SELINUS UNIVERSITY
OF SCIENCES AND LITERATURE

Exploration of Data Science Techniques in Predicting Students' Relevant Courses of Study on Getting to Higher Institutions

By
Ismail Olaniyi MURAINA

A DISSERTATION

Presented to the Department of Data Science
program at **SELINUS UNIVERSITY**

Faculty of Computer Science
in fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
IN DATA SCIENCE

2022

Declaration

I Ismail Olaniyi MURAINA---------- do hereby attest that I am the sole author of this thesis and that its contents are only the result of reading and the research I have done

Student ID: UNISE1431IT

Dedication

I am sincerely dedicated this Thesis to my lovely sister, wife and children, your patience, love, care, encouragement and support are all the source of inspiration to complete this study.

Acknowledgment

My reserved and undiluted appreciation goes to Almighty Allah for His blessings, guidance and success to complete this work in peace and sound health and for accepting my prayers. I am using this opportunity to say a big THANK YOU to all members of my family for the support and love I received from you, I as well appreciate the favorable atmosphere provided for me at SELINUS University of Sciences and Literature, Faculty of Computer Sciences for the opportunity to pursue a Doctorate in Computer Science (Data Science).

Table of Contents

List of Tables

List of Figures

Publications from the Thesis

[1] Muraina, I.O (2021). Partial Least Squares Structural Equation Path Modeling in Determining Influential Factors towards Students' Relevant Courses of Study Selection Challenges. *International Cappadocia Scientific Research Congress Proceedings held on 15-17 December, 2021* /Cappadocia-Nevschir

[2] Muraina, I. O (2021). Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts. *7Th International Mardin Artuklu Scientific Researches Conference Proceedings held 10-12 December, 2021*/Mardin-Turkey

[3] Muraina, I. O & Hamzat, I. O (2021). Academicians' Stress Management with Machine Learning Classifiers. *Istanbul International Modern Scientific Research Congress – II Proceedings held on 23 – 25 December, 2021*/ Istanbul, Turkey

[4] Muraina, I. O; Lesi, B. O; Oladapo, W; Hamzat, I. O (2021). Artificial Neural Network Model to Predict Students' Relevant Courses of Study on Getting into Higher Institutions. *International Siirt Conference on Scientific Research Proceedings held on 5-7 November, 2021*/Siirt University

[5] Muraina, I. O; Ajetunmobi, R. & Adedokun, A (2021). Accelerate Data Science and Data Analytics Optimization with Edge Computing. *4Th International African Conference on Current Studies Proceedings held on 20-22 October, 2021*/ Bani Waleed University, Libya

Abstract

Background: *Data Science is a composite of a number of pre-existing disciplines. It is a young profession and academic discipline. Its popularity has exploded since 2010, pushed by the need for teams of people to analyze the big data that corporations and governments are collecting. The basic ideas underlying the definitions of data science is used to acquire knowledge from data in some relevant fields and to provide support for existing scientific research and management decision-making schema. Data Science is seen as an 'Action Plan' for expanding the technical areas of the field of statistics. Also it was seen as a combination of statistics, computer science, and information design. Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data. In the same vein, data science is the study of where information comes from, what it represents, and how it can be turned into a valuable resource in the creation of business and IT strategies. Data science as the methods and technologies used to conduct scientific research through management and utilization of scientific data. As scientific data have become more accessible, data science has been used to better characterize the data-intensive nature of today's science and engineering. Increasingly, data scientists are also drawn from a variety of different academic and professional backgrounds, including health, information management, computer science, psychology and education. Looking it from this angle it implies that data science and its applications will only continue to grow. Theories are formulated to explain, predict and understand phenomena and in many cases to challenge and extend existing knowledge within the limit of critical bounding assumptions. It is a structure that can hold or support the theory of a research study. The theoretical frameworks underpinning this study are: Decision theory, Risk theory, data theory, probability theory and machine learning or computational learning theory. Data are so important and applicable to all aspect of human life. In data science, data helps to improve quality of life, data provides indisputable evidence to observation that might lead to wasted resources due to taking wrong and incorrect conclusion, data helps to respond to challenges before becoming full-blown crisis. Prediction is about using information you have to produce or project information you do not have. Also, prediction is getting series of information and data to filter, sequence, and sort them into insights that will facilitate decision making. Accurate prediction can potentially transform business, industry, and almost any organizational domains like education, marketing, healthcare, insurance just a few domains seeking for accurate predictions to enhance their decisions. A forecast model is one of the most common predictive analytics models. It handles metric value prediction by estimating the values of new data based on learning from historical data. It is often used to generate numerical values in historical data when there is none to be found. One of the most common predictive analytics models are classification models. These*

models work by categorizing information based on historical data. While classification and forecast models work with historical data, the outlier model works with anomalous data entries within a dataset. The time series model focuses on data where time is the input parameter. The clustering model takes data and sorts it into different groups based on common attributes. Artificial Intelligence (AI) is all about how to make the system as intelligent as human beings. The intelligent systems in question are seen to be conceivable by incorporating the machines (computers) with learning, processing and decision making abilities. Much multi-criteria decision aid software have been developed for the past 30 years. Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The learning ability of the students will be improved by knowing their data and skills they possess, the course of study of students can be predicted using different models and techniques. Thus, there is the need to use data science techniques to predict students' course of study on getting to higher level of their education. By applying various data science techniques (ML & AI) in predicting students course of study in science is a key approach to utilizing large volumes of available WAEC and NECO grades for extracting knowledge of students in the higher institution.

Purpose: *The goal of this research is to provide the novel framework based on data mining, cleaning, and classification for predicting students' course of study in sciences on getting to higher institution. Furthermore, the primary objective of this research is to show the exploration of data science techniques in predicting students' relevant course of study. The effort is made to review the relevant literature articles whose research works are concentrated for making decision in the educational sector. Finally, the focus was also to determine the areas that required more attention to data mining, data selection and data prediction techniques along with machine learning techniques and the use of some applications for prediction and decision making*

Methods and Materials: *The study employed mixed method design approach; with all graduate students of high/secondary schools formed the population. WAEC and NECO grades dataset were used to categorize the content with respect to the student data informatics and classify the data for analyzing their classes, to improve the data prediction strength in course of study by evaluating WAEC/NECO grades with the use of data science (supervised and/or unsupervised machine learning) algorithms, to validate the information with the use of classification algorithms, to evaluate the performance of proposed approach by evaluating series of parameters (Precision, recall, F-measure, the accuracy with classification rate), to compare the performance of different classifiers on WAEC and NECO datasets, and to predict relevant course of study for student using apps, regression algorithm, decision trees*

algorithm and neural network algorithm. The data analysis and interpretation were carried out in three phases. The first phase focused on the analysis of data which sub-divided into three parts – Dataset analysis using data science techniques (Machine Learning Models), Questionnaire analysis using quantitative methods (SmartPLS) and Interview analysis using qualitative approach (Thematic method). The second phase was on presentation of data and the last phase was based on interpretation the results of the three parts already mentioned. Hence, data was collected as a part of the research and the researcher's analysis of the data. Presenting the data collected and its analysis in comprehensive and easy to understand manner of how the research flow in order to have good analysis strength. The major concerned areas include the use of data science techniques to improve the prediction capacity of students selection of relevant courses of study by getting into higher institutions, to find out the impact or effect of personal interest, peers/friends influence, parental influence, and personal performance of students towards selection of courses with respect to moderating factors like gender and age, and qualitative analysis that showed the opinion or responses of sampled participants on how students' interest, peers/friends, parents and performances influenced their selection of courses and the challenges facing them in carrying out such actions. The section also, explained the essence of the use of different classifiers for accurate predictions with support from previous studies.

Results and Discussion: *Principal Component Analysis (PCA) was used by researcher to summarize the information content in large data tables by means of a smaller set or summary indices that could be more easily visualized or analyzed. The use of PCA was to reduce the dimensionality of datasets, to increase the interpretability and at same time with no loss of any information as we had it in the original datasets. In the use of PCA, Kaiser-Meyer-Olkin Measure of Sampling Adequacy was used and got a desirable result of 0.651(Which is greater than .6), at the same time Bartlett's Test of Sphericity was significant at .000 (Which is less than 0.01/0.05). For proper prediction model to be achieved, the dataset was partitioned into two: Train/Test ratios. This was done to avoid over-fitting of the variables used. Hence, the dataset was split into full training data; 10-fold cross-validation, 70/30, and 75/25. From the findings, all split ratios performed well under all the classifiers used such as: Naïve Bayes' Classifier, J48 Pruned Tree Classifier, Multilayer Perceptron (Neural Network) Classifier, K- Nearest Neighbours Classifier, Decision Table Classifier, Support Vector Machine (SVM) Classifier, and on Random Forest Classifier. It showed that all the classifiers used were good algorithms/models to predict students' relevant courses of study on getting into higher institution. The partial least squares path modeling or partial least squares structural equation modeling (PLS-PM; PLS-SEM) was used to analyze the questionnaire instrument with the use of SmartPLS software. This method was employed to predict unobserved variables via the input*

values of instances. The results showed that Parental Influence had direct relationship with the student's age, likewise the students' peers/friends was a strong factor that can influence the student's courses of study selection based on the age of such student at the time of selection, also student's use of personal interest to make choice had impact on the age of the student. Others did not have any significant relationship with the variables. The Q^2 vale > .5 supported the findings as well as r^2, f^2 and coefficient path plots. The results of questionnaire analyzed with PLS-SEM complemented that age of the students was a determinant to parental, peers/friends influence to the students while gender did not contribute statistically significant to the selection of courses towards entering higher institutions. The results from findings also showed that parents of the students could choose the course for them. The findings buttressed the earlier findings on how data science techniques could be used to predict students relevant of courses on getting into higher institutions.

Conclusion: *In total, the ensemble learning method was used to build the models and predict the students' relevant courses of study on getting into higher institutions. It was noted that selection of courses using machine learning algorithms or models showed high predictive accuracy of all the machine learning models used with some minor variances. It was also observed that age of the students during the time of course selection were determined from the influence of parents, peers, and other people around such students. In the same vein, personal perceptions of people towards course selection was that female students were always guided more than male counterparts but these assertions were not statistically proved right. Decision making on course of study by students is a very technical issue that needs experts which at time might be students' parents, teachers or admission officers of the institutions to avoid wrong selection of course which may result to unsatisfactory programme or drop out from school or total failure of the students in the programme.*

CHAPTER ONE

INTRODUCTION

1.1: The Perspective

Data science has received widespread attention in academic and industries worldwide. New data science research institutes and organizations have continued to emerge, from time to time, on the scene. Many universities have launched data science courses and degree programs likewise other companies have established employment positions for data scientists. According to (Thomas & Patil, 2012) the data scientist is "the sexiest job of the 21st century." Currently, there are several viewpoints regarding the definition of data science. However, there is no consensus definition. Researchers believe that, as a new science, the research objectives of data science are different from those of other, more established branches of science. There are several current viewpoints on data science.

1.1.1: Data Science from Scientific Data Perspective

Data science as the methods and technologies used to conduct scientific research through management and utilization of scientific data. As scientific data have become more accessible, data science has been used to better characterize the data-intensive nature of today's science and engineering. Many disciplines use data technology to deal with scientific data from their respective areas. From this, Y-informatics emerged, including bioinformatics, neuro-informatics, and social informatics (Zhu & Xiong, 2015). Data science is the management, processing, and use of scientific data to support scientific research

1.1.2: Data Science from Business Data Perspective

The fundamental concepts of data science are by extracting knowledge from data to solve business problems (Provost & Fawcett, 2013). From this point of view, the acquisition of knowledge from business data in order to make decisions is one aspect of data science. This is similar to what BI scientists work on. For this reason, many BI scientists are also called data scientists. However, compared to BI issues, data science focuses more on common issues in the analysis of various business data, i.e., the issues on BI methodology (Zhu & Xiong, 2015).

1.1.3: Data Science as a Science of Statistics, Computing Technology, and Artificial Intelligence (AI).

This viewpoint often comes up in discussions on what data scientists are. It is generally believed that data scientists should have skills in statistics, computing technology, AI, and related fields and that data scientists are not individual people specializing in one field so much as teams consisting of statisticians, computer scientists, AI experts, and domain experts. This viewpoint is simple: Because statistics, computing technology, and AI are all used to process and analyze data, they are all a natural part of data science.

1.1.4: Data Science from Integrated Perspectives

Data science is defined as the study of the generalizable extraction of knowledge from data. The definition further pointed out that a data scientist needs to have comprehensive skills covering statistics, machine learning, AI, and database management and have a deep understanding of problem design (Dhar, 2013). This viewpoint can be seen as an integration of the first three perspectives

1.1.5: What is Data Science?

Data Science is a composite of a number of pre-existing disciplines. It is a young profession and academic discipline. Its popularity has exploded since 2010, pushed by the need for teams of people to analyze the big data that corporations and governments are collecting. The basic ideas underlying the definitions of data science is used to acquire knowledge from data in some relevant fields and to provide support for existing scientific research and management decision-making schema. However, all works described above are still not enough to establish data science as a new, unique branch of science (Zhu & Xiong, 2015). Data Science is seen as an Action Plan for Expanding the Technical Areas of the Field of Statistics (Cleveland, 2001). Since this definition was given, the concept of Data Science has been further developed and has been increasingly linked to data analytics and big data. Data Science is defined as not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data (Hayashi, 1998). Also it was seen as a combination of statistics, computer science, and information design (Shum, et al., 2013). Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data (Provost & Fawcett, 2013). In the same vein, data science is the study of where information comes from, what it represents, and how it can be turned into a valuable resource in the creation of business and IT strategies (Banafa, 2014). Donoho (2015)

said that data science is the coupling of scientific discovery and practice which involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications. Data science is now widely accepted as the fourth mode of scientific discovery, on par with theory, physical experimentation and computational analysis. Techniques based on Big Data are showing promise not only in scientific research, but also in education, health, policy, and business (MIDAS, 2017). Data science according to SimpliLearn (2021) is the same as data driven science which is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms either structured or unstructured similar to data mining. Hence, data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology (Cao, 2017).

1.2: Background of Data Science

1.2.1: Historical Background of Data Science

Data Science has revolutionized several aspects of human life, let start looking at it from the time it was first mentioned and where it was from.

Table 1: Summary of Historical Background of Data Science

YEAR	SCHOLARS	CONTRIBUTIONS
1962	John W. Tukey	*He wrote "The Future of Data Analysis" and he was the first to introduce the term "bit" as a contraction of "binary digit*
1974	Peter Naur	*He published the Concise Survey of Computer Methods, which surveyed data processing methods across a wide variety of applications. The term "data science" becomes clearer*
1977	IASC	*The International Association for Statistical Computing was founded*
1089	Gregory Piatetsky-Shapiro	*He organized and chaired the first Knowledge Discovery in Databases (KDD) workshop.*
1994	BusinessWeek	*"Database Marketing." Was published*
1996	IFCS	*"Data science" was included in the title of the conference ("Data science, classification, and related methods")*
	Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth	*Published "From Data Mining to Knowledge Discovery in Databases."*

1997	Jeff Wu	*In his inaugural lecture called for statistics to be renamed "data science" and statisticians to be renamed "data scientists."* *The journal Data Mining and Knowledge Discovery was launched*
1999	Jacob Zahavi	*Was quoted in "Mining Data for Nuggets of Knowledge"*
2001	William S. Cleveland	*He published "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. he proposed new discipline in the context of computer science and the contemporary work in data mining*
	Leo Breiman	*He published "Statistical Modeling: The Two Cultures" (pdf): The first culture ➔ data are generated by a given stochastic data model. The second culture ➔data uses algorithmic models and treats the data mechanism as unknown.*
2002	CODATA of the International Council for Science (ICSU)	*Data Science Journal was launched*, the focus of the journal was publishing papers on "the management of data and databases in Science and Technology
2003	Journal of Data Science	*Journal of Data Science was launched which assumed to* provide a platform for all data workers to present their views and exchange ideas
2005	Thomas H. Davenport, Don Cohen, and A. l Jacobson	*They published "Competing on Analytics"*
	The National Science Board	*Published "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century."*
2007	The Research Center for Dataology and Data Science	*The center was established at Fudan University, Shanghai, China*
2008	JISC	*Published the final report of a study it commissioned to "examine and make recommendations on the role and career development of data scientists and the associated supply of specialist data curation skills to the research community*
2009	Yangyong Zhu and Yun Xiong	*They published "Introduction to Dataology and Data Science," in which they stated "Different from natural science and social science, Dataology and Data Science*
	Committee on Science of the National Science and Technology Council	*Harnessing the Power of Digital Data for Science and Society was published*
	Hal Varian	*Said that "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"*
	Kirk D. Borne and his team	*They submitted to the Astro2010 Decadal Survey a paper titled "The Revolution in Astronomy Education: Data Science for the Masses "*
	Mike Driscoll	*Wrote in "The Three Sexy Skills of Data Geeks", that "with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity"*
	Nathan Yau	*Wrote in "Rise of the Data Scientist": that "As we've all read by now, Google's chief economist Hal Varian commented in January that the*

		next sexy job in the next 10 years would be statisticians.
	Troy Sadkowsky	*He created the data scientists group on LinkedIn as a companion to his website, datasceintists.com (which later became datascientists.net).*
2010	Kenneth Cukier	*Wrote in "*The Economist *Special Report Data, Data Everywhere": that a new kind of professionals has come the data scientist do not need to combine series of skills again*
	Mike Loukides	*Wrote in "What is Data Science?" that "Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution.*
	Hilary Mason and Chris Wiggins	*Wrote in "A Taxonomy of Data Science" that "…we thought it would be useful to propose one possible taxonomy… of what a data scientist does, in roughly chronological order: Obtain, Scrub, Explore, Model, and interpret*
	Drew Conway	*Wrote in "The Data Science Venn Diagram" that "…one needs to learn a lot as they aspire to become a fully competent data scientist. He presented the Data Science Venn Diagram… hacking skills, math and stats knowledge, and substantive expertise."*
2011	Pete Warden	*Wrote in "Why the term 'data science' is flawed but useful" that "There is no widely accepted boundary for what's inside and outside of data science's scope. Is it just a faddish rebranding of statistics?*
	David Smith	*Wrote in "'Data Science': What's in a name?" that "The terms 'Data Science' and 'Data Scientist' have only been in common usage for a little over a year, but they've really taken off since then: many companies are now hiring for 'data scientists', and entire conferences are run under the name of 'data science'*
	Matthew J. Graham	*Talked at the Astrostatistics and Data Mining in Large Astronomical Databases workshop about "The Art of Data Science". He said that "To flourish in the new data-intensive environment of 21st century science, we need to evolve new skills… We need to understand what rules [data] obeys how it is symbolized and communicated and what its relationship to physical space and time is."*
	Harlan Harris	*Wrote in "Data Science, Moore's Law, and Money ball" that "'Data Science' is defined as what 'Data Scientists' do. What Data Scientists do have been very well covered, and it runs the gamut from data collection and munging, through application of statistics and machine learning and related techniques, to interpretation, communication, and visualization of the results"*
	D.J. Patil	*Wrote in "Building Data Science Teams" that "meeting was the start of data science as a distinct professional specialization.... we realized that as our organizations grew, we both had to figure out what to call the people on our teams. 'Business analyst' seemed too limiting. 'Data analyst' was a contender; many of the people on our teams had deep engineering expertise. 'Research scientist' was a reasonable job title used by companies like Sun, HP, Xerox, Yahoo, and IBM. However, The term that seemed to fit best was data scientist: those who use both data and science to create something new. "*
2012	Tom Davenport and D.J.	*They published "Data Scientist: The Sexiest Job of the 21st Century" in the* Harvard Business Review. *They also exemplified the modern data*

	Patil	*scientist—that is, one who applies his or her data-savvy expertise in any setting that demands it, including healthcare, e-commerce, social media, and journalism—just to name a few*
… 2021		*The average age of data scientists in 2018 was 30.5, the median was lower. The younger half of data scientists was just entering college in the 2000s, just when all that funding was hitting academia. . There were also quite a few other opinions about what data science actually was. Everybody wanted to be on the bandwagon that was sexy, prestigious, and lucrative. Data science is one of the most high ranked and recognized professions world wide*

1.2.2: Data Science Continues to Grow

Over the last six years, data science has continued to evolve and permeate nearly every industry generates or relies on data. In a 2010 article published in The Economist, Kenneth Cukier says data scientists "combine the skills of software programmer, statistician, and storyteller/artist to extract the nuggets of gold hidden under mountains of data."

Today, data scientists are invaluable to any company in which they work, and employers are willing to pay top dollar to hire them. Also, data science degree programs have emerged to train the next generation of data scientists.

Increasingly, data scientists are also drawn from a variety of different academic and professional backgrounds, including health, information management, computer science, psychology and education. Looking it from this angle it implies that data science and its applications will only continue to grow. That's because big data will become even bigger. For instance, 97 percent of Americans now own a cell phone of some kind, according to the Pew Research Center (PRC, 2021). Nearly eight in ten U.S. adults own desktop or laptop computers, while roughly half now own tablet computers and around one in five own e-reader devices. In addition, 78 percent of healthcare consumers wear—or are willing to wear—technology to track their lifestyle and/or vital signs (Accenture, 2016).

1.2.3: What do we learn from the past history of Data science?

Data science history so far teaches many lessons:

1. Data should not be taken for granted. In the history, data was not as easy to access as we do have it today; also people are not willing to share it freely, as it is now. This doesn't negate the fact that privacy and other ethical concerns remain, and data scientists must know how to operate within an ethical framework as the data grows. And even though data is more accessible, much of it remains unstructured, paving the way for new methods of analyses.

2. Thinking big. Big data requires big analyses, and as technology evolves, data scientists must evolve their high-performance computing skills as well. This includes the ability to perform complex data mining and predictive analytics.
3. Knowing the context. Unlike in the past when data scientists worked primarily in the information technology sector, today's data scientists work in a variety of industries, helping organizations make data-driven decisions that change the way in which they compete in the larger marketplace. To be successful, data scientists must be well-versed in data communication and strategic decision-making.
4. Belonging to scientists. In the past, there was a belief that data manipulations of any kind should be done by only scientist. Today, data science has opened doors for every profession to utilize for different purposes with no prior knowledge of statistics and programming.

1.3: Problem Specification

Data science covers preparing data for analysis, which includes data cleaning, data processing (Data aggregation), data modeling (Data manipulation), data visualization and data presentation techniques to perform advanced data analysis. Data science is a domain of study that deals with vast volume of data using modern tools and techniques to find unseen patterns, drive meaningful information and make good decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis in data science can be from multiple sources and present in various formats. Data science or data-driven science enables better decision making, predictive analysis, and pattern discovery. It makes one finding the leading course of a problem by asking the right questions, performing exploratory study on the data, modeling the data using various algorithms, communicating and visualizing the result using graphs, dashboards and many more. Data science plays a decisive role in the development of all activities involved in education sectors. The learning ability of the students will be improved by knowing their data and skills they possess, the course of study of students can be predicted using different models and techniques. Thus, there is the need to use data science techniques to predict students' course of study on getting to higher level of their education and probably their future job placement after graduation.

1.3.1: Statement of the problem

Studies of this kind in the past had suffered enough robust claimed data and relevant literatures in other to make choice or predict a successful outcome in education domain. By applying various data science techniques (ML & AI) in predicting students course of study in science is a key approach to utilizing large volumes of available WAEC and NECO grades for extracting knowledge of students in the higher institution. The goal of this research is to provide the novel framework based on data mining, cleaning, clustering (grouping) and classification for predicting students' course of study in sciences on getting into higher institution and possibly their future job placement after graduation.

1.3.2: Motivations and Objectives of the Research

The primary objective of this research was to show the exploration of data science techniques in predicting students' relevant course of study on getting to higher institutions. The effort is made to review the relevant literature articles whose research works are concentrated for making decision in the educational sector. The key points (WAEC and NECO grades dataset, courses of studies (Especially in sciences), methodology, results, accuracy, and proper prediction) of the different research works along with the use of tools or techniques was highlighted. Finally, the focus was to determine the areas that require more attention to data mining, data selection and data prediction techniques along with machine learning techniques (Data science techniques) and the use of some applications for prediction and decision making. Therefore, the research objectives are:

- To use appropriate applications and data science techniques to predict students course of study using WAEC and NECO grades dataset.
- To categorize the content with respect to the student data informatics and cluster the data for analyzing their data pattern.
- To improve the data prediction strength in course of study by evaluating WAECand NECO grades information with the use of data science (supervised and unsupervised machine learning) algorithms.
- To validate the information with the use of classification algorithms
- To evaluate the performance of proposed approach by evaluating series of parameters (Precision, recall, F-measure, the accuracy with classification rate).
- To compare the performance of different classifiers algorithms on WAEC and NECO datasets.

- To predict relevant course of study for student using apps, Partial Least Squares Structural Equation Modeling (PLS-SEM), and different classification algorithms such as Support Vector Machine algorithm, decision trees algorithm and neural network algorithm and many more.

1.3.3: Assumptions and Databases

The assumptions and databases considered for this research work are as follow:

- WAEC grades are considered potent for predicting the course of study of students on getting to higher institution taken from institution admission department database
- NECO grades are also considered useful to predict students' course of study and possibly for predicting future job placement of students after graduation

1.3.4: Gaps identified

Evidence from different reviewed literatures like articles, conferences proceedings and books, text books and several posts on magazines online and social media, it was observed that much work of data science and its applications were focused on series of domains such as industries, transportations, health sectors, and businesses, this study focused on educational decision making problems. However, there is still plenty of work needed in conducting researches on exploration of data science techniques for educational predictions in order to exploit all their potentials and usefulness. In the future, it is required that more attention should be paid to the datasets for grade datasets and prediction using the incremental machine learning approaches. Hence, the need to evaluate this method on additional datasets and in particular on large datasets to show the effectiveness of the method for computation time of large data. In addition, investigation on how this method can be extended to be applicable to other types of datasets in educational domain.

1.3.5: Organization of the Thesis

Chapter 1: This part of the study is called Introduction which presents the research perspective, background and problem specifications.

Chapter 2: It is a literature review part of the research study that describes the concepts, definitions, procedures of data science, theories, models, algorithms and analytics of the previous studies along with the previous studies related to predicting student course of study.

Chapter 3: This chapter is a research methodology part of the study which covers the detailed description of the research design, approach, population and sample size, collection, and analysis.

Chapter 4: This is experimental results and analysis section of the study which discusses the analysis and validation of the collected data using different approaches.

Chapter 5: This chapter centers to results, discussions, findings and conclusion. At last, the part also points to direction of the future studies in form of recommendations.

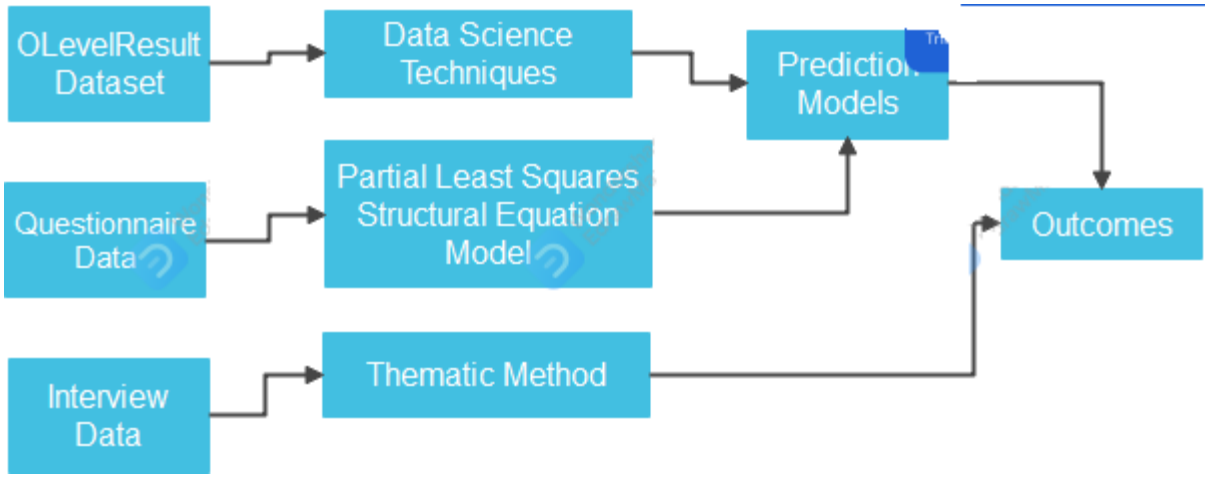

Figure 1: Conceptual framework for the study

CHAPTER TWO

LITERATURE REVIEW

2.1: Introduction

An equitable and flexible mechanism for assigning courses of study in the university is a major concern for many students coming from secondary school into university. The student course of study has great impact on student life, family, neighborhood, distance, transportation cost and eventually future jobs placement. Facing this intricate multi-dimensional optimization problem, the choice of courses of study usually utilize to stable-matching techniques which only produce stable matching that do not incorporate these different objectives; this can be expensive and inequitable. Students all over the world are usually faced with the task of career decision-making (Moleke, 2004). The choice of careers, subjects, and courses of study in schools and of subsequent paths to follow, are always difficult problems facing prospective undergraduates. Careers election is one of many important choices students will make in determining their future plans (Pitan & Adedeji, 2014). Often, choosing the right subject combination leading to the right profession can make the difference between enjoying and detesting the career in future. Dedicating oneself to career choices that are unattainable leads to frustration. Another major consequence of wrong choice of course among university graduates is unemployment (Pitan, 2010). Moleke (2004), reported that choosing the wrong course could be pitiable in which graduates were asked if they would choose the same or different course of study if they were to start again. Good high-school grades do not necessarily translated into good selection of relevant course of study in college or university. There are many factors influencing students' choice of course of study such as parental influence, ignorance of the students, lack of having access to information on time, the extent to which teachers/counselors encourage the students, motivate them and enhance their expectations for continuing education. The study of Geiser & Santelices (2007) finds that high-school grade point average (HSGPA) is consistently the best predictor not only of freshman grades in college, the outcome indicator most often employed in predictive-validity studies, but of four-year college outcomes as well. A previous study, also demonstrated that HSGPA in college-preparatory courses was the best predictor of freshman grades for a sample of almost 80,000 students admitted to the University of California

2.2: Theoretical Framework

Theories are formulated to explain, predict and understand phenomena and in many cases to challenge and extend existing knowledge within the limit of critical bounding assumptions. It is a structure that can hold or support the theory of a research study. The theoretical framework introduces and describes the theories that explain while the research problem understudy exists (USCLibraries, 2021). Students make decisions throughout their college career, and decision-making is an important element in the learning process (Zocco, 2010). Decision making can be defined as "The process of making choices by identifying a decision, gathering information, and assessing alternative resolutions. Using a step-by-step decision-making process can help one make more deliberate, thoughtful decisions by organizing relevant information and defining alternatives. The theoretical frameworks underpinning this study are: Decision theory, Risk theory, data theory, probability theory and machine learning or computational learning theory.

2.2.1: Decision Theory

Decision theory is the theory about making decisions. There are many various ways to theorize about decisions. For instance the question like "What should be my course of study when I get into higher institution?" Also, another question may go this way "Will this course of study provide good job employment after graduation from the university?" Almost everything that a human being does involves decisions. Therefore, to theorize about decisions is almost the same as to theorize about human activities. However, decision theory is not quite as all-embracing as that. It focuses on only some aspects of human activity. In particular, it focuses on how we use our freedom. In the situations treated by decision theorists, there are options to choose between, and we choose in a non-random way. Our choices, in these situations, are goal-directed activities. Hence, decision theory is concerned with goal-directed behaviour in the presence of options. Decision theory according to Bradley (2014) is the study of how choices are and should be made in a variety of different contexts. A decision maker or decision making body has a number of options before them: the actions they can take or policies they can adopt. The exercise of each option is associated with a number of possible consequences, some of which are desirable from the perspective of the decision makers' goals, others are not. Which consequences will result from the exercise of an option depends on the prevailing features of the environment

2.2.2: Risk theory

This is the study of the impact of possible outcomes on the process and consequences of decisions. Students make course selection (CS) decisions with varied return. The development of

risk theory has an interesting background, spanning six centuries, and has found application in the fields of mathematics (Protter, & Yildirim, 2004, Escobar & Seco, 2008) as well as providing combinations of fields leading to a common line of research (Hansson, 2006). The essential fact is that "risk" means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomenon depending on which of the two is really present and operating.

2.2.3: Data Theory

Data theory provides abstract models for understanding the information conveyed by real world observations. This enables the development of general guidelines for determining which analytics procedures are appropriate in any given research context. Data are always the result of a creative step on the part of the researcher. The nature of the data is never automatically determined by any particular set of empirical observations

2.2.4: Probability Theory

Probability is the measure of chance of occurrence of a particular event. The basic concept of probability is widely used in the field of education due to its stochastic nature. The outcome of a random event in probability theory cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by chance. The word probability has several meanings in ordinary conversation. In the society as well as in education we often use probability assessments to plan informally or to make decisions. Formal probability theory is also an important and fundamental tool use by researchers, healthcare providers, insurance company, education and many others to make decisions on context of uncertainty. Probability stands to provide information about the likelihood that something will happen. The researcher can determine the probability of a student to graduate with distinction if such student performs excellently in some subjects while in high school. Also, researcher can as well determine the correct choice of course of study for students from his/her academic activities in the high school

2.2.5: Machine Learning Theory

Machine Learning Theory, also known as Computational Learning Theory, aims to understand the fundamental principles of learning as a computational process and combines tools from Computer Science and Statistics. Theory of computation (TOC) is a branch of Computer Science that is concerned with how problems can be solved using algorithms and how efficiently they

can be solved. Real-world computers perform computations that by nature run like mathematical models to solve problems in systematic ways. The essence of the theory of computation is to help develop mathematical and logical models that run efficiently and to the point of halting. Since all machines that implement logic apply TOC, studying TOC gives learners an insight into computer hardware and software limitations. The goals of this theory according to Blum (2021) are both to aid in the design of better automated learning methods and to understand fundamental issues in the learning process itself. Machine Learning Theory draws elements from both the Theory of Computation and Statistics and involves tasks such as:

- ✓ Creating mathematical models that capture key aspects of machine learning, in which one can analyze the inherent ease or difficulty of different types of learning problems.
- ✓ Proving guarantees for algorithms (under what conditions will they succeed, how much data and computation time is needed) and developing machine learning algorithms that provably meet desired criteria.
- ✓ Mathematically analyzing general issues, such as: "When can one be confident about predictions made from limited data?", "How much power does active participation add over passive observation for learning?", and "What kinds of methods can learn even in the presence of large quantities of distracting information?".

Machine Learning Theory is both a fundamental theory with many basic and compelling foundational questions, and a topic of practical importance that helps to advance the state of the art in software by providing mathematical frameworks for designing new machine learning algorithms. It is an exciting time for the field, as connections to many other areas are being discovered and explored, and as new machine learning applications bring new questions to be modeled and studied. It is safe to say that the potential of Machine Learning and its theory lie beyond the frontiers of our imagination (Blum, 2021)

2.2.6: Prediction Theory

The ability to accurately predict future outcomes of students' course of study is not an easy job and this has been the goal of forecasters for decades. The prediction field is rich with sophisticated methods taken from high technology disciplines. In order to increase accuracy of prediction, one must be able to apply additional information. This additional information does not have to come from observation data. In fact, it should come from knowledge of how a system operates internally, with all the inherent feedback loops, and the external factors that influence it. Specifically, it comes from knowing how factors act as leading forces that can be observed in

advance of the system's response. This implies modeling how the inertial properties of one entity may affect those of another, and how the corresponding feedback effects can produce highly nonlinear responses. Without an approach that can characterize inertial properties whose time constants are sufficiently long, there is effectively no chance of predicting future responses with useful accuracy (Cave, 2020). Furthermore, the researcher added that" a unified prediction theory, supported by advanced tools and a proper education is the only way that problems requiring multi-step prediction can be solved. The ability to accurately predict outcomes depends upon the inherent properties of the system itself. Given that these properties exist, people must be trained and armed to take full advantage of opportunities to accurately predict and control their future". Predictions can only be made when the accuracy of the prediction mechanism can be characterized in terms of historic data used to compare *a priori* predicted outcomes to the actual outcomes. A priori means that once one has seen the outcomes, any changes to the prediction mechanism will generally require re-characterization of the error using data that has not been seen

2.3: Background Study

2.3.1: Concept of Data

Merriam-Webster (2021) defined data as facts or information used usually to calculate, analyze or plan something. At a point in time, we gather facts and numbers which we examine and consider making plans or decisions. It implies that without data it is very hard to make good decisions according to Cambridge Dictionary (2021) facts or numbers are collected to be examined and considered and finally used to help making decisions. Simplilearn (2021) sees data as different types of information that usually formatted in a particular manner. To so many researchers, data, information and knowledge are interrelated in sequential order. Data are the raw material for information while information is the raw material for knowledge. Data is any information in raw or organized from using alphabets, numbers or symbols that refers to or represents preferences, ideas, traits, categories e.t.c (Mihal, 2016). Therefore, data can be seen as facts or figures, or information that is stored in or used by computer, which can be analyzed or used in an effort to gain knowledge or inferred conclusion or make decisions in form of information.

2.3.2: Why data?

Data are so important and applicable to all aspect of human life. Data helps to improve quality of life, data provides indisputable evidence to observation that might lead to wasted resources due to taking wrong and incorrect conclusion, data helps to respond to challenges before becoming full-blown crisis. Data allows organizations to measure the effectiveness of a given strategy. It serves as a good determinant of the cause of problems and proffers the solutions. Data helps to present a strong argument for system change. It helps to make solid decisions and also increases efficiency. So, data is useful to every individual. According to Import.io (2018) whichever industry you work in or whatever your interest, data is important to you.

2.3.4: Different forms of data

- ✓ **Personal Data:** Facts that are specific to one such as demographics, location, mail address e.t.c
- ✓ **Transactional Data:** Facts that require an action to collect such as click on AD, buy an item, visit a place, go to a web page e.t.c
- ✓ **Web Data:** Facts that you might pull from the internet either for research or others.
- ✓ **Sensor Data:** Facts produced by objects which often referred to as internet of things (IoTs).
- ✓ **Big Data:** Facts that have 5VS such as volume, velocity, variety, veracity and value.

2.3.5: Data Types

Generally, data can be categorized as:

[1]. **Structured Data:** This type of data adheres to a pre-defined model and therefore easy to use or analyze. It usually conforms to a tabular format with key relationship between the different attributes and instances (rows and column). Example include: MS Excel sheet, SQL Database.

[2]. **Semi-Structured Data:** In this type of data, it is partly structured and partly unstructured. It does not obey the tabular structure of the data models or tables but uses tags or other makers to differentiate between semantic elements and enforce hierarchies of record and fields with the data. Examples include: E-mail, XML, Zip files, Web pages.

[3]. **Unstructured Data:** This category of data is not arranged in any form of data models or schema and as a result, it cannot be stored in a traditional relational Database or RDBMS Examples include: Audio, Video, Images, and Photos.

2.3.6: Why Data need to be analyzed?

Data are not yet fully useful or might not yet consider as important until is/are used or refined as a new information for further usefulness. So, data need to be analyzed in order to draw helpful conclusions from it because data analyses is the process by which analytical and logical reasoning are applied on data to gain information from the data.

2.3.7: Why data need to be scientifically analyzed?

In science, there are procedures required to follow in order to arrive at a good ends. The science approaches need to be followed properly in analyzing data to make a good decision. In doing these, the data is/are required to be collected, organized and displayed in a manner that would be presented in form of tables, graphs and equations. According to Sihag (2019), data need to be analyzed in order to draw conclusions.

2.3.8: Who are data scientists?

Data scientists are analytical experts who utilize their skills in both technology and social science to find trends and manage data. They are also referred to as those who use the knowledge of statistics, mathematics, programming and domain expertise to solve societal problems. According to Bowne-Anderson (Bowne-Anderson, 2018), the use of industry knowledge, contextual understanding, and skepticism of existing assumptions are to uncover solution to various challenges.

Data scientists are big data wranglers, who are specialized in gathering and analyzing large data set both structured and unstructured and finally models data then interpret the outcome to create actionable plains for society. They combine computer science knowledge with statistic and mathematics to process and analyze data for decision making. They are responsible to discover insight from structured, semi-structured, unstructured data to assist shape or meet specific required needs or goals. Hence, a data scientist embodies the perfect combination of business knowledge, technical skills, expertise and knowledge of statistics.

2.3.9: Data Science Professional Related Careers

At present, data science is a cutting-edge field that allows one to make an important impact on society. The professional careers include:

Table 2: Careers and their Functions

Careers	Functions
Data Analysts	They collect, process and perform statistical data analysis with the goal to help making better decisions
Data Engineers	They are the designers, builders and managers of information or big data infrastructure
Business analysts	They have deep knowledge of different business processes and embody business intelligence
Marketing Analysts	They Study information to better assist organization in making, informed decisions about market opportunities
Data Architects	They create the blueprint for data management system to integrate centralized, protect and maintain data sources
Data and Analytics Managers	They Lead data science teams
Business Intelligence Analysts	They gather data in a variety of ways through software, reviewing competitors data and industry trend to develop an understanding of the direction the company should move towards
Data mining Specialists	Responsible for identifying patterns and relationship to help a company predict future behaviours
Machine Learning Engineers	A subset of AI that works with big data applications with knowledge of advance mathematics and software programming
Database Administrators	Information technology professionals that ensure the optimal storage and access to an organization data
Database Developers	They are also called database engineers or database programmers that are responsible for the design, programming, construction and implementation of new databases for platform updates and changes in user needs
Statistician	One who works with mathematical techniques to help analyze and interpret data to solve real time problem

2.4: Prediction Concept

According to dictionary definition, prediction is seen as what someone thinks will happen. A prediction is a forecast, but not only about the weather. *Pre* means "before" and *diction* has to do with talking. So a prediction is a statement about the future. It's a guess, sometimes based on facts or evidence, but not always. To statisticians, prediction is the process of determining the magnitude of statistical variates at some future point of time. APADictionaryofPsychology, (2021) defined prediction as empirical research concerned with forecasting future events or behavior: the assessment of variables at one point in time so as to predict a phenomenon assessed at a later point in time. To Janssens & Martens (2018), Prediction is the act of forecasting what is going to happen in the future. Prediction is central to many domains especially in education as students' success, progress, performance and future choice of course or job opportunity are prescribed or recommended on implicit or explicit expectations about future students outcomes.

The field of prediction research is gaining importance because of the increasing interest in precision education

2.4.1: Prediction and Decision Making

There are a number of gaps between making a prediction and making a decision (Athey, 2017). The correlation does not imply the causation. This difference is not well observed and marked in the fields of machine learning, Big Data analytics or Data Science. It is emphatically noted from different fields that the sole focus is to increase prediction accuracy with more data and ways to run existing models but less works seem to be dedicated to steps after accuracy of a prediction. Prediction is about using information you have to produce or project information you do not have. Also, prediction is getting series of information and data to filter, sequence, and sort them into insights that will facilitate decision making. According to Mosavi (2015) accurate prediction can potentially transform business, industry, and almost any organizational domains like education, marketing, healthcare, insurance just a few domains seeking for accurate predictions to enhance their decisions.

2.4.2: Predictive Modeling

Predictive modeling, also called predictive analytics, is a mathematical process that seeks to predict future events or outcomes by analyzing patterns that are likely to forecast future results. The major goal in predictive modeling is to give answer to what is most likely to happen in the future. Once data has been collected, the analyst selects and trains statistical models, using historical data. Although it may be tempting to think that big data makes predictive models more accurate, statistical theorems show that, after a certain point, feeding more data into a predictive analytics model does not improve accuracy. The old saying "All models are wrong, but some are useful" is often mentioned in terms of relying solely on predictive models to determine future action. Predictive modeling is defined as the process of taking known results and developing a strategy (model) that can predict values for new occurrences. The modeling uses historical data to predict the future event (Mitchell, 2019) . Making causal relationships between variables when applying predictive analysis technique is not encouraging, because if prediction cannot state that one variable caused another but we can state that a variable had an effect on another and what that effect should be. Therefore, predictive models are designed to assess historical data, discover patterns, observe trends and finally use the information to come up with predictions about the future trends (Selerity, 2021). Ali (2020) defined predictive modeling as a statistical technique

using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modeling can be used to predict just about anything, from subject ratings and a student's next GPA score in a semester to which student can graduate freely next session.

2.4.2.1: Types of Predictive Models

Predictive modeling are of different categories specifically design for a particular function, they include:

2.4.2.1.1: Forecast Models

A forecast model is one of the most common predictive analytics models. It handles metric value prediction by estimating the values of new data based on learning from historical data. It is often used to generate numerical values in historical data when there is none to be found. One of the greatest strengths of this model is its ability to input multiple parameters. For this reason, they are one of the most widely used predictive analytics models. They are used in different industries and business purposes. For example, a teacher can predict how many students will get grade 'A' in a course or a call centre can predict how many support calls they will get in a day or a shoe store can calculate inventory they need for the upcoming sales period using forecast analytics. Forecast models are popular because they are incredibly versatile.

2.4.2.1.2: Classification Models

One of the most common predictive analytics models are classification models. These models work by categorizing information based on historical data. Classification models are used in different industries because they can be easily retrained with new data and can provide a broad analysis for answering questions. Classification models can be used in different industries like education, finance and retail, which explains why they are so common compared to other models.

2.4.2.1.3: Outliers Models

While classification and forecast models work with historical data, the outlier model works with anomalous data entries within a dataset. As explained in literatures, anomalous data refers to data that deviates from the norm. It works by identifying unusual data, either in isolation or in relation with different categories and numbers. Outlier models are useful in industries where identifying anomalies can save organizations millions of dollars, namely in retail, finance also in education

(It can affect student's future reputations). One reason why predictive analytics models are so effective in detecting fraud or malpractices is because outlier models can be used to find anomalies. Since an incidence of fraud or malpractices is a deviation from the norm, an outlier model is more likely to predict it before it occurs. For example, when identifying a fraudulent transaction, the outlier model can assess the amount of money lost, location, purchase history, time and the nature of the purchase. In education, outliers can be used to assess plagiarized works of other scholar, where the source of the work comes from, the original owner of the work and other detailed history of the work. Outlier models are incredibly valued because of their close connection to anomaly data.

2.4.2.1.4: Time Series Model

While classification and forecast models focus on historical data, outliers focus on anomaly data. The time series model focuses on data where time is the input parameter. The time series model works by using different data points (taken from the previous year's data) to develop a numerical metric that will predict trends within a specified period. If organizations want to see how a particular variable changes over time, then they need a Time Series predictive analytics model. For example, if a teacher is interested in knowing students' performance in his/her course over three sessions or a small business owner wants to measure sales for the past four quarters, then a Time Series model is needed. A Time Series model is superior to conventional methods of calculating the progress of a variable because it can forecast for multiple regions or projects simultaneously or focus on a single region or project, depending on the organization's needs.

2.4.2.1.5: Clustering Model

The clustering model takes data and sorts it into different groups based on common attributes. The ability to divide data into different datasets based on specific attributes is particularly useful in certain applications, like marketing, education and many other domains. For instance, marketers can divide a potential customer base based on common attributes; also a teacher can divide students based on series of attributes. It works using two types of clustering – hard and soft clustering. Hard clustering categorizes each data point as belonging to a data cluster or not, while soft clustering assigns data probability when joining a cluster.

2.4.2.2: How to Apply Predictive Analytics Models in Data Science

Predictive analytics models have their strengths and weaknesses and are best used for specific uses. One of the biggest benefits applicable to all models is that they are reusable and can be

adjusted to have common business rules. A model can be reusable and trained using algorithms. The analytical models run one or more algorithms on the data set on which the prediction is going to be carried out. It is a repetitive process because it involves training the model. Sometimes, multiple models are used on the same data set before one that suits target/business objectives is found. It is important to note that predictive analytics models work through an iterative process. It starts with pre-processing, then to data mining in order to understand target/business objectives, next to this is data preparation. After preparation is done, then the data is modeled, evaluated and finally deployed. Once the process is completed, it can be iterated on and on again.

2.4.2.3: Limitations of Predictive Analytics Models

Despite the immense economic and academic benefits of predictive analytics models, yet there are some disadvantages to the models. Predictive models need are specific set of conditions to work, if these conditions are not met, then it is of little value to the organization/domain.

Predictive analytics models needs to be a huge sample size representative of the population. Ideally, the sample size should be in the high thousands to a few million. If datasets are smaller than the predictive analytics models then it will be unduly influenced by anomalies in the data, which will distort findings. The need for massive datasets inevitably locks out a lot of small to medium-sized organizations who may not have this much data to work with.

Predictive analytics models rely on machine learning algorithms, and these algorithms can properly assess data if it is labelled properly. Data labelling is a particularly demanding and meticulous process because it needs to be accurate. Incorrect classification and labelling can cause several problems, like poor performance and accuracy in findings.

Data models have a problem with generalisability, which is the ability to transfer findings from one case to another. While predictive models are effective in their findings for one case, they often struggle to transfer their findings to a different situation. Hence, there are some applicability issues when it comes to the findings derived from a predictive analytics model. However, there is a solution in certain methods, like transfer learning that could help mitigate some of these shortcomings.

The machine’s inability to explain what and why it did what it did makes their computations to be so exceptionally complex that humans have trouble finding. All this makes it difficult for a

machine to explain its work, or for humans to do so. Yet model transparency is necessary for a number of reasons, with human safety chief among them. Promising potential fixes: local-interpretable-model-agnostic explanations (LIME) and attention techniques.

Bias in data and algorithms in form of non-representation can skew outcomes and lead to mistreatment of large groups of humans. Further, baked-in biases are difficult to find and purge later. In other words, biases tend to self-perpetuate.

2.4.2.4: Predictive Model Algorithms

Data algorithms or predictive algorithms play a huge role in data analytics because they are used in data mining and statistical analysis to help determine trends and patterns in data. There are several types of algorithms built into the analytics model incorporated to perform specific functions. They include time-series algorithms, association algorithms, regression algorithms, clustering algorithms, decision trees, outlier detection algorithms and neural network algorithms. Each algorithm performs a specific function. As explained earlier, outlier detection algorithms detect the anomalies in a dataset, whereas regression algorithms predict continuous variables based on other variables present in the dataset. Predictive algorithms use one of two things: machine learning or deep learning. Both are subsets of artificial intelligence (AI). Machine learning (ML) involves structured data, such as spreadsheet or machine data. Deep learning (DL) deals with unstructured data such as video, audio, text, social media posts and images. Some of the more common predictive algorithms are:

- ✓ **Random Forest:** This algorithm is derived from a combination of decision trees, none of which are related, and can use both classification and regression to classify vast amounts of data.
- ✓ **Generalized Linear Model (GLM) for Two Values:** This algorithm narrows down the list of variables to find "best fit." It can work out tipping points and change data capture and other influences, such as categorical predictors, to determine the "best fit" outcome, thereby overcoming drawbacks in other models, such as a regular linear regression.
- ✓ **Gradient Boosted Model:** This algorithm also uses several combined decision trees, but unlike Random Forest, the trees are related. It builds out one tree at a time, thus enabling the next tree to correct flaws in the previous tree. It's often used in rankings, such as on search engine outputs.

- ✓ **K-Means:** A popular and fast algorithm, K-Means groups' data points by similarities and so is often used for the clustering model. It can quickly render things like personalized retail offers to individuals within a huge group, such as a million or more customers with similar features.
- ✓ **Prophet:** This algorithm is used in time-series or forecast models for capacity planning, such as for inventory needs, sales quotas and resource allocations. It is highly flexible and can easily accommodate heuristics and an array of useful assumptions.

2.4.4: Data Science and Artificial Intelligence Relevancies

Artificial Intelligence (AI) is all about how to make the system as intelligent as human beings. The intelligent systems in question are seen to be conceivable by incorporating the machines (computers) with learning, processing and decision making abilities (Russell & Norvig, 2016) The abilities are as a result of vast knowledge that helps the system to train with intelligent behaviour. A.I is on the use of numerous strategies of learning, understanding and processing techniques which can be applied on various problems or domains. The most common A.I techniques are Artificial Neural Networks, Support Vector Machines, Heuristics, and Markov Decision Process (Russell & Norvig, 2016). Artificial Intelligence is well known for its applications like natural language processing, data retrieval by using intelligent systems, expert systems for various domains, theorem proving & game playing, Scheduling and combinatorial problems , robotics and so on (Mujthaba, Abdalla, Manjur, & Mohammed, 2020) and (Nilsson, 2014). Known question rises how the A.I is related to data science, as almost all humans' beings uses the data for their wide variety of applications in day to day life. Data science plays major and noticeable roles from gathering to visualizing data.

2.4.5: Data Science and Machine Learning Relevancies

Machine Learning (M.L) is considered as subset, practical approach and application of Artificial Intelligence based algorithms. As the name indicates machine deals with wide variety of data of various domains and design the system. The system has ability to identify the new set of data by training with the existing data samples or derive the new set of rules. ML makes use of the machine as efficient by the use of supervised, unsupervised and semi-supervised and reinforced algorithms (Bell, 2020). There are numerous techniques proposed by M.L like game analytics, facial recognition, voice recognition, cloud computing, stock trading, and internet of things (IoTs). In this, data science plays an important role by providing the data in good means to have

effective M.L algorithms. Machine learning techniques are used to routinely find the appreciated primary patterns inside complex data that we would otherwise brawl to determine (Mujthaba, Abdalla, Manjur, & Mohammed, 2020)

2.4.6: Data Science and Big Data Relevancies

Big Data has recently gained much attention and reputation in both the popular press and academic circles. It is well established that we have now entered the era of "Big Data" and more of the recent emphases are placed on Data Science because of the explosion in the availability of Big Data, which are usually described as data having the following characteristics:

- ✓ ***Volume***. It is estimated that tens of exabytes of data are gathered worldwide each day and this amount is forecasted to double every 40 months. For instance, it is estimated that Walmart collects more than 2.5 petabytes of unstructured data from 1 million customers every hour.
- ✓ ***Velocity***. This means the speed with which data is generated, the speed of data creation is even more important than its volume. Twitter messages and Facebook posts are examples of social media that generated data at a very high velocity rate
- ✓ ***Variety***. This implies that Big Data includes a wide variety of data types, including WhatApp images, Instagram post of pictures and text messages, Facebook statuses, pictures on Google's Picasa or Flickr, articles in Wikipedia and online journals and periodicals, Tweets on Twitter, readings from various sensors, YouTube movies, and much more. All of these are sources of unstructured data, not suitable to be stored in classical relational databases, which assume that data possess a certain structure.
- ✓ ***Veracity***: It refers to the quality of the data that is being analyzed, it has many records that are valuable and also contribute in a meaningful way to the overall results. If veracity of data is low, it means it has high percentage of meaningless data. Eventually, the non-valuable of data is referred to as a noise

Therefore, because of the above characteristics of the data, the knowledge domain that deals with the storage, processing and analysis of these data sets are labelled as Big Data It would be a mistake, however, to equate Data Science with Big Data. Data does not have to be "big" in order for the extraction of knowledge from it to be challenging.

2.4.7: Data Science and Data Analytics Relevancies

Analytics is a newly term that has been variously defined and has recently increased in usage and popularity. INFORMS (2021), defines analytics as the application of scientific and mathematical methods to the study and analysis of problems involving complex systems. More so, Data analytics examine large data sets to identify trends, develop charts, and create visual presentations to help businesses make more strategic decisions. On the other hand, Data sciences is all about design and construct new processes for data modeling and production using prototypes, algorithms, predictive models, and custom analysis This definition differs from that of Data Science in that it makes explicit the end goal of having the insight to make an informed decision. The data is one input into a cyclic process in which the collection of data drives decisions, which in turn drive the collection of more data.

Although it is easy to collect a large volume of data without first thinking about what decisions these data will be used to make, this indiscriminate approach collection is not likely to lead to meaningful results. The cyclic nature of the analytics process is critical.

Mason and Wiggins (2010) provide article about the data science process as: Obtain, Scrub, Explore, Model, and Interpret. This was later increased to six by Ullivan (2020) as: ask questions, data collection, summarizing data, modeling, inference, and communication of results

One cannot over-emphasize the importance of the underlying model that one adopts after the exploration phase, as it informs what conclusions one can and will eventually reach. Finally, analytics is often seen as consisting of three different types of analysis:

- ✓ *Descriptive*. Summarizing historical data and identifying patterns and trends.
- ✓ *Predictive*. Forecasting what future data will look like if current trends continue.
- ✓ *Prescriptive*. Determining what actions to take in order to change undesirable trends.

Frankly speaking, with respect to the analytics process in, the first two of these types make up the analyze step, while the third is the primary driver of the optimize step. Most current analytics research is focused on the second and third steps, each of which is challenging in its own right. The close relationship between Data Science and Analytics should be evident from the above discussion. Although the Analytics process is data-driven and the focus is rightly on the data as the raw material, viewing Data Science through an Analytics lens highlights important steps in the overall process that can be challenging regardless of the size of the data. Prescriptive

Analytics requires both the development of abstract modeling frameworks and the techniques from computational mathematics required to analyze these models once they are populated with specific data (Technical-Report-15T-009, 2021)

2.5: Data Science Techniques and Methods

Data science uses series of statistical and analytical techniques to analyze data sets. It has taken hold at many enterprises and data scientist is currently becoming one of the most sought professions all over the globe. Data science applications utilize techniques such as machine learning and the power of Big data to develop high deep insights and new capabilities from predictive analytics to image and object recognition, conversational AI systems and beyond (Schmelzer, 2020).

Now, in looking at the various data science techniques and methods, the following techniques and methods are available to carry out proper analysis:

2.5.1: Classification Techniques

The major question data scientists are looking to answer in classification problems is, "What category does this data belong to?" There are many reasons for classifying data into categories.

Perhaps the data is an image of handwriting and you want to know what letter or number the image represents. Or perhaps the data represents loan applications and you want to know if it should be in the "approved" or "declined" category. Other classifications could be focused on determining patient treatments or whether an email message is spam or to find out whether a student passed or failed. The algorithms and methods that data scientists use to filter data into categories include:

- ✓ **Decision trees**. These are a branching logic structure that uses machine-generated trees of parameters and values to classify data into defined categories.
- ✓ **Naïve Bayes classifiers**. Using the power of probability, Bayes classifiers can help put data into simple categories.
- ✓ **Support Vector Machines**. SVMs aim to draw a line or plane with a wide margin to separate data into different categories.
- ✓ **K-nearest neighbor**. This technique uses a simple "lazy decision" method to identify what category a data point should belong to based on the categories of its nearest neighbors in a data set.

- ✓ **Logistic regression**. A classification technique despite its name, it uses the idea of fitting data to a line to distinguish between different categories on each side. The line is shaped such that data is shifted to one category or another rather than allowing more fluid correlations.
- ✓ **Neural networks**. This approach uses trained artificial neural networks, especially deep learning ones with multiple hidden layers. Neural nets have shown profound capabilities for classification with extremely large sets of training data.

2.5.2; Regression Techniques

What if instead of trying to find out which category the data falls into, you'd like to know the relationship between different data points? The main idea of regression is to answer the question, "What is the predicted value for this data?" A simple concept that comes from the statistical idea of "regression to the mean," it can either be a straightforward regression between one independent and one dependent variable or a multidimensional one that tries to find the relationship between multiple variables.

Some classification techniques, such as decision trees, SVMs and neural networks, can also be used to do regressions. In addition, the regression techniques available to data scientists include the following:

- ✓ **Linear regression**. One of the most widely used data science methods, this approach tries to find the line that best fits the data being analyzed based on the correlation between two variables.
- ✓ **Lasso regression**. Lasso, short for "least absolute shrinkage and selection operator," is a technique that improves upon the prediction accuracy of linear regression models by using a subset of data in a final model.
- ✓ **Multivariate regression**. This involves different ways to find lines or planes that fit multiple dimensions of data potentially containing many variables.

2.5.3: Clustering and association analysis techniques

Another set of data science techniques focuses on answering the question, "How does this data form into groups, and which groups do different data points belong to?" Data scientists can discover clusters of related data points that share various characteristics in common, which can yield useful information in analytics applications. The methods available for clustering uses include the following:

- ✓ **K-means clustering**. A k-means algorithm determines a certain number of clusters in a data set and finds the "centroids" that identify where different clusters are located, with data points assigned to the closest one.
- ✓ **Mean-shift clustering**. Another centroid-based clustering technique, it can be used separately or to improve on k-means clustering by shifting the designated centroids.
- ✓ **DBSCAN**. Short for "Density-Based Spatial Clustering of Applications with Noise," DBSCAN is another technique for discovering clusters that uses a more advanced method of identifying cluster densities.
- ✓ **Gaussian Mixture Models**. GMMs help find clusters by using a Gaussian distribution to group data together rather than treating the data as singular points.
- ✓ **Hierarchical clustering**. Similar to a decision tree, this technique uses a hierarchical, branching approach to find clusters.

Association analysis is a related, but separate, technique. The main idea behind it is to find association rules that describe the commonality between different data points. Similar to clustering, we're looking to find groups that data belongs to. However, in this case, we're trying to determine when data points will occur together, rather than just identify clusters of them. In clustering, the goal is to segregate a large data set into identifiable groups, whereas with association analysis, we're measuring the degree of association between data points.

The above methods and techniques in the data science tool belt need to be applied appropriately to specific analytics problems or questions and the data that's available to address them. Good data scientists must be able to understand the nature of the problem at hand -- is it clustering, classification or regression? -- and the best algorithmic approach that can yield the desired answers given the characteristics of the data. This is stand of data science, in fact, a scientific process, rather than to just program in order to arrive at a solution. Using these techniques, data scientists can tackle a wide range of application problems. Data scientists work with different types of datasets for various purposes. Now that Big data is generated every second through different media. The role of data science has become more important.

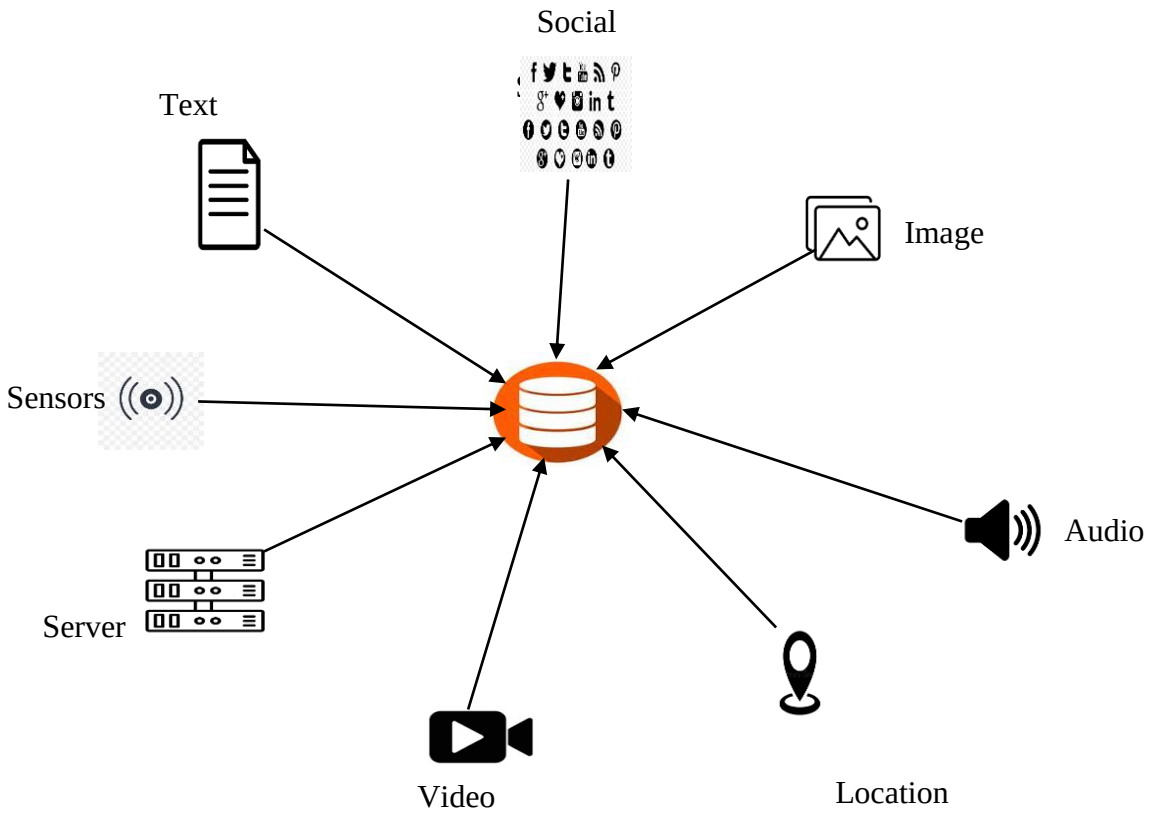

Figure 2: Different datasets for data scientists

2.6: Main Components of Data Science

Plewes (2019) was of the opinion that missing any of the core components of data science can result in the failure of one's efforts to realize any true business value. He highlighted four major components of data science as: 1. Data Strategy, 2. Data Engineering, 3. Data Analysis and Models, 4. Data Visualization and Operationalization. He further said that data science, at its core, is a practice that involves finding patterns within data. From these patterns, insight can be derived and used for business intelligence purposes or as the basis for creating new product features. Both of these outcomes of a data science project can be beneficial to product teams that are looking to differentiate their offerings in the market and provide customers with greater

value. However, before the team can begin to implement data science they should be well versed in the core components of the domain

2.6.1: Data strategy

Data strategy is simply used in determining what data is/are to gather and why. In most cases, this crucial stage is either overlooked, not given enough thought or not formalized. To be clear, this aspect is focusing only about the data one needs to address in a business problem/opportunity and why – the other considerations are important, but they're not the first step. Deciding on a data strategy requires one to make the connection between the data one is going to gather and the business goals. Not all data is created equal. In the end, the effort put into gathering data, as well as formatting it correctly and getting rid of "garbage" (*data that doesn't serve you're the purposes*), will be a reflection of both how hard that is to do, and how valuable it might be. Is the work of the team to identify data that is mission-critical to the business goals, and thus, is worth the time and energy to collect and sort? Then one might identify other data as being "nice-to-have", but won't contribute substantially to meeting the goals, so it may not be worth collecting if it requires a lot of additional time and effort.

2.6.2; Data Engineering

Data Engineering is about the technology and systems that are leveraged to access organize and use the data. It primarily involves the creation of software solutions for data problems. These solutions typically involve establishing a data system then creating data pipelines and endpoints within that system. This can involve bringing together dozens of technologies, often at a vast scale. Data engineering is important to data science overall because you can't actually do any science without it. In the end, data engineering allows data to flow from or to the product and through the ecosystem to various stakeholders. One cannot write an algorithm to improve image scheduling, for instance, unless data from the device can get to the person or "both" this means that the person is going to analyze the data and make recommendations or decisions. Engineering is the "plumbing" that lets one makes use of the data. To understand the difference between who does the data analysis or codes the corresponding algorithms, and who does data engineering, it's useful to look at the skills of a data engineer. A data engineer is a better programmer and is more of an expert in distributed systems than a data scientist. Data engineering requires in-depth understanding of a wide range of data technologies and frameworks, as well as how to combine them to create solutions that enable business processes with data pipelines.

2.6.3: Data Analysis and Mathematical Models

Data Analysis and Mathematical Models is regarded as the "heart" of data science; this is where a lot of what associate with data science happens. In this, one takes data and use Mathematics or an algorithm or both, to model how a "system" works. The data analysis and mathematical modeling aspect of data science is anything that involves the combination of: Computing, Mathematics and/or Statistics, A domain (like Education, healthcare etc), and the application of the scientific method or aspects of it:

- ✓ To further break it down, think of data analysis and mathematical models in terms of how one can use data:
- ✓ To describe, extract insights or make predictions about a service, product, person, business or technology or more likely – a combination of them
- ✓ To create a "tool" that replaces or supplements what a person does

The first use case refers to what science has always done: obtain an understanding and where possible, create a model to make a prediction utilizing data. The second use case, again, refers to what engineers have always done with math and science: find a way to use their knowledge to create a tool that does something to support a human, or is faster/better than a person could do.

What is new in the realm of data analysis and mathematical modeling is the computing power, the incredible amount of data available, and some new algorithms. In addition, now that we have access to more advanced computing power, we've only recently been able to build on many existing mathematics and statistics that could previously could not be utilized because of computational power limitations.

2.6.4: Visualization and Operationalization

Visualization and Operationalization are lumped into one particular category because they occur hand-in-hand so often. Operationalization is the more general notion, though. Simply put, it is the idea that you're going to do something with the data at hand (after analysis and modeling) and eventually draw a conclusion or take an action, in different occasions, after drawing a conclusion or taking an action the next thing is to visualize the data analyzed. The reason for this is that visualization is often the easiest way to convey the meaning of the data or analysis to the person whose job is to interpret the output of the data science.

- ✓ Data Visualization is not just about taking the data analysis and presenting it "correctly"'. Sometimes, it involves going back into the raw data and understanding what needs to be visualized based on the needs and goals of both the user and the operations.
- ✓ Data Operationalization is really about doing something with the data; someone (or occasionally a machine) has to make a decision and/or take an action based on the math and computing that has happened.

Then use this tool to have conversations about what data one is going to gather and why, and how one is either going to optimize or transform a system with the product or service. This will naturally lead to the steps of data strategy, data engineering, data analysis and model and data visualization and operationalization.

2.7: Relationship between Data Science, Artificial Intelligence and Machine Learning

The terms Data science, Artificial Intelligence and Machine learning are from the same domain yet they have specific applications and meaning, data science is an umbrella term that encompasses of data analytics, data mining, machine learning and artificial intelligence. Data science does the work of data gathering and data transformation after gathering and transforming data one needs to make predictions and right insights, then one needs machine learning techniques to do this by using supervised learning and unsupervised learning both are used to extract prediction from given datasets. In order to get insight from prediction being made, one needs the knowledge of data analytics which is the process of data science. Next is to perform some actions this is where AI comes in the picture (Loon, 2021)

Data can be analyzed according to its type like text, statistical, predictive and perspective. Data Science consists of countless statistical practices whereas AI relates how use of computer algorithms in an intelligent way. Data Science deals with notion to tackle big data which includes certain operations like data cleaning, training, analysis, process, modeling and so on. A data scientist collects the data from various ways and incorporated with machine learning algorithms.. Data science is expected to do a lot of innovations in the areas like applied computing, medical sciences, professionals & social life activities, computing paradigms, data management systems and many more to have a better decision making (Schmelzer, 2020).

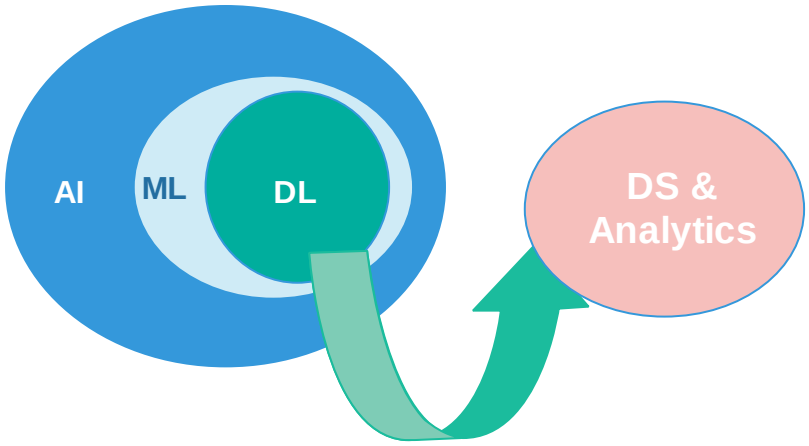

Figure 3: Relationship between AI, ML, DL and Data Science

2.8: Things Required Data Scientist to have and take note of

2.8.1: Basic Skills of a Data Scientist: A Data Scientist should be able to:

- ✓ Ask the right question
- ✓ Understand the data structure
- ✓ Interpret and wrangle data
- ✓ Apply statistical and mathematical methods
- ✓ Visualize data and communicate with stakeholders

2.8.2: Technologies that have helped Data Scientist to:

- ✓ Build and train machine learning models
- ✓ Manipulate data with technology
- ✓ Extract information from data
- ✓ Build data tools applications and services

2.8.3: Common Challenges faced by Data Scientist are summarized below:

- ✓ When data quality does not conform to the set standards
- ✓ Data integration is a complex task
- ✓ Data is distributed into large cluster in HDFS which is difficult to integrate and analyze
- ✓ Unstructured and semi-structured data are harder to analyze

2.9: Decision Making

Decision making is the most important thing we do in life. We decide on whom to marry, what college or University to attend, what course to study, what job offer to accept, what city to live in, what business to invest in, what service our corporation should sell, how best to advertise,

how to resolve a conflict and so on. A complex decision involves many factors all of which play a role in bringing that decision about. But not all the factors are equally important. Our challenge is to find a way to determine their priorities so we can mix them in the right proportion to make a successful decision (Saaty & Peniwati, 2007). We are all fundamentally decision makers. Everything we do consciously or unconsciously is the result of some decision. The information we gather is to help us understand occurrences in order to develop good judgments to make decisions about these occurrences. Not all information is useful for improving our understanding and judgments (Saaty, 2006). To make a decision according to Saaty, (2006) we need to know the problem, the need and purpose of the decision, the criteria of the decision, their sub-criteria, stakeholders and groups affected and the alternative actions to take. We then try to determine the best alternative, or in the case of resource allocation we need priorities for the alternatives to allocate their appropriate share of the resources.

The significance of making decision is as important to every individual persons and organizations as it is crystal clear that everyone makes decision on a particular issue or business or the other. The decision made by one should be void of ingenuity, intuition and personal judgments but rather should be on the basis of scientific and statistics studies. To make a true decision, it is necessary on time and precise information that decision support systems help managers in this area to make true decisions even in unpredicted positions (Khodashahri & Sarabi, 2013). The authors pressed further that the use of decision making systems correctly create better realization on solving problems and developing communication procedure. So, decision support systems are very flexible and interactive computer systems that use to support all decision making process in conditions that problem is semi-structured. The software for decision making are in fact serve as a counselor beside a decision maker and provides this possibility that could run with huge information mass and use them arbitrarily and in suitable models frame to improve decision making. To Alvani (2009) one major method of decision making is the realization of the study of decision structure. Khademi (1993) gave three types of decision making as:

- ✓ ***Independent decision making***: Such decision making is when decision maker has general authority and complete executive ability for made decisions. Decision support system that could do such decision making is known personal support.
- ✓ ***Sequential dependence decision making***: Such decision making is when decision maker only conducts a part of decision making and provide his decision making results another person

for future decision makings. Decision support system that could protect such decision making is known organizational support.

- ✓ ***Convergent dependence decision making***: Such decision making is when a group as a council is decision making responsible. Decision support system that could protect this decision making is known group support.

Regarding to three types of decisions making observes that second decision making is the most important and usual one and therefore a decision support system should support such decision making.

Decision Support Systems are computer based systems that use in all organizational levels and by some individuals to make decision and solve semi-structured problems. These systems really designed to help people especially manager in all decision making stages even they play effective role to know and evaluate displaceable solutions and select the most important ones by manager (Mahmoodi, 2003). The concentration towards the use of software applications to support decision making has been around for years. The past decades have witnessed a tremendous development in the graphical user interface, which facilitates the use of more advanced computational techniques to a wider group of users. As a consequence, several decision analytic tools have emerged in recent years. Decision software based on classical decision theory, which includes Visirule, Expert Choice AHP, Decision Lens, DecideIT, D-sight and many more have successfully been commercialized and are used by various professional decision analysts and decision makers to aid them in their work. However, most classical decision models and software based on them consist of some straightforward set of rules applied to precise numerical estimates of probabilities and values (Danielson, Ekenberg, Johansson, & Larsson, 2004).

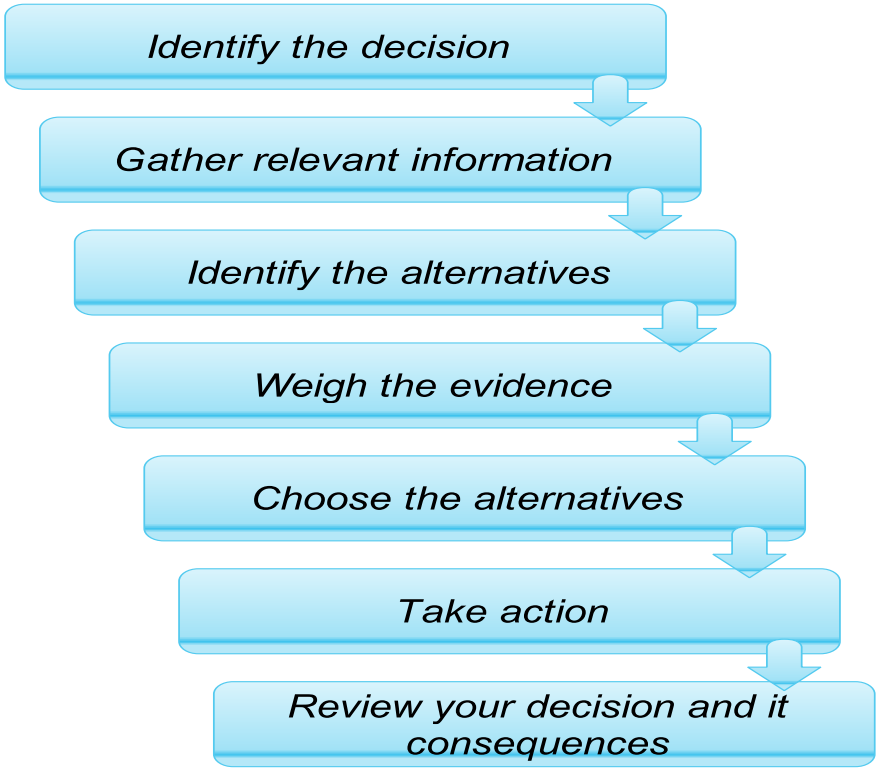

Figure 4: Genera Model of Decision Making Process

2.9.1: Overview of some Decision Making Software

Much multi-criteria decision aid software have been developed for the past 30 years (Weistroffer, Smith, & Narula, 2005). There are two fundamental characteristics that differentiate them: the type of problem that the tool is aimed at resolving and the methodology it is based on. Extensive research' efforts have been channeled towards understanding and formalizing the activity of decision making in the education domain. As a result, a number of systems are qualified to support the activity that has been developed. These systems can be naturally classified into three categories according to (Kryssanov, Abramov, Fukuda, & Konishi, 1997): information retrieval systems, decision support systems and expert systems. It was observed that no clear distinction between the aforementioned categories as found in the literature, hence the suggestion is that a distinction should be made based on the following notions that:

- ✓ an information retrieval system is a computer-based system to capture, manipulate, retrieve and transmit organized data necessary to solve a professional task according to detailed transactions defined by a user;

- ✓ a decision support system is a knowledge-based information system to capture, handle and analyze information which affects or is intended to affect decision making performed by people in the scope of a professional task appointed by a user;
- ✓ An expert system is a knowledge-based system to be used instead of or together with a human operator to make decisions in the framework of a professional task with explanations for users.

Decision support systems are interactive, computer based systems that aid users in judgment and choice activities. They provide data storage and retrieval, but enhance the traditional information access and retrieval functions with support for model building and model-based reasoning. They support framing, modeling, and problem solving (Roger & Marek, 2007). In everyday life decisions often use intuition, despite a lot of shortcomings of this method thus developed a new systematic called decision analysis, namely intelligence, perception and philosophy. After using intelligence, perception and philosophy to create the model, determining the possible value, set a value on the expected results and assessing the preference for time and preference for risk, then to arrive at a decision required logic (Marimin, 2004). One model that can be used as a decision-making process is by using the analytic hierarchy process. Below are brief discussions about some selected decision making software:

2.9.1.1: VisiRule

VisiRule offers multiple interactive and visual tools that help the decision maker to better understand and manage his multi-criteria problem. The aim of using this software is to help a decision maker to structure and to better understand his problem by providing valuable information about the consequences of making choices. VisiRule is a graphical AI software tool which enables business users to build rule-based systems to model and automate their decision-making processes. It offers a simple and familiar flow charting interface for drawing the logic underpinning the personal, educative or business decisions. VisiRule includes a code generator and rule engine, so that models can be immediately compiled and executed locally. In addition, VisiRule charts can be embedded within other applications and architectures using Rest and Json. The Visirule software is always used as a decision supporting tool, in which the rules are basically and precisely presented using Logic Programming Model. The RSA-Expert can be of great use to researchers in making a firm decision in utilizing suitable statistical data analysis in researches (Muraina, Rahman, Adeleke, & Aiyegbusi, 2013). Visirule allows experts / researchers to concentrate on explaining and

establishing the structure of the logic correctly using their chosen tools - those embedded materials that can assist researcher to accomplish his mission (Bilgi, kulkarni, & Spenser, 2010). It is also seen to be a powerful tool that helps to avoid or scrape from some likely errors or bugs which can come into play when trying to code logic in a text based rule language. Identifying the knowledge used in decision making or problem solving is a very crucial component of the expert system design (Giarratano, 2007).

2.9.1.2: Expert Choice AHP (Analytic Hierarchy Process)

The Analytic Hierarchy Process (AHP) with expert choice is a multi-criteria decision making (MCDM) approach that assists the decision-makers facing a complex problem with multiple conflicting and subjective criteria (such as business, personal choice selection, and many more). AHP using expert choice has the advantage of permitting a hierarchical structure of the criteria, which provides users with a better focus on specific criteria and sub-criteria when allocating the weights (Ishizaka & Labib, 2009). One of AHP's strengths is the possibility to evaluate quantitative as well as qualitative criteria and alternatives on the same preference scale of nine levels. These can be numerical, verbal or graphical. The last step of the decision process is the sensitivity analysis, where the input data are slightly modified in order to observe the impact on the results. If the ranking does not change, the results are said to be robust. The sensitivity analysis is best performed with an interactive graphical interface. Expert Choice allows different sensitivity analyses, where the main difference is the various graphical representations.

Multi criteria decision making is one of the most widely used methods in the decision-making area. The objective of multi criteria decision making is to select the best alternative from several mutually exclusive alternatives based on their general performance regarding various criteria [or attributes] decided by the decision maker (Chen, 2005) (Andayan & Mardapi, 2012).

2.9.1.3: D-Sight

D-Sight is designed as a solution dedicated to supporting decision-making processes. It provides a framework allowing one to evaluate alternatives based on several criteria that already defined. The software support one in identifying the best solution. D-Sight is a collaborative decision-making platform that allows one to evaluate alternatives, analyze data and make transparent decisions. It brings people together and integrates their perspectives into the optimal decision.

D-Sight is the third generation of PROMETHEE (**P**reference **R**anking **O**rganization **METH**od for **E**nrichment **E**valuations) and GAIA (**G**eometrical **A**nalysis for **I**nteractive **A**id) based

applications. It offers multiple interactive and visual tools that help the decision maker to better understand and manage his multi-criteria problem (Hayez, De-Smet, & Bonney, 2012). Multi-criteria decision aid (MCDA) addresses problems where choices, alternatives, items, etc are evaluated on several conflicting criteria. The aim is to help a decision maker to structure and to better understand his problem by providing him valuable information about the consequences of his choices, the synergies and redundancies between criteria, the influence of parameters, the comparison of action profiles. D-Sight is regularly used by several universities and research centers all over the world for teaching and research purposes (Macharis, Bernardini, De-Smet, & Hayez, 2010). In addition, D-Sight is used in private companies as well. Some of those collaborations led to the publication of different case studies (Hayez, De-Smet, & Bonney, 2012).

The hierarchy tool in D-sight allows the decision-makers to group the criteria in a tree with several levels in order to structure the problem. In the visual analysis tools such as the GAIA representation, the results can then be analyzed from different point of views, from the highest hierarchy level to the lowest. The PROMETHEE V algorithm for optimization under constraints was present in D-Sight. The so-called decision-maker's brain is also now included in the GAIA plane and shows the uncertainty that one could have on the decision with respect to the criteria weights value. A weights elicitation tool is available to assist the participant in determining the relative importance of the criteria through an interactive procedure. A wizard tool in the software allows anyone to easily define the problem step by step. The PROMETHEE parameters (function and thresholds) are important in the decision process. That is why a uni-criterion score visualization tool was part of the software to help the user identify the impact of his parameters.

2.9.1.4: Decision Lens

Decision Lens is online decision-making software that is based on multi-criteria decision making. To Saaty, (1980) decision Lens implements the Analytic Hierarchy Process (AHP) and the Analytic Network Process (ANP) and is used in fields such as energy (Saaty, 1996) medical research (Cheever, et al., 2009) and group decision-making (Mu and Butler, (2009); Begicevic, Divjak, and Hunjak, (2011)). Decision Lens is a cloud-based prioritization and resource allocation software solution for critical decision-making in many areas including Research and Development, capital planning, IT portfolio planning, and budget allocation. Decision Lens combines experts' judgments with business data to establish priorities in an efficient, collaborative framework. By providing one with a streamlined and automated solution, Decision

Lens saves organizations countless hours and dollars while simultaneously providing the optimal value for investments.

Decision Lens allows one to clearly lay out the process by which one wants to make decision, it gives one the solution to methodically determine how important each piece of information is to one in making decision, it provides the transparency one needs to help uncovering the motivations and thoughts of the people one has included in the process and also builds in the ability to change mind and quickly see the results of the choices.

Decision Lens allows for greater collaboration, transparency, efficiency, consistency, and analysis in decision making, subjectivity is inherent in decision making and although we cannot avoid it, we can make it explicit, group dynamics and facilitation are key in a collaborative decision making process, it supports single and multi winner decision types and can generally be used in any decision where the alternatives are known and strategic evaluation criteria can be developed.

2.10: Data Mining

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, one can say that data mining is mining knowledge from data. Osmar (1999) defined Data Mining as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is a process which finds useful patterns from large amount of data (Ramageri, 2011). Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that’s why data mining is considered one of the most important frontiers in database

and information systems and one of the most promising interdisciplinary developments in Information Technology (Ramageri, 2011). With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Table 3: Phases of Data Mining

Phases		Action Performed in the phase
Data cleaning	➔	*Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.*
Data integration	➔	*At this phase, multiple data sources, often heterogeneous, may be combined in a common source*
Data selection	➔	*At this phase, the data relevant to the analysis is decided on and retrieved from the data collection.*
Data transformation	➔	*Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.*
Data mining	➔	*It is the crucial step in which clever techniques are applied to extract patterns potentially useful.*
Pattern evaluation	➔	*In this phase, strictly interesting patterns representing knowledge are identified based on given measures.*
Data Visualization	➔	*This is the final phase in which the discovered knowledge is visually represented to the user. This essential phase uses visualization techniques to help users understand and interpret the data mining results.*

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. Here are some examples in more detail:

2.10.1: Data Mining Classifications

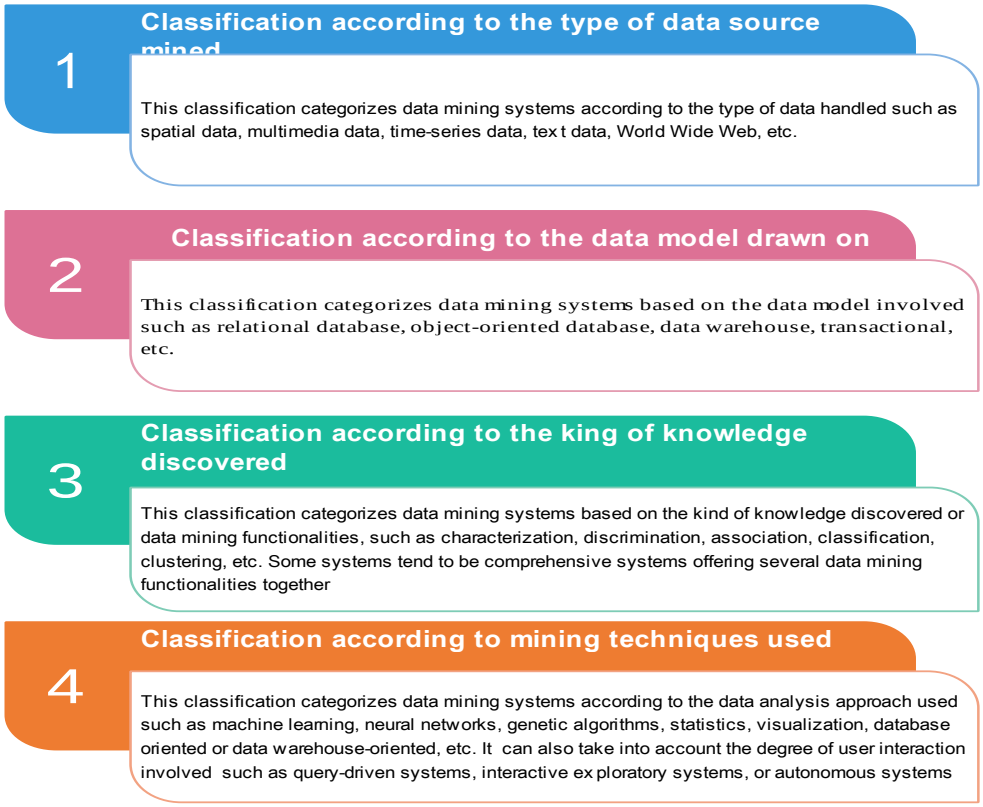

Figure 5: Data Mining Classifications and functions

2.10.2: Data Mining Tasks

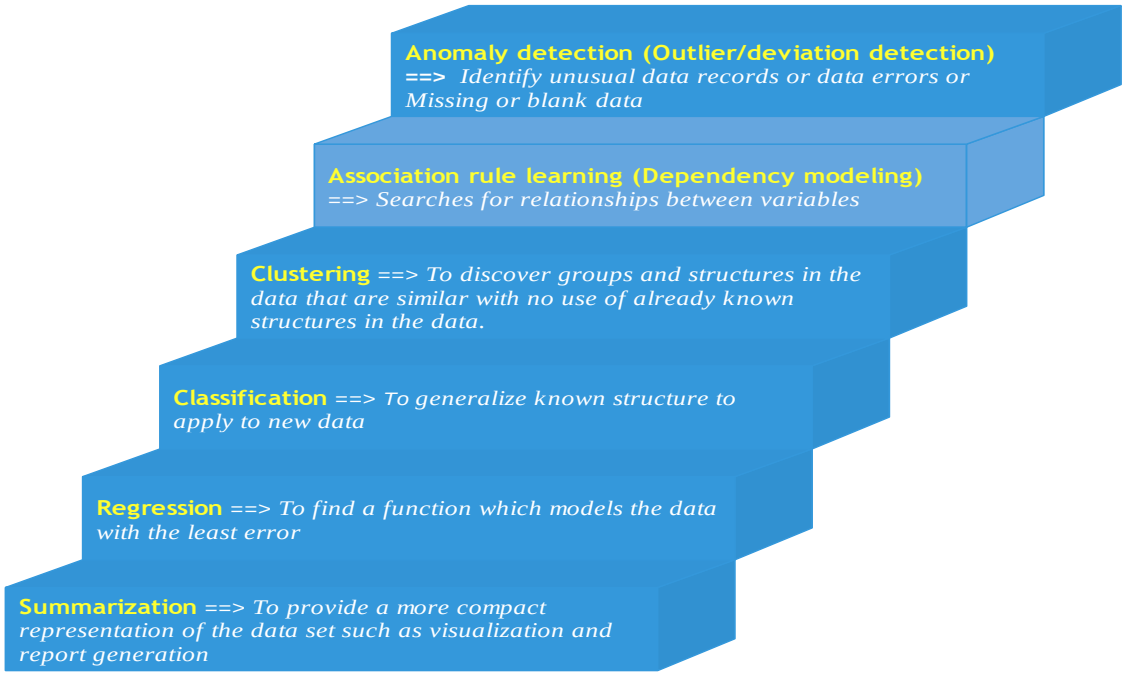

Figure 6: Six common tasks in Data mining

2.11: Summary

In this chapter, researcher discussed firstly the theoretical framework/background of the study where he took a close look at series of theories like decision theory, risk theory, data theory, probability theory, machine learning theory and prediction theory and how all these theories related to the current study. Furthermore, clear concept of prediction and decision making was critically defined and explained with prediction modeling, its types and how all the types would be applied in Data Science. Finally, the researcher stated the limitations of model algorithms and showed the relationships between Data Science, Artificial Intelligence, Machine Learning, Big Data and Data Analytics.

Data Science techniques and methods were explained accordingly by looking at classification technique, regression technique as well as clustering and association techniques. Then, followed were components of Data Science where the researcher explained data strategy, data engineering data analysis and mathematical models, and visualization and operationalization. In conclusion, the researcher outlined the things required the Data Scientist to have and to take note of such as basic skills, technologies and challenges.

Hence, decision making was comprehensively defined with proper looking at decision making software examples and their implications in prediction which include: VisiRule, Expert Choice AHP, D-sight and Decision lens. The chapter ends with meaning, phases, classification, and tasks of Data Mining

CHAPTER THREE

Research Materials and Methods

3.1: Introduction

In this chapter, researcher gave details on how the study designs were systematically arranged to ensure valid and reliable results that addressed the earlier formulated research aims and objectives. The chapter also critically found answers to questions like: What data to collect and what data to ignore? Who to collect data from (Population and Sampling methods)? How to collect the data (Data Collection methods)? And how to analyze the data collected for the study? For proper clarity, the researcher re-iterated the stated objectives as: to use appropriate applications and data science techniques to predict students course of study using WAEC and NECO grades dataset to categorize the content with respect to the student data informatics and cluster the data for analyzing their data pattern, to improve the data prediction strength in course of study by evaluating WAEC and NECO information with the use of data science (supervised and unsupervised machine learning) algorithms, to validate the information with the use of classification algorithms, to evaluate the performance of proposed approach by evaluating series of parameters (Precision, recall, F-measure, the accuracy with classification rate), to compare the performance of different classifiers and clustering algorithms on WAEC and NECO datasets to predict relevant course of study for student using apps, regression algorithm, decision trees algorithm and neural network algorithm and many more.

3.2: Research Design

The research design is an investigator's overall strategy to justify the stated objectives or to answer the research questions or hypotheses formulated. The choice of the design used in this study was underpinned by assumption that learners or students choice of courses while getting into higher institutions open to different dimensions which single approach may not be appropriate. In this thesis, prediction models design was used which are mostly used to predict future occurrences such as course of study of students in the remote future, in this design, retrospective cohort study which involves the use of historical data. Equally, mixed-design of both quantitative experimental and qualitative ethnography designs was established in the study. In a quantitative experimental study, the researcher aimed to produce generalizable knowledge about the causes of a phenomenon while in a qualitative ethnography the researcher aimed to produce contextual real-world knowledge about the behaviors, social structures and shared beliefs of students.

3.3: Target Population and Sampling Method

The target population of this thesis involved all male and female students that have written WAEC or NECO examinations within South-West Geo-political zone of Nigeria. This zone covered six highly populated states which include: Lagos State, Oyo State, Ogun State, Osun State, Ondo State, and Ekiti State respectively. For archive results collected did not require any sampling method. The use of interview approach was based on judgment or purposeful (Selective or subjective) sampling method because of its cost-effective advantage of the sampling approach. The researcher used convenience sampling approach where students were selected based on availability and willingness to take part.

3.4: Instrumentation and Data Collection

WAEC or NCO archive results of students were used. According to Druva (2021), Data archiving is the practice of identifying data that is no longer active and moving it out of production systems into long-term storage system. A data archiving strategy optimizes how necessary resources perform in the active system, allowing users to quickly access data archive storage devices or data archiving plans for easy retrieval and more cost-effective information storage. It also clarifies how users should move data for best performance within applicable regulations and law. It enables long term storage and retention of data and provides secure locations for storing mission-critical information for use when needed. More than 300 WAEC or NECO results were collected for the study. The results were collected from school to school in order to get enough results for the analysis. Schools have been keeping their WAEC and NECO results for years as an archive. The results of students selected were between 2010 till date

Questionnaire instrument was also used to collect data from students already gained admission into the universities or colleges. The questionnaire was designed via Google form and was administered and collected using social media platforms such as WhatApp, FaceBook, Twitter etc. Questionnaire is a series of questions or items asked to individuals to obtain statistically useful information about a given topic. When properly constructed and responsibly administered, questionnaires become a vital instrument by which statements can be made about specific group or people or entire population (Davis, 2014). A questionnaire is a research instrument consisting of a series of questions for the purpose of gathering information from respondents. It can be thought of as a kind of written interview which can be carried out face-to-face, by telephone, computer means or post. It provides a relatively cheap, quick and efficient way of obtaining large amounts of information from a large sample of people (McLeod, 2018). Questionnaire sample

size was selected via the use of simple random sampling where each member of the population has an equal chance or probability of being selected.

At the same time, structured interview was equally used to solicit information from the students with respect to their decisions on choosing course of study on getting into higher institutions. Interview is a powerful tool for obtaining qualitative research data (McMillian, 2008). This strategy is very essential to gather data from participants and also on issues that may not be directly observed. Hence, 20 students (Participants) were sampled for the study. The interview questions were developed in a manner that allowed for maximum flexibility. The main purpose of the use of interview was to solicit the views, and to gather in-depth responses to the issues of making decisions

3.5: Confidentiality

Confidentiality was considered important to this study because the participant (students) status could be at risk if any disclosure was made about their results. Confidentiality meant that I knew who the participants were, and that their identities should not be revealed. I did require the participant permission to get information from them. So, I was obliged to ensure that no harm occurred to those voluntarily participated and they have all made decision to assist me through the study

3.6: Informed Consent

In this study, a code of informed consent has guided the involvement of participants as well as where results used got from. The schools were visited one by one to request for the data (results), likewise the participants were contacted and agreed before sending questionnaire to them. In case of interview conducted, participants were invited to participate in the research and they were fully informed about the nature of the study. Although the direction that the research would take might not entirely predictable, this makes the researcher to inform them the nature and diversity of the study. Hence, earlier informing them made the participants to be enthusiastic in participating in the study.

3.7: Validity and Reliability

Validity and reliability of the instruments archive results selected where from well known bodies: WAEC is West African Examination Council which was established valid. also NECO is Nigeria Examination Council which is an examination body established in Nigeria in 1999 so,

the two bodies were valid and reliable in Africa and worldwide because some university all over have being admitted Nigerian or Africa international student s for enrolments.

Questionnaire instrument items and statements were checked by colleagues in research, computer science and recounted fields to ensure content, face, construct and criterion validities of the instruments. Hence, the questionnaire was subjected to Chronbach' Alpha reliability statistics which yielded 0.89 reliability index. This value implied high reliability capacity of the instrument. In another words, the instrument is highly desirable for the researcher to use.

Interview instrument was validated by colleagues in English language department to look and check the use of the language coupled with the target audience for the research. The instrument was proved reliable after conducting two pilot interviews where the responses were collected and rated and came up with reliable judgment. This implied that the instrument can be used at different occasions to get the similar results.

3.8: Data Analysis Procedures

Data analysis is the most crucial aspect of research work. It summarizes collected data and involves the interpretation of data gathered through the use of analytic and logical reasoning to come up with patterns, relationships or tends. It can also be seen as the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to Shamova and Resnik, (2013) data analysis procedures give researcher a way of drawing inductive inferences from researcher's data and differentiate it from other confusing statistics that present in the data. The analysis covered both analysis design and interpretation of the results

3.8.1: Data Analysis

The analysis was done in phases:

Phase one was done using data science techniques which included the use of pre-processing techniques (checking for missing data, removing outliers, transforming variables). In order to realize this, Python, R, Weka and RapidMiner tools were used, clustering techniques, classification techniques and finally determine the prediction of student's course of study with series of algorithms on datasets.

Second phase was based on the analysis of questionnaire instrument using PLS-SEM (Partial Least Squares Structured Equation Modeling) techniques with the use of SmartPLS

The third phase was done on the data collected by the use of interview instrument using thematic approach.

3.8.2: Interpretation

The interpretations were established using appropriate and relevant tables, charts, boxes and graphical representations with the use of visualization tools.

Second phase was done using PLS-SEM (Partial Least Squares Structured Equation Modeling) with the help of SmartPLS software from the data gotten through Questionnaire

The third phase was based on the use of thematic approach. Thematic analysis involves transcribing or coding and examining with close look into identifying broad themes and patterns. The interviews were transcribed and thematic analysis was conducted. This involved coding all the data before identifying and reviewing key themes. Each theme was examined to gain an understanding of participants' perceptions and motivations.

3.9: Credibility

The term credibility means the degree to which people believe and trust what researcher tells them about (Cambridge-Dictionary, 2021). The source credibility theory states that people are more likely to be persuaded when the source presents itself as credible. In this research study, the data collected were from reliable and valid sources and they were proved credible. Yale University group defined credibility in terms of a researcher's expertise and trustworthiness. Expertise referred to a communicator's or researcher's qualifications or abilities to know the truth about the data collected, whereas trustworthiness was conceptualized as perceptions of the communicator's / researcher's motivations to tell the truth about the data collected O'Keefe (1990), defined the source credibility as "judgment made by a perceiver concerning the believability of a communicator."

3.10: Generalizability/Transferability

SAGE (2014) defined transferability of research findings as the extent to which it can be applied in other contexts and studies. It can also be seen as an alternative term used in place of generalisability and external validity. Transferability also referred to as the degree to which the results of a research can be transferred to other contexts or settings with other data sets/ respondents. WAC (2021), described transferability as the process of applying the results of research in one situation to other similar situations. In this study, the researcher gave the

assurance that result obtained could be applied to solve similar problems in another environment with which has close characteristics as where the research was conducted. The results of this study could be generalized to the entire population. Hence, the results could be used to predict future student course of study on getting into higher institutions

3.11: Dependability

To Sandelowski (2000), dependability in research referred to the consistency and the reliability of the research findings and the degree to which research process or procedures are made/documented, give room for auditing , following and critiques. In this study, dependability was achieved through the proper processes followed in logical, traceable and clear form (Tobin & Begley, 2004).

3.12: Confirmability

This is the last criteria of trustworthiness that a qualitative research design must have. It is the level of confidence that the research study's findings are based on the participants' narratives and words rather than potential researchers biases (Davis, 2014). In this study, the interview instrument results/findings were solely based on participants' personal words which were collated for the analysis. This study is necessary at this moment because past studies were centralized on predicting the performance or feature job placement but this study was based on new direction of predicting the cause of study of student when getting to higher institution.

3.13: Ethical Issues in the Study

According to Ng (2021), a good research work should be well adjusted, well planned, appropriately designed and ethically approved. So in this research work a research protocol was developed and dully followed the concerned students /participants consent were sought with precise roles of every student and researcher. Conclusively, no ethical issues were breached in carrying out this research study

3.14: Conflict of Interest Assessment

A conflict of interest is asset of circumstances that creates a risk that professional judgment or actions regarding a primary interest will be unduly influenced by a secondary interest. In this study the three main element of conflict of interest were looked into: The primary interest, the secondary interest, the conflict itself. The primary interest according to the purpose of a professional activity, So primary interest promote and protect the integrity of research, the participants and their educational quality. The secondary interest was based on professional

advancement in the field of computer science, research recognition is to obtain Ph.D in Data science, colleagues in the same fields. Therefore no part of the interest was neglected or conflicted with one another.

3.15: Position Statement

Casano (2021), defined position statement as ways for researchers to let others know their stance on particular study. Therefore, the following are researcher position statements.

- ✓ The study was conducted to solve the problem students usually encounter when they want to select appropriate course of study on getting to higher institutions
- ✓ The researcher was in support of this, because no research has been conducted before now to look into the problem at hand.

3.16: Summary

Since, it is the duty of researcher to ensure that the research work was conducted in an ethical and responsible manner from planning to the end of the work. The chapter focused on methodologies and procedures required to carry out effective study which include the research design, population, sample and sampling method, instrumentation, validity, and reliability of the instrument, instrument implementation analysis , procedure, generalizability\transferability, credibility, dependability, confirmability, conflict of interest assessment, position statement and finally the ethical issues in the study.

CHAPTER FOUR
Data Analysis and Findings

4.1: Introduction

This chapter covered analysis, presentation, and interpretations of data collected in the preceding chapter. It presents the data using a clear text narrative, supported by tables, graphs and charts. The chapter's results were reflected and aligned with the purpose of this study. It started by presenting demographic data to understand the representativeness of the sample. The analysis and interpretation of data was carried out in three phases. The first phase focused on the analysis of data which sub-divided into three parts – Dataset analysis using data science techniques (Machine Learning Models), Questionnaire analysis using quantitative methods and Interview analysis using qualitative approach (Thematic method). The second phase was on presentation of data and the last phase was based on interpretation the results of the three parts already mentioned. This chapter consists of the data that has been collected as a part of the research and the researcher's analysis of the data. Presenting the data collected and its analysis in comprehensive and easy to understand manner of how the research flow in order to have a good analysis chapter. The Analysis chapter is the foundation on which the researcher draws the conclusion, identifies patterns and gives recommendations. The entire utility of the research work depends on how well the analysis is done. The researcher properly documented the various types of data (quantitative and qualitative) and the relevant approach, tools and conclusion were drawn accordingly.

4.2: Analysis of Dataset used with Machine Learning Algorithms/Models (Data Science Techniques)

4.2.1: Data Cleansing

4.2.1.1: Data Input

OlevelResult (csv) file Dataset was input into the machine learning language such as R, Python, Weka etc. The dataset contained 7 attributes and 304 instances (Observations)

Input variable were gotten from WAEC and NECO results in the past and present. The transformation was given in table1

Table 4: Input Data Transformation

S/N	Input Variables	Domain		
		Subjects	Grade Score	Code
1	O/Level Score	English Language	70 – 100	1
			60 – 69	2
			50 – 59	3
			40 – 49	4
			0 – 39	5
		Mathematics	70 – 100	1
			60 – 69	2
			50 – 59	3
			40 – 49	4
			0 – 39	5
		Biology	70 – 100	1
			60 – 69	2
			50 – 59	3
			40 – 49	4
			0 – 39	5
		Physics	70 – 100	1
			60 – 69	2
			50 – 59	3
			40 – 49	4
			0 – 39	5
		Chemistry	70 – 100	1
			60 – 69	2
			50 – 59	3
			40 – 49	4
			0 – 39	5
			Code	
2	Gender	Male	1	
		Female	2	
3	State of Examination	Lagos	1	
		Ogun	2	
		Oyo	3	
		Osun	4	
		Ondo	5	
		Ekiti	6	
4	Examination Type	WAEC	1	
		NECO	2	

4.2.1.2: Output variables were derived from addition of all the scores divided by five

Table 5: Output Data Transformation

S/N	Output Variables	Domain	
		Score Grade	Code
1	Excellence	75 – 100	1
2	Very Good	65 – 74	2
3	Good	55 – 64	3
4	Average	40 – 54	4
5	Poor	0 - 39	5

Table 6: Output Data Transformation for qualifying criteria

S/N	Output Variable	Domain
1	Highly Qualified for any sciences	1
2	Qualified for less challenging sciences	2
2	Qualified with special consideration	3

4.2.2: Data Preprocessing

Every attribute and instances were checked during preprocessing to ensure that there were no missing data. In the same way, the instances were reweighted using class balance so that each class has the same total weight. The total sum of weights across all instances was maintained. Since the class is numeric, the class was discretized using equal-width discretization to establish pseudo classes for weighting.

Table 7: Showing the attributes, Missing Count, Data Type, Mean and SD

S/N	Attribute Name	Missing Count	Data Type	Mean	SD
1	English_Language_Score	0	Numeric	57.09	6.70
2	Mathematics_Score	0	Numeric	51.92	14.18
3	Biology_Score	0	Numeric	56.16	7.29
4	Physics_Score	0	Numeric	55.44	10.16
5	Chemistry_Score	0	Numeric	59.83	9.39
6	Average_Score	0	Numeric	58.09	6.70
7	Prediction_Criteria				
	Label	Weight	Count	Data Type	
1	Highly Qualified for any	101.333	134	Norminal	

	sciences				
2	Qualified for less challenging sciences	101.333	157	Norminal	
3	Qualified with special consideration	101.333	13	Norminal	

4.2.3: Principal Component Analysis

Principal Component Analysis (PCA) is a member of data science (unsupervised machine learning) techniques, which is a main linear technique for dimensionality reduction; PCA is used to explain the variance-covariance structure of a set of variables through linear combinations. Principal Component Analysis performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. According to Sartorius (2020) PCA is a statistical procedure used by data scientists and analysts that allows them to summarize the information content in large data tables by means of a smaller set or summary indices that can be more easily visualized or analyzed. Large datasets are increasingly common and are often cumbersome to analyze and interpret. The use of PCA, as a technique in data science to reduce the dimensionality of datasets, help to increase the interpretability and at same time with no loss of any information as we have it in the original datasets (Jollife & Cadima, 2016). In practice, the covariance (and sometimes the correlation) matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigenvectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space (with dimension of the number of points) has been reduced (without data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.

Table 8: Correlation Matrix

		English	Mathematics	Biology	Physics	Chemistry
Correlation	English	1.000				
	Mathematics	.210	1.000			
	Biology	.242	.527	1.000		

	Physics	.162	.306	.412	1.000	
	Chemistry	.166	.112	.181	.465	1.000

Table 9: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.651
Bartlett's Test of Sphericity	Approx. Chi-Square	256.955
	Df	10
	Sig.	.000

Kaiser-Meyer-Olkin Measure of Sampling Adequacy should be greater than .6. Then from the table it is 0.651. Likewise Bartlett's Test of Sphericity should be significant. Also it is significant sig. = .000

Table 10: Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.153	43.066	43.066	2.153	43.066	43.066	1.825
2	1.041	20.820	63.886	1.041	20.820	63.886	1.643
3	.870	17.392	81.278	.870	17.392	81.278	1.116
4	.498	9.966	91.244				
5	.438	8.756	100.000				

The total extracted sums of squared loading indicated that the three factors are good for the dimensionality reduction produced by the PCA (Table 10). This is in support of the output of the scree plot suggested in Figure 7 as well as the figure 8

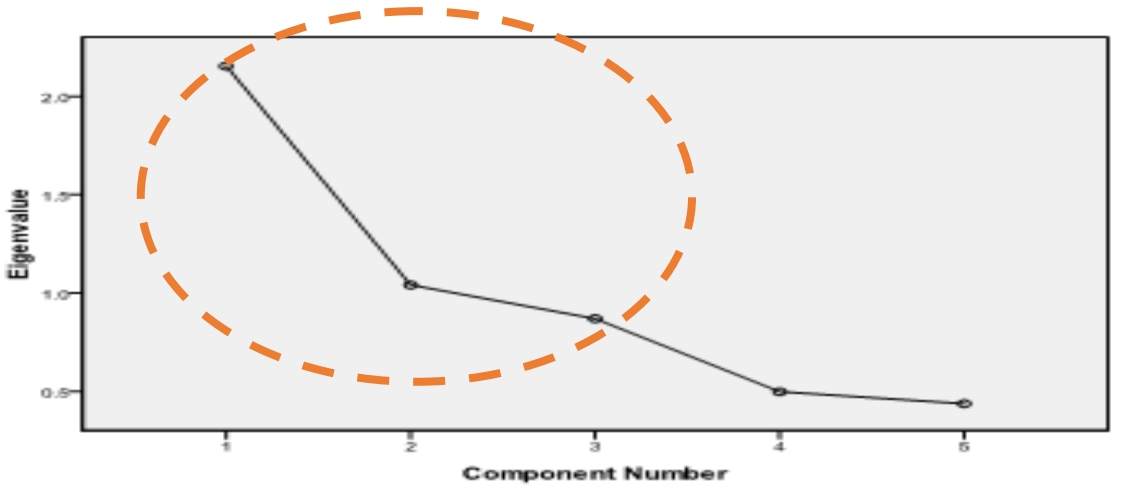

Figure 7: Scree Plot to show the number of factors to retain

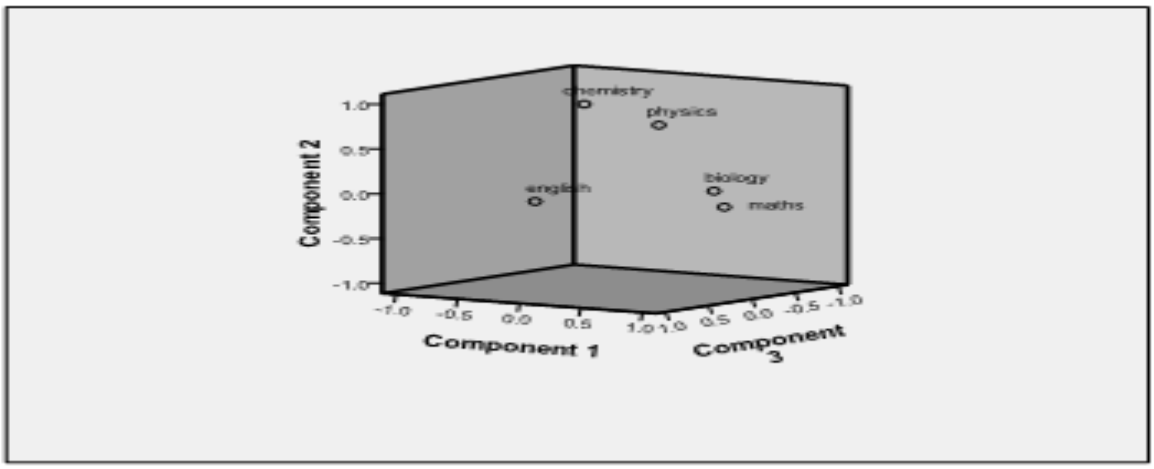

Figure 8: Component Plot in Rotated Space to retain three components

4.2.4: Data Mining Procedure

The below procedures were taken to mine the dataset up to outputting the results:
Input the training data and testing data samples – This was done in four different ways: 1. Using split of 70% of training set and 30% of testing set, 2. Using split of 75% training set and 25% of testing set, 3. Using 10-Fold Cross Validation, and 4. Using evaluation on full training data set.

- ✓ Initialization of the classifiers – In this, 9 classifiers were used to ascertain the high level of accuracy for the prediction. The classifiers include: Naïve Bayes classifier, Multilayer Perceptron classifier (Neural Network), Support Vector Machine (SMO) classifier, K-Nearest Neighbor (IBK) classifier, Decision Table classifier, J48 (Decision Tree) classifier, Random Forest classifier, Random Tree classifier, and Logistic Regression classifier respectively

- ✓ Training samples was done by using the training data to train the models, and then testing the performance of models using the testing samples. At the end, the output of the classification results was gotten.

4.2.4.1: Splitting Dataset principle

The purpose of splitting data into different categories is to avoid over-fitting. To Kumar, (2021) dividing dataset into three parts to avoid over-fitting and model selection bias. He further suggested that training set should be the largest, where cross-validation set or development set and testing set can share the same percentages. It was observed that at times researchers omitted test set and divided the dataset into just two parts usually called the development set and the test set. In doing this, we try to build a model upon training set followed by trying to optimize hyper-parameters on the development test as much as ready; we try and evaluate the training set. As also suggested by Kumar (2021), if the size of dataset is between 100 and 1,000,000 we can split the dataset into the ratios of 60:20:20 which means 60% to training set, 205 to the development set and 20% to the testing set but if the dataset is greater than 1 million datasets then we can split the dataset into the ratios 98:1:1 or 99:0.5:0.5. Joseph and Vakayil, (2021) opined that the training set can be divided into multiple sets and the model can then be trained using cross-validation. The method of splitting dataset into training and testing dataset is not an easy task, so it should be done by applying the method repeatedly on the training set.

Data splitting or train-test split is the partioning of data into subsets for model training and evaluation separately (Weng, 2021). The dataset of 30,805 could be split into 80% of training and 20% of validation set. The dataset can be divided into two parts, with different ratios such as 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10 train/test split in which the training dataset is used to construct the model whereas, the testing dataset is used to assess the model's predictive capability (Nguyen, et al., 2021). Splitting dataset into series of categories seemed easy but needs a careful procedure taken by researchers because the size of the datasets and the train/test split ratios can greatly affect the outcome of the models and as well as the classification performance itself. Hence, data splitting is commonly used in machine learning to split data into a train, test, or validation set. This splitting approach makes the researcher to find the model hyper-parater and also estimate the generalization performance. (Birba, 2020)

4.3: Results and Analysis

4.3.1: Dataset Demographic Information

Table 11: Gender Difference of the students in the dataset

		Frequency	Percent	Cumulative Percent
Valid	Male	212	69.7	69.7
	Female	92	30.3	100.0
	Total	304	100.0	

The table 11 showed the gender difference of the students in the dataset. 212 (70%) were male while 92(30%) were female students

Table 12: Examination Type of the students in the dataset

		Frequency	Percent	Cumulative Percent
Valid	WAEC	295	97.0	97.0
	NECO	9	3.0	100.0
	Total	304	100.0	

The table 12 revealed the examination type of the students in the dataset. 295 (97%) of the students sat for WAEC while only 9(3%) sat for NECO examination type

Table 13: State where Examination took place by the students

		Frequency	Percent	Cumulative Percent
Valid	Lagos State	97	31.9	31.9
	Ogun State	14	4.6	36.5
	Oyo State	139	45.7	82.2
	Osun State	20	6.6	88.8
	Ondo State	15	4.9	93.8
	Ekiti State	19	6.3	100.0
	Total	304	100.0	

The table 13 depicted the state where the students took the examination. Results indicated that 97 (32%) were from Lagos State, 14 (5%) were from Ogun State, 139 (46%) were from Oyo State, 20 (7%) were from Osun State, 15 (5%) were from Ondo State, and 19(6%) that were from Ekiti State respectively.

4.3.2: Definition of terms used in the analysis

4.3.2.1: Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the absolute value of the difference between the forecasted value and the actual value. It tells us how big of an error we can expect from the forecast on average. AME measures the average magnitude of the errors in a set of forecasts, without considering their direction. This error measures accuracy for continuous variables. It is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

4.3.2.2: Root Mean Square Error (RMSE)

To adjust for large rate errors, we calculate the Root Mean Square Error (RMSE) by squaring the errors before we calculate their mean and then taking the square root of the mean. In doing this, we arrive at a measure of the size of the error that gives more weight to the large but infrequent errors than the mean. RMSE is a quadratic scoring rule which measures the average magnitude of the error. It finds the distance between forecast and corresponding observed values are each squared and then averaged over the sample. At the end, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This implies that RMSE is most useful when large errors are particularly undesirable.

4.3.2.3: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)

The MAE and The RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE = MAE, then all the errors are of the same magnitude. The lower values of RMSE and MAE the better the accuracy of the forecast; if the predicted responses are very close to the true responses the RMSE will be small but if the predicted and true responses differ substantially, then the RMSE will be large. A value of zero would indicate a perfect fit to the data.

4.3.2.4: Relative Absolute Error (RAE)

Relative Absolute Error (RAE) is a way to measure the performance of a predictive model. It is primarily used in machine learning and data mining. It is expressed as a ratio, comparing a mean error (residual) to errors produced by a trivial or naïve model. A reasonable model (The model

that gives better result) will result in a ratio of less than one. In another words, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one (Cichosz, 2014; Hill, 2012).

4.3.2.5: Root Relative Square Error (RRSE)

The Root Relative Square Error (RRSE) is relative to what it would have been if a simple predictor had been used. Emphatically, this simple predictor is just the average of the actual values. Thus, the relative square error takes the total squared error of the simple predictor by taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted

4.3.2.6: Kappa Statistics

The value of Kappa statistics is defined as the numerator that represents the discrepancy between the observed probability of success and the probability of success under the assumption of an extremely bad case. Kappa value is a metric that compares an observed accuracy with expected accuracy (Random chance). It is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. A Kappa value of 1 represents perfect agreement, while a value of 0 represents no agreement. It is frequently used to test interrater reliability. It represents the extent to which the data collected in the study are correct representation of the variables measured

Table 14: Standard Values of Kappa Statistics

.0 - .20	None	0 – 4 %
.21 - .39	Minimal	4 – 15 %
.40 - .59	Weak	15 – 35 %
.60 - .79	Moderate	35 – 63 %
.80 - .90	Strong	64 – 81 %
Above .90	Almost perfect	82 – 100%

4.3.2.7: Confusion Matrix

A confusion Matrix is an N X N matrix used for evaluating performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. The rows represent the predicted values of the target variables. According to Chaurasia, (2021) confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true

values are known. It is also called error matrix that has two dimensions: Actual value and predicted value. A confusion matrix is a type of matrix that contains information about actual and predicted classifications done by a classification model/algorithm. The performance of such model/algorithm is commonly enhanced using the data the matrix (Kohavi and Provost, 1998). Sharma, (2019) defined a confusion matrix as a measure used for summarizing the performance of a classification algorithms. Calculating a confusion matrix will give the researcher an idea where the classification model is right and what type of error it is making. Mohajon, (2020) put that a confusion matrix is a tabular way of visualizing the performance of a prediction model. Each, in a confusion matrix denotes the number of predictions made by the model where it classified the classes correctly or incorrectly. The row of confusion matrix indicates the true class; the column means the classifier's output. Each entry, gives the number of instances/observations of row that were classified as column

Table 15: Summary of Measure Terms in Classification Models

Accuracy	It is how close a measured value to the actual (True) value Accuracy = (TP+TN)/Total
Precision	It is how close the measured value are to each other Precision = TP/Predicted Yes
Recall	It is the ratio of all correctly predicted positive predictions Recall = TP/Actual Yes
Error Rate	It is calculated as the number of all incorrect predictions divided by the total number of the datasets. The best error rate is 0.0, while the worst error rate is 1.0 Error Rate = 1- Accuracy = (FN+FP)/Total

4.3.2.8: Precision

Precision is a metric that quantifies the number of correct positive predictions made. It evaluates the fraction of correct classified instances among the ones classified as positive. The value between 0.0 for no precision and 1.0 for full or perfect precision; the formula:

Precision = True Positive/ (True Positives + False Positives)

4.3.2.9: Recall

Recall is a type of metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed

positive predictions. For imbalance learning recall is typically used to measure the coverage of the minority class. The value between 0.0 for no recall and 1.0 for full or perfect recall. The formula:

Recall = True Positive / (True Positive +False Negative)

4.3.2.10: F-Measure

Classification accuracy is widely used because it is one single measure used to summarize model performance. So, F-measure provides a way to combine both precision and recall into a single measure called f-measure that captures both properties. Once precision and recall have been calculated for binary or multiclass classification problems, the two scores can be combined into the calculation of F-Measure. Just like precision and recall, a poor f-measure score is 0.0 and the best or perfect f-measure score is 1.0. Formula:

F-Measure = (2* Precision * Recall) / (Precision + Recall)

4.3.3: Analysis based on the use of Naïve Bayes' Classifier

Table 16: Showing Different Split Ratios with values of Various Statistics (**Naïve Bayes' Classifier)**

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	280	290	84	101
Incorrectly Classified Instances	24	14	4	5
Accuracy	92%	96%	96%	96%
Kappa statistic	0.88	0.93	0.93	0.93
Mean absolute error (MAE)	0.06	0.05	0.05	0.05
Root mean squared error (RMSE)	0.18	0.16	0.16	0.16
Relative absolute error (RAE)	13.66 %	10.78%	10.95%	12.06%
Root relative squared error (RRSE)	38.59 %	33.36%	34.45%	33.41%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error

(RRSE). Results showed that the model built was highly accurate across the different split ratios used, 10-fold cross-validation had 92% accuracy, while evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 had the same 96% accuracy. Also, the Kappa statistic value supported the findings with 0.88 for 10-fold cross-validation and 0.93 for all other splits (evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30); the closeness to 1 indicated perfect agreement. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE and MSE had values from 0.05 – 0.06 errors, and 0.16 – 0.18 errors, which indicated good accuracy of the forecast; RAE and RRSE had range values from 11% - 14% errors and 33% errors - 39% errors, this implied that the model was a good forecasting type(See Table 16 for the details,).

Table 17: Detailed Accuracy by class for Naïve Bayes' Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
0.968	0.970	0.969	Highly Qualified for any sciences
0.890	0.866	0.878	Qualified for less challenging sciences
0.901	0.923	0.912	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
0.975	0.993	0.984	Highly Qualified for any sciences
0.992	0.873	0.928	Qualified for less challenging sciences
0.908	1.000	0.952	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
0.944	1.000	0.971	Highly Qualified for any sciences
1.000	0.857	0.923	Qualified for less challenging sciences
0.938	1.000	0.968	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
0.926	0.970	0.947	Highly Qualified for any sciences
0.975	0.885	0.928	Qualified for less challenging sciences

0.960	1.000	0.980	Qualified with special consideration

Table 17 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 17, it could be inferred that all values generated were very close to 1.0, Precision had range values between 0.890 to 1.000, recall had between 0.857 to 1.000, and f-measure had between 0.878 to 0.984; which signified that the model was highly perfect or the best.

Table 18: Confusion Matrix for Naïve Bayes' Classifier

10-Fold Cross-Validation			
a	b	c	Classification
98.31	3.02	0	a = Highly Qualified for any sciences
3.23	87.78	10.33	b = Qualified for less challenging sciences
0	7.79	93.54	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
100.58	0.76	0	a = Highly Qualified for any sciences
2.58	88.42	10.33	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
21.93	0	0	a = Highly Qualified for any sciences
1.29	23.24	2.58	b = Qualified for less challenging sciences
0	0	38.97	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
24.2	0.76	0	a = Highly Qualified for any sciences
1.94	29.69	1.94	b = Qualified for less challenging sciences
0	0	46.77	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 18 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 98.31 correctly predicted, and 3.23 incorrectly predicted; under 'b' had 87.78 correctly predicted and 3.02 incorrectly predicted; under 'c' had 93.54 correctly predicted and 10.33 incorrectly predicted , for evaluation

using all training data under ‘a’ had 100.58 correctly predicted and 2.58 incorrectly predicted, under ‘b’ had 88.42 correctly predicted and 0.76 incorrectly predicted, under ‘c’ had 101.33 correctly predicted and 10.33 incorrectly predicted, split ratio of 75:25 under ‘a’ had 21.93 correctly predicted and1.29 incorrectly predicted, under ‘b’ had 23.24 correctly predicted and 0 incorrectly predicted, under ‘c’ had 38.97 correctly predicted and 2.58 incorrectly predicted, and split ratio of using 70:30 under ‘a’ had24.2 correctly predicted and 1.94 incorrectly predicted, under ‘b’ had 29.69 correctly predicted and 0.76 incorrectly predicted, under ‘c’ had 46.77 correctly predicted and 1.94 incorrectly predicted. Note that all correct predicted classes were highlighted in “Yellow” colour in table 18.

4.3.4: Analysis based on the use of J48 Pruned Tree Classifier

Table 19: Showing Different Split Ratios with values of Various Statistics (J48 Pruned Tree Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	304	304	88	105
Incorrectly Classified Instances	0	0	0	0
Accuracy	100%	100%	100%	100%
Kappa statistic	1.00	1.00	1.00	1.00
Mean absolute error (MAE)	0.00	0.00	0.00	0.00
Root mean squared error (RMSE)	0.00	0.00	0.00	0.00
Relative absolute error (RAE)	0%	0%	0%	0%
Root relative squared error (RRSE)	0 %	0 %	0%	0%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across the different split ratios used, all the split ratios used showed (10-fold cross-validation, while evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30) had 100% accuracy. Also, the Kappa statistic value supported the findings with 1.00 for all the splits; the closeness to 1 indicated

perfect agreement. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE, MSE, RAE, and RRSE had the same value of 0.0;, this implied that the model was a good forecasting type(See Table 19 for the details,).

Table 20: Detailed Accuracy by class for J48 Pruned Tree Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
1.00	1.00	1.00	Highly Qualified for any sciences
1.00	1.00	1.00	Qualified for less challenging sciences
1.00	1.00	1.00	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
1.00	1.00	1.00	Highly Qualified for any sciences
1.00	1.00	1.00	Qualified for less challenging sciences
1.00	1.00	1.00	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
1.00	1.00	1.00	Highly Qualified for any sciences
1.00	1.00	1.00	Qualified for less challenging sciences
1.00	1.00	1.00	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
1.00	1.00	1.00	Highly Qualified for any sciences
1.00	1.00	1.00	Qualified for less challenging sciences
1.00	1.00	1.00	Qualified with special consideration

Table 20 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 20, it could be inferred that

all values generated were all equal to 1.0, this signified that the model was highly perfect or the best.

Table 21: Confusion Matrix for J48 Pruned Tree Classifier

10-Fold Cross-Validation			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0	101.33	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0	101.33	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
21.93	0	0	a = Highly Qualified for any sciences
0	27.11	0	b = Qualified for less challenging sciences
0	0	38.97	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
25	0	0	a = Highly Qualified for any sciences
0	34	0	b = Qualified for less challenging sciences
0	0	47	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 21 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under ‘a’ had 101.33 correctly predicted, and 0 incorrectly predicted; under ‘b’ had 101.33 correctly predicted and 0 incorrectly predicted; under ‘c’ had 101.33 correctly predicted and 0 incorrectly predicted , for evaluation using all training data under ‘a’ had 101.33 correctly predicted and 0 incorrectly predicted, under ‘b’ had 101.33 correctly predicted and 0 incorrectly predicted, under ‘c’ had 101.33 correctly predicted and 0 incorrectly predicted, split ratio of 75:25 under ‘a’ had 21.93 correctly predicted and 0 incorrectly predicted, under ‘b’ had 27.11 correctly predicted and o incorrectly predicted, under ‘c’ had 38.97 correctly predicted and 0 incorrectly predicted, and split ratio of using 70:30 under ‘a’ had 25 correctly predicted and 0 incorrectly predicted, under ‘b’ had 34 correctly predicted

and 0 incorrectly predicted, under 'c' had 47 correctly predicted and 0 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 21.

4.3.5: Analysis based on the use of Multilayer Perceptron (Neural Network) Classifier

Table 22: Showing Different Split Ratios with values of Various Statistics (Multilayer Perceptron (Neural Network) Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	299	303	78	88
Incorrectly Classified Instances	5	1	10	17
Accuracy	98%	100%	88%	83%
Kappa statistic	0.97	0.99	0.82	0.75
Mean absolute error (MAE)	0.03	0.01	0.07	0.12
Root mean squared error (RMSE)	0.12	0.04	0.20	0.32
Relative absolute error (RAE)	6.83%	2.06%	16.65%	27.22%
Root relative squared error (RRSE)	24.98%	8.22 %	41.97%	66.66%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across the different split ratios used, all the split ratios used showed 10-fold cross-validation had 98% accuracy, evaluation using all training data had 100%, split ratio of 75:25 had 88% accuracy, while split ratio of using 70:30 had 83% accuracy. Also, the Kappa statistic value supported the findings showed that 10-fold cross-validation had 0.97 perfect agreements, evaluation using all training data had 0.99 perfect agreements, split ratio of 75:25 had 0.82 perfect agreements, while split ratio of using 70:30 had 0.75 perfect agreements. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would

happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE and MSE had values from 0.01 – 0.12 errors, and 0.04 – 0.32 errors, which indicated good accuracy of the forecast; RAE and RRSE had range values from 2% - 27% errors and 8% errors - 67% errors, this implied that the model was a good forecasting type(See Table 22 for the details,).

Table 23: Detailed Accuracy by class for Multilayer Perceptron (Neural Network) Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
0.981	0.993	0.987	Highly Qualified for any sciences
0.992	0.955	0.973	Qualified for less challenging sciences
0.975	1.000	0.987	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
0.994	1.000	0.997	Highly Qualified for any sciences
1.000	0.994	0.997	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
0.919	1.000	0.958	Highly Qualified for any sciences
0.759	0.905	0.825	Qualified for less challenging sciences
0.980	0.800	0.881	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
0.951	1.000	0.975	Highly Qualified for any sciences
0.670	0.942	0.783	Qualified for less challenging sciences
0.980	0.667	0.793	Qualified with special consideration

Table 23 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 23, it could be inferred that all values generated were very close to 1.0, Precision had range values from 0.759 to 1.000, recall had between 0.667 and 1.000, and f-measure had between 0.783 and 1.000; which signified that the model was highly perfect or the best.

Table 24: Confusion Matrix for Multilayer Perceptron (Neural Network) Classifier

10-Fold Cross-Validation			
a	b	c	Classification
100.58	0.76	0	a = Highly Qualified for any sciences
1.94	96.82	2.58	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0.65	100.69	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
21.93	0	0	a = Highly Qualified for any sciences
1.94	24.53	0.65	b = Qualified for less challenging sciences
0	7.79	31.18	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
24.96	0	0	a = Highly Qualified for any sciences
1.29	31.63	0.65	b = Qualified for less challenging sciences
0	15.59	31.18	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 24 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 100.58 correctly predicted, and 1.94 incorrectly predicted; under 'b' had 96.82 correctly predicted and 0.76 incorrectly predicted; under 'c' had 101.33 correctly predicted and 2.58 incorrectly predicted , for evaluation using all training data under 'a' had 101.33 correctly predicted and 0.65 incorrectly predicted, under 'b' had 100.69 correctly predicted and 0 incorrectly predicted, under 'c' had 101.33 correctly predicted and 0 incorrectly predicted, split ratio of 75:25 under 'a' had 21.93 correctly predicted and1.94 incorrectly predicted, under 'b' had 24.53 correctly predicted and 7.79 incorrectly predicted, under 'c' had 31.18 correctly predicted and 0.65 incorrectly predicted, and split ratio of using 70:30 under 'a' had24.96 correctly predicted and 1.29 incorrectly predicted, under 'b' had 31.63 correctly predicted and 15.59 incorrectly predicted, under 'c' had 31.18 correctly predicted and 0.65 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 24.

4.3.6: Analysis based on the use of K- Nearest Neighbours Classifier

Table 25: Showing Different Split Ratios with values of Various Statistics (K- Nearest Neighbours Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	227	304	53	78
Incorrectly Classified Instances	77	0	35	27
Accuracy	75%	100%	61%	74%
Kappa statistic	0.62	1.00	0.43	0.62
Mean absolute error (MAE)	0.17	0.00	0.26	0.17
Root mean squared error (RMSE)	0.41	0.01	0.51	0.41
Relative absolute error (RAE)	38.78%	0.93%	58.88%	38.37%
Root relative squared error (RRSE)	86.52%	1.11 %	106.07%	84.73%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across the different split ratios used, 10-fold cross-validation had 75% accuracy, evaluation using all training data had 100% accuracy, split ratio of 75:25 had 61%, and split ratio of using 70:30 had 74% accuracy. Also, the Kappa statistic value supported the findings with 0.62 for 10-fold cross-validation and 1.00 for evaluation using all training data, 0.43 for split ratio of 75:25, and 0.62 for split ratio of 70:30; the closeness to 1 indicated perfect agreement. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will

produce a ratio greater than one. MAE and MSE had values from 0.00 – 0.26 errors, and 0.01 – 0.51 errors, which indicated fairly good accuracy of the forecast; RAE and RRSE had range values from 1% - 59% errors and 1% errors - 106% errors, this implied that the model using KNN was a fairly good forecasting type(See Table 25 for the details,).

Table 26: Detailed Accuracy by class for K- Nearest Neighbours Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
0.904	0.903	0.904	Highly Qualified for any sciences
0.579	0.873	0.696	Qualified for less challenging sciences
0.935	0.462	0.618	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences
1.000	1.000	1.000	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
0.913	0.931	0.922	Highly Qualified for any sciences
0.435	0.929	0.592	Qualified for less challenging sciences
1.000	0.200	0.333	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
0.924	0.939	0.931	Highly Qualified for any sciences
0.560	0.942	0.702	Qualified for less challenging sciences
1.000	0.500	0.667	Qualified with special consideration

Table 26 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 26, it could be inferred that all values generated were very close to 1.0, Precision had range values between 0.435 to 1.000, recall had between 0.200 to 1.000, and f-measure had between 0.333 to 1.000; which signified that the model was weakly perfect considering 75:25 and 70:30 ratios.

Table 27: Confusion Matrix for K- Nearest Neighbours Classifier

10-Fold Cross-Validation			
a	b	c	Classification
91.5	9.83	0	a = Highly Qualified for any sciences
9.68	88.42	3.23	b = Qualified for less challenging sciences
0	54.56	46.77	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0	101.33	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
20.42	1.51	0	a = Highly Qualified for any sciences
1.94	25.17	0	b = Qualified for less challenging sciences
0	31.18	7.79	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
23.44	1.51	0	a = Highly Qualified for any sciences
1.94	31.63	0	b = Qualified for less challenging sciences
0	23.38	23.38	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 27 indicated that the classifier was not as good as earlier ones used; it explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 91.5 correctly predicted, and 9.68 incorrectly predicted; under 'b' had 88.42 correctly predicted and 64.39 incorrectly predicted; under 'c' had 46.77 correctly predicted and 3.23 incorrectly predicted , for evaluation using all training data under 'a' had 101.33 correctly predicted and 0 incorrectly predicted, under 'b' had 101.33 correctly predicted and 0 incorrectly predicted, under 'c' had 101.33 correctly predicted and 0 incorrectly predicted, split ratio of 75:25 under 'a' had 20.42 correctly predicted and1.94 incorrectly predicted, under 'b' had 25.17 correctly predicted and 32.69 incorrectly predicted, under 'c' had 7.79 correctly predicted and 0 incorrectly predicted, and split ratio of using 70:30 under 'a' had 23.44 correctly predicted and 1.94 incorrectly predicted, under 'b' had 31.63 correctly predicted and 24.89 incorrectly predicted, under 'c' had 23.38 correctly predicted and 0.65 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 27

4.3.7: Analysis based on the use of Decision Table Classifier

Table 28: Showing Different Split Ratios with values of Various Statistics (Decision Table Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	303	304	88	105
Incorrectly Classified Instances	1	0	0	0
Accuracy	100%	100%	100%	100%
Kappa statistic	1.00	1.00	1.00	1.00
Mean absolute error (MAE)	0.02	0.01	0.02	0.02
Root mean squared error (RMSE)	0.04	0.01	0.02	0.02
Relative absolute error (RAE)	3.54%	2.90%	4.15%	4.56%
Root relative squared error (RRSE)	8.48%	2.90%	4.167%	4.58%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across all the split ratios used, (10-fold cross-validation, while evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30) with 100% accuracy. Also, the Kappa statistic value supported the findings with 1.00 for all the splits; all equal to 1 indicated perfect agreements. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE, and MSE had 0.01 – 0.02 and 0.01 – 0.04, while RAE, and RRSE had 3% - 5%, and 3% - 9%;, this implied that the model was a good forecasting type(See Table 28 for the details,).

Table 29: Detailed Accuracy by class for Decision Table Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences
1.000	0.994	0.995	Qualified for less challenging sciences
0.994	1.000	0.995	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences
1.000	1.000	1.000	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences
1.000	1.000	1.000	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences
1.000	1.000	1.000	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration

Table 29 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 29, it could be inferred that all values generated were between 0.994 – 1.000, this signified that the model was highly perfect or the best for the classifier used.

Table 30: Confusion Matrix for Decision Table Classifier

10-Fold Cross-Validation			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0	100.69	0.65	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification

101.33	0	0	a = Highly Qualified for any sciences
0	101.33	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
21.93	0	0	a = Highly Qualified for any sciences
0	27.11	0	b = Qualified for less challenging sciences
0	0	38.97	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
25	0	0	a = Highly Qualified for any sciences
0	34	0	b = Qualified for less challenging sciences
0	0	47	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 30 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 101.33 correctly predicted, and 0 incorrectly predicted; under 'b' had 100.69 correctly predicted and 0 incorrectly predicted; under 'c' had 101.33 correctly predicted and 0 incorrectly predicted , for evaluation using all training data under 'a' had 101.33 correctly predicted and 0 incorrectly predicted, under 'b' had 101.33 correctly predicted and 0 incorrectly predicted, under 'c' had 101.33 correctly predicted and 0 incorrectly predicted, split ratio of 75:25 under 'a' had 21.93 correctly predicted and 0 incorrectly predicted, under 'b' had 27.11 correctly predicted and o incorrectly predicted, under 'c' had 38.97 correctly predicted and 0 incorrectly predicted, and split ratio of using 70:30 under 'a' had 25 correctly predicted and 0 incorrectly predicted, under 'b' had 34 correctly predicted and 0 incorrectly predicted, under 'c' had 47 correctly predicted and 0 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 30

4.3.8: Analysis based on the use of Support Vector Machine (SVM) Classifier

Table 31: Showing Different Split Ratios with values of various Statistics (Support Vector Machine (SVM) Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	282	283	84	101
Incorrectly Classified Instances	22	21	4	4
Accuracy	93%	93%	95%	96%

Kappa statistic	0.89	0.90	0.92	0.93
Mean absolute error (MAE)	0.24	0.24	0.23	0.23
Root mean squared error (RMSE)	0.30	0.30	0.29	0.29
Relative absolute error (RAE)	53.51%	53.38%	51.95%	51.27%
Root relative squared error (RRSE)	63.55%	63.32%	61.16%	60.06%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across the different split ratios used, 10-fold cross-validation and evaluation using all training data had 93% accuracy, split ratio of 75:25 had 95%, while split ratio of 70:30 had 96% accuracy. Also, the Kappa statistic value supported the findings with 0.89 for 10-fold cross-validation and 0.90 for evaluation using all training data, 0.92 for split ratio of 75:25, while split ratio of 70:30 had 0.93; the closeness to 1 indicated perfect agreement. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE and MSE had values from 0.22 errors, and 0.30 errors, which indicated good accuracy of the forecast; RAE and RRSE had range values from 51% - 54% errors and 60% errors - 64% errors, this implied that the model was a good forecasting type(See Table 31 for the details,).

Table 32: Detailed Accuracy by class for Support Vector Machine (SVM) Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
0.944	0.963	0.953	Highly Qualified for any sciences
0.957	0.822	0.884	Qualified for less challenging sciences

0.892	1.000	0.943	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
0.950	0.963	0.956	Highly Qualified for any sciences
0.957	0.834	0.892	Qualified for less challenging sciences
0.897	1.000	0.946	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
0.919	1.000	0.958	Highly Qualified for any sciences
1.000	0.833	0.909	Qualified for less challenging sciences
0.938	1.000	0.968	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
0.906	1.000	0.951	Highly Qualified for any sciences
1.000	0.865	0.928	Qualified for less challenging sciences
0.960	1.000	0.980	Qualified with special consideration

Table 32 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 32, it could be inferred that all values generated were between 0.892 – 1.000 for precision, 0.822 – 1.000 for recall , and 0.884 – 0.980 for f-measure respectively, this signified that the model was highly perfect or the best for the classifier used.

Table 33: Confusion Matrix for Support Vector Machine (SVM) Classifier

10-Fold Cross-Validation			
a	b	c	Classification
97.55	3.78	0	a = Highly Qualified for any sciences
5.81	83.26	12.26	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
97.55	3.78	0	a = Highly Qualified for any sciences
5.16	84.55	11.62	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification

22	0	0	a = Highly Qualified for any sciences
2	23	3	b = Qualified for less challenging sciences
0	0	39	c = Qualified with special consideration
Split 70.0% train, 30% test			
a	b	c	Classification
24.96	0	0	a = Highly Qualified for any sciences
2.58	29.04	1.94	b = Qualified for less challenging sciences
0	0	46.77	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 33 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 97.55 correctly predicted, and 5.81 incorrectly predicted; under 'b' had 83.26 correctly predicted and 3.78 incorrectly predicted; under 'c' had 101.33 correctly predicted and 12.26 incorrectly predicted , for evaluation using all training data under 'a' had 97.55 correctly predicted and 5.16 incorrectly predicted, under 'b' had 84.55 correctly predicted and 3.78 incorrectly predicted, under 'c' had 101.33 correctly predicted and 11.62 incorrectly predicted, split ratio of 75:25 under 'a' had 22 correctly predicted and 2 incorrectly predicted, under 'b' had 23 correctly predicted and 0 incorrectly predicted, under 'c' had 39 correctly predicted and 3 incorrectly predicted, and split ratio of using 70:30 under 'a' had 24.96 correctly predicted and 2.58 incorrectly predicted, under 'b' had 29.04 correctly predicted and 0 incorrectly predicted, under 'c' had 46.77 correctly predicted and 1.94 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 33.

4.3.9: Analysis based on the use of Random Forest Classifier

Table 34: Showing Different Split Ratios with values of various Statistics (Random Forest Classifier)

Test mode:	10-fold cross-validation	Evaluate on training data	Split 75.0% train, 25% test	Split 70.0% train, 30% test
Correctly Classified Instances	303	304	87	97
Incorrectly Classified Instances	1	0	1	8
Accuracy	100%	100%	99%	92%
Kappa statistic	1.00	1.00	0.98	0.88
Mean absolute error (MAE)	0.04	0.01	0.10	0.09
Root mean squared error (RMSE)	0.11	0.03	0.19	0.18

Relative absolute error (RAE)	10.45%	1.41%	21.67%	20.45%
Root relative squared error (RRSE)	25.19%	5.35%	38.73%	37.48%

The analysis was done considering different split ratios such as 10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30 against number of correct classified observations/instances, number of incorrect classified observations/instances, Accuracy value, Kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Results showed that the model built was highly accurate across the different split ratios used, 10-fold cross-validation and evaluation using all training data had 100% accuracy, split ratio of 75:25 had 99%, while split ratio of 70:30 had 92% accuracy. Also, the Kappa statistic value supported the findings with 1.0 for 10-fold cross-validation and evaluation using all training data, 0.98 for split ratio of 75:25, while split ratio of 70:30 had 0.88; the closeness to 1 indicated perfect agreement. Four error measures were used to further test the accuracy of the model as well as the performance of the model built. In order to find the differences between forecasted value and the actual value and to adjust for large rate errors; MAE and RMSE were used. From literatures, the lower values of RMSE and MAE the better the accuracy of the forecast. Likewise, to measure the performance of the predictive model and to see what would happen when a simple prediction was used; RAE and RRSE were used. From studies, a good forecasting model will produce a ratio close to zero while a poor model will produce a ratio greater than one. MAE and MSE had values from 0.01 – 0.10 errors, and 0.03 – 0.19 errors, which indicated good accuracy of the forecast; RAE and RRSE had range values from 1% - 22% errors and 5% - 39% errors, this implied that the model was a good forecasting type(See Table 34 for the details,).

Table 35: Detailed Accuracy by class for Random Forest Classifier

10-Fold Cross-Validation			
Precision	Recall	F-Measure	Class
0.994	1.000	0.997	Highly Qualified for any sciences
1.000	0.994	0.997	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Evaluate on training data			
Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Highly Qualified for any sciences

1.000	1.000	1.000	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 75.0% train, 25% test			
Precision	Recall	F-Measure	Class
0.944	1.000	0.971	Highly Qualified for any sciences
1.000	0.952	0.976	Qualified for less challenging sciences
1.000	1.000	1.000	Qualified with special consideration
Split 70.0% train, 30% test			
Precision	Recall	F-Measure	Class
0.975	1.000	0.987	Highly Qualified for any sciences
0.809	0.981	0.886	Qualified for less challenging sciences
1.000	0.833	0.909	Qualified with special consideration

Table 35 depicted the results from precision, recall, and the f-measure of the analysis across all the four split ratios (10-fold cross-validation, evaluation using all training data, split ratio of 75:25, and split ratio of using 70:30). From the past studies and evidences, it was established that the value between 0.0 for no precision, no recall, and no f-measure while 1.0 for full or perfect precision, recall, and f-measure respectively. By inspecting the table 35, it could be inferred that all values generated were between 0.809 – 1.000 for precision, 0.833 – 1.000 for recall , and 0.886 – 0.980 for f-measure respectively, this signified that the model was highly perfect or the best for the classifier used.

Table 36: Confusion Matrix for Random Forest Classifier

10-Fold Cross-Validation			
a	b	c	Classification
101.33	3.78	0	a = Highly Qualified for any sciences
0.65	100.69	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Evaluate on training data			
a	b	c	Classification
101.33	0	0	a = Highly Qualified for any sciences
0	101.33	0	b = Qualified for less challenging sciences
0	0	101.33	c = Qualified with special consideration
Split 75.0% train, 25% test			
a	b	c	Classification
21.93	0	0	a = Highly Qualified for any sciences
1.29	25.82	0	b = Qualified for less challenging sciences
0	0	38.97	c = Qualified with special consideration
Split 70.0% train, 30% test			

a	b	c	Classification
25	0	0	a = Highly Qualified for any sciences
1	33	0	b = Qualified for less challenging sciences
0	8	39	c = Qualified with special consideration

The confusion matrix (Matrices) in Table 36 explained the number of classes that were correctly and incorrectly predicted. This was a step further to prove that the model built was perfect for the study. From the findings, 10-fold cross-validation under 'a' had 101.33 correctly predicted, and 0.65 incorrectly predicted; under 'b' had 100.69 correctly predicted and 3.78 incorrectly predicted; under 'c' had 101.33 correctly predicted and 0 incorrectly predicted , for evaluation using all training data under 'a' had 101.33 correctly predicted and 0 incorrectly predicted, under 'b' had 101.33 correctly predicted and 0 incorrectly predicted, under 'c' had 101.33 correctly predicted and 0 incorrectly predicted, split ratio of 75:25 under 'a' had 21.93 correctly predicted and 1.29 incorrectly predicted, under 'b' had 25.82 correctly predicted and 0 incorrectly predicted, under 'c' had 38.97 correctly predicted and 0 incorrectly predicted, and split ratio of using 70:30 under 'a' had 25 correctly predicted and 1 incorrectly predicted, under 'b' had 33 correctly predicted and 8 incorrectly predicted, under 'c' had 39 correctly predicted and 0 incorrectly predicted. Note that all correct predicted classes were highlighted in "Yellow" colour in table 36.

4.4: Analysis of Questionnaire Instrument

4.4.1: Demographic Information for Questionnaire Data

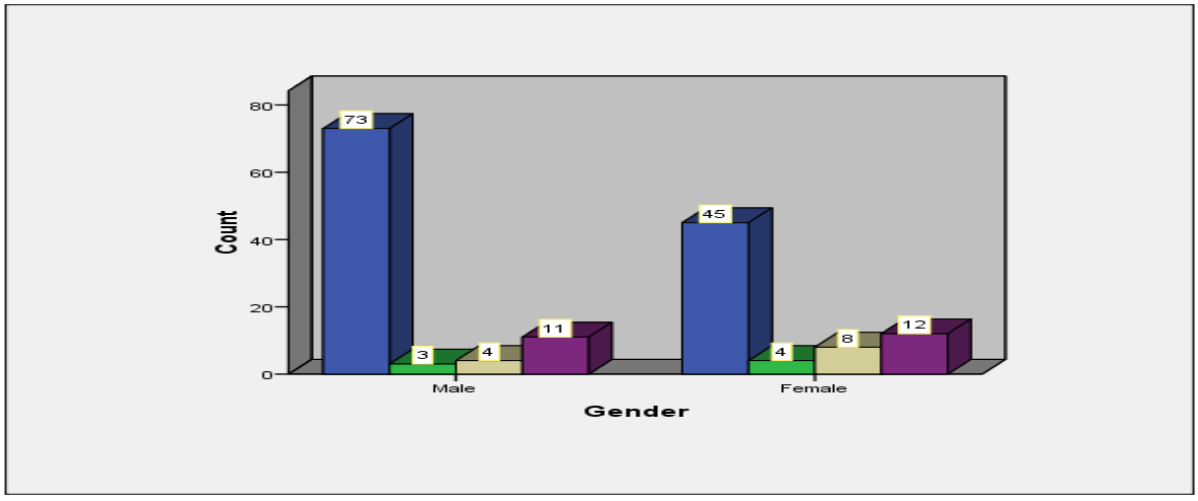

Figure 9: Choice of Course of study Influential Bar Chart

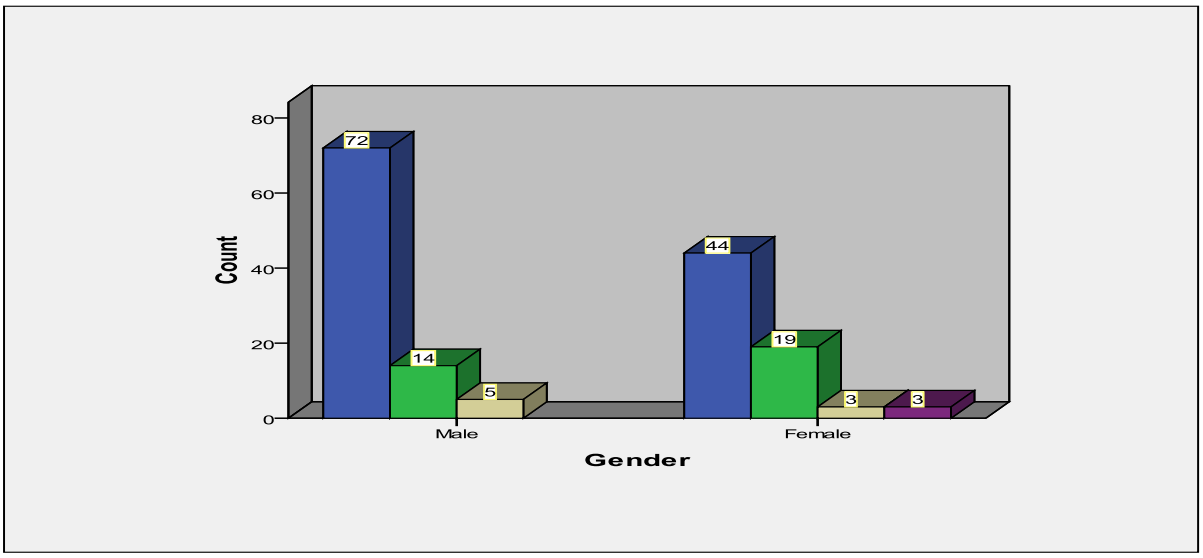

Figure 10: Satisfaction with Course of Study bar Chart

Gender demographic information of respondents was shown in figure 9 and 10 based on choice of course of study influential as well as satisfaction with the course of study. It showed that student that selected their courses of study based on their interest were the most among male and female respondents (73 males, 45 females), likewise highest number of males and females were satisfied with their choices made with 72 and 44 counts of male and female respondents. Note that: Personal Interest is in Blue, Peers/friends is in Green, Parental Influence is in Brown, and Personal Performance is in Purple.

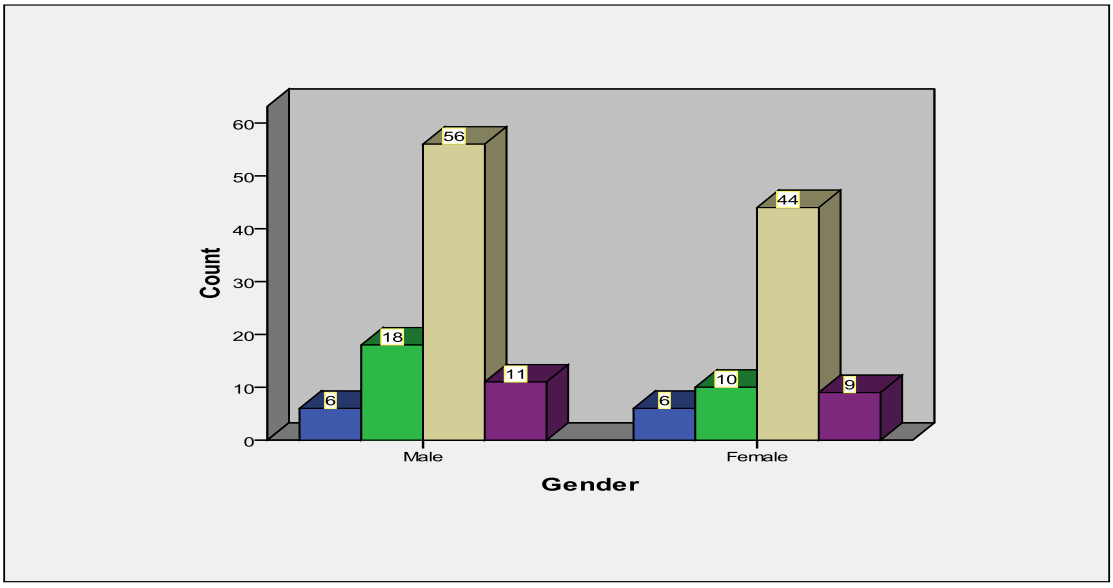

Figure 11: Any Opportunity to Change Course of Study bar Chart

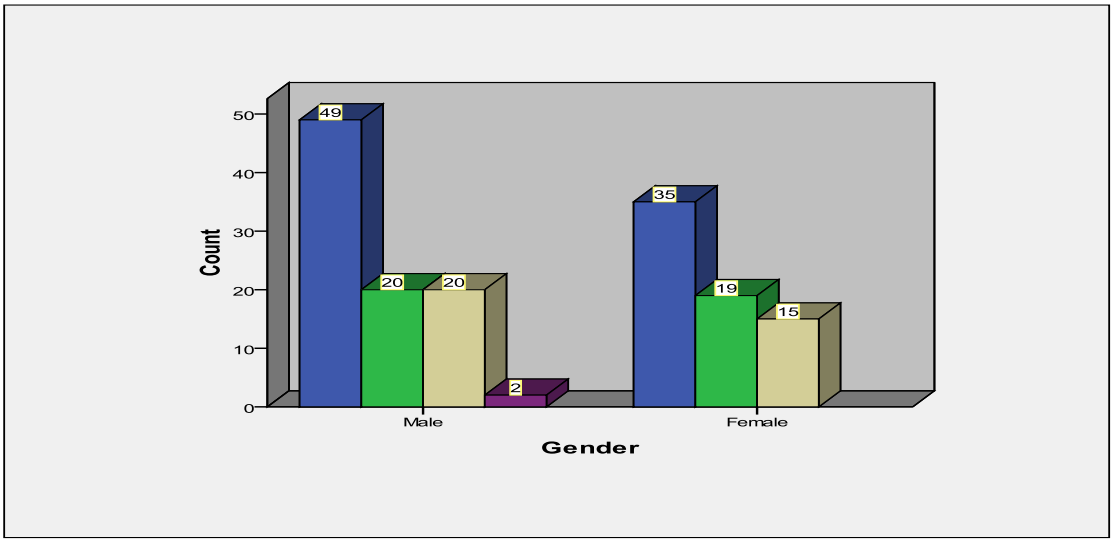

Figure 12: Rating the Challenges Encountered towards Course of Study Selection bar Chart

Gender demographic information of respondents shown in figure 11 and 12 was based on if there was any opportunity for the students to change their initial courses of study and the challenges they mostly encountered in the act of selecting relevant courses of study. It showed that most students said that they could not changed their courses again that they had used to the course selected for both male and female respondents (56 males, 44 females). In the same vein, it was revealed from figure 4 that majority of the respondents rated challenges faced in the selection of course of study high with 49 and 35 for both male and female respondents. Note that for figure 3:

Yes, I will change immediately is in Blue, Yes, but I cannot change again is in Green, No, I am getting used to it is in Brown, and No, because of time wasting is in Purple. For figure 4: Highly challenging is in Blue, Partially challenging is in Green, Somehow challenging is in Brown, and Not challenging is in Green respectively.

4.4.2: The Use of Partial Least Squares Structural Equation Modeling (PLS-SEM)

The partial least squares path modeling or partial least squares structural equation modeling (PLS-PM; PLS-SEM) is a method for structural equation modeling that gives an estimation of complex cause-effect relationships in form of path models with latent variables. In the same vein, SmartPLS was used which finds the relationship between the observed variables or indicators and the latent variables which is called "Measurement Models or Outer Models" whereas, the relationship structure between the latent variables of the model is regarded as the "Structure Models or Inner Models". The major direction of using PLS-SEM is to predict unobserved variables via the input values of instances (Hair et al, 2014). SmartPLS is one of the prominent software for implementing partial least squares structural equation modeling (Wong, 2013). Sarstedt (2019) opined that PLS-SEM has become a popular tool for analyzing relationships between variables (Observed and latent). The PLS-SEM is mainly used to estimate complex cause-effect relationship models with latent variables as the most silent research methods across a variety of disciplines (Capeda-Carrion, et al, 2019). PLS-SEM is a method used according to Rönkkö and Evermann, (2013); Sarstedt; Hair; Ringle; Thiele; & Gudergan, (2016), to estimate path models with latent variables and extend the principal component and canonical correlation analysis in statistics. Partial Least Squares is a powerful tool that has capability to adapt with the use of a very small sample sizes than structural equation modeling (SEM) for the similar size and model complexity, and also more robust to easily specify formative constructs (Jackson, 2003; MacKenzie; Podsak.; Podsako, 2011). In the formative model, it was necessary to assess the indicator weights and loads, and perform redundancy analyses. Chin (1998) in his study provided redundancy analysis, in which each formatively specified construct correlated with its alternative measure. The SmartPLS software has been used severally by scholars to solve cause-effect problems of different domains, Rungle, Wende & Becker, (2015) used it because of its embedded graphical user-interface that could be used to estimate the PLS-SEM models. The first PLS software was published nearly ten years after LISREL III (Lohmöller, 1989). Huang, (2021) categorized PLS-SEM as a good predictive model methods that aims at determining the relationship between variables

PLS-SEM is mainly designed to detect whether the causal relationship has a statistically significant mutual linear relationship. It is rather suitable for the construction of theoretical models. The use of PLS-SEM as a method to explore the relationship between the research variables with the help of PLS Algorithm and Bootstrapping to perform the repetitive sampling "n" times in order to derive path coefficients and significance as was discussed in Henseler & Chin, (2010). It can also handle the correlation and influence between the dimensions (Hou ,Lo & Lee , 2020). What is called PLS-SEM does not require the data to have a normal distribution, since it uses a non-parametric bootstrap technique to test the significance of the coefficients (Valls, Martín-Cervantes, Sánchez & Martínez, 2021)

Table 37: Showing Bootstrapping Results

Bootstrapping	P-Value
Parent al Influence➔ Age	**0.016****
Parental Influence ➔ Gender	0.407
Peers/friends Influence ➔ Age	**0.000****
Peers/friends Influence ➔ Gender	0.177
Personal Interest ➔ Age	**0.000****
Personal Interest ➔ Gender	0.272
Personal Performance ➔ Age	0.479
Personal Performance ➔ Gender	0.616

The result from table 37 indicates that Parental Influence has direct relationship with the student's age, likewise the students' peers/friends is a strong factor that can influence the student's courses of study selection based on the age of such student at the time of selection, also student's use of personal interest to make choice has impact on the age of the student. Others do not have any significant relationship with the variables

Table 38: Showing Q-Square Values

Variables	Q-Square Values
Age	0.50**
Gender	-0.01

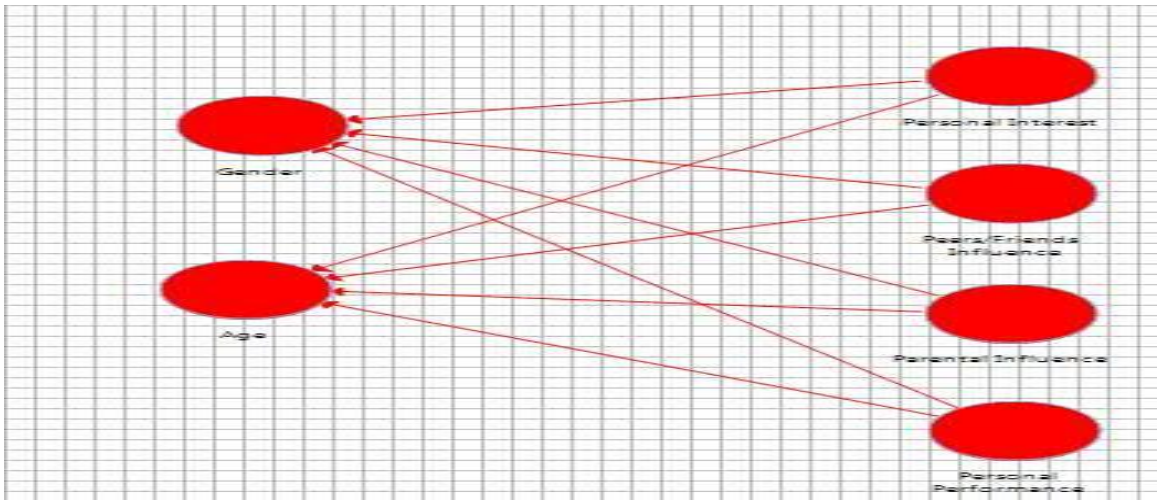

Figure 13: Hypothesized/hypothetical Path Model

In order to propose the model for the cause and effect of problems attached to course of study selection by the students, a hypothetical model path was created as seen in Figure 13. It showed the flow of observed and intent variables

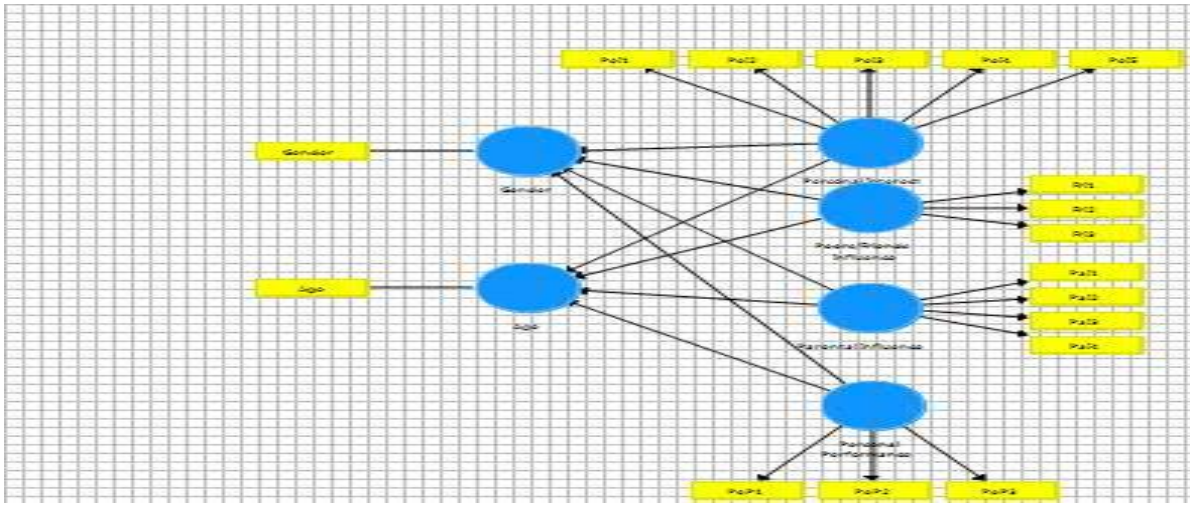

Figure 14: Initial Path Model

The figure 14 showed complete construct of the path model using all the intent variables. This is an initial path model

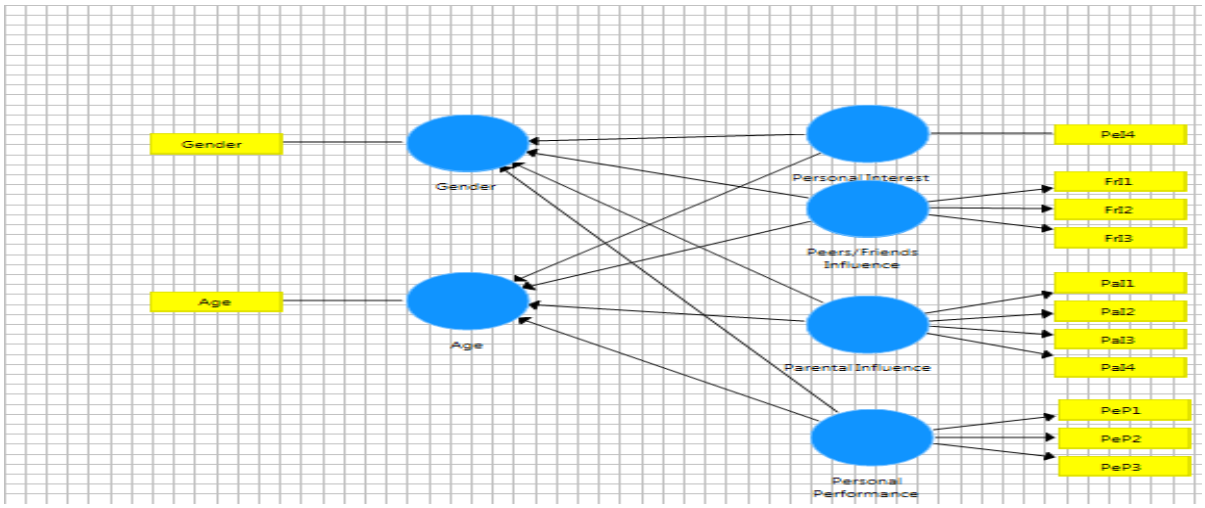

Figure 15: Semi-Final Path Model

In the next figure as can be seen from figure 15, four intents variables were dropped to enhance the model path. This was done because they did not contribute significantly to the model

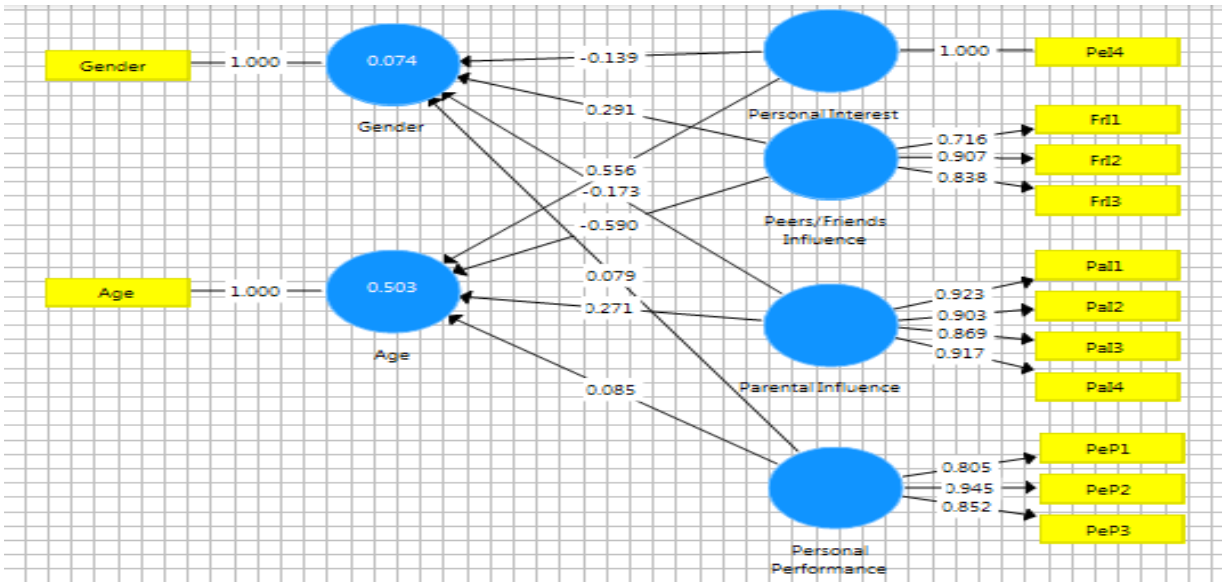

Figure 16: Final Path Model

At the end, fig.4 depicted the final path model for the study with coefficient values between observed variables (Gender and Age) and intent variables (Personal Interest, Peers/Friends Influence, Parental Influence, and Personal Performance). In complementing the analysis detailed, figure 17 –figure 19 were presented to establish the r square, f-square, and the path coefficients already displayed in the models.

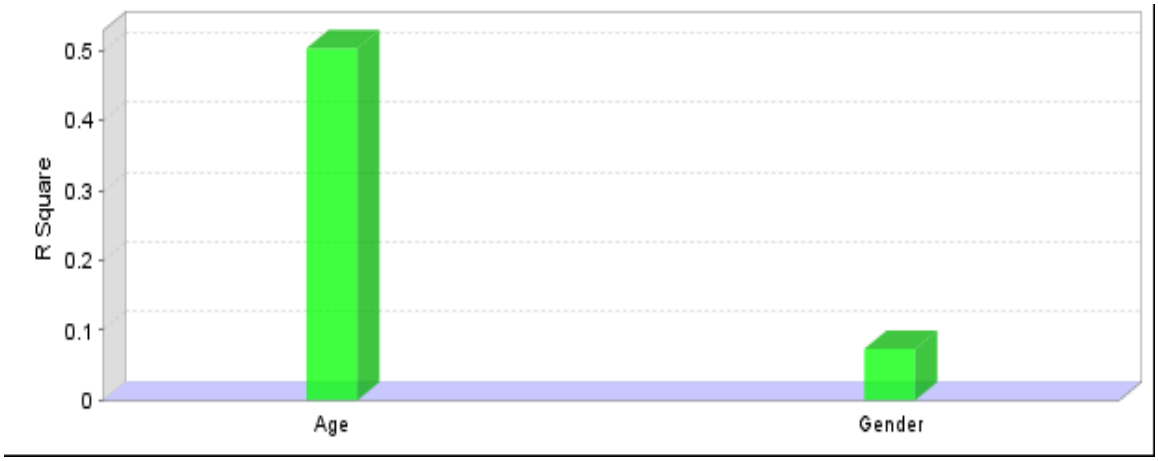

Figure 17: R-square plot

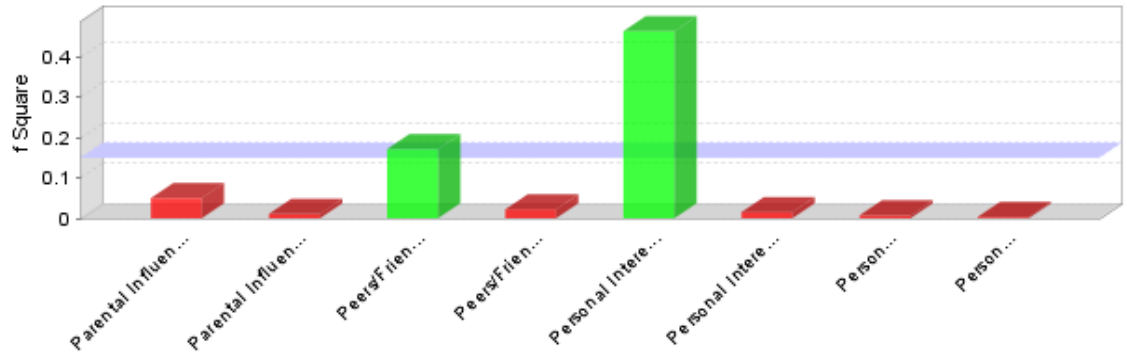

Figure 18: F-square plot

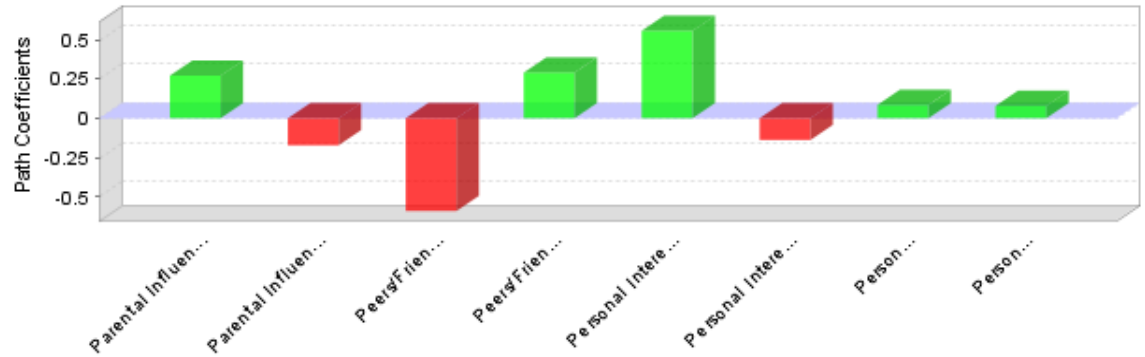

Figure 19: Path Coefficient plot

4.4.3: Analysis of Interview Instrument

Thematic analysis method was used which always used for analyzing qualitative data such as interview in form of audio, video recordings that entails searching across a dataset to identify, analyze, and report repeated patterns (Braun & Clarke, 2006). It is a method for describing data, but it also involves interpretation in the processes of selecting codes and constructing themes (Kiger & Varpio, 2020).

4.4.3.1: Introduction

An interview was conducted on phone asking selected participants questions relating to how the participants selected their courses when they wanted to enter higher institutions.

4.4.3.2: Audience

Current students in higher institutions as well as those that have graduated were interviewed for this study. 20 students formed the interviewees

4.4.3.3: Coding

The participants interviewed were anonymously coded as P1 to P20 (Participant 1 to participant 20) not to directly mentioned their names. Three questions were structured for the interviewees which are stated below:

Question1: How did you select the course you studied/studying and was it based on your personal interest or parental influence or peers/friends influence or personal performance?

Question2: Do you ever regret of being read/reading the course?

Question3: Do you think selection of course of study is brain storming and confusion exercise?

4.4.3.4: Report and Interpretation

Question1: How did you select the course you studied/studying and was it based on your personal interest or parental influence or peers/friends influence or personal performance?

In answering this first question, the P1, P2, said that they made their selection with the guidance from their parents. P6, P9, P17, P19, and P20 claimed that their parents chose the course for them. One of them said “My Dad chose the course for me , because I was very young then to decide which was based on my past performances, I think” , while some of participant were of the claim that their selection were based on their personal interest with indirect or direct guidance of their parents. In contrary, P2, P6, P8 said that their selection or choice was done at

the entry into the institutions through their tutors or peers. The last four participants P6, P13, P15, and P18 chose their courses based on their personal performance right from their primary school days

Question2: Do you ever regret studying/reading the course?

To this question, 85% of the total sampled for the interviewed (P1, P3, P4, P5, P7, P8, P9, P10, P11, P12, P13, P14, P15, P17, P18, P19, and P20) were of opinion that they read the courses they were interested in; 10% of the sampled participants (P2 and P16) claimed that initially they did not like the course given to them but later developed very strong interest in the courses. The remaining 5% of the participants (P6) was only of the opinion that he regreted of being among the graduate of the course he read with the reason that the course was forced on him and eventually performed bad in the course.

Question3: Do you think selection of course of study is brain storming and confusion exercise?

The response to question3 showed that all the participants agreed and believed that selection or making choice of course of study was a brain storming and confusion exercise especially when they were very young to decide independently.

CHAPTER FIVE

Discussion, Conclusion, and Recommendations

5.1: Introduction

This chapter of the thesis discusses the findings and results obtained from research covering all three division analyses adopted in this study. The major concerned areas include the use of data science techniques to improve the prediction capacity of students selection of relevant courses of study by getting into higher institutions, to find out the impact or effect of personal interest, peers/friends influence, parental influence, and personal performance of students towards selection of courses with respect to moderating factors like gender and age, and qualitative analysis that showed the opinion or responses of sampled participants on how students' interest, peers/friends, parents and performances influenced their selection of courses and the challenges facing them in carrying out such actions. The section also, explained the essence of the use of different classifiers for accurate predictions with support from previous studies.

5.1.1: Machine Learning (Data science) Techniques Analysis with Dataset Discussion

To start with, the current study dataset was subjected to principal component analysis (PCA) which is mostly used by data scientists and analysts to summarize the information content in large data tables by means of a smaller set or summary indices that can be more easily visualized or analyzed. The use of PCA, is a technique in data science to reduce the dimensionality of datasets, to increase the interpretability and at same time with no loss of any information as we have it in the original datasets (Jollife & Cadima, 2016). In the use of PCA, Kaiser-Meyer-Olkin Measure of Sampling Adequacy was used and got a desirable result of 0.651(Which is greater than .6), at the same time Bartlett's Test of Sphericity was significant with .000 (Which is less than 0.01/0.05). For proper prediction model to be achieved, the dataset was partitioned into two: Train/Test ratios. This was done to avoid over-fitting of the variables used. In support of this, Kumar, (2021) claimed that dividing dataset into three parts is to avoid over-fitting and model selection bias. As also suggested by Kumar (2021), if the size of dataset is between 100 and 1,000,000 we can split the dataset into the ratios of 60:20:20 which means 60% to training set, 205 to the development set and 20% to the testing set but if the dataset is greater than 1 million datasets then we can split the dataset into the ratios 98:1:1 or 99:0.5:0.5. Joseph and Vakayil, (2021) opined that the training set can be divided into multiple sets and the model can then be trained using cross-validation. The dataset of 30,805 according to Nguyen, et al., (2021) could be

split into 80% of training and 20% of validation set. The dataset can be divided into two parts, with different ratios such as 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10 train/test split in which the training dataset is used to construct the model whereas, the testing dataset is used to assess the model's predictive capability (Nguyen, et al., 2021). Hence, the dataset was split into full training data; 10-fold cross-validation, 70/30, and 75/25 were used. From the findings, 70/30 and 75/25 performed well under Naïve Bayes' Classifier, all splits performed perfectly on J48 Pruned Tree Classifier, 10-fold cross-validation and evaluate on training set performed exceptionally on Multilayer Perceptron (Neural Network) Classifier, evaluate on training data, 10-fold cross-validation, and 70/30 performed good on K- Nearest Neighbours Classifier, All splits performed accurately on Decision Table Classifier, likewise all splits performed wonderfully and accurately on Support Vector Machine (SVM) Classifier, and at the time all splits performed better on Random Forest Classifier. It showed that all the classifiers used were good algorithms/models to predict students' relevant courses of study on getting into higher institution. This is in support of Nguyen, et al., (2021) opined that the dataset can be divided into two parts, with different ratios such as 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, and 90:10 train/test split in which the training dataset is used to construct the model whereas, the testing dataset is used to assess the model's predictive capability.

5.1.2: Partial Least Squares Path Modeling with the use of Questionnaire Discussion

The partial least squares path modeling or partial least squares structural equation modeling (PLS-PM; PLS-SEM) was used to analyze the questionnaire instrument. This approach was used for structural equation modeling that gives an estimation of complex cause-effect relationships in form of path models with latent variables; concentrated on the use of SmartPLS software. This method was employed to predict unobserved variables via the input values of instances (Hair et al, 2014). SmartPLS is one of the prominent software for implementing partial least squares structural equation modeling (Wong, 2013). PLS-SEM was designed, in this study, to detect whether the causal relationship has a statistically significant mutual linear relationship. Its use was mainly to explore the relationship between the research variables with the help of PLS Algorithm and Bootstrapping to perform the repetitive sampling "n" times in order to derive path coefficients and significance and also to handle the correlation and influence between the dimensions (Hou ,Lo & Lee , 2020). The results showed that Parental Influence has direct relationship with the student's age, likewise the students' peers/friends is a strong factor that can influence the student's courses of study selection based on the age of such student at the time of

selection, also student's use of personal interest to make choice has impact on the age of the student. Others do not have any significant relationship with the variables. The Q^2 vale > .5 supported the findings as well as r^2, f^2 and coefficient path plots. The results of questionnaire analyzed with PLS-SEM complemented that age of the students' was a determinant to parental, peers/friends influence to such students while gender did not contribute statistically significant to the selection of courses towards entering higher institutions.

5.1.3: Thematic Analysis with the use of Interview Discussion

The results from findings showed that parents of the students chose the course for them. One of them said "My Dad chose the course for me , because I was very young then to decide the decision was based on my past performances, I think" , while some of participant were of the claim that their selection were based on their personal interest with indirect or direct guidance from their parents. In contrary, a few students interviewed said that their selection or choice were done at the entry into the institutions through their tutors or peers. The last four participants claimed that they chose their courses based on their personal performance right from their primary school days. In answering the question on whether they regretted choosing the course of study or not 85% of the total sampled for the interviewed were of opinion that they read the courses they were interested in; 10% of the sampled participants claimed that initially they did not like the course given to them but later developed very strong interest in the courses. The remaining 5% of the participants was only of the opinion that he regretted of being among the graduate of the course he read with the reason that the course was forced on him and eventually performed bad in the course. On the last question asked, the interviewee responses to question showed that all the participants agreed and believed that selection or making choice of course of study was a brain storming and confusion exercise especially when they were very young to decide independently. The findings buttressed the earlier findings on how data science techniques could be used to predict students relevant of courses on getting into higher institutions

5.2: Conclusion

In this thesis, the ensemble learning method was used to build the model and predict the students' relevant courses of study on getting into higher institutions. It was noted that selection of courses using machine learning algorithms or models showed high predictive accuracy of all the machine learning models used with some minor variances. It was also observed that age of the students during the time of course selection were determined from the influence of parents,

peers, and other people around such students. In the same vein, personal perceptions of people towards course selection was that female students were always guided more than male counterparts but these assertions were not statistically proved right. Decision making on course of study by students is a very technical issue that needs experts which at time might be students' parents, teachers or admission officers of the institutions to avoid wrong selection of course which may result to unsatisfactory programme or drop out from school or total failure of the students in the programme.

5.3: Recommendations

The following recommendations were provided:

1. Professional counselors should always guide students towards selection of relevant courses of study on getting into higher institutions
2. Parents duty towards students' selection of relevant courses of study should be based on interest of the students not by coercing them
3. Ability of a child should be monitored right from childhood to easily predict relevant course of study by such child before getting out of hands
4. Curriculum upgrading from time to time to incorporate the new technologies that will enhance the adoption of proper models for prediction
5. Provision of series of training that will include the solution of courses of study selection problems from elementary schools to higher institutions

References

Accenture. (2016). Digital Health Technology Vision 2016. *Accenture* .

Ali, R. (2020, August Sunday). *Predictive Modeling: Types, Benefits, and Algorithms.* Retrieved August Sunday, 2021, from Net Suite: http://www.netsuite.com

Alvani, S. M. (2009). *General management.* Tehran: Ney publication.

Andayan, i. S., & Mardapi, D. (2012). Andayani S, Mardapi D, editors. Performance Assessment Dalam Perspektif Multiple Criteria Decision Making. . *International Journal of Engineering & Technology* , 118.

APADictionaryofPsychology. (2021, August Tuesday). *APADictionaryofPsychology*. Retrieved August Tuesday, 2021, from APADictionaryofPsychology: http://www.apadictionaryofPsychology

Athey, S. (2017). Beyond Prediction: Using Big Data for Policy Problems. *Science* , 483-485.

Banafa, A. (2014). *What is Data Science?* Retrieved July Saturday, 2021, from works.bepress.com: http://works.bepress.com/ahmed-banafa/15/

Begicevic, N., Divjak, B., & Hunjak, T. (2011). AHP-based group decision making using keypads. *International Journal of Economics and Business Research. 3 (4). doi:10.1504/IJEBR.2011.040953.* , 443.

Bell, J. (2020). *Machine learning: hands-on f or developers and technical professionals.* . John Wiley & Sons.

Bilgi, N. B., kulkarni, R. V., & Spenser, C. (2010). An expert system using of Decision Logic charting Approach For Indian legal Domain with Specific reference to transfer of property Aet. *International Journal of Artificial Intelligence and Expert system (IJAE), Volume (1): issue (2)* .

Birba, D. E. (2020, December Thursday). *A Comprehensive Study of Data Splitting Algorithms for Machine Learning Model Selection.* Retrieved December Thursday, 2021, from Machine Learning Model Selection: http:///www.google.com

Blum, A. (2021). *Machine Learning Theory.* Department of Computer Science Carnegie Mellon University.

Bowne-Anderson, H. (2018, October Monday). *What Data Scientists really do, According to 35 Data Scientists.* Retrieved October Monday, 2021, from HBR: http://www.hbr.org

Bradley, R. (2014). *Decision Theory: A Formal Philosophical Introduction.* London: London School of Economics and Political Science.

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qual Res Psychol. 3(2)* , 77 - 101.

Cambridge_Dictionary. (2021, October Friday). *Meaning of Data*. Retrieved October Friday, 2021, from Cambridge Dictionary: http://www.dictionary.cambridge.org/data

Cambridge-Dictionary. (2021, October Friday). *Source Credibility.* Retrieved October Friday, 2021, from Cambridge Dictionary: http://www.dictionary.cambridge.org

Cao, I. (2017). Data science: challenges and directions. *The Communications of ACM* , 59 – 68.

Casano, A. (2021, October Friday). *Position statement: definition and examples.* Retrieved October Friday, 2021, from Study: http://www.study.com

Cave, W. C. (2020). *Prediction theory of control systems.* 309 Morris Avenue – Suite J Spring Lake.

Cepeda-Carrion, G., Cegarra-Navarro, J., & Cillo, V. (2019). Cepeda-Carrion, G; Cegarra-NavarrTips to Use Partial Least Squares Structural Equation Modeling (PLS-SEM) in Knowledge Management. *Journal of Knowledge Management (23)1* .

Chaurasia, P. K. (2021, December Thursday). *Confusion Matrix.* Retrieved December Thursday, 2021, from mgcup.ac.in: http://www.mgcup.ac.in

Cheever, M. A., Allison, J. P., Ferris, A. S., Finn, O. J., Hastings, B. M., Hecht, T. T., et al. (2009). The Prioritization of Cancer Antigens: A National Cancer Institute Pilot Project for the Acceleration of Translational Research. *Clinical Cancer Research. 15 (17). doi:10.1158/1078-0432.CCR-09-0737. PMC 5779623. PMID 19723653.* , 5323–5337.

Chen, Z. (2005). *Consensus in group decision making under linguistic assessments:*. Manhattan: Kansas State University Manhattan.

Chin, W. W. (1998). The partial least squares approach for structural equation modelling. In G. E. Marcoulides, *Modern Methods for Business Research.* USA: Hillsdale, NJ.

Cichosz, P. (2014). *Data Mining Algorithms: Explained Using R.* John Wiley Sons.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *ISI Review* , 21–26.

Danielson, M., Ekenberg, L., Johansson, J., & Larsson, A. (2004). *The Decideit Decision Tool* . Sweden: Mid Sweden University.

Davis, B. (2014, June Friday). *What does confirmability mean in qualitative research?* . Retrieved October Friday, 2021, from M V Organizing: hhtp://www.mvorganizing.org

Dhar, V. (2013). Data Science and Prediction. *Commun. ACM* , 64-73.

Donoho, D. (2015). *50 years of Data Science". unpublished.* . Retrieved July Saturday, 2021, from dl.dropboxusercontent.com: https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience

Druva. (2021, October Friday). *Data Archiving Definition*. Retrieved October Friday, 2021, from Druva: http://www.data.text

Geiser, S., & Santelices, M. V. (2007). *VALIDITY OF HIGH-SCHOOL GRADES IN PREDICTING STUDENT SUCCESS BEYOND THE FRESHMAN YEAR: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes*.* BERKELEY: Research & Occasional Paper Series: CSHE.6.07.

Giarratano, J. C. (2007). *Principles and programming Fourth* . India: Indian Edition Thomson.

Hair, F. J., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Hair, F. J; Sarstedt, M; Hopkins, L & Kuppelwieser, V. G (2014). Partial Least Squares Structural EquationAn Emerging Tool in Business Research. *Hair, F. J; Sarstedt, M; Hopkins, L & Kuppelwieser, V. G (2014). Partial LeaEuropean Business Review.(26)2* , 106-121.

Hayashi, C. (1998). What is Data Science ? Fundamental Concepts and a Heuristic Example. In Y. K. Hayashi C., *Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization.* Tokyo: Springer.

Hayez, Q., De-Smet, Y., & Bonney, J. (2012). D-Sight: a new decision making software to address multi-criteria problems. . *International Journal of Decision Support System Technology* .

Hill, A. (2012). *The Encyclopedia of Operations Management: A Field Manual and Glossary of Operations Management Terms and Concepts.* FT press.

Hsing-Yu Hou, Y.-L. L., & Chin-Feng, L. (2020). Hsing-Yu Hou , Yu-Lung Lo and Chin-Feng Lee (2020). Predicting Network Behavior Model of E-Learning Partner Program in PLS-SEM. *Journal of Applied Science* , 4656.

Huang, C. -H. (2021). Using PLS-SEM Model to Explore the Influencing Factors of Learning Satisfaction in Blended Learning. *Educ. Sci.* , 249.

Import.io. (2018, June Monday). *What is Data, and why is it important?* Retrieved from Data Text: http://www.data.text

INFORMS. (2021, August Monday). *Operation Research and Analytics.* Retrieved August Monday, 2021, from INFORMS: http://www.informs.org

Ishizaka, A., & Labib, A. (2009). Analytic Hierarchy Process and Expert Choice: Benefits and limitations . *ORInsight, 22(4)* , 201-220.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the n:q hypothesis. *Jackson, D.L. (2003). Revisiting sample size and number of parameter esStruct. Equat. Model. Multidiscip. J.* , 128–141.

Janssens, C. J., & Martens, F. K. (2018). *Prediction Research : An Introduction.* Atlanta: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Jolliffe, I. T., & Cadima, J. (2016, December Thursday). *Jolliffe, Ian T and Cadima, Jorge (2016). Principal Component Analysis: A rReview and Recnt Developments.* . Retrieved December Thursday, 2021, from RSTA: https://doi.org/10.1098/rsta.2015.0202

Joseph, V. R., & Vakayil, A. (2021, December Thursday). *Split: An Optimal Method for Data Splitting.* Retrieved December Thursday, 2021, from Split: An Optimal Method for Data Splitting: https://doi.org/10.1080/00-401706.2021.1921037

Khademi, A. (1993, August Monday). *DSS.* Retrieved August Monday, 2021, from Familiarity to decision support system concepts: Management knowledge,: http://www.decisionsupportsystem.net

Khodashahri, N. G., & Sarabi, M. M. (2013). Decision Support System. . *Singaporean Journal of Business, Economics and Management Studies Vol.1, No.6* , 95-102.

Kiger, M. E., & Varpio, L. (2020). Thematic Analysis of Qualitative Data: AMEE Guide. Medical Teacher. *Medical Teacher.*

Kohavi, R. P. (1998). *Kohavi, R & Provost, F (1998). Glossary of terms, Machine Learning – Spcial Isuue on Applications of Machine Learning and the Knowledge Discovery Process.*

Retrieved from Kohavi, R & Provost, F (1998). Glossary of terms, Machine Learning – Spcial Isuue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning, 30.https://doi.org/10.1023/A:1017181826899: https://doi.org/10.1023/A:1017181826899

Kryssanov, V. V., Abramov, V. A., Fukuda, Y., & Konishi, K. (1997). *A Decision-Making Support System Based On Know-How.* Tokyo, 203 Japan.

Kumar, S. (2020, December Thursday). *Data Splitting Techniques to fit any Machine Learning Model.* Retrieved December Thursday, 2021, from Towards Data Science: Kumar, Sachin (2020). Data Splitting Techniques to fit any Machine Learning Model. www.towardsdatascience.com

Lohmöller, J. B. (1989). *Latent Variable Path Modeling with Partial Least Squares.* Germany: Physica: Heidelberg.

Loon, R. V. (Composer). (2021). Relationship between AI,ML and Data Science. [SimpliLearn, Performer, & Simplilearn, Conductor] online.

Macharis, C., Bernardini, A., De-Smet, Y., & Hayez, Q. (2010). PROMETHEE in a multi actors setting: the use of the Multi actor Multi Criteria Analysis (MAMCA) methodology with D-SIGHT. *In the proceedings of the OR52 conference.* London .

Mahmoodi, B. (2003). *Information Systems Role in Crisis Management.*

Marimin, P. T. (2004). *Aplikasi Pengambilan Keputusan Kriteria Majemuk. Jakarta:* . Indonesia: Gramedia Widiasarana Indonesia.

Mason, H., & Wiggins, C. (2010, August Monday). *Towards Data Science.* Retrieved August Monday, 2021, from Taxonomy of Data Science: http://www.towardsdatascience.com

McLeod, S. (2018, October Friday). *Questionnaire: Definition, Examples, Design and Types.* Retrieved October Friday, 2021, from Simply Psychology: http://www.simplypsychology.org

McMillian, J. H. (2008, January Friday). *Interview.* Retrieved October Friday, 2021, from Educational Research: Fundamentals for the Consumer: http://www.researchgate.net

Merriam-Webster. (2021, October Friday). *Definition of Data*. Retrieved October Friday, 2021, from A Datum: http://www.merriamwebster.com/data

MIDAS. (2017). *About MIDAS* . Retrieved July Saturday, 2021, from midas.umich.edu: http://midas.umich.edu/about/

Mihal, A. (2016, January Monday). *Data, Collection of Data, Organization of Data, Presentation of Data, Analysis of Data.* Retrieved October Monday, 2021, from Slide Share: http://www.slideshare.net

Mitchell, M. (2019, August Sunday). *Selecting the Correct Predictive Modelling Technique.* Retrieved August Sunday, 2021, from Towards Data Science: http:/www.towardsdatascience.com

Mohajon, J. (2020, December Thursday). *Confusion Matrix for your Multi-Class Maching Learning Mode.* Retrieved December Thhursday, 2021, from Towards Data Science: http:// www.towardsdatascience.com

Moleke, P. (2004). *Employment experiences of graduates. Employment and economic policy research.* Abuja: Human Science Research Council.

Mosavi, A. (2015, August Sunday). *Predictive Decision Making.* Retrieved August Sunday, 2021, from Research Gate: http:/www.researchgate.net

Mu, E., & Butler, B. S. (2009). The Assessment of Organizational Mindfulness Processes for the Effective Assimilation of IT Innovations. *Journal of Decision Systems. 18. doi:10.3166/JDS.18.27-51. ,* 27.

Mujthaba, G. M., Abdalla, A.-A., Manjur, K., & Mohammed, R. (2020). Data Science Techniques, Tools and Predictions. *International Journal of Recent Technology and Engineering, 8(6)* .

Muraina, I. O., Rahman, M. A., Adeleke, I. A., & Aiyegbusi, A. E. (2013). Use of a Rule Tool in Data Analysis Decision Making. *Information and Knowledge Management Vol.3, No.3* , 131-142.

Nguyen, Q. H., Ly, H., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, Y. Q., et al. (2021). Nguyen, Q H; Ly, H; Ho, L S; Al-Ansari, N; Le, H V; Tran, Y Q; Prakash, Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problem in Engineering* , 1-15.

Nilsson, N. J. (2014). *Principles of artificial intelligence.* Morgan: Kaufmann.

Ny, C. J. (2021, October Friday). *Common Ethicai Issues in Research and Publication* . Retrieved October Friday, 2021, from NEBI: http://www.nebi.nlm.nih.com

O'keefe, D. (1990). *Pasuation: Theory and Research.* Newbury ParkCA: SAGE.

Osmar, R. Z. (1999). Principles of Knowledge Discovery in Databases. *CMPUT690: University of Alberta* , 15.

Pitan. (2010). *Assessment of skills mismatch among employed university.* Ibadan: University of Ibadan.

Pitan, O. S., & Adedeji, S. O. (2014). Students' choice of courses: Determining factors, sources of information, and relationship with labour market demands in Nigeria. *Africa Educattion Review* , pp. 445-458.

Pitan., O. S. (2010). *Assessment of skills mismatch among employed university graduates in Nigeria labour market.* Ibadan: University of Ibadan.

Plewes, S. (2019, August Thursday). *4 Components of a Data ScienceProject.* Retrieved August Thursday, 2021, from Macadamian: http://www.macadamian.com

PRC. (2021). *Mobile Fact Sheet*. Retrieved August Wednesday, 2021, from data:text: http://www.data:text/html:charset=utf-8;base64,

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data. *N* , 51-59.

Provost, F., & Fawcett, T. (2013). *Data science and its relationship to big data and data-driven decision making: Big Data.* China: liebertpub.

Racz, A., Bajusz, D., & Heberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecles, 26* , 1111.

Ramageri, B. M. (2011). Data Mining Techniques and Applications. . *Indian Journal of Computer Science and Engineering Vol. 1 No. 4* , 301-305.

Roger, R. F., & Marek, J. D. (2007). *Decision Support Systems Encyclopedia of Library and Information Science.* Pittsburgh: University of Pittsburgh.

Rönkkö, & Evermann. (2013). Organ. Res. Methods. *Appl. Sci.* , 182–209.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach.* Malaysia: Pearson Education Limited.

Saaty. (1996). *Decision Making with Dependence and Feedback: The Analytic Network Process,*. Pennsylvania: Pittsburgh, Pennsylvania: RWS Publications, ISBN 0-9620317-9-8.

Saaty, T. L. (2006). *How to Make a Decision Scientifically: The Analytic Hierarchy Process.* Pittsburgh: University of Pittsburgh | Pittsburgh, PA, USA.

Saaty, T. L. (1980). *The Analytic Hierarchy Process.* New York: McGraw-Hill: New York.

Saaty, T. L., & Peniwati, K. (2007). *This is an Excerpt from Group Decision Making: Drawing Out and Reconciling Differences.* . Decision Lens, p34-64.

SAGE. (2014, October Friday). *Transferability in research.* Retrieved October Friday, 2021, from SAGE: http://www.sage.com

Sandelowski. (2000, october Friday). *Determining the Key success Factors that Influence the Brand Loyality.* Retrieved October Friday, 2021, from Repository: http://www.repository.nwu.ac.za

Sarstedt, M., & Cheah, J. H. (2019). SarstPartial Least Squares Structural Equation Modeling Using SmartPLS: A Software Review. *Journal of Marketing Analysis,7* , 196-202.

Sarstedt, M., Hair, J. F., Ringle, C. M., & Thiele, K. O. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Gudergan J. Bus. Res.* , 3998–4010.

Sartorius. (2020, December Thursday). *What is Principal Component Analysis (PCA) and How it is Used?* Retrieved December Thursday, 2021, from sartorius: http://www.sartorius.com

Schmelzer, R. (2020, August Tuesday). *Research Business Analytics.* Retrieved August Tuesday, 2021, from 15 common data science techniques to know and use: http://www.researchbusinessanalytics.techtarget.com

Selerity. (2021, August Sunday). *Types of Predictive Analytics Model and how they work.* Retrieved August Sunday, 2021, from Selerity: http://www.seleritysas.com

Shamova, A. E., & Resnik, B. R. (2013). *Responsible conduct of research.* USA: Oxford University Press.

Sharma, P. (2019, December Thursday). *Decoding the Confusion Matrix,.* Retrieved December Thuesdat, 2021, from Towards Data Science: www.towardsdatascience.com

Shum, S. B., Hall, W., Keynes, M., Bake, R. S., Behrens, J. T., Hawksey, M., et al. (2013). *Educational Data Scientists: A Scarce Breed* . Retrieved July Saturday, 2021, from simon.buckinghamshum.net: http://simon.buckinghamshum.net/wp content/uploads/2013/03/LAK13Panel-Educ Data Scientists.p

Sihag, P. (2019, September Monday). *Scientific Method for Data Analysis that can also be applied to other aspect of life.* Retrieved October Monday, 2021, from Medium: http://www.medium.com

SimpliLearn. (2021, July Saturday). Data Science.

Simplilearn. (2021, July Monday). *What is Data:Types of Data and how to Analyze Data?* . Retrieved October Monday, 2021, from Simpli Learn: http://www.simplilearn.com

Subahi, A. F. (2018). Data Collection for Career Path Prediction Based on Analysing Body of Knowledge of Computer Science Degrees. *Journal of Software , 13* (10), 533-546.

Technical-Report-15T-009. (2021). *Data Science and Analytics.* Lehigh: Department of Industrial and System Engineering Lehigh Univrsity.

Thomas, H. D., & Patil, D. J. (2012, October). Data Scientist the sexest Job of the 21st Century. *Harvard Business Review* .

Tobin, G., & Begley, C. (2004, October Friday). Mthodological rigor within a qualitative framework. *Journal of Advanced Nursing* , 388-396.

Ullivan, M. (2020, August Monday). *Towards Data Science.* Retrieved August Monday, 2021, from Evolution of Data Science: What is next?: http://www.towardsdatascience.com

USCLibraries. (2021). Organizing your social sciences research paper. *Research Guides* .

Valls Martínez, M., Martín-Cervantes, P., Sánchez Pérez, A., & Martínez Victoria, M. (2021). An Assessment with Partial Least Squares Structural Equation Modeling. *Valls Martínez,*

M.d.C.; Martín-Cervantes, P.A.; Sánchez Pérez, A.M.; Martínez Victoria, M.d.C. Learning Mathematics of Financial Operations during the COVID-1Mathematics , 9.

WAC. (2021, October Friday). *Generalizability and Transferability* . Retrieved October Friday, 2021, from WAC: http://www.wac.colostate.edu

Weistroffer, H. R., Smith, C. H., & Narula, S. C. (2005). Multiple Criteria Decision Analysis. In J. Figueira, S. Greco, & M. Ehrgott, *State of the Art* (pp. 989-1018). Springer.

Weng, J. (2021, Decmber Thursday). *Data Splitting for Model Evaluation.* Retrieved December Thursday, 2021, from Towards Data Science: http://www.towardsdatascience.com

Wong, K. K. (2013, December Thursday). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Techniques Using SmartPLS.* Retrieved December Thursday, 2021, from Research Gate: http://www.researchgate.net

Xu, J., Moon, K. H., & Van, D. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in SignalProcessing* , 742-753.

Zhu, Y., & Xiong, Y. (2015). *Defining Data Science.* China: Fudan University.

Zocco, D. (2010). Risk Theory and Student Course Selection. *Research in Higher Education Journal – Volume 3* , 1-29.