 (The Descriptive Statistics)

Descriptive Statistics				
	groups	Mean	Std. Deviation	N
H_school	Excellent	84.778	1.9047	40
	good	80.745	4.0371	149
	pass	78.460	4.6526	35
	V.good	82.724	3.3207	102
	Total	81.613	4.0724	326
F_grade	Excellent	2.9613	.18966	40
	good	2.3903	.38995	149
	pass	2.1189	.58243	35
	V.good	2.7065	.18513	102
	Total	2.5301	.42370	326
Math	Excellent	81.25	6.130	40
	good	76.70	7.757	149
	pass	75.03	7.374	35
	V.good	78.29	7.348	102
	Total	77.58	7.566	326
Programming	Excellent	85.70	5.145	40
	good	67.62	10.125	149
	pass	47.23	16.861	35
	V.good	77.95	6.194	102
	Total	70.88	14.119	326

The first important one is the **Descriptive Statistics** table shown above. This table is very useful as it provides the mean and standard deviation for the two different dependent variables, which have been split by the independent variables. In addition, the table provides "Total" rows, which allows means and standard deviations for groups only split by the dependent variable to be known.

Table No: 4.19 (The Multivariate Tests)

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.998	32606.536[b]	4.000	319.000	.000	.998
	Wilks' Lambda	.002	32606.536[b]	4.000	319.000	.000	.998
	Hotelling's Trace	408.859	32606.536[b]	4.000	319.000	.000	.998
	Roy's Largest Root	408.859	32606.536[b]	4.000	319.000	.000	.998
group	Pillai's Trace	.692	24.064	12.000	963.000	.000	.231
	Wilks' Lambda	.330	36.569	12.000	844.286	.000	.309
	Hotelling's Trace	1.959	51.846	12.000	953.000	.000	.395
	Roy's Largest Root	1.923	154.348[c]	4.000	321.000	.000	.658

a. Design: Intercept + groups

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

The **Multivariate Tests table** is where we find the actual result of the **MANOVA**. You need to look at the second Effect, labelled "groups", and the Wilks' Lambda row (highlighted in red). To determine whether the one-way MANOVA was statistically significant you need to look at the "Sig." column. We can see from the table that we have a "Sig." value of .000, which means $p < .0005$. Therefore, we can conclude that the Academic student academic level upon graduation was significantly dependent on these four subjects ($p < .0005$).

The above table shows the F values for the independent variables in the model.

To determine how the dependent variables, differ for the independent variable, we need to look at the Tests of Between-Subjects Effects table (highlighted in red):

Table No: 4.20 (The Tests of Between-Subjects Effects)

Tests of Between-Subjects Effects							
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	H_school	986.575[a]	3	328.858	24.048	.000	.183
	F_grade	19.441[b]	3	6.480	53.637	.000	.333
	Math	932.772[c]	3	310.924	5.665	.001	.050
	Programming	35043.291[d]	3	11681.097	126.447	.000	.541
Intercept	H_school	1522923.994	1	1522923.994	111365.152	.000	.997
	F_grade	1477.714	1	1477.714	12230.812	.000	.974
	Math	1382480.976	1	1382480.976	25189.135	.000	.987
	Programming	1106690.031	1	1106690.031	11979.790	.000	.974
group	H_school	986.575	3	328.858	24.048	.000	.183
	F_grade	19.441	3	6.480	53.637	.000	.333
	Math	932.772	3	310.924	5.665	.001	.050
	Programming	35043.291	3	11681.097	126.447	.000	.541
Error	H_school	4403.366	322	13.675			
	F_grade	38.904	322	.121			
	Math	17672.655	322	54.884			
	Programming	29746.279	322	92.380			
Total	H_school	2176798.640	326				
	F_grade	2145.241	326				
	Math	1980675.000	326				
	Programming	1702764.000	326				
Corrected Total	H_school	5389.941	325				
	F_grade	58.345	325				
	Math	18605.426	325				
	Programming	64789.571	325				

a. R Squared = .183 (Adjusted R Squared = .175)

b. R Squared = .333 (Adjusted R Squared = .327)

c. R Squared = .050 (Adjusted R Squared = .041)

d. R Squared = .541 (Adjusted R Squared = .537)

We can see from this table above that group has a statistically significant effect on H_school (F (3, 322) = 24.05; $p < .0005$; partial $\eta^2 = .18$) and F_grade (F (3, 322) = 53.64; $p < .0005$; partial η^2 =

.33) and Math (F (3, 322) = 5.67; p < .0005; partial η^2 = .05) and programming (F (3, 322) = 126.45; p < .0005; partial η^2 = .54).

in this case, we accept statistical significance at p < .025.

Table No: 4.21 (The Multiple Comparisons)

Multiple Comparisons							
Sidak							
			Mean Difference			95% Confidence Interval	
Dependent Variable	(I) groups	(J) groups	(I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
H_school	Excellent	Good	4.033*	.6585	.000	2.289	5.776
		Pass	6.318*	.8559	.000	4.052	8.583
		V.good	2.054*	.6899	.019	.228	3.880
	Good	Excellent	4.033*	.6585	.000	5.776	-2.289
		Pass	2.285*	.6946	.007	.446	4.124
		V.good	1.979*	.4752	.000	3.237	-.720
	Pass	Excellent	6.318*	.8559	.000	8.583	-4.052
		Good	2.285*	.6946	.007	4.124	-.446
		V.good	4.264*	.7244	.000	6.181	-2.346
	V.good	Excellent	2.054*	.6899	.019	3.880	-.228
		Good	1.979*	752	.000	.720	3.237
		Pass	4.264*	.7244	.000	2.346	6.181
F_grade	Excellent	Good	.5710*	.06190	.000	.4071	.7348
		Pass	.8424*	.08045	.000	.6294	1.0554
		V.good	.2548*	.06485	.001	.0831	.4264
	Good	Excellent	.5710*	.06190	.000	.7348	-.4071
		Pass	.2714*	.06529	.000	.0986	.4443
		V.good	.3162*	.04467	.000	.4345	-.1979
	Pass	Excellent	.8424*	.08045	.000	1.0554	-.6294
		Good	.2714*	.06529	.000	.4443	-.0986
		V.good	.5876*	.06809	.000	.7679	-.4074
	V.good	Excellent	.2548*	.06485	.001	.4264	-.0831
		Good	.3162*	.04467	.000	.1979	.4345
		Pass	.5876*	.06809	.000	.4074	.7679
Math	Excellent	Good	4.55*	1.319	.004	1.05	8.04
		Pass	6.22*	1.715	.002	1.68	10.76
		V.good	2.96	1.382	.183	.70	6.61

	Good	Excellent	4.55*	1.319	.004	8.04	-1.05
		Pass	1.68	1.392	.790	2.01	5.36
		V.good	1.59	.952	.454	4.11	.93
	Pass	Excellent	-6.22*	1.715	.002	10.76	-1.68
		Good	-1.68	1.392	.790	5.36	2.01
		V.good	-3.27	1.451	.142	7.11	.58
	V.good	Excellent	2.96	1.382	.183	-6.61	.70
		Good	1.59	.952	.454	.93	4.11
		Pass	3.27	1.451	.142	.58	7.11
programming	Excellent	Good	18.08*	1.712	.000	13.54	22.61
		Pass	38.47*	2.225	.000	32.58	44.36
		V.good	7.75*	1.793	.000	3.00	12.50
	Good	Excellent	18.08*	1.712	.000	22.61	-13.54
		Pass	20.40*	1.805	.000	15.62	25.18
		V.good	10.33*	1.235	.000	13.60	-7.06
	Pass	Excellent	38.47*	2.225	.000	44.36	-32.58
		Good	20.40*	1.805	.000	25.18	-15.62
		V.good	30.72*	1.883	.000	35.71	-25.74
	V.good	Excellent	7.75*	1.793	.000	12.50	-3.00
		Good	10.33*	1.235	.000	7.06	13.60
		Pass	30.72*	1.883	.000	25.74	35.71

Based on observed means.

The error term is Mean Square (Error) = 92.380.

*. The mean difference is significant at the .05 level.

From the table above, we notice there are is star sign and the most of the significance values are less than 0.05, that mean there are statistically significant differences between all groups.

Through all the results of the above analysis, we can say that there is significance effect for student's result in high school, the result of the first year, and the result of mathematics and programming methods at the student's final level upon graduation.

Through this analysis, we figured the answer to the third question of this research: Can we identify courses that most influence the student's graduation level through available academic data?

Was reached by demonstrating the existence of an effect and relationship between the student's academic grades as independent variables and his graduation level in the final year as a dependent variable.

These results are preliminary results that confirm the possibility of using the available academic data to predict the level of the student's graduation rate by building and developing a predictive model.

CHAPTER FIVE DESIGN AND IMPLEMENTATION

5.1 Overview

The building dynamic predictive model for using for expecting the student's level upon graduation, starting from early stages in his academic career, helps the educational administration to develop different methodologies to improve the level of students expected to be poorly performing or to support students who are expected to perform well.

Experiments have proven that graduates with weak rates find it difficult to employ and find good university admission to complete their studies, and therefore the need has emerged to use ICT for developing model that can help the limit students expected to graduate to a critical level to help them correctly and timely, by securing an appropriate academic plan after discovering deficiencies and weaknesses, and whenever it was. The early detection, the better the chance to correct the path. In this chapter, we will work to answer the following question: Is it possible to suggest a dynamic model to predict the Student's level at graduation?

The aim of this research is to use data mining techniques to design a dynamic model for predicting graduation performance for final year students at the university using pre-university and examination marks in the first years at the university only, and no socioeconomic or demographic features are used. In this chapter, we used the machine learning regression algorithms were used to create the model.

To construct the predictive model, we used a linear regression algorithm that takes multiple inputs and gives as output the student's level at graduation measured in average.

Before applying the system, it must first determine the appropriate inputs and then set the necessary settings for a system. Therefore, a practical model must be found that begins with the preparation of the requirements for the linear regression algorithm and then builds the model and tests it to be ready for use. The practical model that was followed started from processing the requirements of the decision tree algorithm (inputs, settings) and then building the system and testing it to be ready for use. We will first explain the proposed model and then we will apply the model and test and evaluate its performance.

Before entering the stage of design, the model, there is a preparatory stage that must be carried out, we used in this stage the Feature Selection. Feature Selection technique is the process where we can automatically select the features which contribute most in prediction variable or output in which we are interested in. Feature Selection and data cleaning should be the first step and most important step of the model designing.

The proposed of the dynamic predictive model consists of eight successive steps, as shown below:

5.1 Details of the proposed of the dynamic predictive model:

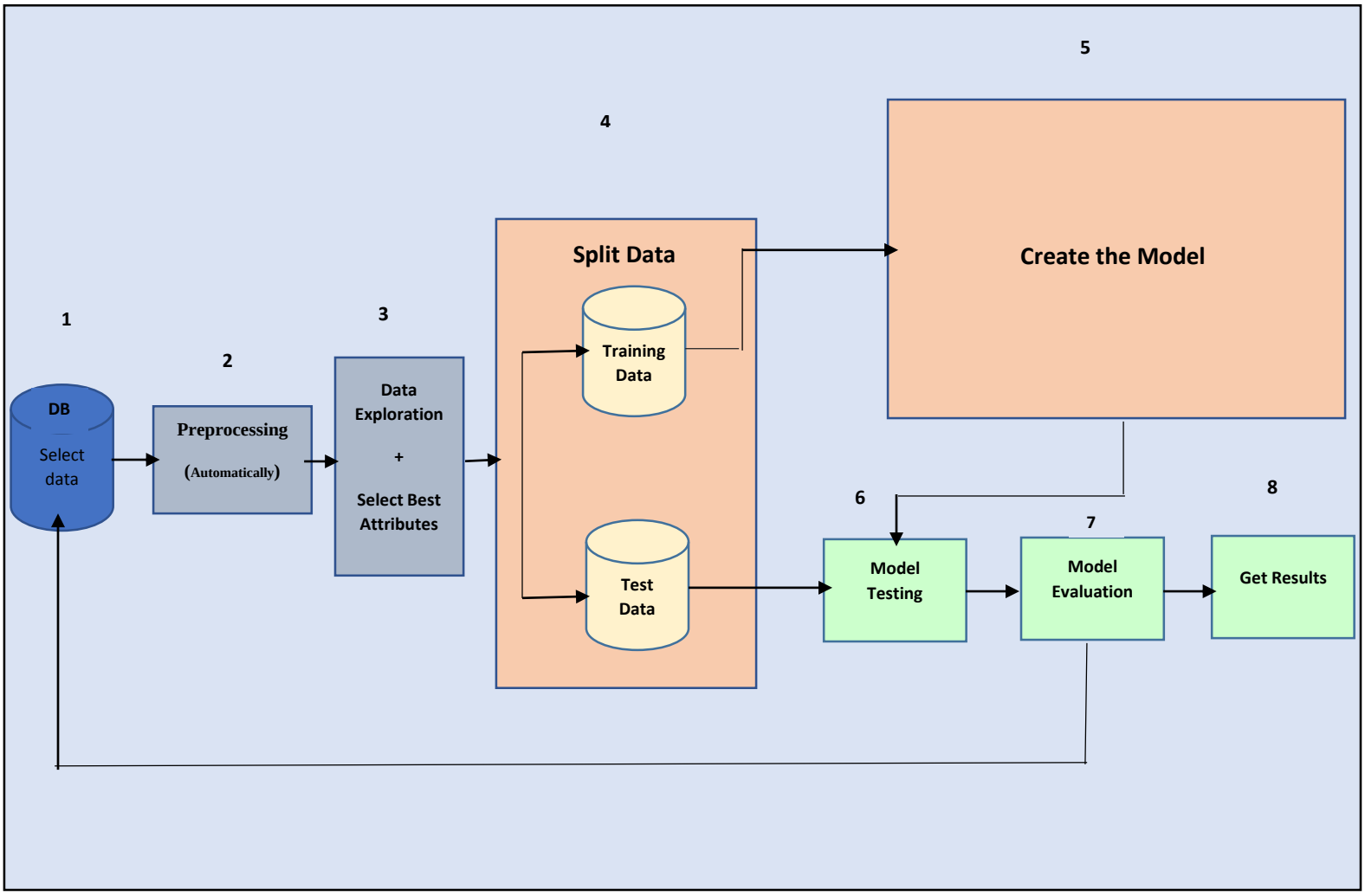


Figure No: 5.1 (Dynamic Predictive Model)

5.3 Explanation of proposed of the dynamic predictive model:

Step 1: Select Data

In this step, we collected the data for our database. Now in this step, we have to mark and select the most relevant data which we need to carry forward.

Step 2: Preprocessing Data

In this step we can Data Preprocessing handles the problems on the data, like missing values, outliers, noisy or dirty data and so on automatically by using python .code like the following:

Data cleaning: - Here, some data that contains impurities, errors, noise and so on is eliminated.

Missing data: - One of the most important requirements of the data mining process is that the data is completely configured that does not contain missing values Of course, there are ways to correct this data, such as the arithmetic mean and other operations.

Step3: Data Exploration and Feature selection

After the pre-processing phase, features selection process was started. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

To simplification of models to make them easier to interpret by researchers and users, to shorter training times, to avoid the curse of dimensionality and to enhanced generalization by reducing over fitting.

This becomes even more important when the number of features are very large. We need not use every feature at your disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are really important for getting best attributes.

The objective of feature selection is improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data."

The summarize importance of feature selection:

It enables the machine learning algorithm to train faster.

It reduces the complexity of a model and makes it easier to interpret.

It improves the accuracy of a model if the right subset is chosen.

It reduces over fitting.

For feature selection techniques the correlation statistic we used the f_regression () function. This function can be used in a feature selection strategy, such as selecting the top k most relevant features (largest values) via the SelectKBest class.

Step4: Split the Data

After determining the data set and getting the best attributes, we divide the data into two parts, the first part is training present 80% of dataset for training the model, second part is test data present 20% of dataset the model training data for testing the model and get the results.

Step 5: Create the Model

Once the data have been divided into the training and testing sets, the next step is using machine learning algorithm on this data to build the dynamic prediction model, we used the regression algorithm for develop the model. Then we can learn the model by using training data.

Step 6: Test the Model

At this stage, we test the model that was built by applying it to predict the level of students who actually graduated and then compare it with the actual results.

Step 7: Model Evaluation

The identifying the related features from a set of data and removing the irrelevant or less important features with do not contribute much to our target variable in order to achieve better accuracy for our model the accuracy results were evaluated to see which variable type can work efficiently with the algorithm in training the datasets of interest.

Here we can use accuracy estimation methods such as cross-validation, R2 score and accuracy_score.

Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model. R2. Score is made for continuous variables, such as for regression problems.

Step **8: Get the Results**

The stage of obtaining results comes after the evaluation process of the model and obtaining the b est results of the evaluation by comparing the current evaluation result with the results of the prev ious evaluation. It is the step in which the final results of the model are approved.

5.4 Implementation and Design the Model

We applied the proposed model to the 326 students who graduated in 2015, 2016 and 2017, and t hen we tested the model to predictive the level of student graduation. And to test the effectivenes s of the model by comparing the results that will be obtained with the actual results.

Table No: 5.1 (sample of data set)

NO	FRMNO	Home	SMATH	CHEMISTRY	PHYSICS	H_S_Result	F_Y_Result	Final
1	**157020**	**0**	**87**	**75**	**77**	**83**	**2.75**	**2.71**
2	**151189**	**0**	**90**	**77**	**78**	**83.1**	**2.8**	**2.48**
3	**25761**	**1**	**63**	**66**	**67**	**71.9**	**0**	**2.22**
4	**187203**	**1**	**82**	**78**	**84**	**84.9**	**2.6**	**2.51**
5	**188795**	**1**	**79**	**78**	**80**	**82.7**	**2.55**	**2.42**
6	**2073**	**1**	**70**	**70**	**66**	**71.7**	**2.48**	**2.42**
7	**163678**	**1**	**76**	**85**	**78**	**83.1**	**2.9**	**2.67**
8	**151911**	**1**	**76**	**79**	**79**	**83**	**2.82**	**2.81**
9	**58980**	**1**	**76**	**71**	**76**	**83.6**	**2.66**	**2.71**
10	**170584**	**1**	**77**	**84**	**78**	**84**	**2.61**	**2.57**
11	**181596**	**1**	**84**	**78**	**84**	**83**	**2.55**	**2.51**
12	**19270**	**1**	**61**	**61**	**68**	**70.9**	**2.51**	**2.4**
13	**147585**	**1**	**77**	**83**	**85**	**84**	**3.02**	**3.08**
14	**97038**	**0**	**88**	**80**	**82**	**82.9**	**2.41**	**2.41**
15	**116521**	**1**	**75**	**84**	**83**	**83.4**	**2.3**	**2.47**
16	**23411**	**1**	**63**	**68**	**75**	**76.9**	**2.8**	**2.67**
17	**23588**	**1**	**70**	**67**	**71**	**73.4**	**2.26**	**2.22**
18	**182555**	**1**	**86**	**77**	**82**	**82.7**	**2.93**	**2.66**
19	**96884**	**1**	**90**	**79**	**81**	**84.4**	**2.8**	**2.73**
20	**187298**	**1**	**79**	**85**	**84**	**84.7**	**2.58**	**2.69**
21	**159287**	**1**	**79**	**78**	**89**	**83.6**	**2.73**	**2.56**
22	**3778**	**1**	**67**	**67**	**73**	**74.1**	**2.36**	**2.35**
23	**17023**	**1**	**80**	**81**	**78**	**82.1**	**2.69**	**2.71**
24	**105787**	**0**	**74**	**82**	**78**	**84.1**	**2.85**	**2.64**

5.4.1 The Preparation to apply the model

The preparation to apply the model passes the following stages as show in the figure below:

```
In [855]: import pandas as pd
          import numpy as np
          from sklearn.model_selection import train_test_split
          from sklearn.feature_selection import SelectKBest
          from sklearn.feature_selection import f_regression
          import matplotlib.pyplot as plt
          from sklearn.feature_selection import mutual_info_regression
          from sklearn.linear_model import LinearRegression
```

```
In [930]: dataset = pd.read_csv('dynamic5.csv')
          dataset.head()
```

Out[930]:

	MATH	CHEMISTRY	PHYSICS	F.grade	S.grade	H.school	Final
0	82.0	78.0	84.0	2.60	2.50	84.9	2.51
1	88.0	75.0	73.0	2.62	2.64	79.4	2.63
2	83.0	76.0	80.0	2.91	2.79	84.9	2.79
3	79.0	78.0	80.0	2.55	2.32	82.7	2.42
4	77.0	77.0	80.0	2.42	2.46	83.4	2.33

Figure No: 5.2(The Preparation Stage)

It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis. It is the stage of removing data that contains interference or noise from the dataset to get that clean database. Before we start apply the model, it is important to preprocess the dataset. In most cases, the raw data needs to be preprocessed before it can be used as input to train a predictive dynamic model.

The data set was divided into two training data sets and test data %80 for training data and 20% for test data as required by the technology used in building the model(X, y) The training data was divided into two input and output sections (X_train) and (y_train) The test data was also divided into two parts (X_test) and (y_test) as the flowing:

1). X_train - This includes all independent variables, these will be used to train the model, also as we have specified the test_size = 0.2, this means 80% of observations from our complete data will be used to train/fit the model and rest 20% will be used to test the model.

2). X_test - This is remaining 20% portion of the independent variables from the data which will not be used in the training phase and will be used to make predictions to test the accuracy of the model.

3). y_train - This is our dependent variable which needs to be predicted by this model, this includes category labels against your independent variables, we need to specify our dependent variable while training/fitting the model.

4). y_test - This data has category labels for your test data, these labels will be used to test the accuracy between actual and predicted categories.

By using Python techniques as shown below:

```
In [883]: X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=0)

In [884]: difference3 = pd.DataFrame(X_train)
          difference3
```

Out[884]:

	MATH	CHEMISTRY	PHYSICS	F.grade	S.grade	H.school
184	93.0	82.0	63.0	2.49	2.47	82.7
260	0.0	0.0	0.0	2.39	2.49	78.0
208	73.0	79.0	80.0	3.01	2.82	84.6
89	79.0	71.0	68.0	2.78	2.68	78.0
247	87.0	76.0	73.0	3.13	3.20	83.1
...	...	...	...	...	...	...
251	84.0	80.0	83.0	2.57	2.68	82.1
192	84.0	80.0	80.0	2.70	2.76	84.6
117	90.0	76.0	81.0	3.12	3.14	85.1
47	93.0	91.0	86.0	2.91	2.76	87.6
172	65.0	70.0	71.0	2.04	2.47	77.0

222 rows × 6 columns

Figure No: 5.2 (The sample of inputs)

```
In [1046]: difference2 = pd.DataFrame(y_train)

           difference2
```

Out[1046]:

	Final
184	2.66
260	2.57
208	3.00
89	2.89
247	3.21
...	...
251	2.68
192	2.84
117	3.03
47	2.85
172	2.53

222 rows × 1 columns

Figure No: 5.3 (The training data)

5.4.2 The Feature Selection technique

Feature selection or feature pruning is a very crucial step in the pipeline of building a good prediction model and to understand the connections among the features and the target. The goal of feature selection is two-fold:

1. Identify and remove features with little or no predictability of the target to prevent over fitting.
2. Identify highly correlated or redundant features and suppress the negative impacts towards the model without losing critical information.

More than often, the feature selection is a rather iterative process, and each step has unique pruning targets. Hence, it's important to understand how each algorithm fits in different scenario in order to achieve the ideal results. (https://medium.com/ro-data-team-blog/feature-selection-strategies-for-regression-models-9147a361bae7)

In [1049]:
```
# feature selection
f_selector = SelectKBest(score_func=f_regression, k=3)
# learn relationship from training data
sfs1 = f_selector.fit(X_train, y_train)
# transform train input data
X_train_fs = f_selector.transform(X_train)
df = pd.DataFrame(X_train_fs,columns=['F-grade', 'S-grade', 'H-school'],dtype=float)
df
```

Out[1049]:

	F-grade	S-grade	H-school
0	2.49	2.47	82.7
1	2.39	2.49	78.0
2	3.01	2.82	84.6
3	2.78	2.68	78.0
4	3.13	3.20	83.1
...	...	...	...
217	2.57	2.68	82.1
218	2.70	2.76	84.6
219	3.12	3.14	85.1
220	2.91	2.76	87.6
221	2.04	2.47	77.0

222 rows × 3 columns

Figure No: 5.4 (Feature Selection technique)

In [927]:
```
# Plot the scores for the features
plt.bar([i for i in range(len(f_selector.scores_))], f_selector.scores_)
plt.xlabel("feature index")
plt.ylabel("F-value (transformed from the correlation values)")
plt.show()
```

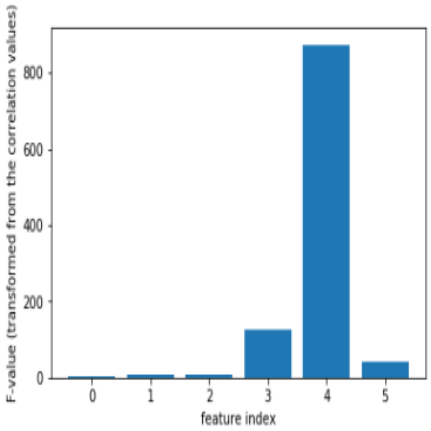


Figure No: 5.5 (Plot the scores for the features)

5.4.3 Build the Model with Features Selection

In this section we built the model using features selection by correlation statistical, via linear regression algorithm as shown in figure below:

```
In [1052]: Linear_regressorc = LinearRegression()
           Linear_regressor.fit(X_train_fs, y_train)
           y_pred = Linear_regressor.predict(X_test_fs)
           dY = np.float16(y_pred)
           dY = np.round(y_pred,2)
           difference = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': dY})
           difference
```

Out[1052]:

	Actual Value	Predicted Value
250	2.67	2.60
256	2.85	2.85
15	2.58	2.72
65	2.86	2.93
213	2.40	2.41
...	...	...
216	2.71	2.57
29	2.81	2.61
97	3.11	3.09
20	2.39	2.19
46	2.33	2.21

96 rows × 2 columns

Figure No: 5.6 (The Results of the Model)

5.4.4 Increase the Effectiveness of the Model's Performance

To increase the effectiveness of the performance of the proposed dynamic model, the features selection technique was used by adding or deleting some features, and each time the model's performance is measured and compared with the previous measurements to obtain the optimal results expected from the application of the model.

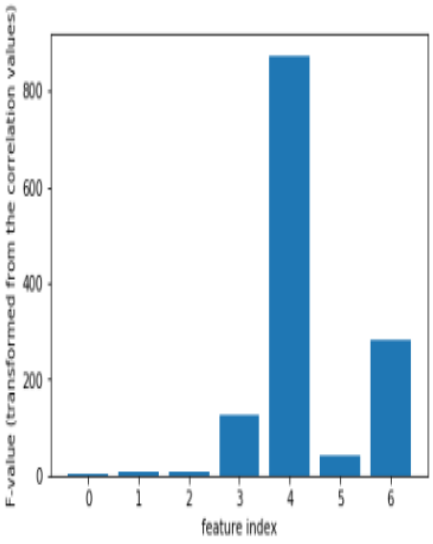
To achieve this, student results in the programming methods course were added, and upon applying the model, better results were obtained than the previous results, as shown in the figure below:

In [1136]:
```
# feature selection
f_selector = SelectKBest(score_func=f_regression, k=3)
# learn relationship from training data
sfs1 = f_selector.fit(X_train, y_train)
# transform train input data
X_train_fs = f_selector.transform(X_train)
df = pd.DataFrame(X_train_fs,columns=['f-grade', 's-grade', 'program method'],dtype=float)
df
```

Out[1136]:

	f-grade	s-grade	program method
0	2.49	2.47	63.0
1	2.39	2.49	57.0
2	3.01	2.82	90.0
3	2.78	2.68	88.0
4	3.13	3.20	95.0
...	...	...	...
217	2.57	2.68	64.0
218	2.70	2.76	80.0
219	3.12	3.14	91.0
220	2.91	2.76	86.0
221	2.04	2.47	53.0

222 rows × 3 columns




```
In [1139]: Linear_regressorc = LinearRegression()
           Linear_regressor.fit(X_train_fs, y_train)
           y_pred = Linear_regressor.predict(X_test_fs)
           dY = np.float16(y_pred)
           dY = np.round(y_pred,2)
           #difference = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': dY})
           #difference
           print('Accuracy:', r2_score(y_test,y_pred))

           Accuracy: 0.8576935479494783
```

Figure No: 5.7 (Improve the performance of Model)

5.4.5 Evolution the Model

At this stage, we tested the model that was built by applying it to predict the level of part of the students who actually graduated, and then compare the results with the actual results.

Model Evaluation is an integral part of the model development process. It helps to find the best techniques that represents the data, by calculate accuracy we can evaluate the model.

Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.

R-squared is a statistical measure of how close the data are to the fitted regression line. ... 0.0 indicates that the model explains none of the variability of the response data around its mean. 1.0 indicates that the model explains all the variability of the response data around its mean. R2. score is made for continuous variables, such as for regression problems.

Calculate the R2 value as:

R2 = 1- sum (residual squared)/ (N*variance of data)

if R-squared value 0.3 < r < 0.5 this value is generally considered a weak or low effect size, - if R-squared value 0.5 < r < 0.7 this value is generally considered a Moderate effect size, - if R-squared value r > 0.7 this value is generally considered strong effect size.

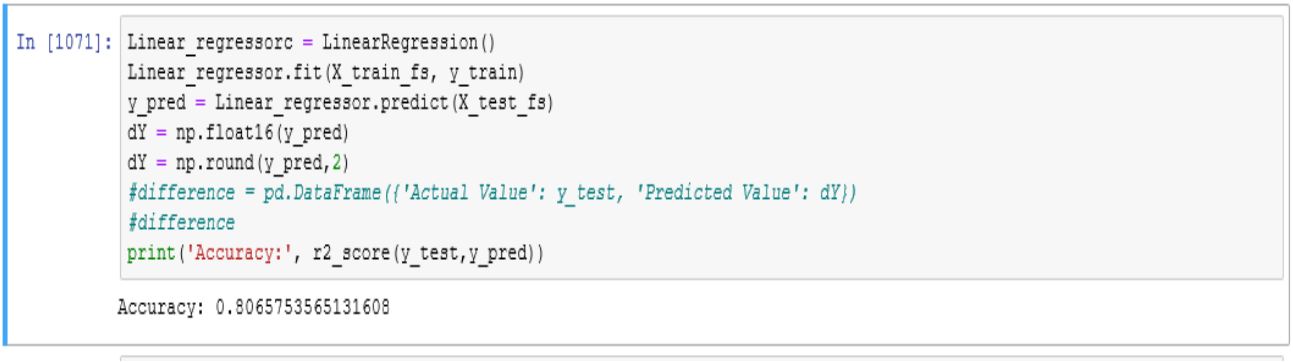


Figure No: 5.8 (Evolution the Model with features selection)

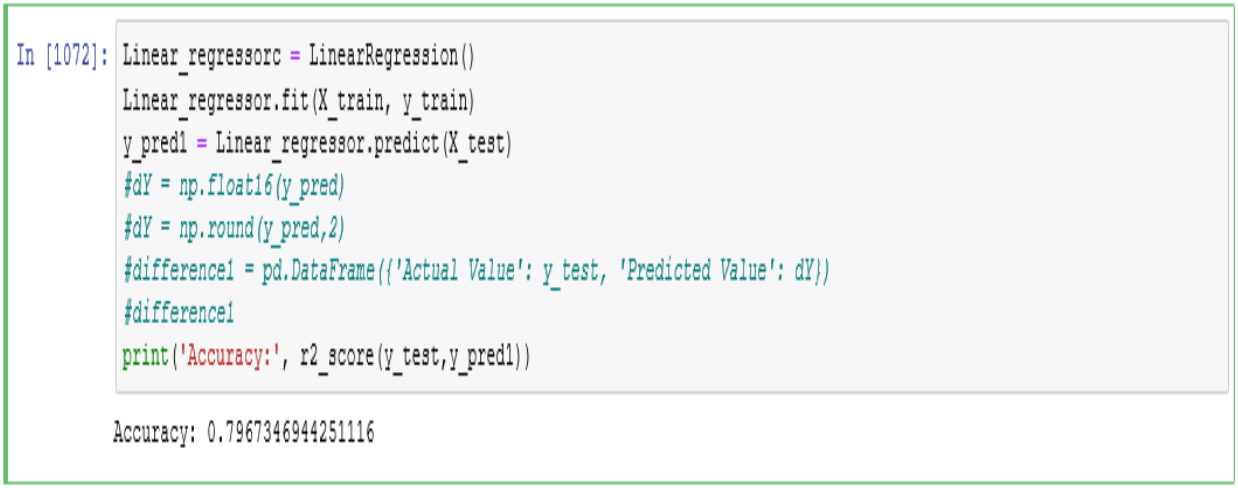


Figure No: 5.9 (Evolution the Model without features selection)

```
In [1139]: Linear_regressorc = LinearRegression()
           Linear_regressor.fit(X_train_fs, y_train)
           y_pred = Linear_regressor.predict(X_test_fs)
           dY = np.float16(y_pred)
           dY = np.round(y_pred,2)
           #difference = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': dY})
           #difference
           print('Accuracy:', r2_score(y_test,y_pred))

           Accuracy: 0.8576935479494783
```

Figure No: 5.10 (Evolution the Model after change the features)

CHAPTER SIX RESULTS AND DISCUSSIONS

6.1 Overview

To develop a dynamic model that can use for prediction the student performance, is an important matter in education. Because we can predict future performance of a student after being enrolled into a university; thus, determining who would do well and who would receive poor scores. Those results would with help making admission decisions more efficient and improve the quality of academic services. Specifically, administrators can use predictive results to evaluate performance of students in the next semesters. Lecturers can select suitable learning strategies for students depending on their scores and estimate how they would help the students improve within a certain of extent. Such benefits impulse the development of computerized methods that could predict the results with highly reliable accuracy. In this note, the multi-input multi-output problem, which aims to construct a forecast model for the prediction of future performance of students, is examined. In this research, the data mining and Machine learning approaches are used to build a dynamic predictive model.

6.2. Results:

By referring to the objectives and questions of the research and through design and implementation that was carried out in the previous chapter, we reached the following results:

6.2.1 Building a dynamic model to predict a student's graduation level

We got a dynamic model that helps predict student level from early on. Where the model was applied and tested on the database of the College of Computer Science, Sudan University of Science and Technology for the years 2012, 2013 and 2014, and we obtained results with high accuracy as shown in the figure below:

```
In [29]: Linear_regressor = LinearRegression()
         Linear_regressor.fit(X_train_fs, y_train)
         y_pred = Linear_regressor.predict(X_test_fs)
         dY = np.float16(y_pred)
         dY = np.round(y_pred,2)
         difference = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': dY})
         difference
```

Out[29]:

	Actual Value	Predicted Value
250	2.67	2.59
256	2.85	2.83
15	2.58	2.75
65	2.86	2.95
213	2.40	2.40
...	...	...
216	2.71	2.59
29	2.81	2.62
97	3.11	3.09
20	2.39	2.26
46	2.33	2.24

96 rows × 2 columns

Figure No: 6.1 (Dynamic Model Results)

6.2.2 Investigate the possibility of predicting the student's graduation level through the student's academic record

The results in chapter 4 by using the statically analysis confirm the possibility of using the available academic data to predict the level of the student's graduation rate, by finding the correlation coefficient and proving the effect of the independent variables with the dependent variable as shown in the figure below:

Table No: 6.1 (Dynamic predictive Model Results)

NO	Candidate variables	Coefficient of correlation	ANOVA Testing	Evaluation
1	The student's First Year Grade	0.625	.000	Good correlation and a strong impact
2	The student's mark in programming	0.561	.001[b]	Good correlation and a strong impact
3	The student's mark in high school	0.463	.000	Good correlation and a strong impact
4	The student's mark in mathematics	0.465	.015	Good correlation and a strong impact
5	The student's mark in chemistry	0.318	.000	Weak correlation and a strong impact
6	The student's mark in Physic	0. 318	.000	Weak correlation and a strong impact
7	The student's mark in Enter__computer	0.403	.021	Weak correlation and a Weak impact
8	The student's home place	0.000	.998	Weak correlation and a Weak impact

6.2.3 Know the reasons for the student's poor performance through the dynamic predictive model

Through we used the features selection technology, we were able to access the features most influencing the final outcome of the student by adding and removing some features in order to reach the best results with acceptable accuracy.

```
In [46]: # feature selection
f_selector = SelectKBest(score_func=f_regression, k=3)
# learn relationship from training data
sfs1 = f_selector.fit(X_train, y_train)
# transform train input data
X_train_fs = f_selector.transform(X_train)
df = pd.DataFrame(X_train_fs,columns=['F-grade', 'S-grade', 'Program method'],dtype=float)
df
```

Out[46]:

	F-grade	S-grade	Program method
0	2.49	2.47	63.0
1	2.39	2.49	57.0
2	3.01	2.82	90.0
3	2.78	2.68	88.0
4	3.13	3.20	95.0
...	...	...	...
217	2.57	2.68	64.0
218	2.70	2.76	80.0
219	3.12	3.14	91.0
220	2.91	2.76	86.0
221	2.04	2.47	53.0

222 rows × 3 columns

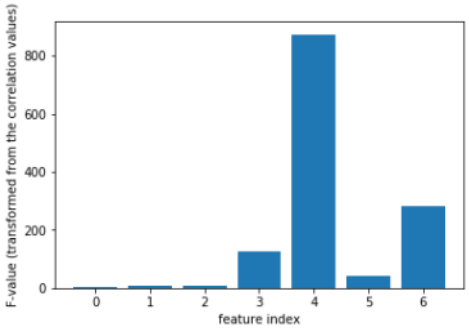


Figure No: 6.2 (Features most influencing of the student level)

6.2.4 Using the features selection technique increases the model efficiency

It became clear through the application and testing of the model, that the use of features selection technology increases the efficiency of the model's performance and makes the model work in a dynamic way through which the largest number of variables can be selected and tested, where the most influential variables are selected for the expected results. This is done by calculating the level of accuracy for each addition or deletion of the variables as shown in the accuracy results below:

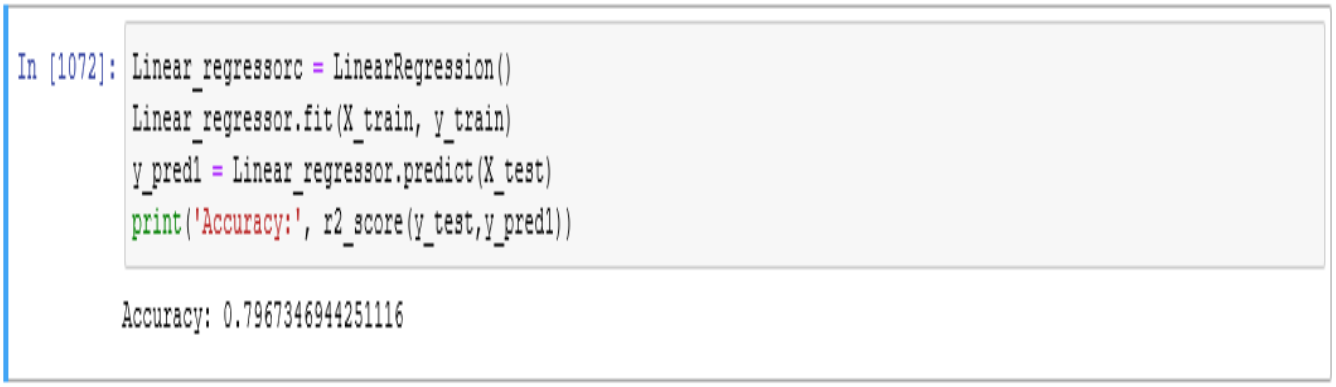


Figure No: 6.3 (Accuracy before add program method score)

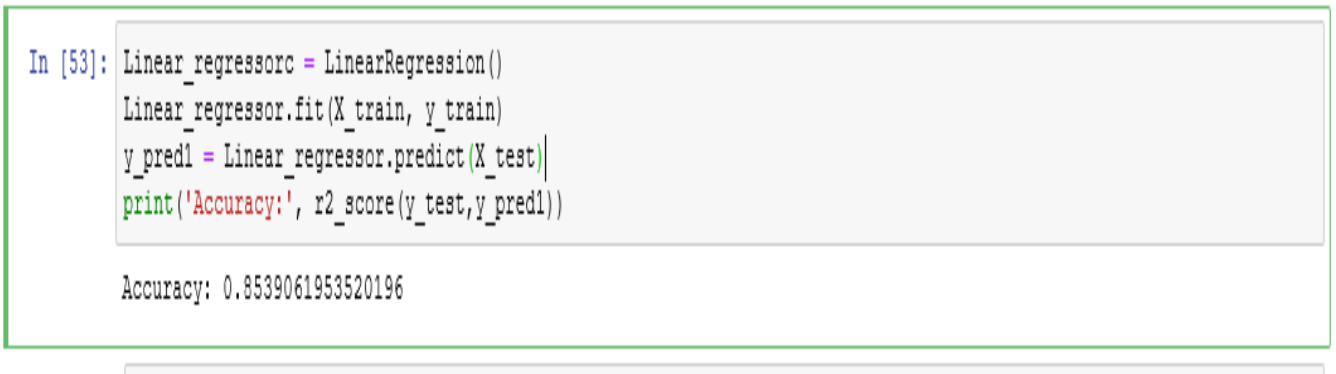


Figure No: 6.4 (Accuracy add program method score)

6.2.5 The results obtained are very close

The results obtained through the use of the converging linear regression algorithm, and this is shown by Using the graph to compare the expected results of the model with the actual results, gives a clearer picture of the extent of the convergence of the actual results with the expected results, in which the points appeared in the form of a straight line as shown in the figure below:

```
In [37]:  # plotting the points
plt.plot(y_test, dY)
plt.plot( dY,y_test, color='green', linestyle='dashed', linewidth = 3,
         marker='o', markerfacecolor='blue', markersize=10)
  # naming the x axis
plt.xlabel('Actual - axis')
# naming the y axis
plt.ylabel('Predicted - axis')

# giving a title to my graph
plt.title('Prediction model!')
  # function to show the plot
plt.show()
```

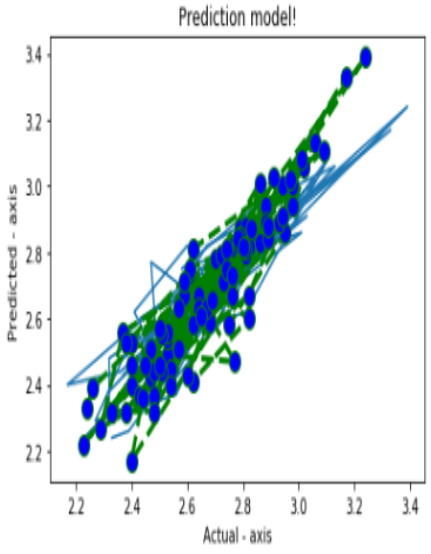


Figure No: 6.5(Diagram to illustrate the relationship between actual results and predictive results)

6.2.6 Build Web application For Applying the Model

To facilitate the process of using the proposed model, we designed web application to enable users to upload new data set to and trained on the model and then can use it to predict Student graduation results as shown in the designs below:

Figure No: 6.6(home page of the web application)

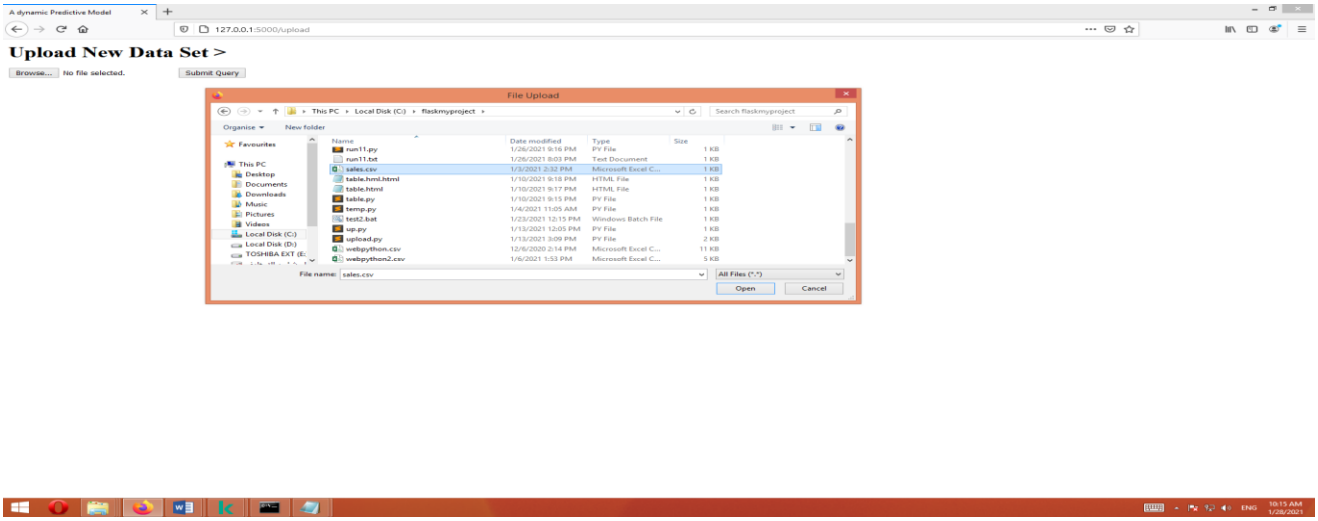


Figure No: 6.7(upload new data set)

Figure No: 6.8 (summit upload new data set)

Figure No: 6.9 (page of the predictive model)

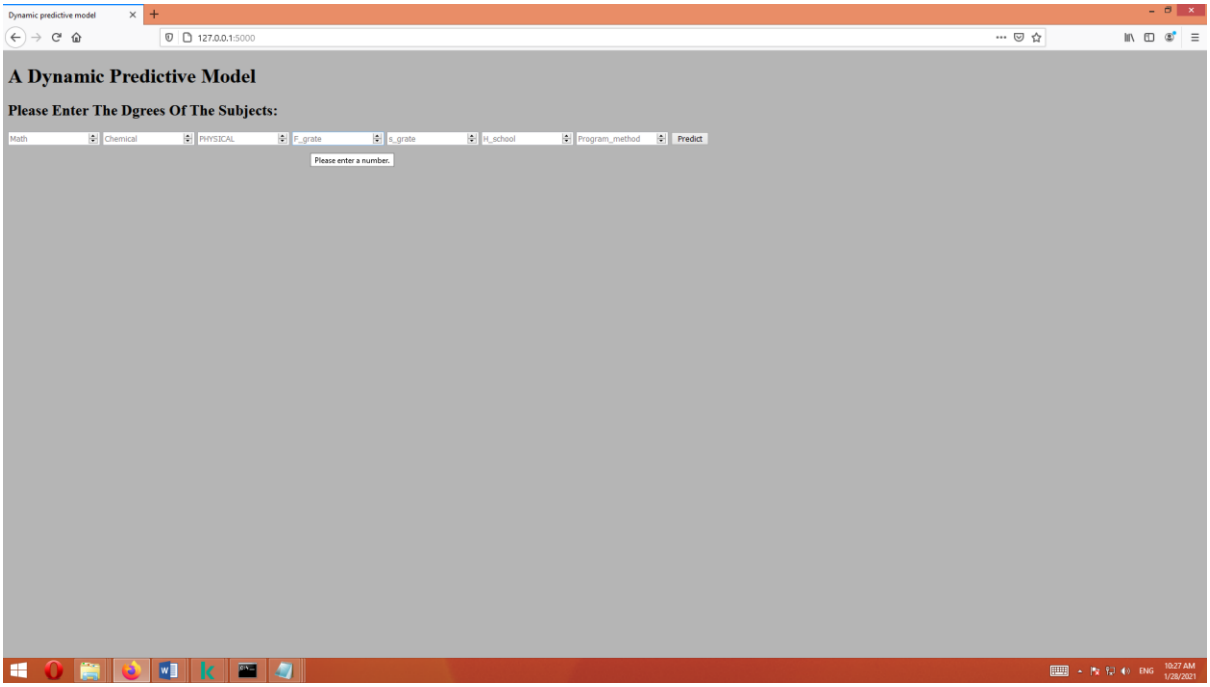


Figure No: 6.10 (page of the enter the subject of the student)

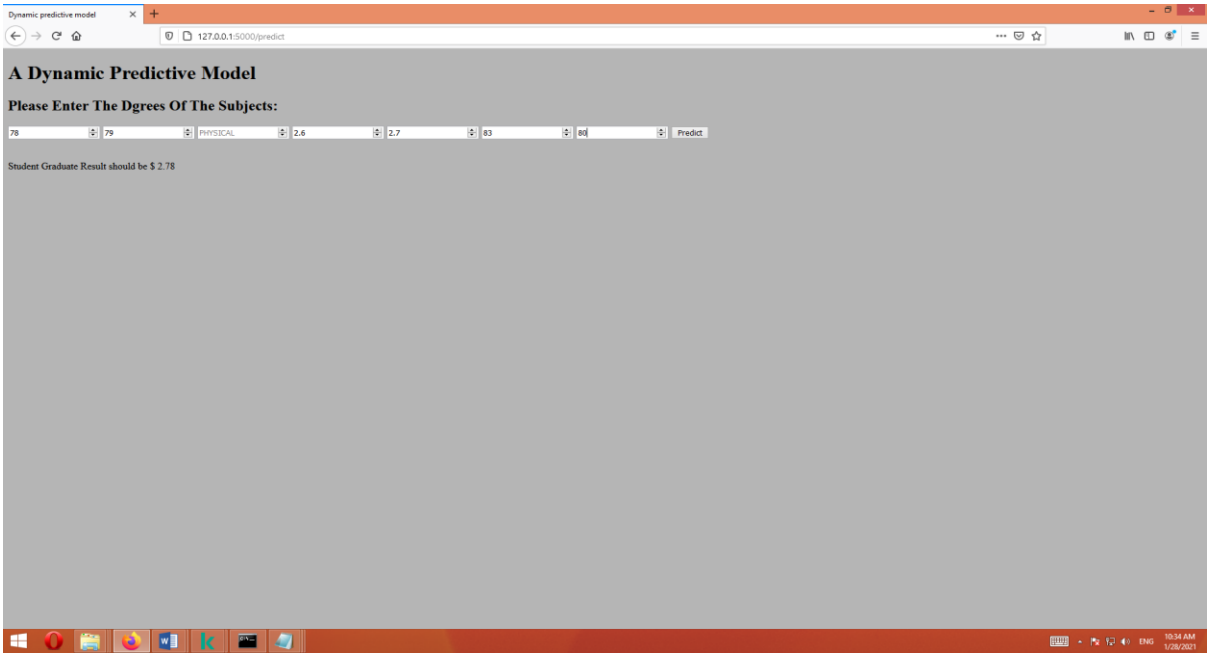


Figure No: 6.11 (predictive result of student)

6.3 Discussion the Results

After building and applying the Dynamic predictive model using the data of students of the College of Computer Science, Sudan University of Science and Technology for three batches 2012, 2013, 2014, we can discuss the results as following:

1-By analyzing the available academic data of student's records, we predicted the level of students 'graduation in the final year, which can use by the educational administration in providing guidance and follow-up to the groups expected to graduate at a critical level.

2- We found that the results of the first year and second year of the students have a major potential for determining the student's level upon graduation, and this indicates the possibility of early prediction of the student's graduation level, which gives sufficient time for the educational administration in the students' orientation and evaluation process.

3- We found that the student's final result in the high school affects the level of the student's graduation, which means that guidance and follow-up can also be focused on the students with lower grades in the high school.

4- We also notice that the student's result in the programming methods subject has an impact on the student's graduation level, because it is a specialized subject closely linked to most of the subjects in computer science field.

5-We found some high school subjects are not significantly influencing the student's performance in the field of computer science, such as physics and chemistry, and this indicates these subjects are not necessary to be among the conditions for admission this field.

6-Implementation of Data Mining concept has made us enable to make calculated decisions about student admissions based on student's available data. Advantages such as lower drop out and higher enrollment ratio, which have been made possible with the help of data science, cannot be overlooked.

The developing a dynamic predictive model for Predicting the outcome of students graduating in the final year of university study is important for any educational institution. Especially for those who aim to give students opportunities to do something useful in their field of study, and those who aim to manage the teaching resources needed for excellent learning experiences, such as the

Sudan University of Science and Technology aim to improve their reputation and classification by selecting high-performing students to involve them in solving real-world issues . Therefore, the result of student graduation is very beneficial. In addition, anticipating an early student graduation score is necessary to help students at risk by alleviating the challenges they face in their careers and helping them excel in the learning process.

The biggest challenge was not having enough data set records to analyze. However, the results demonstrated that this could be done with reasonably large accuracy rates. The main reason this prediction model can attribute to success is the model training method it uses, which relies only on data or samples to build its prediction model.

We also concluded that a set of student performance data is valuable to predict the level of student graduation in the final year, this process is important to improve the quality of education that is vital to attract students to stay in university in general, since the education and evaluation system is very important, we assume to a large extent that it is still More impressive statistics can be extracted from the Student Performance Data Set, which is available for free. We will recommend it as part of future work.

The use of features selection technology gave the model high flexibility by making it a dynamic model that could deal with different numbers of variables in different data set in addition to the ease of control to obtain the best results through flexibility in calculating the accuracy of the model, which made the model more effective and quality.

The accuracy rate resulting from using the model after application is 0.85, and it is an acceptable rate because the goal of actually applying the dynamic predictive model is not to predict the student's rate in an accurate numerical order but rather to help for using in predict the student's level, so we find it acceptable to give the approved accuracy scale this value, taking into account that the forecast is to be made from The first year or second year is aimed at correcting the student's course from an early date, which is relatively long.

CHAPTER SEVEN CONCLUSION

7.1 Conclusion

Data science has become extremely important. This has strengthened the role of the database and the information industry. Through the availability of a large number of databases and information repositories in many institutions and organizations, this has increased the need to analyze and benefit from them. It is not possible without powerful tools. Data mining tools analyze data from different perspectives and summarize the results as useful information. They are employed to find hidden patterns and connections that can be useful in decision-making.

The method of exploration varies, depending on the types of data that can be applied. Also, data mining systems use mathematical, statistical, and smart methods to build future expectations and explore behavior and trends, allowing for the right decisions to be estimated and taken at the right time. The essence of decision support systems is data mining and prediction, early warning and scenario formulation based on simulation models, where decision support systems collect available data with the personal visions of the decision maker, through a set of mathematical models for forecasting and simulation.

There are several types of data extraction: correlation analysis, decision tree, genetic algorithms, Bayesian networks, raw group pathway, neural network, statistical analysis and prediction. Some of the traditional tools used in prediction, for example, regression and differential analysis. New methods include association rules, decision trees, neural networks, and genetic algorithms.

This study focused on data mining and machine learning capabilities for developing a dynamic predictive model in higher education institutions to study educational data with the aim of predicting the academic performance of students at the end of a four-year Bachelor's degree program and identifying effective indicators for students at risk in the first years of their studies and providing the educational institution with the necessary information through which measures can be identified to improve the quality.

This study examined three research questions it aims to find ways to build a dynamic predictive model that helps teachers and program managers with information that might help them motivate

students and improve the education process at the Sudan University of Science and Technology. The first question relates to the possibility of building a dynamic predictive model can use for early prediction of the level of student graduation. Results indicate that the possibility of building the model for predicting the graduation performance in a four-year undergraduate program can be predicted using only pre-university and first-year and second year grades with reasonable accuracy. Further, the model created is dynamic for using for generalizing purpose.

The second question seeks to know the ability of the predict performance of students accurately using academic scores only. For answering this question, we conducted an analytical study of the candidate variables, to obtain preliminary results that clarify the extent of the relationship and effect between the independent variables and the dependent variable, we used many statistical tools like Pearson correlation coefficient analysis, Analysis of variance (ANOVA) and Multivariate analysis of variance (MANOVA).

As for the third question, its axis revolves around the possibility of determining courses that can serve as indicators of the new performance or the student's performance. The course was used for programming methods and the first-year result and second year rate that could serve as these indicators.

There are more variables and factors which may be effective on students' performance in the university such as place of birth, health status, marital status, study of staff data the teaching staff, and the extent of their influence on the students' level, It may be considered in the future work of this research.

Feature selection is the procedure of selecting a subset (some out of all available) of the input variables that are most relevant to the target variable. Target variable here refers to the variable that we wish to predict. It can use for selecting out the most significant features from a given dataset. In many of the cases, Feature Selection can enhance the performance of a machine learning model as well.

The most famous feature selection techniques that can be used for numerical input data and a numerical target variable are the following:

Correlation (Pearson, spearman)

Mutual Information (MI, normalized MI)

In this research we have used one way in order to perform feature selection.

Correlation is a measure of how two variables change together. The most widely used correlation measure is the Pearson's correlation that assumes a Gaussian distribution of each variable and detects linear relationship between numerical variables.

7.2 Important results

1. The research presents a generalizable dynamic predictive model, to limit students expected to graduate to a critical level from early study stages.

2. The possibility of preprocessing the datasets in an automatic way through the model, which made it more flexible and reliable.

3. The process of automatic selection using the features of the variables that most affect the expected outcome of the model gave the model a dynamic and generalizable characteristic and made us able to benefit from more in predicting the results of students in other colleges.

4. The model can be used to discover the relationships between the educational courses and the influence on each other.

5. The use of the model's dynamic feature, made it easier to deal with the different inputs of the model, which change from one batch to another, according change of the surrounding factors affecting academic achievement.

6. The research presented a set of conclusions that serve as a nucleus for designing a system that can be generalized to predict all students' results in different fields.

7. The ability to predict student performance based on the academic data available in his record.

7.3 Recommendations and Future Work

An important future work focus on developing the model. This study was focused to make the model dynamic, based on this approach, we can investigate the possibility of generalizing this approach, which can predict the performance of graduation in any other university studies

programs at the same university or other universities. Thus, giving the university another way to improve educational outcomes.

In the future, we seek to expand the scope of the implementation of the model to include all Sudanese universities in order to see the full picture and help the Ministry of Higher Education to take strategic decisions to improve academic reality in Sudan. For the significant benefit of this research and its development, we recommend the following:

1. It is also useful to build a software tool for data analysis integrated with the electronic systems of universities so that teachers can easily analyze the performance of their students and reduce the need for data analysis experts as possible.

2. Re-apply the model in parallel with the student's academic progress and psychological state, meaning to re-apply the model at the end of the second year to reach more accurate expectations and discover factors that may be found in the student's level upon graduation and trying to address them.

3. Improving the Sudan University of Science and Technology database by building a model data warehouse based on modern foundations.

4. There are several aspects that this research did not address, due to the lack of the required resources, or the lack of them within the missing data for students in the university database, which may be effective in infer more patterns and relationships such as place of birth, health and social status.

Reference

1. *Abu Tair, M. M.-H. (2012). Mining educational data to improve students' performance: a case study. Mining educational data to improve students' performance: a case study, 2.*
2. *Al-shargabi, A. A. (2010). Discovering vital patterns from UST student's data by applying data mining techniques. In 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE) (pp. 547--551). IEEE.*
3. *Asif, R. a. (2017). Analyzing undergraduate students' performance using educational data mining. Computers \& Education, 113, 177--194.*
4. *Asif, R. a. (2017). Analyzing undergraduate students' performance using educational data mining. Computers \& Education, 177--194.*
5. *Baradwaj, B. K. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.*
6. *Borkar, S. a. (2013). Predicting students' academic performance using education data mining. International Journal of Computer Science and Mobile Computing, 2, 273--279.*
7. *Brief, E. t. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In T. E. Mining, Proceedings of conference on advanced technology for education (pp. 1-46).*
8. *Costa, E. B. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247--256.*
9. *Donlevy, J. (2005). Envisioning the future: The US Department of Education's national technology plan. International Journal of Instructional Media, 107-110.*
10. *Fujita, H. a. (2019). Neural-fuzzy with representative sets for prediction of student performance. Applied Intelligence, 49, 172--187.*
11. *Golding, P. a. (2006). Predicting academic performance. In Proceedings. Frontiers in Education. 36th Annual Conference (pp. 21--26). IEEE.*
12. *Han, J. a. (2011). Data mining: concepts and techniques. Elsevier.*
13. *Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13, 61--72.*
14. *Kabakchieva, D. (n.d.). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13, 61--72.*
15. *Kapur, B. a. (n.d.). Comparative study on marks prediction using data mining and classification algorithms. International Journal of Advanced Research in Computer Science, 8, 2017.*
16. *Kularbphettong, K. a. (2012). Mining educational data to analyze the student motivation behavior. World Acad. Sci. Eng. Technol, 6, 1036--1040.*
17. *Kushwah, S. P. (2012). Analysis and comparison of efficient techniques of clustering algorithms in data mining. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 1, 2278--3075.*
18. *Liao, S. N. (2019). A robust machine learning technique to predict low-performing students. ACM Transactions on Computing Education (TOCE), 19, 1--19.*

19. *Moscoso-Zea, O. a.-M. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. Australasian Journal of Engineering Education, 24, 4--13.*
20. *Moscoso-Zea, O. a.-M. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. Australasian Journal of Engineering Education, 24, 4--13.*
21. *Mythili, M. a. (2014). An Analysis of students' performance using classification algorithms. IOSR, Journal of Computer Engineering, 16, 1.*
22. *Nghe, N. T. (2007). A comparative analysis of techniques for predicting academic performance. In 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports (pp. TG2--7). IEEE.*
23. *Nurafifah, M. S.-R. (2019). Review on predicting students' graduation time using machine learning algorithms. International Journal of Modern Education and Computer Science, 11, 1.*
24. *Omolewa, O. T. (2019). Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions. Journal of Engineering and Applied Sciences, 14, 8254--8260.*
25. *Oskouei, R. J. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies. Computer Engineering and Applications Journal, 3, 79-88.*
26. *Pal, S. a. (2017). Performance analysis of students consuming alcohol using data mining techniques. International Journal of Advance Research in Science and Engineering, 6, 238--250.*
27. *Patil, P. (2017). Predicting instructor performance using na $\{$$\backslash$"$\backslash$i$\}$ ve bayes classification algorithm in data mining technique: A survey. International Journal of Advanced Electronics and Communication Systems, 6, 1.*
28. *Padhy, N. a. (2012). he survey of data mining applications and feature scope. arXiv preprint arXiv:1211.5723.*
29. *Ramaswami, M. (2014). Validating predictive performance of classifier models for multiclass problem in educational data mining. International Journal of Computer Science Issues (IJCSI, 11, 86.*
30. *Romero, C. a. (n.d.). Data mining in course management systems: Moodle case study and tutorial. Computers \& Education, 51, 368--384.*
31. *Salal, Y. a. (2019). Educational Data Mining: Student Performance Prediction in Academic. IJ of Engineering and Advanced Tech, 8, 54--59.*
32. *Silva, C. a. (2017). Educational Data Mining: a literature review. In Europe and MENA Cooperation Advances in Information and Communication Technologies (pp. 87--94). Springer.*
33. *talend.com. (2020, 9 12). Retrieved from talend.com: https://www.talend.co.*
34. Yadav, S. K. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv: 1203.3832.
35. *Yehuala, M. A. (2015). Application of data mining techniques for student success and failure prediction (The case of debre_Markos university. International Journal of Scientific \& Technology Research, 4, 91--94.*

36. *Zimmermann, J. a. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. Journal of Educational Data Mining, 7, 151--176.*
37. *Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education, 16, 27.*
38. *Fox, J. (2019). Regression diagnostics: An introduction (Vol. 79). Sage Publications.*
39. *Gon{\c{c}}alves, D. B. (2019). Machine learning in analytical chemistry: applying innovative data analysis methods using chromatographic techniques. gonccalves2019machine.*
40. *Hamza, H. A. (2018). A review of educational data mining tools \& techniques. International Journal of Educational Technology and Learning, 3, 17--23.*
41. *Jain, P. a. (2019). INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES \& RESEARCH TECHNOLOGY A SURVEY OF ISSUES AND CHALLENGES OF DEVELOPING SMART DEVICES USING MACHINE LEARNING ALGORITHMS. jaininternational,, 1.*
42. *Kotsiantis, S. B. (2017). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3--24.*
43. *Nurafifah, M. S.-R. (2019). Review on predicting students' graduation time using machine learning algorithms. International Journal of Modern Education and Computer Science, 11, 1.*
44. *Rouse, M. (2018). Machine learning (ML). WhatIs. com,[Online]. Available: https://searchenterpriseai. techtarget. com/definition/machine-learning-ML.[Accessed 3 May 2019.*
45. *Smolinska, A. a.-C. (2014). Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis. Journal of breath research, 8, 027105.*
46. *talend.com. (2020, 9 12). Retrieved from talend.com: https://www.talend.com*
47. *Zou, K. H. (2003). Correlation and simple linear regression. Radiology, 227, 617--628.*

APPENDIX A

Preparation for building the model

This appendix is a code program used to preparation the predictive model

In [855]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_regression
import matplotlib.pyplot as plt
from sklearn.feature_selection import mutual_info_regression
from sklearn.linear_model import LinearRegression
```

In [930]:
```
dataset = pd.read_csv('dynamic5.csv')
dataset.head()
```

Out[930]:

	MATH	CHEMISTRY	PHYSICS	F.grade	S.grade	H.school	Final
0	82.0	78.0	84.0	2.60	2.50	84.9	2.51
1	88.0	75.0	73.0	2.62	2.64	79.4	2.63
2	83.0	76.0	80.0	2.91	2.79	84.9	2.79
3	79.0	78.0	80.0	2.55	2.32	82.7	2.42
4	77.0	77.0	80.0	2.42	2.46	83.4	2.33

APPENDEIX B

Create the predictive model

This appendix is a code program used to create the dynamic predictive model

In [1049]:
```
# feature selection
f_selector = SelectKBest(score_func=f_regression, k=3)
# learn relationship from training data
sfs1 = f_selector.fit(X_train, y_train)
# transform train input data
X_train_fs = f_selector.transform(X_train)
df = pd.DataFrame(X_train_fs,columns=['F-grade', 'S-grade', 'H-school'],dtype=float)
df
```

Out[1049]:

	F-grade	S-grade	H-school
0	2.49	2.47	82.7
1	2.39	2.49	78.0
2	3.01	2.82	84.6
3	2.78	2.68	78.0
4	3.13	3.20	83.1
...	...	...	...
217	2.57	2.68	82.1
218	2.70	2.76	84.6
219	3.12	3.14	85.1
220	2.91	2.76	87.6
221	2.04	2.47	77.0

222 rows × 3 columns

APPENDEIX C

Testing the predictive model

This appendix is a code program used to test the predictive model

In [1052]:
```
Linear_regressorc = LinearRegression()
Linear_regressor.fit(X_train_fs, y_train)
y_pred = Linear_regressor.predict(X_test_fs)
dY = np.float16(y_pred)
dY = np.round(y_pred,2)
difference = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': dY})
difference
```

Out[1052]:

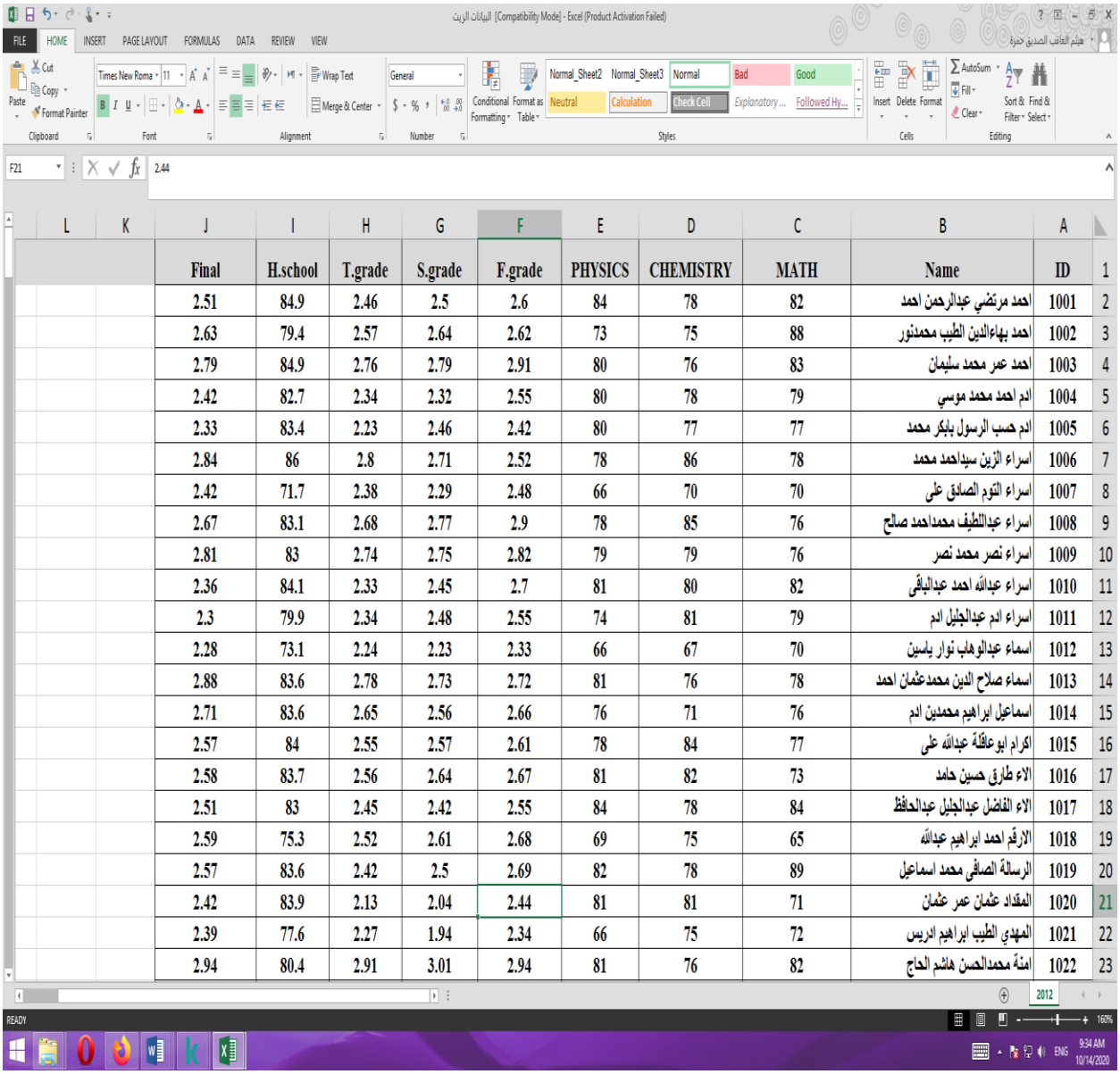
	Actual Value	Predicted Value
250	2.67	2.60
256	2.85	2.85
15	2.58	2.72
65	2.86	2.93
213	2.40	2.41
...	...	...
216	2.71	2.57
29	2.81	2.61
97	3.11	3.09
20	2.39	2.19
46	2.33	2.21

96 rows × 2 columns

APPENDEIX D

Sample of collection data

This appendix is a sample of data which collected for building the model

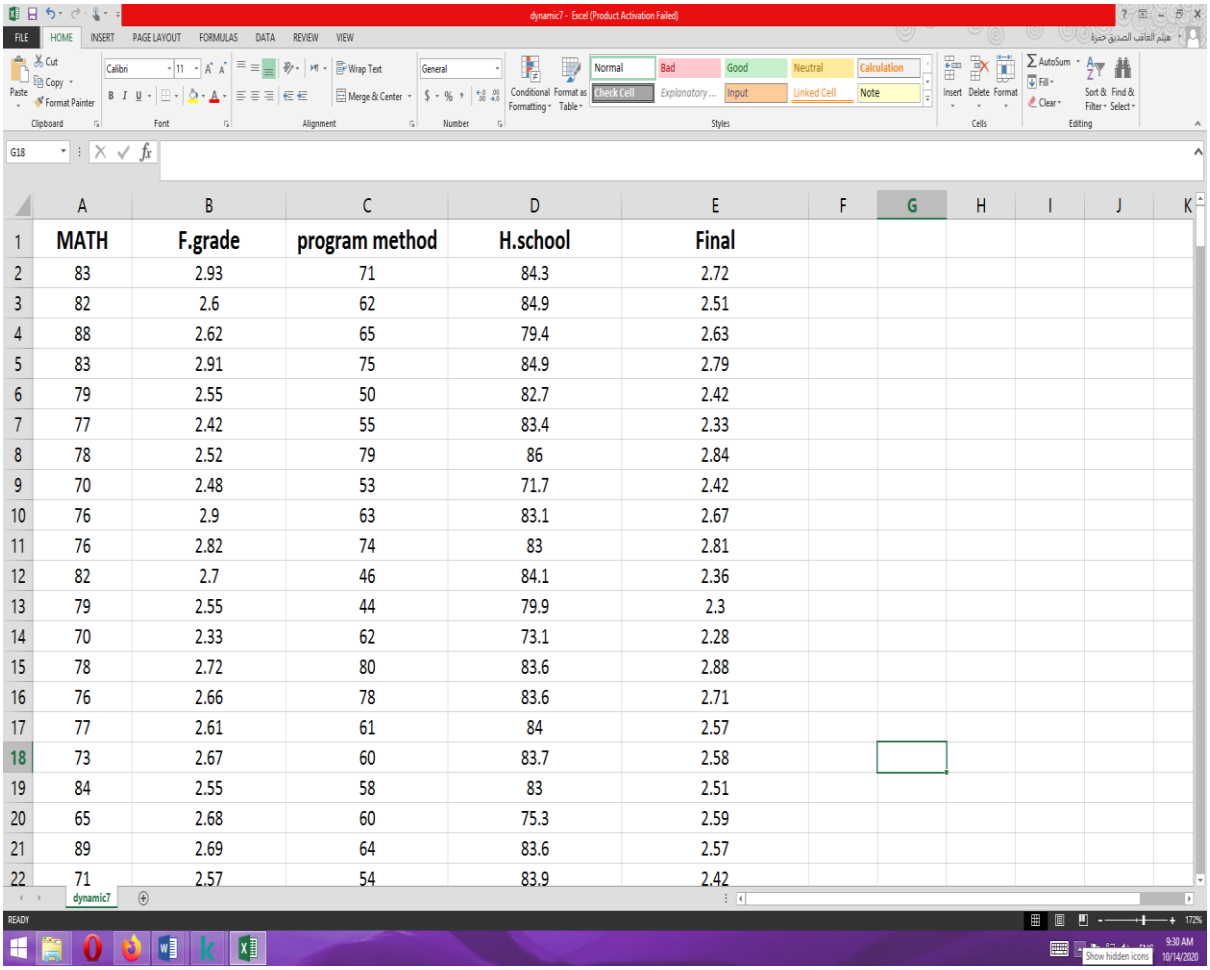


ID	Name	MATH	CHEMISTRY	PHYSICS	F.grade	S.grade	T.grade	H.school	Final
1001	احمد مرتضي عبدالرحمن احمد	82	78	84	2.6	2.5	2.46	84.9	2.51
1002	احمد بهاءالدين الطيب محمدنور	88	75	73	2.62	2.64	2.57	79.4	2.63
1003	احمد عمر محمد سليمان	83	76	80	2.91	2.79	2.76	84.9	2.79
1004	ادم احمد محمد موسي	79	78	80	2.55	2.32	2.34	82.7	2.42
1005	ادم حسب الرسول بابكر محمد	77	77	80	2.42	2.46	2.23	83.4	2.33
1006	اسراء الزين سيداحمد محمد	78	86	78	2.52	2.71	2.8	86	2.84
1007	اسراء التوم الصادق علي	70	70	66	2.48	2.29	2.38	71.7	2.42
1008	اسراء عبداللطيف محمداحمد صالح	76	85	78	2.9	2.77	2.68	83.1	2.67
1009	اسراء نصر محمد نصر	76	79	79	2.82	2.75	2.74	83	2.81
1010	اسراء عبدالله احمد عبدالباقي	82	80	81	2.7	2.45	2.33	84.1	2.36
1011	اسراء ادم عبدالجليل ادم	79	81	74	2.55	2.48	2.34	79.9	2.3
1012	اسماء عبدالوهاب نوار ياسين	70	67	66	2.33	2.23	2.24	73.1	2.28
1013	اسماء صلاح الدين محمدعثمان احمد	78	76	81	2.72	2.73	2.78	83.6	2.88
1014	اسماعيل ابراهيم محمدين ادم	76	71	76	2.66	2.56	2.65	83.6	2.71
1015	اكرام ابوعاقلة عبدالله علي	77	84	78	2.61	2.57	2.55	84	2.57
1016	الاء طارق حسين حامد	73	82	81	2.67	2.64	2.56	83.7	2.58
1017	الاء الفاضل عبدالجليل عبدالحافظ	84	78	84	2.55	2.42	2.45	83	2.51
1018	الارقم احمد ابراهيم عبدالله	65	75	69	2.68	2.61	2.52	75.3	2.59
1019	الرسالة الصافي محمد اسماعيل	89	78	82	2.69	2.5	2.42	83.6	2.57
1020	المقداد عثمان عمر عثمان	71	81	81	2.44	2.04	2.13	83.9	2.42
1021	المهدي الطيب ابراهيم ادريس	72	75	66	2.34	1.94	2.27	77.6	2.39
1022	امنة محمدالحسن هاشم الحاج	82	76	81	2.94	3.01	2.91	80.4	2.94

APPENDEIX E

Sample of dataset

This appendix is a sample of Candidate data which selected to be as input of the model



	A	B	C	D	E
1	MATH	F.grade	program method	H.school	Final
2	83	2.93	71	84.3	2.72
3	82	2.6	62	84.9	2.51
4	88	2.62	65	79.4	2.63
5	83	2.91	75	84.9	2.79
6	79	2.55	50	82.7	2.42
7	77	2.42	55	83.4	2.33
8	78	2.52	79	86	2.84
9	70	2.48	53	71.7	2.42
10	76	2.9	63	83.1	2.67
11	76	2.82	74	83	2.81
12	82	2.7	46	84.1	2.36
13	79	2.55	44	79.9	2.3
14	70	2.33	62	73.1	2.28
15	78	2.72	80	83.6	2.88
16	76	2.66	78	83.6	2.71
17	77	2.61	61	84	2.57
18	73	2.67	60	83.7	2.58
19	84	2.55	58	83	2.51
20	65	2.68	60	75.3	2.59
21	89	2.69	64	83.6	2.57
22	71	2.57	54	83.9	2.42