Awarding Body: **Arden University**
Programme Name: **Data Analytics and Information System Management**
Module Name (and Part if applicable): **Research Project**
Assessment Title: **A Recommendation of Machine Learning Model to Improve Nigerian E Learning Platforms for Student Performance Monitoring**
Student Number: **STU101620**
Tutor Name: **Mohammed Nour Albaarini**
Word Count: **15450**

Declaration

I hereby declare that I carried out the work reported in this thesis in the School of Computing, Arden University, under the supervision of Mohammed Nour Albaarini. I also solemnly declare that to the best of my knowledge no part of this report has been submitted here or elsewhere in a previous application for the award of a degree. All sources of knowledge used have been duly acknowledged.


(Signature and date)….27/06/2021….

ANUCHA SHAMMAH (STU101620)

Acknowledgement

I want to thank my project supervisor, Mohammed Nour Albaarini, for being very patient and supportive through the difficult times in the process of this project. Thank you to my loving mother Elizabeth Asemota for your moral support and giving me life. Thank you, Ebubechi Onyemachi, for all your love and support throughout the phases of this project. Also, I want to thank my colleague Lucain Pouget, a Senior Data Scientist, for all the advice and for putting me through this project. Finally, I want to thank God Almighty for the gift of life, wisdom and understanding.

Abstract

E-learning has made it easy for teachers to share study materials in various media such as video, text, and audio with students. Many schools have implemented the use of E-learning platforms to aid the transfer of this electronic media. In Nigeria, several E-learning platforms have been built and are currently being used by different schools. However, these platforms are not intelligent. As more students join the platform, a lot of data is being generated making it difficult to keep track of student activities. Hence, a system needs to be created to assist teachers in monitoring the students' performance. Using Artificial Intelligence and Machine Learning, an application can be made smart because it can learn from the history of previous data and make decisions by itself. In this research, a machine learning algorithm was proposed to predict student performance at the end of a study year. The dataset used was developed by Open University consisting of 32,593 students. The research project was governed by the CRISP-DM methodology to ensure that no step was overlooked. A multinomial classification technique was needed to predict four classes: "Distinction", "Fail", "Pass" and "Withdrawn". The chosen machine learning algorithm was Artificial Neural Network. The model was trained and had an accuracy of 86%. XGBoost and Logistic Regression algorithms were also tested but the ANN model performed best. The trained model was deployed to the web using Streamlit. The application will serve as a useful tool in the hands of the teachers as it will give them insights into the students will perform so that they can improve their methods of teaching.

Keywords: Machine Learning, Artificial Neural Networks, CRISP-DM, E-learning, Model Deployment.

Table of Contents

Tables

Table of Figures

Chapter 1 - Introduction

1.1 Background

E learning is defined as a framework consisting of electronic devices and technology that allows learning to take place even from a remote location (Moubayed *et al.*, 2018). With E-learning, it is possible to share visual, audio, and text content for the learners to use. The spread of information and technology round the world has increased the popularity of distance learning (Moubayed *et al.*, 2018). With the advent of COVID 19, more people have willingly adopted learning virtually. The huge growth of e-learning has created a huge repository of knowledge with unlimited access to information (Gazzawe *et al.*, 2022). This in turn has attracted an audience of various backgrounds (Gazzawe *et al.*, 2022). The increasing number of audiences on these E-learning platforms makes it difficult for tutors to monitor the engagements of learners. Hence, it is important for these E-learning platforms to have some form of intelligence that enables the tutors to easily monitor the activities of the learners so that they take actions to improve their mode of teaching and motivate the learners (Hussain *et al.*, 2018).

According to (Mishra and Tyagi, 2022), it is very hard for smart systems to make intelligent decisions without the inclusion of Artificial Intelligence (AI) and Machine Learning (ML). Machine learning has caused a disruption in various industries including health, manufacturing, transportation, education and so on. Artificial Intelligence is designed to behave like or totally replace human activities (Shaikh *et al.*, 2022). IT companies like Google, Microsoft, Amazon, and Facebook have integrated AI and Machine Learning in their businesses (Shaikh *et al.*, 2022). Artificial Intelligence and Machine Learning has found its way into institutions. Even Bill Gates also believes that AI and ML will improve human learning in different ways (Shaikh *et al.*, 2022). There are so many possibilities with machine learning in the educational sector. Classification and forecasting are two major applications of machine learning. With student performance assessments, schools can help failing students to improve their grades.

Educational Data Mining models are methods that are developed for exploring large data from educational systems (Albreiki *et al.*, 2021). These unique models have played a vital role in improving the learning system. Researchers have discovered many opportunities to advance technology that provides solutions to schools and to meet the needs of students. The EDM process involves exploring, researching and the application of Data Mining (DM) methods. With the help of techniques from multiple disciplines such as Machine Learning and statistics, analysis of the student data can be done and significant patterns from historical data can be (Albreiki *et*

al., 2021). This knowledge can be useful for predicting the performance of new students (Arcinas, 2022). Insights like knowing the students that are diligently participating in school activities can be derived (Arcinas, 2022). And with a system powered by machine learning, these functionalities can be made possible (Arcinas, 2022).

Educational institutions today are very competitive and have become very complex. Therefore, it has become very challenging to analyze student performance, make high-quality education available, develop strategies for evaluating student performance, and recognize what will be needed in the future (Albreiki *et al.*, 2021). It is very important that educational institutions have a plan that will intervene when students are facing challenges during their studies. Prediction of student performance is best done when they enter new and at different levels, this will give the student management body useful insights to invent useful intervention plans that will benefit the management, tutors, and the students (Albreiki *et al.*, 2021).

E-learning platforms such as Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS) and Massive Open Online Courses (MOOC) have greatly utilized EDM to develop an automatic grading system, recommender systems and an adaptive system (Albreiki *et al.*, 2021). These platforms use smart tools to extract relevant information like the number of times the student accessed a material in the virtual learning environment, the results of student assessments and the total time a student spent on a lecture material. This information collected with time is processed and analyzed with various machine learning techniques to improve the usability of the platforms and develop interactive tools. According to a study done at the University of Montreal, machine learning which is a subset of AI aims to provide computers with knowledge derived from data, findings, and interactions with the world (Albreiki *et al.*, 2021). With this knowledge obtained, the computer can adjust correctly to a new scenario. Machine learning systems learn from the data, analyze the pattern formed and make predictions. The steady increase in the volume of data, less expensive storage systems, and vast computation systems have given rise to the performance of pattern recognition and deep learning techniques at a larger scale. Today, the technology has been made available to allow machine learning models to analyze big and small complex data sets accurately (Albreiki *et al.*, 2021).

E-learning platforms are cheaper and less rigid than the traditional schooling system. However, According to (Albreiki *et al.*, 2021), there are three certain drawbacks to using this system and they include: (i) inconsistency in the grading system making it difficult to benchmark with other virtual learning platforms and thus making it hard to measure how effective the platform is; (ii)

because there is no motivation and students learn at their own pace, there is higher dropout rate of students as compared to the traditional system; (iii) it is difficult to predict the specific needs of students due to the long distance and indirect or no communication between the tutor and the student.

1.2 Problem Statement

According to (Pedro *et al.*, 2019), the scope of artificial Intelligence not only involves adopting the rapidly developing technologies for learning, but it requires rethinking content and various methods to deliver teaching at different levels of education (Pedro *et al.*, 2019). While Nigeria is falling behind, countries like South Africa adopted artificial intelligence in the platform Dapito to understand the performance of the registered students and recommend relevant courses to them (Pedro *et al.*, 2019). Kenya also uses an AI technology in M-Shule, a mobile platform that adapts to the skills and competencies of each student and delivers personalized standard national curriculum to the students through SMS (Pedro *et al.*, 2019). Many researchers focus on the adoption of E-learning into Nigeria because of poor internet connectivity, unstable power supply, lack of ICT skills and high data tariff but no one is focusing on how limited the Nigerian E-learning platforms are in terms of data analytics, data-driven decision making, prediction models and automating processes. It is worth noting that the Nigerian E-learning platforms such as National Open University of Nigeria (NOUN), Roducate, 9ija kids, ULessons, Easyprep, StudyLab360 Ubongo, Nexford University, Noun, and University of Africa to name a few all have received a huge number of enrollments by students after the advent of the pandemic. However, these platforms have not been reported to have implemented machine learning to improve the learning and teaching system. National Open University of Nigeria (NOUN) which has over 600,000 students enrolled, over 90 programs and over 50 programs will greatly benefit from machine learning and data analytics if a machine learning model is implemented on the E-learning platform to perform complex analysis and predict student performance (NOUN DICT, 2022).

According to Kantigi, a data scientist in MIT, all the stakeholders of the Nigerian Education System need to embrace machine learning for improvement and to compete globally with the latest trends (Gbadebo, 2021). This will increase the efficiency of the E-learning portals by eliminating human interaction in analyzing student data and examinations making it easier to learn and teach. Many countries have advanced in the Education sector by including machine learning in their E-learning platforms, but Nigeria is still falling behind (Gbadebo, 2021). The current platforms do not

encourage a personalized learning process for the students, no content analysis and have a lot of time-consuming repetitive tasks that could easily be automated with machine learning (Gbadebo, 2021). Machine learning is very useful for handling complex data analysis and various calculations at high speed without the intervention of human beings.

The benefits of machine learning on Nigerian E-learning platforms for student performance prediction includes an upgrade in planning and improved strategies for making accurate changes in the education management system for better program learning outcomes (Brahim, 2022). Another important benefit is to quickly identify students who are having challenges with the classroom activities to improve their learning outcomes (Brahim, 2022). One way to achieve this is by tuning the model to classify students who pass with low, average, or high grades. Depending on the results, the school can take action to assist those students who are performing low (Brahim, 2022). The third benefit is to suggest better learning methods for students according to their predicted performance for even better results (Brahim, 2022). For example, the prediction of high performing students will assist the school in estimating the number of scholarships to be awarded. Last but not the least, machine learning for student performance prediction will help to reduce the dropout rate in Nigeria which in turn will increase the number of graduating students, improve quality, and increase ranking of schools (Brahim, 2022).

1.3 Aim and Objective

The aim of this research project is to improve E learning platforms in Nigeria by giving teachers insights of the behavioral patterns of students through the prediction of students' performance and hence reducing the dropout rate of students in Nigeria. This aim will be executed by delivering the following objectives:

1. To compare E learning platforms used in Nigeria and other countries and recommend artificial intelligence use cases to be implemented in the Nigerian E-learning platform.
2. To visually analyze the OULAD dataset to spot strong feature correlations with student performance.
3. To create a model of student interest to improve E-learning system personalization.
4. To train Artificial Neural Networks to predict student performance with the OULAD dataset.
5. To deploy the machine learning model to the web for easy use by teachers.

1.5 Research Questions:

1. How can we monitor the performance of students that use Nigerian E learning platforms to reduce dropout rate?
2. How can we use Machine Learning techniques to improve the Nigerian E-learning platform for better personalization?
3. How can we find possible causes of failures in the E-learning platform through visualization of the OULAD dataset?
4. What are the best methods of analyzing a dataset to achieve accurate results?

1.6 How it is Novel:

This research paper recommends a machine learning implementation to the Nigerian E learning Platforms such as Roducate, 9ija kids, ULessons, Easyprep, StudyLab360, Prepclass, Ubongo, Nexford University, Noun, and University of Africa to name a few. Most researchers stop at the training and testing of the model, but in this paper, an extra step which is model deployment is implemented for the trained model to be used in a well-designed web interface.

1.7 Dissertation Structure

In chapter 1, the goal of the project which is to mine the available data on E-learning platforms and feed it to a machine learning model to predict the outcomes of the student result was stated. A problem statement discussing what area this study is providing a solution to was provided along with the novelty of the research. In chapter 2, a literature review was given, and it focused on various related research done in Machine Learning in the Education system, E-learning and Massive Open Online Courses, Attention to the application of Data Science in the Education System and Artificial Intelligence in Predicting Student Performance. In chapter 3, the methodology of the research was given. The chapter presented the concept of the methods used in this research. The chapter gives an answer to the research questions that have been asked in this chapter. Chapter 4 presented a design of the project, the implementation of the data analysis steps presented in chapter 3, the results of the data analysis and the outcome of the model deployment on the web. The last chapter, chapter 5, summarized the findings of the research, proposed recommendations, stated the limitations of the project, and future work to be done.

Chapter 2 – Literature Review

This chapter will contain a review of previous works that have been done relating to Machine Learning, Artificial Intelligence, the Education System and how they have been implemented in E-learning platforms over the years.

2.1 Machine Learning in the Education System

According to (Jalil et al, 2019), machine learning has widely been used in the education system to improve teaching and learning experiences and reduce costs. The paper claims that every opportunity for an advancement in technology has been used to attempt replacing the traditional way of teaching and even teachers entirely. The paper concludes that as machine learning evolves, so will the education sector as it will be used by instructors to understand how students adapt in school and prevent failure. Despite the fear of machine learning ruling the world, it is very advantageous especially in higher education levels. But the problem is that there are not many data analysts or scientists to implement these machine learning tools in these universities. The following paragraphs will review journals that have implemented these machine learning tools in the education sector.

(Kotsiantis et al, 2004), carried out two experiments to determine the best learning algorithm to predict student performance in distance learning. The paper does not only focus on the accuracy of the algorithm but also how it can be an effective tool in the hands of the teachers to help the students. The data used to train the model was given by Hellenic Open University (HOU). The paper focuses on data from the *informatics* course consisting of twelve modules to obtain a bachelor's degree. The sample data of 354 students that chose the course were used in the first experiment. The information collected from each student include age, sex, residence, marital status, number of children, occupation, computer literacy, job associated with computers, attendance, and marks for the assignments given. This was done to predict if the students will fail or pass a given module. Meanwhile in the second experiment, only 28 sample data which represented the maximum number of students in a teacher's class was used. The overall aim in the paper was to provide the teachers with firsthand knowledge of each student's performance before the event of failure so that the teacher can change his approach to teaching. From the two experiments carried out, it was discovered that the best algorithm was Naïve Bayes with an overall accuracy of 72.48% in the first experiment and 66.49% in the second experiment. As fantastic as the idea of this experiment is, some problems could arise from the fact that the dataset is small. This could affect the performance of the model. One of the major problems with a small data size

is overfitting and this occurs when the model has overstudied the training dataset thereby seeing patterns that are not there leading to poor accuracy in predictions (Alencar, 2019). Another problem with this experiment is that the ratio of males to females in the school was 72/28. This also could lead to a bias in the performance of the model as the model has more instances of male than female to learn from (Telus Communications Inc, 2022). Finally, the paper does not mention anything about splitting the dataset for training and testing. It is important to set aside a portion of the dataset for testing so that the model can be tested on how well it predicts new data (Myrianthous, 2021).

(Minaei-bidgoli *et al.*, 2003) attempted to predict the performance of the students based on certain behaviors. The paper grouped the students into different categories to know how each student with a certain attribute performed. The data was extracted from two databases of the online education system, the *Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA)* that was invented in Michigan State University (MSU). The first database contained educational resources like quizzes, exams, practice questions, and examples. The second database contained information about the interaction of the students with these educational materials, not limited to how long they spent on each material, what time they opened it, how many times they were visited and so on. The two relevant research questions asked were: 1) Is there a group of students that react to the resources in the same way? and if so, how can the class of an individual be determined? 2) Can the assignments given by the lecturer be grouped into types? And if so, how can the tutors be helped to give assignments that are more effective? The research focused on 227 students enrolled in a Physics course who already had a final grade for the course. They excluded 34 students that dropped the course after taking some assignments. This information would have been relevant to knowing why and what type of students drop the course. The paper tested 3 classification techniques for the grading attributes: Pass or Fail, High, Medium, or Low, and a 9-point grading system. The paper makes use of Genetic Algorithms (GA) as it is known as a good method for data mining and pattern recognition therefore, increasing accuracy. A comparison of different classifiers was done to determine the best fit for the prediction. The table below shows the results of the experiment.

Table 1 Comparison of Different Classifiers without Genetic Algorithms (GA) (Minaei-bidgoli *et al.*, 2003)

Classifier		Performance %		
		2-Classes	3-Classes	9-Classes
Tree Classifier	C5.0	80.3	56.8	25.6
	CART	81.5	59.9	33.1
	QUEST	80.5	57.1	20.0
	CRUISE	81.0	54.9	22.9
Non-tree Classifier	Bayes	76.4	48.6	23.0
	1NN	76.8	50.5	29.0
	kNN	82.3	50.4	28.5
	Parzen	75.0	48.1	21.5
	MLP	79.5	50.9	-
	CMC	86.8	70.9	51.0

Table 2 Comparison of Different Classifiers with Genetic Algorithms (GA) (Minaei-bidgoli *et al.*, 2003)

Classifier	Performance %		
	2-Classes	3-Classes	9-Classes
CMC of 4 Classifiers without GA	83.87 ± 1.73	61.86 ± 2.16	49.74 ± 1.86
GA Optimized CMC, Mean individual	94.09 ± 2.84	72.13 ± 0.39	62.25 ± 0.63
Improvement	10.22 ± 1.92	10.26 ± 1.84	12.51 ± 1.75

For the model training, the paper only makes use of student data in the homework assignments. Other features like the quizzes, continuous assessments and the student's attendance would have increased the performance of the model. There was no mention about data cleaning to remove duplicates and noise. Like in the previous paper, the size of the data was too small which could lead to overfitting. The researchers normalized the features which is good practice to avoid uneven weight assignment. The paper does not cover how the teachers can use the information derived from the prediction to assist the students in their studies.

Lykourentzou *et al.*, (2009) applied three different combinations of machine learning (ML) techniques to predict student dropout rate. The three machine learning techniques used were Feed Forward Networks, Support Vector Machines and Probabilistic ensemble simplified fuzzy ARTMAP. The Feed-forward Neural network tries to copy the ability of the human brain to learn using a set of interconnected neurons to process information (Kubat, 1999). The success of the machine learning (ML) techniques was determined in terms of "total accuracy, sensitivity, and precision". The paper suggests splitting the dataset into training and testing sets which is considered best practice according to (Xu and Goodacre, 2018) which is called cross validation. The paper presents a two-class output; 1 representing a dropout scenario and 0 representing

completion of the program. Interestingly, the paper proposes the combination of these three ML techniques into one network using three distinct decision tactics since a single technique on its own may not be accurate. The first tactic was to map a student as dropout if one of the identified ML techniques classified the student as dropout meaning the student's level is 1 or greater than 1. The second tactic was to label the student as dropout if two or more techniques classify the student as dropout, meaning that the dropout level of the student is 2 or more. The third tactic was to identify the student as a dropout out if and only if all three techniques classify the student as a dropout out meaning that the student's dropout out level is 3.

According to Xu et al, 2017 some challenges in predicting student performance in degree programs are since students come from different backgrounds in terms of education and culture, data from the courses taken alone do not actually represent fundamental issues and that the students' development stages are not integrated into the process. The paper proposes a method to address these problems. The method has two features: a two layered architecture consisting of multiple base and cascade ensemble predictors to make predictions of the students' development processes and the second feature is the use of latent factor models and probability matrix factorization to identify and cluster based on the importance of the course to construct a base predictor. The clustering system to rate the importance of the courses applies the same principles as a recommender system (Kordik, 2018), therefore Xu et al, (2017) reports that similar challenges experienced in recommender systems were experienced in predicting the students' performance. In recommender systems, there is a problem of users rating only a small segment of items (Idrissi and Zellou, 2020) while in the course relevance cluster, the students don't register for all courses but a small set. This problem is widely recognized as sparsity challenges in recommender systems (Idrissi and Zellou, 2020). Recommender systems address this problem using latent factors to make recommendations of items based on similar preferences of other users Xu et al, (2017). The paper also addresses this challenge using latent factors to cluster relevant courses together. Another method that would have been used for recommending items to users is collaborative filtering algorithm (Thai-Nghe *et al.*, 2010). The paper focuses on predicting the GPA of students; the model can also be used to predict the general performance of the students. The simulation was done on data of 1169 undergraduates over 3 years at the department of Mechanical and Aerospace Engineering. The paper used four ML algorithms namely: linear regression, logistic regression, random forest, and K Nearest Neighbors as base predictors. The results show that random forest performed the best and KNN was reported to have the worst performance. The outcome of this paper was to provide valuable insights to tutors

about their student's performance so that the tutors can recommend courses for the students to take. However, further studies can be done to use the prediction results to create a recommender system that recommends relevant courses to the students.

2.2 E-Learning and Massive Open Online Courses (MOOCs)

According to Xing *et al.*, (2016), because of the flexibility of study time, choice of courses and where and how students learn in Massive Open Online Courses (MOOCs), it would not be best to rely on the predictions of machine learning algorithms such as decision trees and neural networks because they are unstable learners, and they are sensitive to deviations in data. Also, for stable learners such as Support Vector Machines (SVM) and Naïve Bayes, their accuracy is greatly reduced due to the imbalance in the classes and if a small training set is used. This is a problem that Xu et al, (2017) solved using course recommender systems as earlier stated that courses taken by students do not accurately reflect the fundamental issues. However, Xing *et al.*, (2016) proposed using stacked generalization, an ensemble learning approach to account for the variations. According to (Wolpert, 1992), "stacked generalization works by deducing the biases of the generalizers from a given learning set. The deduction is followed by generalizing in a second space the inputs that are guesses of the initial generalizer when taught with part of the learning set and trying to guess the rest of it, and whose output is the correct guess". Xing *et al.*, (2016) restricted their research to a project management course that was hosted by Canvas in 2014. A total of 3617 students registered for the course and the duration was 8 weeks with a total of 11 modules. Clickstream data which are digital imprints left behind by users on a Webpage (Bucklin and Sismeiro, 2009) were collected from the Canvas web page containing information on the most visited pages by the students and how many clicks per source also, information on the test scores and discussion forums were collected in JSON format. Of a total of 1667 students who participated in the discussion forum, a total of 1379 students dropped out of the course. The model was designed to predict students who were to drop out by the next week based on the data from the previous week. Log transformation was performed for each feature of the dataset to remove the skewness and outliers in the dataset (Htoon, 2020). Principal Component Analysis (PCA) which is used to reduce the dimensions of a dataset and increase the interpretation ability while still preventing loss of information (Jollife and Cadima, 2016) was used in the paper to distinguish the dropout student from the remaining students coming from the remaining weeks. The table below shows the predicted results from over 7 weeks.

Table 3 Student Dropout Over Seven Weeks and Stacking Prediction Results (Xing *et al.*, 2016)

Week	First	Second	Third	Fourth	Fifth	Sixth	Seventh
# Student	502	226	146	128	142	129	106
AUC	0.807	0.877	0.889	0.928	0.942	0.942	0.961
Precision	0.807	0.897	0.931	0.942	0.933	0.949	0.958

For this research, further studies can be done to use other feature engineering techniques and include other features like information on the background of students and the history of their learnings online including the test scores attained.

Hussain *et al.*, (2018) applied different machine learning algorithms to the Open University (OU) dataset to identify courses with low student interactions and hence predict the outcome of each student. The researchers stated that the problem of E-learning platforms is the absence of face-to-face meetings which enables teachers to physically monitor the students' engagement with online resources. The researchers also mentioned that up to 78% of students who register for online courses fail to finish it (Simpson, 2010). This raised a concern to predict the likelihood of students dropping out. Therefore, the tutors can only understand what is going on with the students based on their data on the E-learning platform The features collected include the student's highest education level, their final examination results, test scores, and clicks on the online E-learning system. The paper presents a wide range of machine learning techniques applied in this paper include Decision Tree, J48, Classification and Regression Tree (CART), JRIP Decision Rules, Gradient Boosting Trees (GBT), and Naïve Bayes Classifier (NBC). The lifecycle of the prediction process was as follows: Preprocessing, Data Training, Model building, Testing and Evaluation (DSPA, no date). The outcome of the process was to predict the student's performance using the best model. It was problematic for the researchers to prepare the data because they were in different tables, so they wrote an algorithm in MATLAB for Feature Extraction. However, a simple SQL query would have been beneficial to extract the needed features from the multiple tables into a single table (Bhiwoo, 2019). The researchers ensured to remove missing values, label appropriately and select relevant features which are all important stages in data cleaning approach to enhance the quality and accuracy of the data (Dey, 2021). A careful selection of clicks that were relevant for the performance prediction was done using Spearman's correlation in SPSS. The input data was also visualized with charts to identify any anomalies that could reduce the quality of the data (Freitas, 2002). The results of the experiment show that J48 had the highest accuracy of 88.52% while CART performed the least with 82.25% accuracy for predicting low engagement during assessments. Further studies can be done to build a recommender system to advise students with low engagement on the materials to make use of.

2.3 Attention to the Application of Data Science in the Education System

Gradient Boosting Decision Trees (GBDT), a machine learning technology for regression, classification and ranking tasks was proposed by Zeebaree *et al.*, (2021) to predict the performance of university students in the final examinations. The paper compares the performance using different algorithms: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Gradient Boosted Trees). In the paper, the stages involved in the selected method includes data collection and creation of dataset, data preprocessing to remove missing data and errors, training, and testing of the model for classification and finally implementing the model on the Weka platform. The paper follows the Cross Industry Standard Process for Data Mining (CRISP-DM) which is a well-known standard that explicitly identifies the six stages of the life cycle of data science. These six stages are: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment (DSPA, no date). The dataset comprised 450 students enrolled for a bachelor's degree and had written the final exam. The paper presents and exhaustive list of data collected including Age, College, Department, stage in school, sex, home address, home to school distance, number of family members, father's and mother's job, Father and mother's educational achievement, number of years repeated, frequency of internet use, additional work, middle school degree, fatherless or not, choice of college and label. Feature selection would be the best practice for this model. Most often than not, using all the available features in a dataset does not give the most accurate result of the model prediction which reduces the performance (Vickery, 2020). Feature selection can be done manually by removing the features that are highly correlated with each other because there are algorithms that are sensitive to correlated features (Vickery, 2020). A comparison of the four algorithms with respect to the F measure, Precision, AUC, and Recall was done and the result of the paper shows that the GBDT algorithm had the highest overall accuracy of 89.1% and it claims that GBDT solved the problem of overfitting.

Yakubu and Abubakar, (2022) in the paper applied machine learning techniques to predict the performance of higher-level education students in Nigeria. The paper also strictly follows the life cycle of data science. Secondary data of the students in a Nigerian University were collected. A total of 978 undergraduate students from the American University of Nigeria (AUN) was used for the analysis. Data preprocessing which involves cleaning the data to remove errors, missing values, and duplicates was done (Alasadi and Bhaya, 2017). The Data Exploration phase which involves visualizing data with charts, graphs, and tables to find patterns *(Bikakis, Papastefanatos and Papaemmanouil,* 2019) was done and the following insights were discovered: Majority of the

students were between 16-20 years, the percentage of male to female was 51.9% and 48.1% respectively, majority of the students originated from the school's location in the Northeastern part of Nigeria with only 34.4% of students from the Southern parts of Nigeria. Also, filtration was done to select important features for the modelling (Alasadi and Bhaya, 2017), and the writers selected age, gender, high school examination scores, region, and CGPA as the most important. The researchers used dummy variables to encode the features which is an important step for logistic regression (Gupta, 2019). The training and testing data was split with a ratio of 70/30. Using logistic regression, the trained model correctly predicted the CGPA results with an accuracy of 84.7% and had an accuracy of 83.5% when applied to the testing data. The results of the analysis revealed that age was not a factor that influences the success of a student which is also supported by (Glawala *et al*, 2016), female students tend to succeed 1.2 times better than the male students, students who scored high points in their JAMB were likely to have high CGPA results and vice versa, students that come from developed parts in Nigeria were likely to have higher CGPA and students in the third and fourth year were more likely to have better CGPA results. Further research can be done to include the WAEC results of the students, personality traits, and motivation to the model to draw better conclusions. Finally, other machine learning models such as Artificial Neural Network, decision trees, and random forest can be used to compare with the proposed logistic regression method.

2.4 Artificial Intelligence in Predicting Student Performance

Vijayalakshmi and Venkatachalapathy, (2019) in their paper used Deep Neural Networks to predict student performance and not just machine learning algorithms as other researchers had used in the past. The researchers also trained the model using five other machine learning algorithms namely Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbor (KNN) for comparison. The researchers used a dataset from Kaggle for this purpose. The dataset comprised 500 students with 16 features. The features were grouped into three categories: demographics, educational background, and behavioral features. The target variable classes were distributed fairly: 1 representing "High" with 142 occurrences, 2 representing "Medium" with 211 occurrences, and 3 representing "Low" with 127 occurrences in the dataset making a total of 480. The programming was done in the R language. For cross validation, the dataset was split using a ratio of 70/30 for training and testing respectively. The researchers performed deep learning using Keras and TensorFlow packages in R. A neural network has three layers, the inner layer, the hidden layer, and the outer layer (Pyo *et al.*, 2017). These layers are clearly identified in the paper. The researchers specified one input layer

containing 16 neurons, two hidden layers, one having 10 and the other 5 neurons, and the output layer having three classes. Activation functions are an integral part of the neural network because they specify the shape of the neural network output (Lederer, 2021). The researchers have specified “relu” and “softmax” activation functions for the hidden and outer layer respectively. The result of the experiment shows that amongst the six algorithms used, Deep Neural Network performed the best with an accuracy of 84% while Decision Tree and KNN performed the worst with an accuracy of 69%.

Imran *et al.*, (2019) in the paper also proposed a deep neural network for predicting student performance but in Massive Open Online Courses (MOOCs). The paper presents a model with 7 inputs, 2 output nodes and combined 15 different architectures of hidden layers by alternating different numbers of hidden layers and the neurons on each layer. The dataset comprises 20 different attributes of 641138 students who enrolled for a course in edX platform in 2012 fall, 2013 spring and summer 2013. The researchers ensured to preprocess the data for modelling by removing missing values. Out of the 20 attributes, only 7 were identified as important for monitoring and predicting the dropout rate and they include: Viewed, No of Events, Days active, Number of Video played, Number of chapters and Number of Forum posts. The data was split into 60% for training, 25% for validation and 15% for testing. This technique is supported by (Genç and Tunç, 2019). The major difference between Validation and Test set is that while the validation set is labeled, the test set is unlabeled (Brownlee, 2017). A validation set is used to verify the model while it is being trained (Brownlee, 2017). A high training accuracy and low validation accuracy means that the model is overfitted and hence the weights would need to be adjusted but if both the training and validation sets are relatively high, then the model is set for testing (Genç and Tunç, 2019). The model tests the data without labels so that the model can predict the output of the test set (Genç and Tunç, 2019).The model was trained over 10 epochs and the paper reports that the best results was the hidden layer set to 7 with a total of 64 neurons. The prediction accuracy, Recall and classification accuracy were 99.8%, 97.62 and 88.76% respectively. The researchers discovered that adding more layers increased the performance but adding more neurons reduced it.

The problem of wrong course selection was highlighted by (Sood and Saini, 2021) as a factor in reducing student performance. To solve this problem, the paper proposes a hybrid approach consisting of Cluster-based Linear Discriminant Analysis (CLDA) and Artificial Neural Network that recommends motivational speeches and videos to enable the students select the right course for them thereby reducing the rate of dropout. The paper employs Educational Data Mining as an

approach to build a prediction model by exploring and extracting unseen patterns in the features with machine learning techniques. The CLDA was used to cluster students with similar performance together. Then the ANN algorithm specifically, K-Nearest Neighbors was used to classify the student performance to be either dropping out, passed, or failed. The results of the prediction are sent to the tutor to identify and help the weaker students. The researchers use an inclusion and exclusion criteria as defined by (Patino and Ferreira, 2018). Inclusion criteria meaning selecting the most relevant factors relevant to a study and exclusion criteria meaning removing the not so important features that could hamper the results of the study. This is like the feature selection process in data science (Hara *et al.*, 2018). Using this criterion, the researchers were able to select scores obtained, age, gender, subject marks, and highly correlated features to the performance of the students as part of the features for training the model. While records with null and missing values were excluded. However, to account for the missing values, the researchers used the highest occurring values to replace them. This technique is supported by (Kumar, 2021) as one of the methods in handling missing values and he also suggests using the mean and mode for the same purpose. The classification accuracy was compared using CLDA, Naïve Bayes, Neural Network, and Hierarchical clustering and CLDA performed the best with an average of 93% accuracy testing with 5 datasets. This performed best because the LDA grouped the dataset according to the scores of the students. The researchers also performed a 10-fold cross-validation by dividing the dataset into training, test, and validation set (Genc and Tunc, 2019). Further studies can be done to validate the results using real time data to predict students that are liable to dropout by extracting students' data online with Iota wearable devices.

Hossen *et al.*, (2021) used a four-part architectural system to locate the exact learning abilities of the students that will categorize them. The first part is collecting data from students that had obtained a certificate from their high school, The second part is storing the data in a drive for future reference, the third part is testing the trained model with newly saved data and the last part conveys the result of the model to the user. This system is powered by different supervised machine learning techniques such as Decision Tree, Gaussian Naïve Bayes, Random Forest, K-Nearest Neighbors, and Artificial Neural Network to determine the best for predicting student performance. For this experiment, a survey was conducted with 518 randomly selected students enrolled in Daffodil International University and 18 features were collected. The data was grouped into Demographics, Educational Background and Behavioral attributes. The categorical data in the dataset were encoded by one Hot encoding method using Pandas get dummies method (Lukic, 2021). One Hot Encoding method is an import data preprocessing step because it converts

categorical data into numbers which can be understood by the computer to enable seamless computations (Lukic, 2021). Highly correlated features were also removed like discussed in (Zeebaree *et al.*, 2021). The researchers used random forest importance to select the most important features according to ranking (Kursa and Rudnicki, 2011). For this purpose, the Scikit-Learn Random Forest library was implemented. Chi Square is another feature selection technique that was used to determine if the input values significantly affect the output value to be predicted and is supported by (Gajawada, 2019). The dataset was split into 70% and 30% for training and testing respectively. The model was trained by alternating the applied weights to identify the ones giving the best accuracy. Also, the model was tested before and after applying the feature selection techniques and it was discovered that the model performed better in the later. The proposed model with the best accuracy of 92% after training was an Artificial Neural Network which was saved and integrated into a web server.

Summary

From the literature review, researchers have identified that there are massive student dropout and failure rates in high schools and universities. The researchers noted that it was easier to detect when students are slacking off while in the four corners of a classroom because they can easily monitor the behaviors of the students to know if they are distracted, absent, or not understanding the lectures. However, with the advent of COVID 19, the education system is rapidly becoming digitalized and hence very difficult to monitor the attitude of students towards studies. To solve this problem, the researchers have made use of machine learning algorithms and artificial intelligence to create models that can predict the performance of students. Some of the machine learning algorithms that have been implemented include K-Nearest Neighbor (KNN), random forest, Naïve Bayes, GBDT, decision trees, Support Vector Machine and Artificial Neural Network. Over the years, there have been improvements on the accuracy of the different machine learning techniques. Some researchers have experimented combining different machine learning techniques to improve the performance of the model for better predictions. The researchers highlighted that the performance of students could be affected by Age, Educational background, place of origin, interaction with academic resources, sex, parents background and other behavioral attributes. Hence, data on these features were collected through surveys, log clicks on E-learning Platforms, and school database systems. In recent times, there has been a great attention toward the life cycle of a data science project. Most of the researchers followed the CRISP DM life cycle of a data science project which includes Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The researchers agree

that it is best to pre-process the dataset before modelling to improve the quality of the data and increase the accuracy of the prediction. Also, it is best practice to split the dataset into training, validation, and testing to check for overfitting of the trained model. The researchers that used Artificial Neural Networks experimented with different combinations of hidden layers and neurons on each layer to determine which combination performs best. The researchers all conclude that the machine learning model will serve as a useful tool in the hands of the teachers to make data-driven decisions accordingly by changing their teaching styles or recommending materials and resources to the students that are predicted to fail during a semester.

Chapter 3 – Methodology

This chapter highlights the methods used in this research project. Each method is categorized under the research questions that were stated in chapter 1.

3.1 Research Methods

A research method is a particular tool that a researcher uses to collect and analyze data. The research method could either be qualitative, quantitative, or mixed.

This research method is governed by quantitative research and is focused on quantifying the sample to get results. The sample size selected in this project is a good representation of the entire large population, therefore the conclusions made on this sample was taken as the true for the entire population (Queirós *et al.,* 2017). This research method involved the analysis of numerical data with various statistical techniques to answer questions about the sample. In this research, the quantitative research method started with a statement of problem, stated the research questions, literature review and a quantitative data analysis (Queirós *et al.,* 2017). This method was chosen because there was a collection of numbered measures of features and verified conclusions from a population sample. The method followed a standard process and used formal tools for collecting data. It followed an objective and systematic approach of collecting data (Queirós *et al.,* 2017). Data analysis was done using statistical methods and software like Python, Pandas, Numpy and the Sklearn module used (Queirós *et al.,* 2017).

There are four main types of quantitative research, and they include survey research, correlational research, experimental research, and causal-comparative research. Survey research involves the researcher passing out questionnaires to obtain measurements from the characteristics of the population with the aid of statistical methods. Correlational research involves attempting to find existing relationships between features using statistical data (Apuke, 2017). This type of research only attempts to discover trends and patterns in the data but does not try to prove the cause-and-effect factors. Causal-comparative research, also known as quasi experimental research, tries to find the cause-and-effect relationship within variables in a dataset (Apuke, 2017). The researcher does not try to alter the independent variables but measures the effects of the independent variables on the dependent variables. Experimental research involves using scientific methods to find cause and effect relationships between variables (Apuke, 2017). Unlike the name implies, it does not actually involve a laboratory, rather the researcher manipulates all the variables except one to find the effect of the independent variable on the dependent variable (Apuke, 2017).

In this research the goal was to try to find to what extent the features of student (independent variables) are related to their result (dependent variable) and create a model that predicts their final result. Therefore, the Correlational research method is used. The correlation between these features do not prove causality but they are helpful pointers that can help to improve the accuracy of the model prediction. The experimental research method was also used because a variable was manipulated to have a direct relationship to the final result.

The type of data collected was qualitative data and quantitative data. The qualitative data were the demographics of the students including the Highest Education Level, Region, disability status, gender, final result status, and code module. The quantitative data in the dataset include the number of previous attempts, assessment score, Index of Multiple Deprivation (IMD) band, age band of the students, and studied credits. Even though the dataset has some qualitative data in it, quantitative methods were used to encode the variables for the machine learning model to process it. The dataset was a secondary data prepared by Open University. It was collected from the schools Virtual Learning Environment (VLE), and it contained a total of 32,593 student data between the years 2013 and 2014.

3.2 Research Questions

RQ1

Regarding the Experimental research method, some manipulations were done on the assessment scores so that they directly correlate with the final results of the students. This step partially answers RQ1. The assessment scores in the table were not organized in a single row for each student, rather, the scores were spread across all presentations done by the students in different modules and in different years. Hence it was difficult to directly relate the assessment scores to each student. Using statistical methods such as groupby and windows function with Python and POSTGRESQL respectively, the aggregate average of the scores was computed for both years in 2013 and 2014. The scores were calculated over 100%.

RQ2

Because qualitative research is empirical in nature, there is a wide range of analytical methods that can be used. A machine is said to have learnt from its experiences if it continually improves in its performance as more tasks are assigned to it and it is able to make its own decisions and make predictions based on the data that has been fed to it (Ray, 2019). Machine learning methods are quantitative methods, and they are models that systematically analyze qualitative data (Ray, 2019). Machine learning algorithms need to be trained with a very large set of data for it to

discover patterns (Ray, 2019). Machine learning algorithms are grouped into Supervised learning and Unsupervised learning.

In contrast to unsupervised learning, supervised learning is useful for training a machine with the presence of a label (Johnson, 2022). This means that every row point in the dataset already has a column that describes the rows. It is likened to a classroom with the presence of a teacher or supervisor (Johnson, 2022). The algorithms used in supervised learning can learn from the labeled rows and can forecast the outcomes of future data. In supervised learning, the output of data can be determined from past experiences, the performance can be optimized based on past experiences and it can solve various real-world problems. The algorithm takes in training data with the input labeled x and the output y and creates a function that maps the future unknown inputs (Siddiq, 2017).

In this project, the supervised learning technique will be used as the aim of this project is to predict the performance of students. This means that the model will learn from the past student's performance which have been classified as Distinction, Pass, Fail or Withdrawn and predict the outcome of new students. Supervised machine learning becomes extremely challenging when attempting to classify big data. Supervised machine learning techniques are of two types, and they include classification and regression (Siddiq, 2017).

RQ3

This research question was answered using the correlational research method. The Pearson correlation matrix was used to find correlations between the features in the dataset that could possibly lead to failure of the student. Essentially, the correlations do not actually mean that the feature resulted in the failure of the student. It was used to find out what features would be significant for the model to perform better.

Pandas is a python library that is well equipped with tools and data structures for working with structured datasets in the fields of statistics, finance, social sciences and so on (McKinney, 2011). The library has a built-in function that handles everyday data analysis and manipulations on datasets. The goal of this package is to set the foundation for statistics in Python (McKinney, 2011). The pandas library which was founded in 2008 has been evolving to bridge the gap between the current data analysis tools in Python which is a language widely used for scientific computations and many other database languages (McKinney, 2011). This library is used by data scientists and analysts to process large chunks of data. It reads data from various sources and formats and processes it (Hagedorn *et al.*, 2021). Because of the increased application of

machine learning models, many frameworks in Python have been developed including Pandas to create, train and apply different artificial neural networks on the incoming data (Hagedorn *et al.*, 2021). The pandas package is very useful for the pre-processing of the data when it is initially loaded (Hagedorn *et al.*, 2021).

RQ4

The methodology for this research project is governed by the Cross Industry Standard Process for Data Mining (CRISP-DM). According to Ribeiro *et al.*, (2020) the CRISP-DM is an analytical standard to ensure the success of a data mining project. The methodology is described by a hierarchical process consisting of four levels: phase, generic task, specialized task, and process instance (Ribeiro *et al.*, 2020). The CRISP-DM shows the typical life cycle of any data mining project and the repetition of different stages. There is an assumption that there is communication between the business experts and the Data Mining Analysts (Ribeiro *et al.*, 2020). The life cycle of a data mining project is divided into six major parts as seen in the diagram below. According to (Wirth and Hipp, 2000) the order in which the phases occur is not fixed. The arrows reflect the interdependencies between the phases and their importance in a specific project. The next phase is determined by the outcomes of the previous phases. The outer circle represents the cyclic nature of a data mining project. The outcomes of the model deployment could bring about new research questions that would initiate the entire lifecycle of the data mining project.

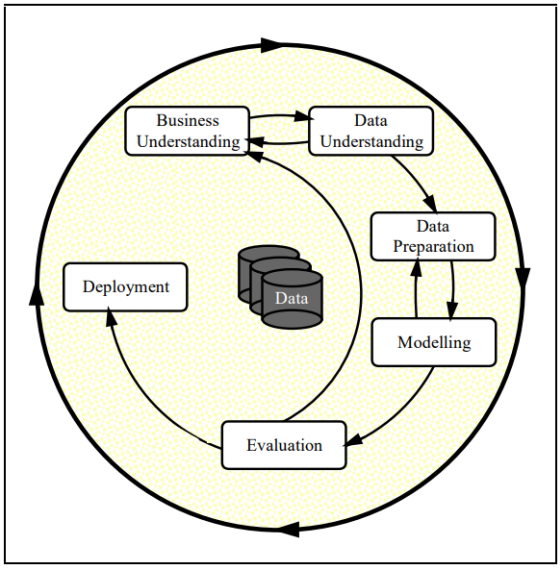

Figure 1 CRoss Industry Standard Process for Data Mining (CRISP-DM) Lifecycle (Wirth and Hipp, 2000)

The six stages of the CRISP-DM lifecycle are highlighted below:

1. Business Understanding: The aim of this first stage is to state the objectives of the research project to have a clear picture of the requirements of the project from a business standpoint (Wirth and Hipp, 2000). This understanding is converted into a business plan that aligns to the stated objectives of the defined business problem (Wirth and Hipp, 2000).
2. Data Understanding: This phase begins with the collection of data followed by a series of processes to understand by exploring and describing the dataset with the goal of checking the quality of the dataset (Schröer *et al.*, 2021). This step requires investigations into the dataset to spot potential hidden information that can widen the research scope. There is a close relation between business understanding and data understanding; a certain level of understanding of the dataset gives rise to business questions and plan formulation.
3. Data Preparation: In this stage, the data is processed to a standard that is suitable to be used as input for the model (Wirth and Hipp, 2000). The type of model to be built specifies what processing is required for the dataset. It also involves inclusion and exclusion criteria to select what relevant features are necessary for the final dataset (Schröer *et al.*, 2021). There is no order to which these processes follow. It can include feature selection, data cleaning, reformation of features or including new features (Wirth and Hipp, 2000).

4. Modeling: This stage involves selecting from a variety of modelling techniques to find the ones that would be suitable for solving the problems identified in the data mining project. It is important to set parameters to build the model (Schröer *et al.*, 2021).
5. Evaluation: After building the model, it is important to verify the model to ensure that it aligns with the business objectives of the data mining project (Wirth and Hipp, 2000). The constructed models can seem to have high quality in terms of data analytics but in fact, it may not be solving the problems the research questions asked. The outcome of this evaluation should be a decision that has been reached regarding the research question (Wirth and Hipp, 2000).
6. Deployment: When the model is built, it is required for the data analyst to carefully document and present to the user either by reports or implementing an iterative data mining process (Wirth and Hipp, 2000).

According to Solano *et al.*, (2021) the CRISP-DM methodology follows a goal focused approach and is widely accepted in data mining projects using machine learning algorithms. The CRISP-DM methodology consists of many attributes that make it very useful for evidence mining (Solano *et al.*, 2021). It maintains a general process that supports the overall structure and depth of the methodology. Research shows that only 18% of data science projects follow the CRISP-DM process model even though this model has existed since the year 2000 (Schröer *et al.*, 2021). It is possible that it is difficult for data scientists to strictly follow this process model. From the literature review, not all studies reflect on the data selection, transformation, and cleaning phases. Some limitations in applying the CRISP-DM methodology are the fact that it does not explicitly state pre-processing steps like data sensors, it cannot handle certain requirements of recent technology like machine learning algorithms (Schröer *et al.*, 2021). Because of these limitations, other methodologies like APREP-DM (a Framework for Automating the Pre-Processing of a Sensor Data Analysis) and Lean Design Thinking Methodology for Machine Learning and Modern Data Projects (LDTM) have been developed to enhance the CRISP-DM methodology (Schröer *et al.*, 2021). The CRISP-DM methodology is flexible for all technologies as it is an organizational process model, and the technologies support this process (Schröer *et al.*, 2021). The CRISP-DM process model clearly states the deployment phase but interestingly, most studies do not mention it. The deployment phase is important as it explains the actions to be taken by the users based on the model that has been trained (Schröer *et al.*, 2021). This could mean that the CRISP-DM process model does not give a satisfactory explanation of the deployment of the trained model to the productive environment where it can be regularly monitored and controlled to improve its

performance (Schröer *et al.*, 2021). The CRISP-DM methodology does not sufficiently cover the entire life cycle of a data science project including the machine learning approaches. However, there are already available methods to deploy a machine learning model, but they are not yet included in the CRISP-DM methodology (Schröer *et al.*, 2021).

3.3 Data Selection and Collection

Data Collection

There are several information systems at the Open University that exist and support the students and the modules. Because of the different sources of information, OU implemented a data warehouse that collects all the information from the different systems available. The data warehouse is based on SAS technology (Kuzilek *et al.*, 2017). Three different data types were collected, and this includes Demographic, Performance, Learning Behavior. The demographics contains the age, gender, region, their previous education and so on. The performance shows the results of the students during the school year (Kuzilek *et al.*, 2017). While the learning behavior represents the clickstreams of the student interactions on the VLE.

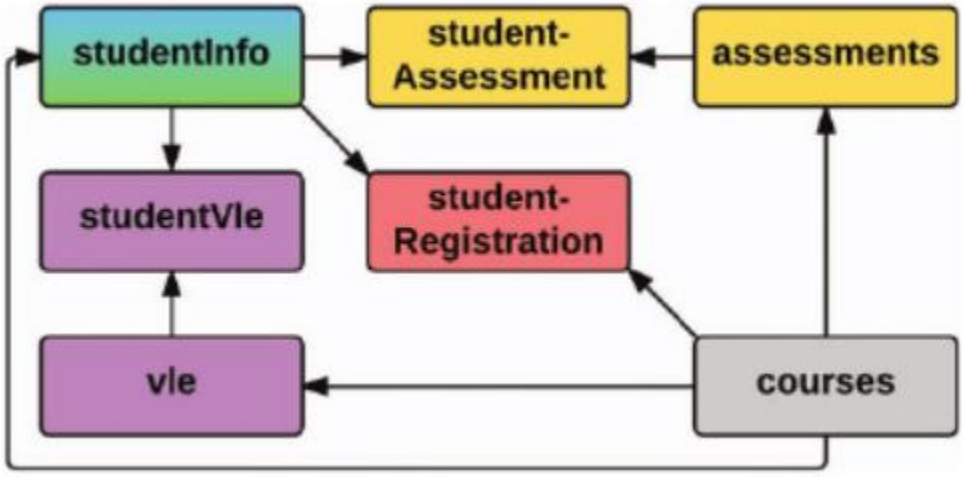

Figure 2 Detailed Dataset Structure (Kuzilek *et al.*, 2017)

Data Selection

The information on the student's demographics, module and VLE interactions have been collected in OU's data warehouse since 2012 (Kuzilek *et al.*, 2017). OU only selected modules that were taught in the years 2013 and 2014 and the selection process was governed by the following rules:

(I) They ensured that the module had over 500 students registered in it; (ii) They ensured that there were two or more presentations in the modules; (iii) They ensured that VLE data had corresponding student data available; (iv) Finally, they ensured that the chosen module had a significant amount of students failing (Kuzilek *et al.*, 2017). Seven out of all the modules that met the criteria were chosen and they include Science, Technology, Engineering, Mathematics module and 3 Social Science modules. The total number of students registered for these modules were 38,239 (Kuzilek *et al.*, 2017).

3.4 Ethics/ Data Anonymization

Because of ethical and privacy concerns, a data anonymization process was designed and applied to the dataset. The whole process was overseen by the Open University management and approval was given from the Vice-Chancellor Executive committee. The following steps were taken in the anonymity of the data (Kuzilek *et al.*, 2017). Firstly, the personal student information like social security number, birth dates and unique identifiers that OU assigns to the students. The module taken were removed and replaced with sematic-free symbols (Kuzilek *et al.*, 2017). The numeric identifiers such as student_id and code_module was both randomized and reassigned. The following were identified as quasi-identifying attributes and they include gender, IMD band, Highest education level, age, student's region, and disability (Kuzilek *et al.*, 2017). Using the ARX anonymization tool, an additional anonymization step was added for these attributes because they still could be traced back to the student with other publicly available sources. With k-anonymity measure, k was set to 5 and the suppression limit was set to 0.7. The maximum number of outliers to be removed was set to 20% of all the records (Kuzilek *et al.*, 2017). Finally, the average re-identification risk criterion was set to 0.05. After the anonymization process, the new number of students became 32,593. The age and the IMD band features were generalized (Kuzilek *et al.*, 2017).

Chapter 4 – Design and Implementation

4.1 Design of the Process

The diagram below shows the design of the process flow for the machine learning model from data collection to the visualization of the predicted outcomes.

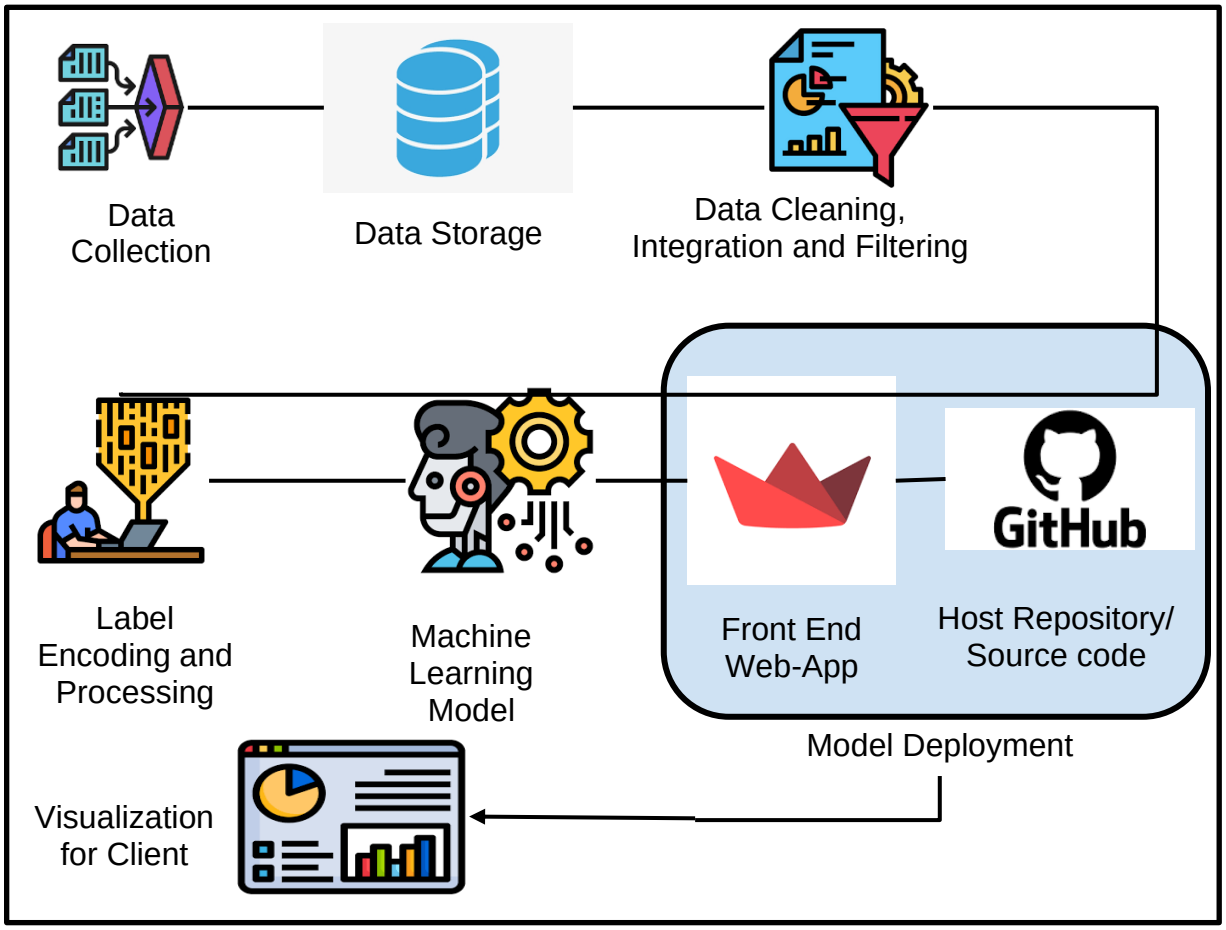

Figure 3 Design of Methodology

The data used in this project is available on Kaggle and the Open University website (Appendix 4) and it is available for all to use for analysis purposes. Two important csv files containing the demographics of the student and the assessment scores were used for this project. It is assumed that the demographics of the students were collected from the registration forms of the students. The assessment scores were recorded after each test was taken. The file size of the data is not large; hence the data was stored on the hard drive of a personal computer for easy importing, accessibility, and usage. Using Google Colab, the datasets were imported and converted to a

data frame. Data Integration, cleaning and filtration was done to improve the quality of the dataset. The two csv files were merged using the student id number as a reference in order to have the score of each student. A database in POSTGRESQL was created containing the student data and scores. Since there were multiple assessment scores, The partition by and order by method were used to calculate the average assessment scores of each student. The scores were grouped according to the different course modules and the year they were taken. After the data integration process was completed, the scores which were missing were filled with the average value of a particular class. For instance, for a student which had a “Distinction”, but no score was recorded, the average score of the “Distinction” class was recorded as the student. This way more data is available rather than deleting the entire row. However, for students that had scores but no registration data, the rows were deleted completely because it would be difficult to compute the data and could be misleading for the model.

Most of the dataset consisted of categorical data which is difficult for a classification model to process. Hence, they were encoded using the pandas “get_dummies” function. The pandas get_dummies function was preferred over the LabelEncoder method because the model would not be able to distinguish the categories. While the get_dummies function converts the categorical variables to binary variables, the LabelEncoder assigns numbers in alphabetical order creating some form of ranking system and this can be misleading for the model. The pandas get dummies function creates more inputs but is said to be more effective. The continuous data in the dataset were normalized with MinMax Scalar.

With the prepared dataset, the dataset was split into training, testing, and validation data with a ratio of 0.5, 0.3 and 0.2 respectively also another variation was tested with 0.7, 0.2 and 0.1. The train_split_method was used for this purpose and the training data was used to train the Artificial Neural Network model. The target data was extracted out and set to Y, while the remaining columns were set to X. Using tensorflow library and keras module, three dense layers were created; One input layer with an input shape of 47 and 12 neurons, one hidden layer with 8 neurons and one output layer with 4 neurons. The output layer has 4 neurons because the classes to be predicted are four. The relu activation function in the hidden layer and the used and the softmax activation function was used in the output layer. The model training was done over 150 epochs. After the model training was completed, the model was evaluated using precision, recall and f1 score metrics. The confusion matrix was also used to show the number of correctly matched classes. The model was saved in a .h5 format and set for deployment.

A GitHub repository containing the trained model, the source code and testing data was created. Streamlit, an open-source app framework for deploying machine learning models, was used for this project. The application built operates on a batch processing method. The user uploads a single csv file containing rows of student data and then clicks the “Make Prediction” button. The application encodes the categorical data and normalizes the continuous data. The saved model makes predictions on the newly received data and returns visualizations of the count of students that will get distinction, pass, fail or get withdrawn from school. A grouped bar chart of the student's results and what IMD band they are in is shown along with a group bar chart of the gender and their result. These charts will give the teachers insights into how their students will perform. A data frame of all the students with their predicted results will be displayed as well. The sample of this application is shown in the result chapter.

4.2 Description of Dataset

The dataset was developed by Open University, and it contains 7 tables about the data of the course presentations at Open University. In this paper, the tables containing information on the student’s information and assessment scores were used for the analysis. The tables contain the information on 22 courses, 32,593 students, assessment scores and logs of student interactions with the course materials in the Virtual learning Environment (VLE) and the demographics of the students. The demographics include Age band, Region, IMD band, Disability, Gender, and Highest Education Level. The Index of Multiple Deprivation (IMD) band is used to describe how deprived an area is. It makes use of the social, economic, and housing information and a single score is derived that describes the living condition of the area. On a scale of 0-100%, 10% means that the area is really deprived and 100% means that the area is well off. From the pie chart below, 9.7% of the students are disabled and 90.3% are not disabled. The results of the students are grouped into four: Pass, Fail, Withdrawn and Distinction. The ratio of male to female students is 54.8% and 45.2%. The highest education level of the students is grouped into five categories: A level or Equivalent, Lower than A Level, HE Qualification, No Formal Qualification, and Post Graduate Qualification. The age bands of the students are 0-35, 35-55, and above 55. Most of the students’ ages fall between the range of 0-35. Other columns present in the table include the “*code_module*” which represents a tag given to the module that the students registered, “*code_presentation”* identifying the presentations taken by the students in each module, “*id_student*” representing the unique student number, “*num_of_prev_attempts*” meaning the number of times a student attempted a module, and “*studied_credits*” representing the number of

module credits the student is studying. The student assessment table contains 173,912 rows of all the scores for every assessment done by the students. The student Assessment table has the following columns an id for each assessment, the student id, the day the student submitted an assessment, a flag representing a result transferred from a previous presentation and the scores of each assessment done by the student. The student scores range from 0 to 100 and scores below 40 mean a failure.

4.3 Data Cleaning and Preparation

The visualization of the dataset revealed some errors within the dataset. The IMD Band had a range called 20-Oct. It is assumed that this is supposed to be 10-20% as the month of October is the 10th month and was not listed amongst the ranges, this could have been because of a computational error. Still on the IMD band, there is a range named 0 and a range named 0-10%. This could create duplicates; therefore, the two ranges were combined into one range 10-20%.

Data Integration

The table below is a snippet of the studentAssessment table. It contains the different assessments in a module taken by each student and their corresponding scores.

Table 4 Student Assessment table

id_assessment	id_student	date_submitted	is_banked	score
1752	11391	18	0	78
1752	28400	22	0	70
1752	31604	17	0	72
1752	32885	26	0	69
1752	38053	19	0	79
1752	45462	20	0	70
1752	45642	18	0	72
1752	52130	19	0	72

The assessment Table as shown below contains the code_module, code_presentation, id_assessment, assessment type, date submitted and the weight of each assessment. The code presentation represents the month and year in which the student took the assessment. The B represents February, and the J represents October. For each year, the total weight of the assessment in a module summed up to a 100 and the marks for the Exam was 100. The exam

was optional, so some students could decide to take it. However, the criteria to take the exam was not given.

Table 5 Assessment table

code_module	code_presentation	id_assessment	assessment_type	date	weight
AAA	2013J	1752	TMA	19	10
AAA	2013J	1753	TMA	54	20
AAA	2013J	1754	TMA	117	20
AAA	2013J	1755	TMA	166	20
AAA	2013J	1756	TMA	215	30
AAA	2013J	1757	Exam		100
AAA	2014J	1758	TMA	19	10
AAA	2014J	1759	TMA	54	20
AAA	2014J	1760	TMA	117	20
AAA	2014J	1761	TMA	166	20
AAA	2014J	1762	TMA	215	30
AAA	2014J	1763	Exam		100

The information from both these tables is important to know what year the students took the assessment and what module it was under. This helped in accurately calculating the total result that the student attained in a single module over 100%. The assessment table and the studentAssessment table were merged to have the information in a single table. An additional column called “score_weight” was calculated. Below shows the result of the merged tables.

Table 6 Data Integration of Student Assessment Table and Assessment Table

id_assessment	id_student	date_submitted	is_banked	score	code_module	code_presentation	assessment_type	date	weight	score_weight
1752	11391	18	0	78	AAA	2013J	TMA	19	10	7.8
1752	28400	22	0	70	AAA	2013J	TMA	19	10	7
1752	31604	17	0	72	AAA	2013J	TMA	19	10	7.2
1752	32885	26	0	69	AAA	2013J	TMA	19	10	6.9
1752	38053	19	0	79	AAA	2013J	TMA	19	10	7.9
1752	45462	20	0	70	AAA	2013J	TMA	19	10	7
1752	45642	18	0	72	AAA	2013J	TMA	19	10	7.2
1752	52130	19	0	72	AAA	2013J	TMA	19	10	7.2
1752	53025	9	0	71	AAA	2013J	TMA	19	10	7.1

Calculated Variables

Score_weight: For a specific module, there are several assessments, and the weights of these assessments sum up to 100 percent. The assessment scores in the score column is over 100. Each id_assessment has their respective weights. The score_weight column was calculated as follows to get the score over the weight of the assessment.

$$Score\ weight = \frac{score}{100} * weight$$

Using POSTGRESQL, a query was written to calculate the score_weight_sum and weight_sum as shown in the table below.

Table 7 Calculated Variables

	id_assessment	id_student	date_submitted	is_banked	score	code_module	code_presentation	assessment_type	date	weight	score_weight	score_weight_sum	weight_sum
0	1760	6516	116	0	63	AAA	2014J	TMA	117.0	20.0	12.60	63.5	100.0
1	1759	6516	51	0	48	AAA	2014J	TMA	54.0	20.0	9.60	63.5	100.0
2	1758	6516	17	0	60	AAA	2014J	TMA	19.0	10.0	6.00	63.5	100.0
3	1762	6516	210	0	77	AAA	2014J	TMA	215.0	30.0	23.10	63.5	100.0
4	1761	6516	164	0	61	AAA	2014J	TMA	166.0	20.0	12.20	63.5	100.0
1739707	15024	2698588	202	0	95	BBB	2014J	TMA	201.0	35.0	33.25	92.4	100.0
1739708	15023	2698588	152	0	95	BBB	2014J	TMA	152.0	35.0	33.25	92.4	100.0
1739709	15022	2698588	109	0	87	BBB	2014J	TMA	110.0	20.0	17.40	92.4	100.0
1739710	15021	2698588	53	0	85	BBB	2014J	TMA	54.0	10.0	8.50	92.4	100.0
1739711	15020	2698588	18	0	100	BBB	2014J	TMA	19.0	0.0	0.00	92.4	100.0

Weight sum: This value was calculated by summing all total weights of each assessment per module and per student. For example, the student id “6516”, the total weights of assessments taken in the module “AAA” summed up to 100.

$$ie\ weight\ sum = 20 + 20 + 10 + 30 + 20 = 100$$

It is important to note that one student can take multiple modules and it is possible that the student repeats the module in another year, hence, when using a groupby method with pandas, the

id_student, id_assessment and code_module was used to uniquely identify the scores of each student.

Score_weight_sum: This value was calculated in a similar manner to the weight_sum. The total "score_weight" across the module was summed up. For example, the student id "6516", the score_wieght sum was calculated as follows:

$$score\ weight\ sum = 12.6 + 9.6 + 6 + 23.1 + 12.2 = 63.5$$

The studentRegistration table containing the student demographics and the studentAssessment table containing the scores of the student were merged using the *inner join* function in pandas. An additional column called score_100 was calculated, and this was done by prorating all scores above 100 to 100%.

Class Determination:

It was noticed that there were some discrepancies in the results of the students. Therefore, a rule was created. Score between 0-40 were mapped to the class Fail, scores between 40-80 were mapped to the class Pass, scores between 80-100 were mapped to the class Distinction and scores less than but equal to 50% that were already mapped to the class Withdrawn were left as Withdrawn, while scores above 50% were mapped to Pass, Fail or Distinction depending on their ranges. There were some missing scores, and this was because there were no records of the student's scores in the studentAssessment table, hence for these missing values, the average score for the respective classes were used rather than removing the entire row. The table below shows a snippet of the entire dataset after integration. The total number of rows after integration became 25,843.

Table 8 Full Training Data Before Encoding

	code_module	code_presentation	id_student	gender	region	highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability	final_result	score_100
0	AAA	2013J	74372	M	East Anglian Region	A Level or Equivalent	20-Oct	35-55	0	150	N	Fail	32.800
1	AAA	2013J	111717	F	East Anglian Region	HE Qualification	90-100%	35-55	0	60	N	Pass	48.900
2	AAA	2013J	147756	M	North Region	Lower Than A Level	60-70%	0-35	0	120	N	Pass	51.300
3	AAA	2013J	185439	M	London Region	HE Qualification	20-Oct	0-35	0	120	N	Fail	7.400
4	AAA	2013J	205719	M	East Anglian Region	A Level or Equivalent	40-50%	0-35	0	90	Y	Fail	27.300
25838	DDD	2014J	2231512	M	South East Region	HE Qualification	90-100%	35-55	0	60	N	Distinction	87.900
25839	DDD	2014J	2339789	M	Scotland	HE Qualification	20-Oct	0-35	0	150	N	Distinction	82.775
25840	DDD	2014J	2471673	M	London Region	HE Qualification	90-100%	55<=	0	100	N	Pass	77.500
25841	DDD	2014J	2516939	M	Scotland	Post Graduate Qualification	90-100%	35-55	0	60	N	Distinction	82.900
25842	DDD	2014J	2560520	M	North Western Region	HE Qualification	50-60%	0-35	0	60	N	Distinction	87.375

Label Encoding

The Dataset set contains categorical data. Categorical data is not understood by classification models. Using One Hot Label Encoding, the categorical data was encoded with numbers to enable the classification model to understand it. For example, the results which had Pass, Fail and Distinction classes were encoded to 2, 1 and 0 respectively.

Normalization

The MinMax scaler was used to scale the dataset from 0-1 because they had wide ranges from each other. Also, the activation function of the output Layer of the classification model is Sigmoid. Sigmoid works with values from 0 to 1 so it is necessary to scale the result labels from 0 to 1.

Feature selection

The following features were selected for the model: number_of_prev_attempts, studied_credits, score, code_module, code_presentation, imd_band, gender, disability, highest_education, age_band, region, and, final_result. The correlation matrix between these features is shown in the results chapter.

4.4 Dataset Splitting

Using the *train_test_split* module, the dataset was split into training, testing, and validation sets with a ratio of 55%, 30% and 15%. Another variation was tested as well 70%, 20% and 10%. The independent variables (X) included all the features except the final_result and the dependent variable (Y) was the final_result.

4.5 Modelling

Three different modelling techniques were tested: Artificial Neural Network, XGBoost and Logistic Regression

Artificial Neural Network (ANN): The TensorFlow library and Keras module were imported for the model training. Various numbers of neurons in the hidden layer were used to get optimal accuracy results. The ratio of the training, testing and validation sizes were also changed to get the optimum result. The hidden layer and output layer activation function used were *relu* and *sigmoid* respectively. The different optimizers used were *adam* and *adagrad*. The diagram below shows the structure of the ANN model that was trained.

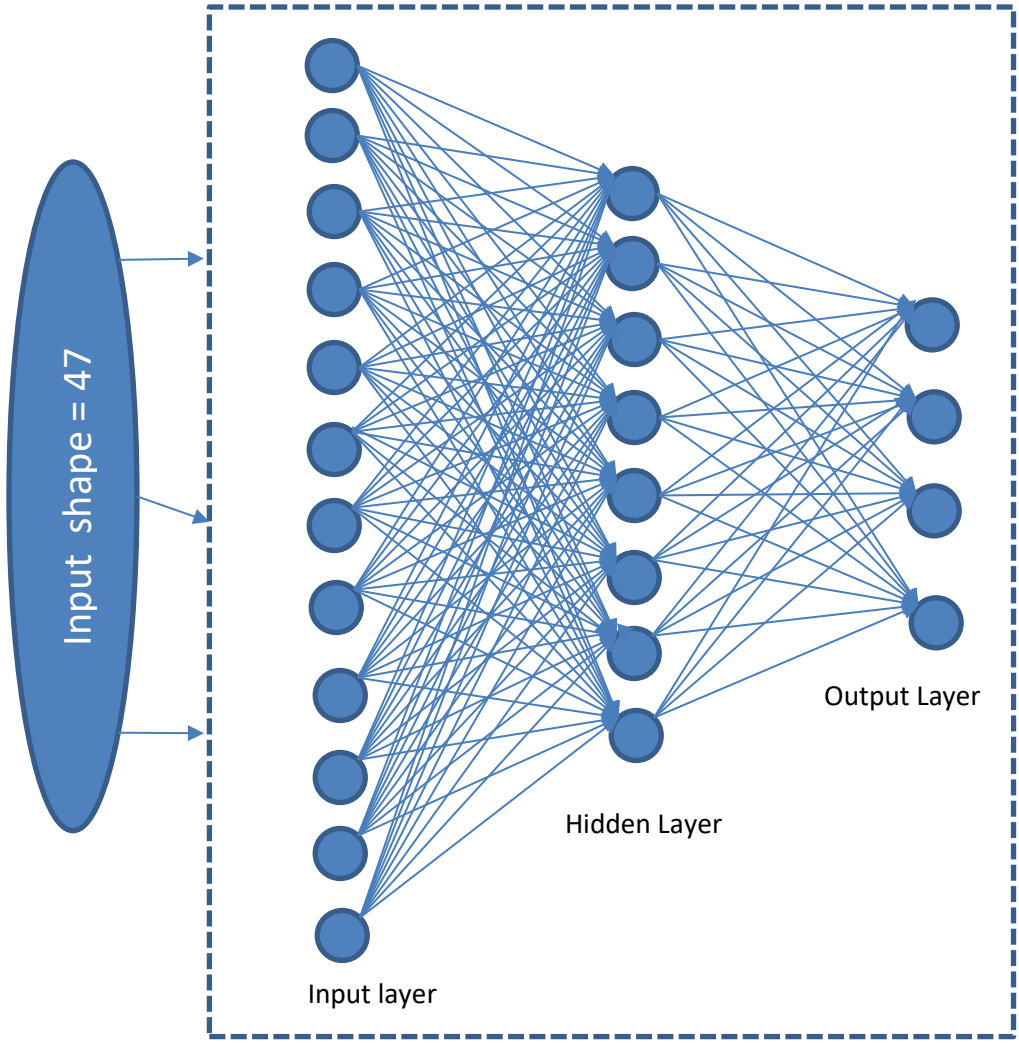

Figure 4 Artificial Neural Network Architecture

XGBoost:

The model was also trained using the XGBoost machine learning algorithm. The tuning parameters used are as follows:

- Learning_rate: the learning rate was set to 0.02.
- N_estimators: the n_estimator was set to 600.
- Silent: Silent was set to True
- Nthread: Nthread was set to 1.

Using Regular expressions column names with “<” were replaced with “_”

Logistic Regression:

Using sklearn.linear_model, the LogisticRegression library was imported. The model was trained with the following tuning parameters:

- Multi_class was set to multinomial.
- The solver was set to lbfgs.
- Max_iter was set to 5000.

4.6 Model Deployment

The model was deployed to the web using Streamlit. The benefit of using streamlit includes the following (Thetechwriters, 2021):

1. Creating applications with simple code
2. It allows live updates of the application version i.e., hot reloading.
3. A widget can be created easily by declaring a variable.
4. Writing a backend or various routes for HTTP request handling is not needed.

The following steps explain the model deployment processes taken. The files used here can be found on the Github page (Appendix 2)

1. The trained model for predicting the classes was saved as .h5 format named **new_ann_model.h5**.
2. The source code was written in the **main.py** file.
3. The libraries to be installed both locally and by Streamlit for this application were listed in **requirements.txt**. These libraries include (streamlit, pandas, numpy, sklearn. Scikit-learn, seaborn, matplotlib, tensorflow and plotly)
4. The page configuration was set as follows:
   a. **page_title**: This is the page title which is displayed at the browser tab, and it was set to “Student Prediction App.”
   b. **page_icon**: This is the favicon of the webpage, and it was set to “**favicon.ico**” which is found in the Github page (Appendix 2)
   c. **layout**: The page content of the page was set to “wide.”
5. The header of the page was set to "Student Performance Prediction App” using **st.header**
6. Using **st.file_uploader**, a user can upload the csv file which contains student data.
7. Using try and except blocks, KeyError and UnicodeDecodeError were handled in case a wrong file type or missing columns are detected. A warning using **st.warning** is displayed

for the user to upload the correct file if a UnicodeDecodeError occurs and if A KeyError Occurs, it displays the names of the columns for the users.

8. Using **pd.get_dummies**, the categorical data are encoded. The id_student, disability_N and gender_F are dropped because the model was not trained with these features.
9. The continuous data are Normalized using MinMaxScaler
10. A button called “Make Prediction” is created using **st.button**. On click of the button the remaining steps occur.
    a. Predictions are made for each row.
    b. Using **np.argmax**, the index of the cell with the highest value is returned
    c. The result is converted to a numpy array.
    d. The numpy array of the prediction results are converted to a dataframe and are concatenated to the original dataset that was uploaded.
    e. A function is created that converts the indices to the predicted results.

Table 9 Final Results Decoding

Indices	Value
0	Distinction
1	Fail
2	Pass
3	Withdrawn

    f. A sub header for the visualizations was created using **st.subheader.**
    g. Four different visualizations are created. One is a bar chart that shows the count of the predicted classes, the second one shows a group bar chart of the predicted classes by gender, the third one is a group bar chart of the predicted classes by IMD band and the fourth one is the confusion matrix of the predicted classes.
    h. Finally, the full dataset with the predicted classes is displayed using **st.dataframe.**

Chapter 5 - Results and Discussions

In this section all the results of the data analysis and model deployment are shown. It comprises all charts that show the relationship between features, the outcome of the machine learning model and a screenshot of the web application.

5.1 Exploratory Data Analysis and Data Visualization

To get more understanding of how the data are related to each other, exploratory data analysis and data visualization was done. This was also useful to discover hidden relationships, spot outliers and select features for the model.

- Relationship between IMD band and Final Result

From the bar plot below, there is a relationship between the IMD band and the result of the students. Students living in the lower IMD band regions have the highest failure rate and this should be a concern to the institution.

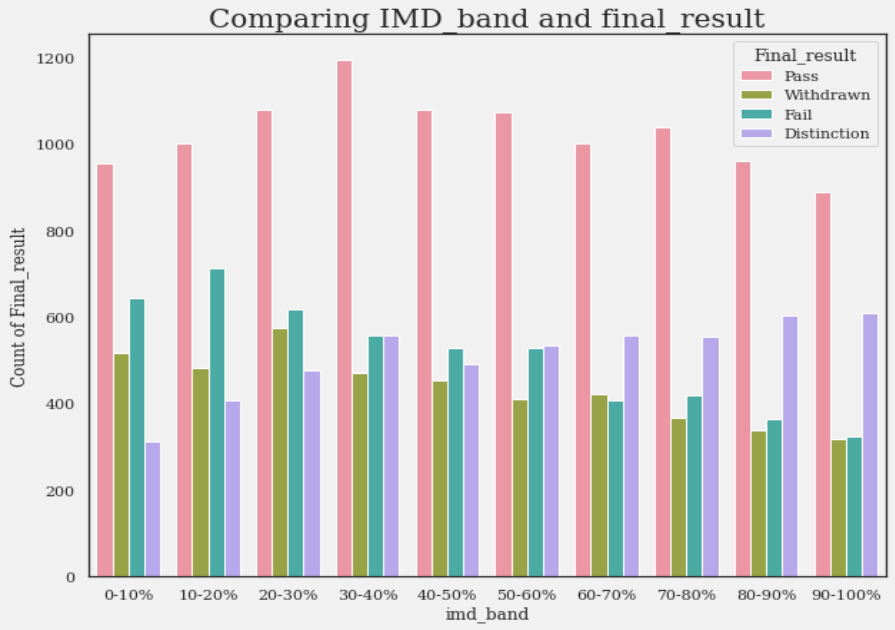

Figure 5 Relationship between IMD band and Final Result

- Relationship between Previous attempt and Final Score

From the box plot below, students that take the course between 0-3 times are more likely to have higher scores than students that retake the course from 3 to 6 times.

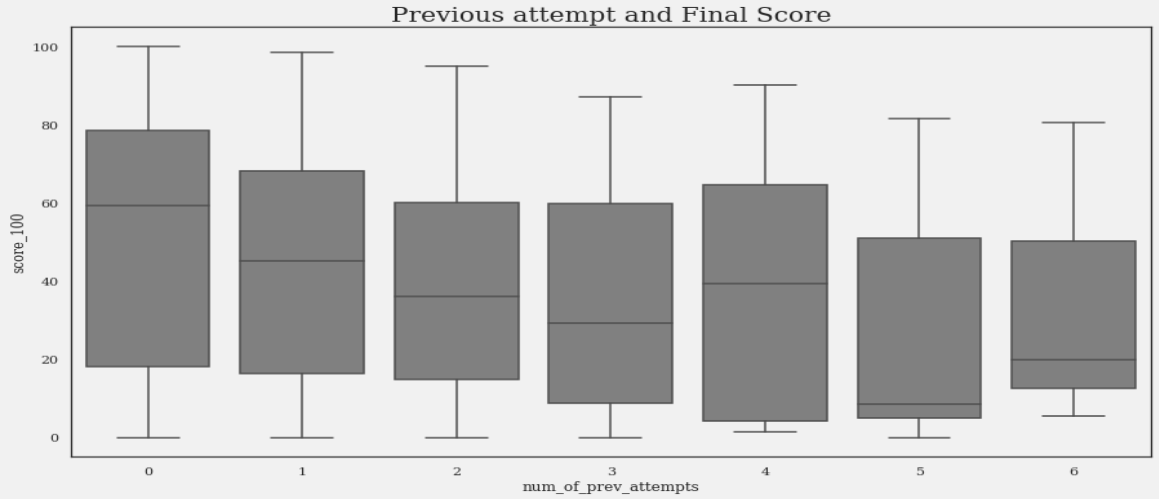

Figure 6 Relationship between Previous attempt and Final Score

- Relationship between Gender and Final Result

From the group bar chart below, the more female students fail. However, the gender is not balanced. The male students are more than female students.

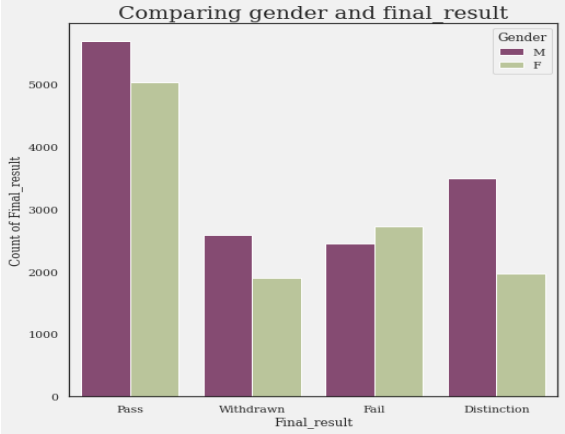

Figure 7 Relationship between Gender and Final Result

- Relationship between Final Result and Scores

The box plot shows that the range of the class Fail is 0-40%, the class Pass is >40% - 80%, the class Distinction is >80% - 100%, and the class Withdrawn is <50%. Because of the overlap between Fail and Withdrawn, the model found it difficult to distinguish the two classes.

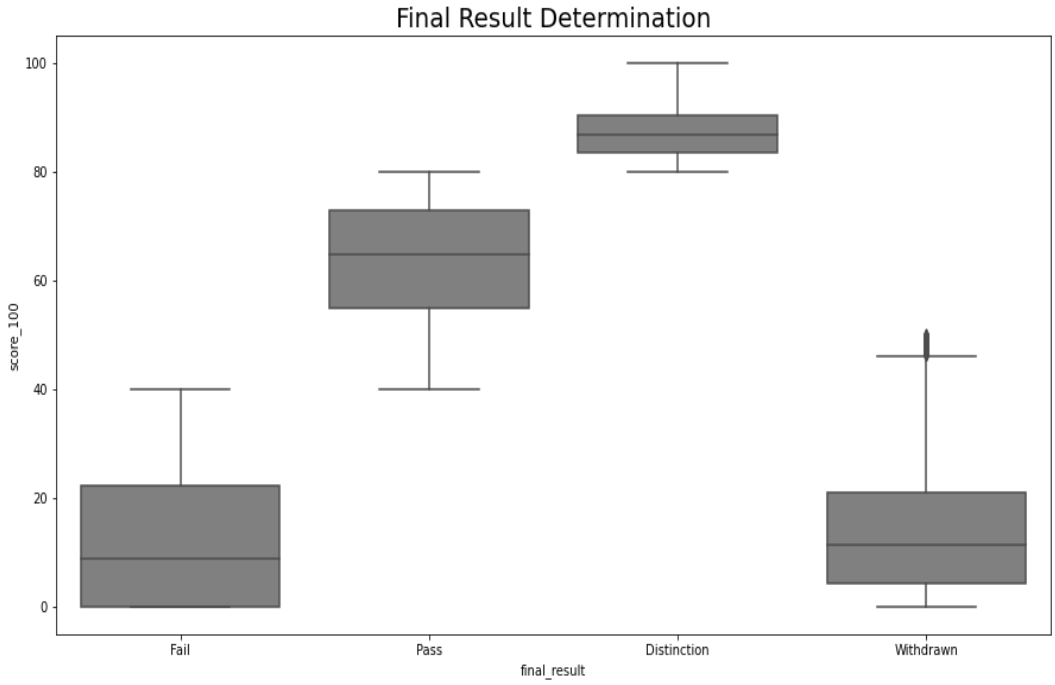

Figure 8 Final Result Determination

- Correlation Matrix

From the correlation matrix, we can see a negative correlation between the Fail class, withdrawn class and the score and a positive correlation between the Distinction class, Pass class and the score.

Figure 9 Correlation Matrix

5.2 Model Summary

The summary of the trained kera sequential model is shown below. It has 3 dense layers, the input layer with 12 neurons, the hidden layer with 8 neurons and the output layer with 4 neurons. It also shows the number of parameters (weights) used on each layer.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 12)                576

dense_1 (Dense)              (None, 8)                 104

dense_2 (Dense)              (None, 4)                 36

=================================================================
Total params: 716
Trainable params: 716
Non-trainable params: 0
_________________________________________________________________
```

Figure 10 ANN Model Summary

5.3 Model Evaluation

The model was evaluated using two different techniques. The classification report and the confusion matrix.

Classification Report

	precision	recall	f1-score	support
0	0.99	0.98	0.99	1109
1	0.73	0.65	0.69	1031
2	0.97	0.99	0.98	2125
3	0.64	0.69	0.66	904
accuracy			0.87	5169
macro avg	0.83	0.83	0.83	5169
weighted avg	0.87	0.87	0.87	5169

Figure 11 Classification Report

The recall value of the classification report means that 98% of the class 0, 65% of the class 1, 99% of the class 2, and 69% of class 3 were found and interpreted. In general, this means that the classifier can find class 0 and 2 better than class 1 and 3.

The precision value means that of the found class, the classifier can accurately predict 99% of the class 0, 73% of class 1, 97% of class 2, and 64% of class 3.

The f1-score means that the harmony of the precision value and the recall value of the class 0 type is 99%, that of class 1 is 69%, 98% for class 2, and 66% for class 3.

The support means that class 0 has a total of 1109 elements in the dataset, class 1 has a total of 1031 elements in the dataset, class 2 has 2125 and class 3 has 904 elements. The overall accuracy of the model is 87%.

Confusion Matrix

Figure 12 Confusion Matrix

The confusion matrix shows that the model predicted 1090 of class 0 of the testing data correctly, and 8 of class 0 incorrectly as class 2. It predicted 667 correctly as class 1, 4 incorrectly as class 2 and 240 incorrectly as class 3. The model predicted 2110 of class 2 correctly, 19, 7 and 39 of class 2 incorrectly as class 0, class 1, and class 3 respectively. Lastly, the model predicted 625 of class 3 correctly, 355 wrongly as class 1 and 3 wrongly as class 3.

Table 10 Accuracy scores for different tested classifiers

Classifier	**Test Ratio (55, 30, 15)**	**Test Ratio (70, 20, 10)**	**Cross Validation**
Artificial Neural Network	86%	87%	Train test split
	N splits = 10		
XGBoost	78.5%		Stratified K-fold
Logistic Regression	78.1%		Stratified K-fold

From the results of the model training, Artificial Neural Network had the highest accuracy of 87%

5.4 Model Deployment

The picture below shows the layout of the web application powered by the trained Machine Learning Model and hosted on Streamlit.

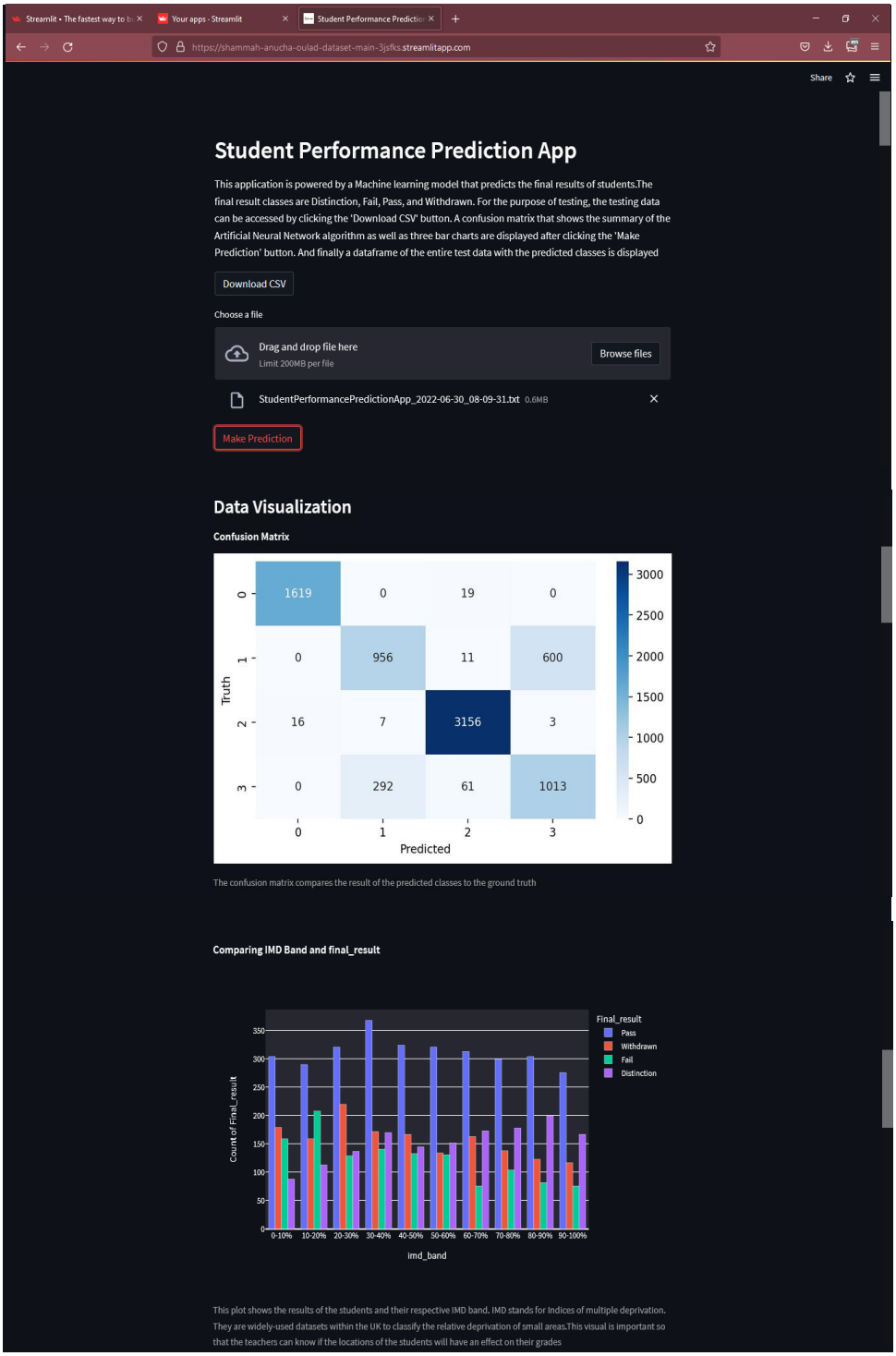

Figure 13 Interface of the Machine Learning Model Website

Chapter 6 – Conclusion and Recommendations

This chapter concludes the entire research. It points out general conclusions arrived at, gives direct answers to the research questions asked in chapter 1, makes recommendations on how to maintain the machine learning model, points out all the limitations and errors encountered throughout the research and finally suggests further studies to embark on. In conclusion, this research focused on using a machine learning algorithm, Artificial Neural Networks (ANN), to help improve the Nigerian E-learning system. This algorithm was chosen because it was implemented in Vijayalakshmi and Venkatachalapathy, (2019) and Hossen *et al.*, (2021) and they had good accuracy scores of 84% and 92% respectively. However, the model was also trained with XGBoost and Logistic Regression algorithms to test for a better accuracy score, but Artificial Neural Network surpassed them. The accuracy score of the Artificial Neural Network (ANN) model was 87% while that of XGBoost and Logistic Regression was 78.5% and 78.1% respectively. The trained model was deployed to the web using Streamlit, an open-source application for deploying machine learning models. The student data used in this research was collected from the Open University website. It was collected from the university's virtual learning platform consisting of 32,593 student data and 7 tables. However only the tables containing the demographics of the students and their scores were used.

6.1 Research question Conclusion

Some answers to the research question stated in chapter were discovered in this study.

RQ1 & RQ4

With proper data mining techniques and model deployment to the web, it is possible to monitor the performance of students. Using the CRISP-DM methodology in conjunction with APREP-DM, data analysts and data scientists can follow a standardized method for analyzing data without skipping any important step as this is crucial to the accuracy of the model predictions. Quite often the model deployment stage is not talked about as research shows in the literature review, but this step is very crucial because it helps users to implement the model that has been trained on a daily or periodical basis. There is no point in spending time cleaning data and creating models that do not end up in production. These machine learning models make the websites more intelligent. In this research, Streamlit was used to deploy the model to the web in a very simple way. Many users can access the webpage and make batch predictions on the student data and get visualizations that can assist them in making informed decisions on the next step to take in preventing student failure or withdrawal from school.

RQ2

This study has highlighted that through machine learning, E-learning platforms can be personalized by creating recommender systems and adaptive systems that monitors the behavior of the user and gives suggestions on what resources the user would be interested in. Personalization was not covered in depth in this study because of the data limitation and time duration of this research. There were features such as the sum clicks that could have been used to model the behavior of student's engagements with the study materials, but this study focused on the classification of students' results instead. However, researchers that have embarked on this topic have been highlighted in the literature review. The machine learning technique used in this project was Artificial Neural Network (ANN). It was used because the prediction problem that was to be solved was a classification problem. There were four classes to be predicted, "Distinction", "Pass", "Fail", and "Withdrawn".

RQ3

To find the possible causes of failure, pandas, numpy, matplot and seaborn library were used to create visualizations of the features. Once again it is important to stress the importance of exploratory data analysis because it helps to spot anomalies in the dataset, discover patterns and to test theories with the aid of a graphical representation. In chapter 4, exploratory data analysis of the features was carried out and some relationships were spotted. One key relationship was the Index of Multiple Deprivation (IMD) and the final result, the graph showed that people living in deprived areas were more likely to fail than people living in well off regions. Features like disability, age band and number of previous attempts were also worth visualizing to spot possible outliers and relationships. The correlation matrix was a very helpful tool to spot strongly related features.

6.2 Recommendations

Retraining of the model should be done at intervals once the model is deployed to production. The model when trained on the observed data had some errors in mapping the input variables to the output variables. Imagine the model trying to predict real world data. Data scientists can assume that the error margins will be similar but in fact, the world changes quickly and the outcomes might be entirely different (Patruno, 2019). The interests of the students can change and hence the model will start predicting wrongly. In order to prevent this from happening, the model must be retrained at regular intervals (Patruno, 2019). One way to achieve this is by checking the distribution of the new dataset and if it has changed significantly from the training data, then the model should be retrained (Patruno, 2019).

Monitoring of the machine learning model in production is very important. Software applications are constantly monitored in production, the same should be done as well for the deployed model. Without monitoring, it will be hard to find out issues in the model on the spot. It might take months before discovering issues with the model, therefore constant monitoring is important. To successfully monitor the model, key data can be recorded while the predictions are being made. These key data include the unique user id assigned by the system using the machine learning; this data is important to keep track of the prediction path before and after it goes through the Machine Learning model, the input features before and after they get processed by feature engineering, the output probabilities, and the predicted value (Verre, 2019). It would be best to have the ground truth label to compare the predicted outcomes but in fact, the ground truth label can only be made available at the end of the session (Verre, 2019). Therefore, the monitoring of the machine learning model would be done after the study session to see how well the model performed. But before the session ends the model can be tracked as explained in the first paragraph above. A couple of metrics should be tracked, and they include, the number of predictions made, how long the model took to make the predictions and the changes in the input and output variable distribution (Verre, 2019).

It is possible to use Streamlit for model deployment in production, but this is best used as a prototype to demonstrate the interface and how the model can be interacted with (Shaji, 2021). FastAPI can be used to serve the model to a production-ready website. It supports data validation using pydantic and automatic API documentation (Shaji, 2021). Finally, Streamlit and FastAPI can be containerized using Docker (Shaji, 2021). Docker is better to use than the built-in capabilities of Linux and other operating systems because it makes containerization easier and faster (IBM Cloud Education, 2022).

6.3 Errors and Limitations

One of the major limitations of this project was that the data used was secondary data from Open University. A real-world data set would have been more practical because there would have been more features that would contribute to the outcome of the student results. Also, there was no access to a functional website, and that is the main reason that Streamlit was selected for the model deployment. A realistic approach would be to deploy the model to the E-learning platform where real time student data can be used to retrain the model at intervals and give updates on the outcome of the predictions.

Although the Artificial Neural Network model had an accuracy of 87%, the classification report shows that the model was not able to distinguish between the Fail and Withdrawn classes. Therefore, more features would be needed to help the model to distinguish between the two classes.

From the correlation matrix, it is seen that of all the independent variables, only the score had a significant correlation score between the dependent variables (Final_result). Having more independent features that correlate with the dependent feature would have strengthened the model to make more accurate predictions. It is also more practical that the score is not the only defining factor because a simpler method will suffice instead of a machine learning method. However, it is possible that more analysis and feature selection can be done to improve the accuracy of the model.

The train_test_split method was used in the Artificial Neural Network (ANN) model. A better approach would be to use the Stratified K-Fold cross validation method because it splits the data into training and testing several times.

There were limitations in using data only from the E-learning platform. This is because some students may not like to use the materials available on the website, but they make use of resources from other locations, hence the data retrieved from the E-learning platform could be misleading and not accurately represent the behavior of the student. Social media platforms would also contain information about the students but because of data privacy, access to this information is restricted (Mengoni *et al.*, 2018).

Computational limitations also hindered the progress of this research. The studentVle table was very large. It contained about 10 million rows of log data; hence the file size was large. It was very difficult to process the large amount of data from a local computer. The loading and importing time took very long hours on tools like POSTGRESQL and Google Colab. Some ways this problem could have been handled are by (i) chunking the data using chunksize library in pandas; (ii) dropping columns that are not needed; (iii) using a parallel computing library called Dask; Dask scales Numpy, pandas and scikit module for fast computation and low memory usage (AskPython, no date). All the listed methods help in reducing the file size.

6.4 Future Works

In a later study, a recommender system that recommends materials for students who are predicted to fail or get withdrawn can be built. This will reduce the failure rate even with little or no

intervention of the teachers. It will serve as a useful tool in the hands of not only the teachers but the students especially because they immediately proffer solutions that can help the students to pick up the pace in their studies. Not only will it be beneficial for students that are slacking off in school, but it can serve as a good tool for students that are doing well to even get better in their study journey. Just like Netflix recommends movies based on the preferences of similar viewers, the recommender system can recommend tutorial videos that have been viewed by students with distinction to students that are predicted to fail. Encouragement notes could also be displayed during the period of their studies to motivate the student to finish the course.

In the future, more machine learning methods can be explored. Machine learning methods like Decision tree, Naïve bayes classifier, Bayesian Network, Support Vector Machine, K-Nearest Neighbour (KNN) can be used. Just like other research in the literature review, a comparison of these algorithms can be made to find out which algorithms perform best. Also, like in (Lykourentzou *et al.*, 2009), this study can be replicated to combine multiple machine learning methods to have even better precision accuracy. The research covers a binary classification technique, but an improvement will be a multinomial classification technique as covered in this study. Because the model developed in this study was not able to properly distinguish the Fail and Withdrawn classes, a technique that allows the combined results of two different machine learning algorithm results would be paramount.

Classification techniques were used in this research; other supervised learning techniques can be used like Regression analysis to predict the actual scores of the students. Unsupervised learning techniques like clustering can be used to group the students that have similar behaviors together. Hidden relationships between the students can also be discovered (Mengoni *et al.*, 2018). Social activities of the students will be relevant features to include in the model to get better clusters that represent the student behaviors and attributes.

References

Alasadi, S.A. and Bhaya, W.S. (2017) 'Review of Data Preprocessing Techniques.pdf', *Journal of Engineering and Applie Sciences*, pp. 4102–4107.

Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A systematic literature review of student' performance prediction using machine learning techniques', *Education Sciences*, 11(9). Available at: https://doi.org/10.3390/educsci11090552.

Alencar, R. (2019) *Dealing With Very Small Dataset*. Available at: https://www.kaggle.com/rafjaa/dealing-with-very-small-datasets (Accessed: 15 February 2022).

Apuke, O.D. (2017) 'Quantitative research methods: A synopsis approach', *Kuwait Chapter of Arabian Journal of Business and Management Review*, (33(5471)), pp. 1–8.

Arcinas, M.M. (2022) 'Design of Machine Learning Based Model to Predict Students Academic Performance', *ECS Transactions*, (107(1)), p. 3207.

AskPython (no date) *Handling Large Datasets for Machine Learning in Python*. Available at: https://www.askpython.com/python/examples/handling-large-datasets-machine-learning (Accessed: 25 June 2022).

Bhiwoo, C. (2019) *Querying Multiple Tables in SQL Server*. Available at: https://www.pluralsight.com/guides/querying-multiple-tables (Accessed: 19 February 2022).

Bikakis, N., Papastefanatos, G. and Papaemmanouil, O. (2019) 'Editorial : Big Data Exploration , Visualization', (December).

Brahim, G. Ben (2022) 'Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features', *Arabian Journal for Science and Engineering*, 47(8), pp. 10225–10243. Available at: https://doi.org/10.1007/s13369-021-06548-w.

Brownlee, J. (2017) *What is the Difference Between Test and Validation Datasets?* Available at: https://machinelearningmastery.com/difference-test-validation-datasets/ (Accessed: 19 February 2022).

Bucklin, R.E. and Sismeiro, C. (2009) 'Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing', *Journal of Interactive Marketing*, 23(1), pp. 35–48. Available at: https://doi.org/10.1016/j.intmar.2008.10.004.

Dey, V. (2021) *Understanding the Importance of Data Cleaning and Normalization*. Available at: https://analyticsindiamag.com/understanding-the-importance-of-data-cleaning-and-normalization/ (Accessed: 19 February 2022).

DSPA (no date) *What is CRISP-DM?* Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: 18 February 2022).

Freitas, A. (2002) 'Data mining and knowledge discovery with evolutionary algorithms', *Springer Science & Business Media* [Preprint].

Gajawada, S. (2019) *Chi-Square Test for Feature Selection in Machine Learning*. Available at: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223 (Accessed: 21 February 2022).

Gazzawe, F. *et al.* (2022) 'The Role of Machine Learning in E-Learning Using the Web and AI-Enabled Mobile Applications', *Mobile Information Systems*, 2022. Available at: https://doi.org/10.1155/2022/3696140.

Gbadebo, B. (2021) *Kantigi Advocates Machine Learning to Boost Nigeria's Education Sector*. Available at: https://celebritygig.com/kantigi-advocates-machine-learning-to-boost-education/ (Accessed: 13 March 2022).

Genç, B. and Tunç, H. (2019) 'Optimal training and test sets design for machine learning', *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(2), pp. 1534–1545. Available at: https://doi.org/10.3906/elk-1807-212.

Glawala Amuda, B., Kidlindila Bulus, A. and Pur Joseph, H. (2016) 'Marital Status and Age as Predictors of Academic Performance of Students of Colleges of Education in the Nort-Eastern Nigeria', *American Journal of Educational Research*, 4(12), pp. 896–902. Available at: https://doi.org/10.12691/education-4-12-7.

Gupta, A. (2019) *One Hot Encoding-Method of Feature Engineering*. Available at: https://medium.com/analytics-vidhya/one-hot-encoding-method-of-feature-engineering-11cc76c4b627 (Accessed: 19 February 2022).

Hagedorn, S., Kläbe, S. and Sattler, K.U. (2021) 'Putting Pandas in a Box'.

Hara, S. *et al.* (2018) 'Feature Attribution As Feature Selection', pp. 1–23.

Hossen, A. *et al.* (2021) 'A Web Based Four-Tier Architecture using Reduced Feature Based Neural Network Approach for Prediction of Student Performance', *International Conference on Robotics, Electrical and Signal Processing Techniques*, pp. 269–273. Available at: https://doi.org/10.1109/ICREST51555.2021.9331003.

Htoon, K. (2020) *Log Transformation: Purpose and Interpretation*. Available at: https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9 (Accessed: 19 February 2022).

Hussain, M. *et al.* (2018) 'Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores', *Computational Intelligence and Neuroscience*, 2018. Available at: https://doi.org/10.1155/2018/6347186.

IBM Cloud Education (2022) *What is Docker?* Available at: https://www.ibm.com/cloud/learn/docker#toc-why-use-do-OyDVNwwD (Accessed: 25 June 2022).

Idrissi, N. and Zellou, A. (2020) 'A systematic literature review of sparsity issues in recommender systems', *Social Network Analysis and Mining*, 10(1). Available at: https://doi.org/10.1007/s13278-020-0626-2.

Imran, A.S., Dalipi, F. and Kastrati, Z. (2019) 'Predicting student dropout in a MOOC: An evaluation of a deep neural network model', *ACM International Conference Proceeding Series*, pp. 190–195. Available at: https://doi.org/10.1145/3330482.3330514.

Jalil, N.A., Hwang, H.J. and Dawi, N.M. (2019) 'Machines learning trends, perspectives and prospects in education sector', *ACM International Conference Proceeding Series*, pp. 201–205. Available at: https://doi.org/10.1145/3345120.3345147.

Johnson, D. (2022) *Supervised vs Unsupervised Learning: Key Differences*. Available at: https://www.guru99.com/supervised-vs-unsupervised-learning.html#:~:text=well “labeled.”-,Unsupervised learning is a machine learning technique%2C where you do,of unknown patterns in data (Accessed: 30 March 2022).

Kordik, P. (2018) *Machine Learning Algorithms for Recommender Systems Part 1(algorithms, evaluation and cold start)*. Available at: https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683d0ed (Accessed: 18 February 2022).

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2004) ‘Predicting students’ performance in distance learning using machine learning techniques’, *Applied Artificial Intelligence*, 18(5), pp. 411–426. Available at: https://doi.org/10.1080/08839510490442058.

Kubat, M. (1999) ‘Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994’, *The Knowledge Engineering Review*, (13(4)), pp. 409–412.

Kumar, A. (2021) *Python – Replace Missing Values with Mean, Median and Mode*. Available at: https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/ (Accessed: 20 February 2022).

Kursa, M.B. and Rudnicki, W.R. (2011) ‘The all relevant feature selection using random forest’. Available at: https://doi.org/arXiv preprint arXiv:1106.5112.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) *Open University Learning Analytics dataset*. Available at: https://doi.org/doi: 10.1038/sdata.2017.171.

Lederer, J. (2021) ‘Activation functions in artificial neural networks: A systematic overview’. Available at: https://doi.org/arXiv preprint arXiv:2101.09957.

Lukic, M. (2021) *One-Hot Encoding in Python with Pandas and Scikit-Learn*. Available at: https://stackabuse.com/one-hot-encoding-in-python-with-pandas-and-scikit-learn/ (Accessed: 21 February 2021).

Lykourentzou, I. *et al.* (2009) ‘Dropout prediction in e-learning courses through the combination of machine learning techniques’, *Computers and Education*, 53(3), pp. 950–965. Available at: https://doi.org/10.1016/j.compedu.2009.05.010.

McKinney, W. (2011) ‘pandas: a Foundational Python Library for Data Analysis and Statistics’, *International Journal of RF and Microwave Computer-Aided Engineering*, 19(5), pp. 583–591. Available at: https://doi.org/10.1002/mmce.20381.

Mengoni, P., Milani, A. and Li, Y. (2018) *Clustering students interactions in eLearning systems for group elicitation*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-95168-3_27.

Minaei-bidgoli, B. *et al.* (2003) ‘Predicting student performance: an application of data mining methods with an educational web-based system’, *33rd Annual Frontiers in Education*, 1, pp. T2A-13.

Mishra, S. and Tyagi, A.K. (2022) *The role of machine learning techniques in internet of things-based cloud applications. In Artificial Intelligence-based Internet of Things Systems*. Edited by P.

Souvik, D. Debashis, and R. Buyya. Springer International Publishing.

Moubayed, A. *et al.* (2018) 'E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics', *IEEE Access*, 6(c), pp. 39117–39138. Available at: https://doi.org/10.1109/ACCESS.2018.2851790.

Myrianthous, G. (2021) *How to Split a Dataset into Training and Testing Sets with Python*. Available at: https://towardsdatascience.com/how-to-split-a-dataset-into-training-and-testing-sets-b146b1649830 (Accessed: 15 February 2022).

NOUN DICT (2022) *NOUN DICT*. Available at: https://nou.edu.ng/ (Accessed: 15 March 2022).

Patino, C.M. and Ferreira, J.C. (2018) 'Inclusion and exclusion criteria in research studies: definitions and why they matter', *Jornal Brasileiro de Pneumologia*, (44), pp. 84–84. Available at: https://doi.org/http://dx.doi.org/10.1590/S1806-37562018000000088.

Patruno, L. (2019) *The Ultimate Guide to Model Retraining*. Available at: https://www.kdnuggets.com/2019/12/ultimate-guide-model-retraining.html (Accessed: 25 June 2022).

Pedro, F. *et al.* (2019) 'Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development Education Sector United Nations Educational, Scientific and Cultural Organization', *Ministerio De Educación* [Preprint]. Available at: https://en.unesco.org/themes/education-policy-.

Pyo, S. *et al.* (2017) 'Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets', *PLoS ONE 12(11)* [Preprint]. Available at: https://doi.org/ 10.1371/journal.pone.0188107.

Queirós, A., Faria, D. and Almeida, F. (2017) 'Strengths and limitations of qualitative and quantitative research methods', *European journal of education studies* [Preprint].

Ray, S. (2019) 'Introduction to Machine Learning and Different types of Machine Learning Algorithms', *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, pp. 35–39.

Ribeiro, R. *et al.* (2020) 'Predicting the tear strength of woven fabrics via automated machine learning: An application of the CRISP-DM methodology', *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, 1, pp. 548–555. Available at: https://doi.org/10.5220/0009411205480555.

Schröer, C., Kruse, F. and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', *Procedia Computer Science*, 181(2019), pp. 526–534. Available at: https://doi.org/10.1016/j.procs.2021.01.199.

Shaikh, A.A. *et al.* (2022) 'The Role of Machine Learning and Artificial Intelligence for making a Digital Classroom and its sustainable Impact on Education during Covid-19', *Materials Today: Proceedings*, (56), pp. 3211–3215.

Shaji, A. (2021) *Serving a Machine Learning Model with FastAPI and Streamlit*. Available at: https://testdriven.io/blog/fastapi-streamlit/ (Accessed: 25 June 2022).

Siddiq, S. (2017) 'Machine Learning in E-Learning Computer Science 4490Z Final Report Machine Learning in E-Learning Department of Computer Science London , Ontario , Canada

Author : Samirah Siddiq Email : ssiddi62@uwo.ca Supervisor : Dr . Hanan Lutfiyya', (May), pp. 0–39. Available at: https://doi.org/10.13140/RG.2.2.36213.37605.

Simpson, O. (2010) '- The CWP Retention Literature Review Final report', (July 2010). Available at: https://doi.org/10.13140/RG.2.2.15450.16329.

Solano, J.A. *et al.* (2021) 'Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test', *Procedia Computer Science*, 198(2020), pp. 512–517. Available at: https://doi.org/10.1016/j.procs.2021.12.278.

Sood, S. and Saini, M. (2021) 'Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation', *Education and Information Technologies*, 26(3), pp. 2863–2878. Available at: https://doi.org/10.1007/s10639-020-10381-3.

Telus Communications Inc (2022) *Seven Types of Data Bias in Machine Learning*. Available at: https://www.telusinternational.com/articles/7-types-of-data-bias-in-machine-learning (Accessed: 15 February 2022).

Thai-Nghe, N. *et al.* (2010) 'Recommender system for predicting student performance', *Procedia Computer Science*, 1(2), pp. 2811–2819. Available at: https://doi.org/10.1016/j.procs.2010.08.006.

Thetechwriters (2021) *Machine Learning Model Deployment Using Streamlit*. Available at: https://www.analyticsvidhya.com/blog/2021/10/machine-learning-model-deployment-using-streamlit/ (Accessed: 22 June 2022).

Verre, J. (2019) *Life of a Model After Deployment*. Available at: https://towardsdatascience.com/life-of-a-model-after-deployment-bae52eb83b75 (Accessed: 25 June 2022).

Vickery, R. (2020) *The Art of Finding the Best Features for Machine Learning*. Available at: https://towardsdatascience.com/the-art-of-finding-the-best-features-for-machine-learning-a9074e2ca60d (Accessed: 18 February 2022).

Vijayalakshmi, V. and Venkatachalapathy, K. (2019) 'Comparison of Predicting Student's Performance using Machine Learning Algorithms', *International Journal of Intelligent Systems and Applications*, 11(12), pp. 34–45. Available at: https://doi.org/10.5815/ijisa.2019.12.04.

Wirth, R. and Hipp, J. (2000) 'CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39', *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), pp. 29–39. Available at: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining.

Wolpert, D. (1992) 'Stacked Generalization (Stacking)', *Neural Networks*, 5, pp. 241–259.

Xing, W. *et al.* (2016) 'Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization', *Computers in Human Behavior*, 58, pp. 119–129. Available at: https://doi.org/10.1016/j.chb.2015.12.007.

Xu, J., Moon, K.H. and Van Der Schaar, M. (2017) 'A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs', *IEEE Journal on Selected Topics in*

Signal Processing, 11(5), pp. 742–753. Available at: https://doi.org/10.1109/JSTSP.2017.2692560.

Xu, Y. and Goodacre, R. (2018) 'On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning', *Journal of Analysis and Testing*, 2(3), pp. 249–262. Available at: https://doi.org/10.1007/s41664-018-0068-2.

Yakubu, M.N. and Abubakar, A.M. (2022) 'Applying machine learning approach to predict students' performance in higher educational institutions', *Kybernetes*, 51(2), pp. 916–934. Available at: https://doi.org/10.1108/K-12-2020-0865.

Zeebaree, D.Q. *et al.* (2021) 'Machine Learning Semi-Supervised Algorithms for Gene Selection: A Review', *2021 IEEE 11th International Conference on System Engineering and Technology, ICSET 2021 - Proceedings*, (November), pp. 165–170. Available at: https://doi.org/10.1109/ICSET53708.2021.9612526.

Appendix

1. Data Analysis in Google Colab. Available at: https://colab.research.google.com/drive/1-f9j8u_OytktooC4xDiYFzKarEIi-Odu#scrollTo=hhinmacz5a61
2. Private repository on Github for source code. Available at: https://github.com/shammah-anucha/Oulad_dataset Please note that permission needs to be granted for access.
3. Student Performance Prediction Application on Streamlit. Available at: https://shammah-anucha-oulad-dataset-main-3jsfks.streamlitapp.com/
4. OULAD Dataset: Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171