PREDICTION OF STUDENT ACADEMIC PERFORMANCE

BY

JOSHUA DAVID YAKUBU

BHU/16/04/05/0026

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF SCIENCE AND TECHNOLOGY, BINGHAM UNIVERSITY, KARU NASARAWA STATE. IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE DEGREE (B.Sc.) IN COMPUTER SCIENCE

SEPTEMBER, 2020

CERTIFICATION

This is to certify that this project report titled "**PREDICTION OF STUDENT ACADEMIC PERFORMANCE"** was carried out and written by **JOSHUA DAVID YAKUBU** with matriculation number **BHU/16/04/05/0026** in partial fulfilment for the award of Bachelor of Science (B.Sc.) in computer science, Faculty of Science and Technology, Department of Computer Science, Bingham University, Karu, Nassarawa state.

Signature Adelaiye Date 2nd October, 2020

Mr. Adelaiye Oluwasegun

(Supervisor)

Signature ____________________________ Date ____________________

Dr. Faki Silas

(Head of Department)

DEDICATION

This project is dedicated to the Lord God Almighty for his sufficient grace, love, wisdom and support during the course of this project.

ACKNOWLEDGEMENT

With a deep sense of appreciation, respect and gratitude, I want to say a big thank you God Almighty who has made it possible for the entire project to be a success.

Mr. Adelaiye, Oluwasegun for his support, guidance, and making this project a success and the lecturers of the department. My deepest appreciation goes to the H.O.D. of computer science, Dr. Faki Silas. A special thanks to my supervisor.

Last but not the least, let me express my sincere gratitude to my ever loving, caring, selfless parents Mr. and Mrs Yakubu for their love and support, my siblings Marvellous and Wonderful for being there for me all the times.

ABSTRACT

Educational Data Mining techniques are used to collect useful information from raw data. Any educational institution goals are to ensure students or people attain knowledge or skills in the duration of their stud but factors such as the number of students' who fail that negatively affect the reputation of the institution. This project employed the using of machine learning techniques in order to predict the final (G3) score of students in two secondary schools in Portugal. Of the regression and classification techniques tested, the gradient boosting technique yielded the best results and was deployed on a student management system.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER ONE

1.0 INTRODUCTION

A school is an education institution that awards academic degrees in various academic disciplines. One of the factors that determine a good university is the number of people who graduate from university. In regards to this, it is important to be able to find factors that affect students' performance and this can be done with Data Mining. It involves the method of discovering new aspects and trends from a large data collection utilizing machine learning, analytics, and computer structures.

Its application in the academic sector is known as Educational Data Mining (EDM). It is increasingly in demand and draws further interest because of the growth in academic information of learning systems and even the development in traditional education. Probing traditional database records can provide answers to problems such as "finding students who have failed the tests," while Educational Data Mining provides answers to additional problems such as "predicting students who are more likely to pass". Various machine learning techniques such as Linear Regression, Artificial Neural Networks, Support Vector Regression, and Clustering will be used to predict students' future performance.

1.1 BACKGROUND OF THE STUDY

Historically, EDM is a relatively new field in science. While researchers have been collecting and analysing data from educational software for a long time, conferences have only recently established EDM as a field of its own (Scheuer & McLaren, 2011). Nowadays many tools and techniques are available to us that can change or improve the education system. The increasing digitisation of educational data has helped the researcher easily capture these available data and extract meaningful information in order to take corrective decisions. Educational data is growing rapidly as more and more education system store information online. New areas have been opened like new computer-supported interactive learning methods and tools-intelligent tutoring systems, simulation games which have opened up opportunities to collect and analyze student data, discover patterns and trends in those data and to make new discoveries and evaluating ideas on how students learn through classes. The data collected from online learning systems can be aggregated over a large number of students and can include many variables that can be explored by data mining algorithms for model building. In today's era, educational systems try to offer a customized learning method, by understanding the individual's goals, attitude, and knowledge. Educational Data Mining can be seen as an iterative cycle of hypothesis formation, testing, and refinement. The educators are responsible for the design, planning, installation and management of education systems, while students use and interact with them. The application of data mining is different for educators and students. For students, The aim is to identify activities , resources and learning tasks that will improve their learning, based on their attitude and interests while for educators(), the goal is to have more feedback from students for evaluating the structure of the course content and its effectiveness on the learning process, to classify students based on their needs, to discover information to improve the adaptation and customization of the course, etc. Educational Data Mining (EDM) is an upcoming field in Knowledge discovery. Due to the widespread growth of higher education, predictions related to student's performance can be accurately done through EDM. Not only predictions, classification, associations, and grouping can also be done with perfection using statistical and software tools. The Education system can be equipped with more information relating to the future drop out of students and their success in enrolled courses. Not only students but other stakeholders could be benefitted from EDM. Nowadays interactive e-learning methods and tools have opened an opportunity to collect and scrutinize student data. In the educational field, data mining techniques can generate useful patterns that can be used both by educators and learners. Not only may EDM assist educators to improve the instructional materials and to establish a decision process that will modify the learning environment or teaching approach, but it may also provide recommendations to learners to improve their learning and to create individual learning environments (Srivastava & Sirvastava., 2015).

1.2 STATEMENT OF THE PROBLEM

One of the problems in education institutions is the dropout students which are students who leave the institution without completion of the course. This could be as a result of their bad performance in academic activities or unruly behavior in the academic environment. In the case of bad performance, the number of failures in certain subjects or courses can prevent the students from continuing their classes.

1.3 AIM AND OBJECTIVES

The main aim of the project is to build a prediction model to predict the future performance of student using certain data features.

At the end of this project the following objectives must be completed:

1. Review of previous literature and systems.
2. Planning, appraoch and design of proposed systtem with a well detailed shematic.
3. Analysis and performance results of the proposed system against existing systems.

4. Summary of work done, conclusion and recommendations.

1.4 SCOPE OF THE STUDY

This project is focused on predicting the students' score using scores of secondary students in Alentejo region of Portugal. The dataset was collected from Gabriel Pereira and Mousinho da SIlveria secondary schools and contains 649 records of these students with 33 attributes. Machine learning techniques especially classification and regression models would to built to predict the final score. The model will be deployed on a student management application and predict student's scores. .

1.5 MOTIVATION

I have decided to carry out this project to show the importance of Education Data Mining in schools and show how it can be applied in these institutions. I will also like show how different machine learning techniques can be utilized in the field of education.

1.6 SIGNIFICANCE OF STUDY

This research is very important as it serves as a gateway to bigger studies on different techniques involving the use of machine learning techniques to improve the quality of education an institution can provide. It focuses on institutions using the data at their disposal to provide answers to questions needed to be answered such as why the students fail, why the students succeeded, what teacher technique works best etc.

1.7 ORGANIZATION OF THE STUDY

1. Chapter One – Introduction: This chapter has to do with the introduction, background of the study, scope, aims and objectives, and motivation.

2. Chapter Two - Literature Review: This chapter deals with the research on the related literature based on the prediction of student's performance using machine learning techniques.
3. Chapter Three - System Design, Methodology, and its Application: This chapter has to do with the methodology used to design and implement the student's performance prediction system.
4. Chapter Four - System Development: This chapter deals with the processes involved in designing and implementing the student's performance prediction system.
5. Chapter Five - Evaluation, Conclusion, and Recommendation: This chapter deals with the summary and conclusion of the study project.

1.8 DEFINITION OF TERMS

a. Education Data Mining: It is an emerging discipline that seeks to develop methods for exploring unique types of data from educational settings and using those methods to better understand students and the settings they learn in. https://edtechreview.in/dictionary/394-what-is-educational-data-mining

b. Data Mining: It is looking for unknown, relevant, and potentially useful patterns in huge data sets. https://www.guru99.com/data-mining-tutorial.html

c. Machine Learning: Machine learning is an artificial intelligence (AI) technology that provides systems with the ability to learn and develop automatically from experience without explicit programming. https://expertsystem.com/machine-learning-definition/

d. Linear Regression: is a machine learning algorithm used for predicting values or numbers based on independent variables or factors

e. Artificial Neural Network: is a computational model designed after neuronal activity in the human brain. https://searchenterpriseai.techtarget.com/definition/neural-network

f. Clustering: It is a Machine Learning technique that involves the grouping of data points. https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

g. Django: It is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. https://www.djangoproject.com/

h. CSS: Cascading Style Sheet is used to determine standards for your web pages, including layout, format and interface variations to different devices and screen sizes. https://www.w3schools.com/css/css_intro.asp

i. PostgreSQL: is an open-source, object-relational database management system (ORDBMS) that is not owned or controlled by one company or individual. https://www.techopedia.com/definition/3499/postgresql

j. HTML: HyperText Markup Language is the language used to create webpages. https://techterms.com/definition/html

k. JavaScript: is a scripting language primarily used on the web to enhance HTML pages and render these pages in a dynamic and interactive fashion. https://www.techopedia.com/definition/3929/javascript-js

CHAPTER TWO

LITERATURE REVIEW

2.0 INTRODUCTION

The literature review reveals that these problems have been of interest for various researchers during the last few years. The development of data mining models for predicting student performance at various levels, and comparison of those models, are discussed in a number of research papers.

2.1 KNOWLEDGE DISCOVERY IN DATABASES

Knowledge Discovery in Databases (KDD) is concerned with various means of making meaning out of data or discovering or extracting facts or information out of the data. KDD can be defined as the overall process of discovering useful knowledge from data (Vanessa & Paska, 2020). KDD is necessary in any sector as the traditional means of turning data into knowledge can be slow and expensive and as data grows in that particular sector, these factors get worse to a point that it becomes unrealistic to search through all the data. KDD is applied in various areas, these includes: education, marketing, investments, fraud detection, manufacturing, medicine, telecommunications etc.

2.2 DATA MINING

Data mining is the use of specific algorithms to extract new aspects and patterns from a large dataset using machine learning, database systems and statistics (Sumit, 2020). Educational Data Mining (EDM) is an evolving discipline that aims to establish methods to explore unique and increasingly large-scale data from educational settings and to use these methods to better understand students and the settings they learn in (Rajni & Malaya, 2013). EDM is data mining that is focused on educational data. EDM focuses on taking the educational data and performing several Data Mining techniques to perform various tasks such as prediction e.g. prediction of scores or grades of students. EDM can also be used for discovering groups of students that have similar traits (Berland, Baker, & Blikstein, 2014).

2.3 CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

Cross Industry Standard Process for Data Mining (CRISP-DM) is a repetitive process that provides a structured approach to data mining processes (Rob, 2018). This reliable data mining model consisting of six phases. Fig 2.1 shows these six phases

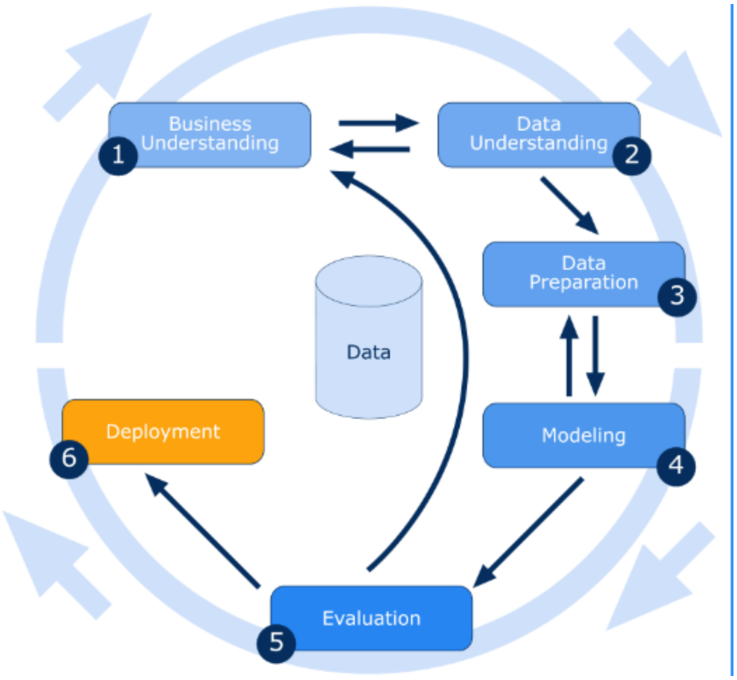


Fig 2.1 Data mining processes (Rob, 2018)

The CRISP-DM six phases are:

1. Business Understanding – the purpose of this phase is to understand the business objectives. It also involves understanding the current situation of the business by finding out the data available, constraints, assumptions made concerning the business objectives. Then, defining an acceptable criterion for the model to be created. (Sumit, 2020)**.**

2. Data Understanding – the phase involves acquiring the data available in the business and studying the data to determine if facts or information can be attained. This can include finding out the relationship between various attributes, finding the mean, median, mode of the data etc. It also includes determining the quality of data which includes finding out how many records have missing data, how the features available correlate to the business objectives etc (Sumit, 2020).
3. Data Preparation – this involves getting the data suitable as an input to the machine learning model. The format of data can determine how well the machine learning model works in performing its task. It involves making sure that the data is free from outliers and that records do not contain missing data. This also involves changing categorical variables to the appropriate format and sometimes data transformation which involves changing the format of the data (Sumit, 2020).
4. Modelling – this involves putting the prepared data into a machine learning model in order to see results. The machine learning model chosen could be a regression or classification model. The prepared dataset is split into training and testing data sets and the records are split randomly. The model is fed with the training data set to try and get a generalization view of the dataset and the test data tests how good or accurately the model performs. The model is tuned by changing its hyper parameters which can help improve how accurately the model performs (Sumit, 2020)**.**
5. Evaluation – the process involves analysis of the results of the model and determining if it is good enough to solve the business objective. It also involves determine if the features chosen as input in the model would be relevant or available in the future of the business (Sumit, 2020)**.**

6. Deployment – this involves choosing the best model and incorporating it to the production systems such as web, desktop or mobile applications where predictions can be made in real time. It also involves the maintenance of the model and checking to verify that the model is performing the task accurately (Sumit, 2020).

2.4 PREDICTION OF STUDENT ACADEMIC PERFORMANCE USING NEURAL NETWORK, LINEAR REGRESSION AND SUPPORT VECTOR REGRESSION: A CASE STUDY

This research presents the prediction of the academic performance of students from the educational database using their scores in particular, without economic, social and psychological factors. The dataset used was obtained from the Student Information System (SIS) of Hawassa University for the School of Computer Science (Obsie & Adem, 2018). It comprises 134 undergraduate students who graduated from the university in the year 2015, 2016 and 2017 which consisted of 52(38.81%), 39(29.10%) and 43(32.09%) students respectively. Each student record has the students' ID, sex, mobile number, section, entry year, nationality, University Entrance Examination Results (UEER), Course scores, Grade Point Average (GPA) at the end of each semester and the final Cummulative Grade Point Average (CGPA). The attributes used in the prediction of the final CGPA are the scores of each course and the GPA at the end of each semester. The dataset was cleaned by removing students who didn't complete their education in the institution and dealing with missing values. The dataset is transformed in order to meet the requirements of the data mining algorithms i.e. Neural Network (NN), Linear Regression (LR) and Support Vector Regression (SVR) which were implemented using Waikato Environment for Knowledge Analysis (WEKA) toolkit. To evaluate the data mining algorithms on how well it works, the Root Mean Square Error (RMSE) is used, small RMSE values give an indication of good prediction of the target values. The correlation coefficient is used to show if the linear relationship between the scores of the student and final CGPA is positively or negatively correlated, if it is positively correlated, the increase in the scores will cause an increase in the final CGPA of the student but if it is negatively correlated, the increase in the scores will cause a decrease in the final CGPA. The Statistical Package for the Social Sciences (SPSS) was used for the correlation analysis. There were three scenarios used in predicting the final CGPA, these include:

1. The students' university course scores from the first 2 years: in this scenario, SVR produced the best results with correlation coefficient of 0.9305 and 0.1608 RMSE values, seconded by LR with correlation coefficient of 0.9239 and 0.1675 RMSE values and NN was the least accurate with correlation coefficient of 0.9089 and 0.1900 RMSE values (Obsie & Adem, 2018)..
2. The students' university course scores from the first 3 years: in this scenario, LR produced the best results with correlation coefficient of 0.9758 and 0.0954 RMSE values, seconded by SVR with correlation coefficient of 0.9742 and 0.0992 RMSE values and NN was the least accurate with correlation coefficient of 0.9511 and 0.146 RMSE values.
3. The students' Semester GPA at the end of each semester from the first 3 years: in this scenario, LR and SVR produced equal prediction results of the correlation coefficient of 0.9805 and RMSE values of 0.0857 and 0.0862 respectively. The least accurate is NN with a correlation coefficient of 0.9763 and a RMSE of 0.100 (Obsie & Adem, 2018)..

2.5 STUDENT'S PERFORMANCE PREDICTION USING DEEP LEARNING AND DATA MINING METHODS

This research is focused on the prediction of student performance. The dataset was collected from the Learning Management System from a Saudi University database with 1100 student records (Sultana, Rani, & Farquad, 2019). It has eleven different features which are: the student's ID, the number of times hands were raised in class, the number of times resources were visited, the number of announcements viewed, the number of times discussions were made, the parent answering survey, the parent school satisfaction, student absence days, internal assessment, external assessment and the total marks. The total marks were classified into three classes which are:

1. LOW - scores between 0 - 69
2. MEDIUM - scores between 70 and 84
3. HIGH - scores between 85-100

Various models were implemented which includes: Multilayer perceptron (MLP), Multi Class Classifier, Support Vector Machine (SVM), Naive- Bayes, Instance Based Learner (IBK), Lazy Locally Weighted Learning (Lazy LWL), Random Forest and Decision Tree which are implemented using the WEKA toolkit (Sultana, Rani, & Farquad, 2019). These models are tested over a variety of metrics and the results are shown in Table 2.1.

Table 2.1: the various models implemented with their various metrics used to show the accuracy of the models (Sultana, Rani, & Farquad, 2019).

Methods	Accuracy	TP	FP	Kappa Statistics	ROC
MLP	99.45	1.00	0.00	0.99	1.00
Multi class classifier	99.81	1.00	0.00	0.99	1.00
SVM	93.90	1.00	0.10	0.89	0.94
Naive-Bayes	97.45	0.98	0.00	0.95	0.99
IBK	79.81	0.91	0.16	0.63	0.87
Lazy LWL	86.72	1.00	0.00	0.75	1.00
Random Forest	**100**	1.00	0.00	1	1.00
Decision Tree	**100**	1.00	0.00	1	1.00

Table 2.1 shows that the models that performed the best is the random forest algorithm and the decision tree algorithm.2.6 PREDICTING STUDENTS' RESULTS IN HIGHER EDUCATION USING A NEURAL NETWORK

This research is focused on the prediction of the GPA of students after the first year of study in order to prevent or reduce the number of students who leave the school due to bad grades (Oancea et al, 2017). The dataset consists of 1000 samples from the previous three graduates' generation of Nicolae Titulescu University in Bucharest. The input data used for the prediction model are: the type of study program(part time/full time education), the gender of the student, high school graduation GPA, age of the student, difference in years from the moment the students graduate high school until he/she enrolls at university. Students were classified in three classes according to their GPA, these classes are:

1. POOR RESULTS - those students with GPA lower than 6
2. MEDIUM RESULTS - those students with GPA between 6 and 8
3. GOOD RESULTS - those students with GPA greater than 8

A deep learning model implemented with encog framework using java programming language was implemented with one input layer, two hidden layers and an output layer. The input layer comprises seven neurons which comprises the input data. The first hidden layer comprises 50 neurons, the second hidden layer consists of 400 neurons and the output layer consists of three neurons, each representing the classes of the student GPA. 800 of the samples were used as training data while 200 were used as test data. After training the model, it was discovered that the model had 86.6% of accurately predicting if a student would get poor results, 94.2% accuracy in predicting a student with medium results and 85.7% accuracy in predicting a student with good results (Oancea et al, 2017)

2.7 DATA MINING USING ENSEMBLE CLASSIFIERS FOR IMPROVED PREDICTION OF STUDENT ACADEMIC PERFORMANCE

This research aims to predict student academic performance using classification, filtering and association rule mining. The dataset used was collected from two secondary schools from the alentejo region in Portugal (Satyanarayana & Nuckowski, 2016). During the preprocessing stage, 111 students were removed from the dataset due to incomplete data entry. The dataset contains 395 students for mathematics and 649 students for Portuguese. There were 33 attributes/features used as input for the models.

In this work, the goal was to create models that could classify scores for mathematics and Portuguese grades based on the European Exchange program (Erasmus) grade conversion system where:

1. A = 16-20
2. B = 14-15
3. C = 12-13
4. D = 10-11
5. F = 0-9

These models include the decision tree, online bagging and ensemble filtering

Table 2.2 : Predictive accuracies after using the different classification techniques

Dataset	Predictive accuracy of student academic performance		
	Decision Tree (j48)	Online Bagging	Ensemble Filtering
Mathematics	0.78	0.82	**0.95**
Portuguese	0.71	0.79	**0.94**

In Table 2.2, using association mining techniques such as apriopi, filtered associator and tertius. It was found that ensemble voting provided stronger factors that determine student achievement than using any individual algorithm.

Another classification model was created for a different dataset which was the First year Computer Systems Technology students from the New York City College of Technology (CUNY). The attributes of each record in the database comprises two test scores, mid-term score and a final. The goal was to predict the final grade given two test scores and mid-term score. The five-level classification of the final grade is:

1. A = >=80
2. B = 60-79
3. C = 40-59
4. D = 30-40
5. F = <30

The same models were used as shown in Table 2.6

Table 2.3: Predictive accuracies after using the different classification techniques

Dataset	Predictive accuracy of student academic performance		
	Decision Tree (j48)	Online Bagging	Ensemble Filtering
CST Course	0.63	0.75	**0.91**

From Table 2.3, it was found that ensemble classifier was the best classifier and that ensemble filters show a huge improvement in predictive accuracy.

2.8 PREDICTING STUDENTS' ACADEMIC PERFORMANCES – A LEARNING ANALYTICS APPROACH USING MULTIPLE LINEAR REGRESSION

The research involved the prediction of students' academic performance in CS201 using the dataset that was obtained from the university of Jos, Nigeria (Oyerinde & Chia, 2017). Multiple linear regression was the algorithm used in this study. The dataset comprises of students' scores in Math 103, Math 203, Math 205 and CS201. Statistical Package for Social Scientists (SPSS) was used to implement the model. The metrics used to evaluate the accuracy of the model are: R, R square, Adjusted R square and standard error of the estimate. After preprocessing and evaluating the model, it was found that the R square value is 0.890. This means that the scores in Math 103, Math 203 and Math 205 comprises 89% of what determines the scores of CS 201.

2.9 MODELLING, PREDICTION AND CLASSIFICATION OF STUDENT ACADEMIC PERFORMANCE USING ARTIFICIAL NEURAL NETWORKS

This research focuses on the prediction of the students' performance using the dataset from University Q in China (Lau et al, 2019). The Artificial Neural Networks (ANN) is the algorithm used for the prediction of their performance . It comprises 1000 students within three departments from the year 2011 - 2013, with 249 being male and 751 being female. Each record comprises gender, location, whether or not repeating student, previous school area, parents' occupation and entrance exam results in Chinese, English, Maths, comprehensive science and proficiency test. In the preprocessing stage, it was found that female students tend to get higher CGPA than male students, also it was found that repeating students tend to score lower CGPA than current students and that the location of the student either urban or rural does not affect the CGPA of the students. ANN was modelled using the MathWorks MATLAB software.

The ANN comprised of eleven input neurons, two hidden layers consisting of 30 neurons and a single output neuron which is the students' predicted CGPA. The metrics used in evaluating the ANN model are: The Mean Square Error (MSE), regression analysis, error histogram, confusion matrix and Receiver Operating Characteristics Curve (ROC). The MSE was approximately equal to 0.27, the R-value was 0.64, in the error histogram, most errors occur near the zeroth point. The confusion matrix shows that 95% of girls and 55% of boys were correctly predicted in terms of what sex the student was and lastly the ROC evaluates the effectiveness of an ANN's accuracy in prediction and classification, the AUC value is 0.86 is achieved which is quite successful as an

AUC with the value of 1 represents a perfect test. Overall, the ANN achieved a good prediction accuracy of 84.8%.

2.10 A COMPARATIVE STUDY TO PREDICT STUDENT'S PERFORMANCE USING EDUCATIONAL DATA MINING TECHNIQUES

The research aims to compare Bayesian networks and decision trees as classification methods to predict whether students will drop out or not in the department of Industrial Engineering (Khasanah & Harwati, 2017). The dataset was obtained from the Universitas Islam Indonesia's Information System (UNISYS) of 2007. Initially, the dataset contained 178 students, after cleaning the data, 104 students with 13 (12.5%) classified as dropouts and 91 (87.5%) classified as not. The dataset has 12 features/attributes which include: gender (GE), origin(OR), father education (FE), father occupation (FO), mother education (ME), mother occupation (MO), senior high school type (ST), senior high school department (SD), senior high school final grade (SF), first semester attendance (AT), first semester GPA (GPA) and drop out or not. Table 2.4 shows the feature selection methods and the features selected based on certain criteria of each method.

Table 2.4 Feature selection result

Feature selection method	Selected attributes/features
Correlation-based Attribute evaluation	MO, AT, GPA
Gain-Ratio Attribute evaluation	GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF
Information-Gain Attribute evaluation	GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF
Relief Attribute evaluation	AT, GPA, FO, SD, FE, MO, GE, ME, OR, ST, SF
Symmetrical Uncertainty Attribute	GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF

Table 2.5 shows that were three scenarios chosen for the classification analysis, these are:

1. Using all the attributes.
2. Using selected attributes from the feature selection result.
3. Using the most selected attributes.

Table 2.5 Classification results

Scenario	Attributes Used	Classification Algorithm	Accuracy Rate	Number of Incorrectly Classified Instance
1	All	Bayesian Network	95.19%	5
		Decision Tree	93.27%	7
2	GPA, AT, SD, GE, FO, MO, ME, OR, FE	Bayesian Network	97.11%	3
		Decision Tree	94.23%	6
3	GPA, AT, SD, GE, FO, MO, ME, OR	Bayesian Network	**98.08%**	2
		Decision Tree	94.23%	6

From Table 2.5, it can be seen that the most accurate predictor is the Bayesian Network which uses the most selected features.

2.11 STUDENT PERFORMANCE PREDICTION

The research is focused on predicting the performance of regular and Direct Second Year students (DSE - students who joined after completing their diploma) of the engineering institute of Mumbai University (Shetty, Shetty, & RoundHal, 2019). The Student Grade Prediction Index (SGPI) is a value used to determine the academic performance of the students. The dataset contains 6050 instances and 55 attributes. Some of these attributes are: parents' name, date of birth, address, student type (regular or DSE), SSC Board, SSC Percentage, HSC Board, HSC Percentage, DSE Board, DSE percentage, Semester Marks (Graduation), Number of KT's, Education Gap, Branch and Admission type (CAP or minority). The dataset was preprocessed and was discovered that some of the attributes had null values such as the semester pointers, the null values were replaced with the medians of the columns. Due to the unbalanced nature of the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was introduced to handle the imbalance. The classification of the dataset was achieved using several models such as Random forest, Bagging, Boosting, Naive Bayes, Decision Tree, Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP). The results for the classification of 3-category and 5-category for the regular and diploma students are shown in table 2.6 and table 2.7.

Table 2.6 Results of the algorithms for regular students

Algorithms	Accuracy	
	3 category (*100)	5 category (*100)
Random Forest	**0.99**	**0.99**
Gaussian Naive Bayes	0.78	0.87
Support Vector Machine	0.70	0.81
Bagging	**0.99**	**0.99**
Boosting	0.98	0.99
Decision Tree	0.98	0.98
Linear Regression	0.73	0.78
MLP	0.85	0.77
CNN	0.92	0.80

Table 2.7 Results of the algorithms for diploma students

Algorithms	Accuracy	
	3 category (*100)	5 category (*100)
Random Forest	**0.99**	**0.99**
Gaussian Naive Bayes	0.97	0.95
Support Vector Machine	0.93	0.88
Bagging	**0.99**	**0.99**
Boosting	0.98	**0.99**
Decision Tree	0.98	0.98
Linear Regression	0.73	0.79
MLP	0.62	0.77
CNN	0.88	0.78

From the table 2.6, it is seen that Random Forest and Bagging performed the best for both the 3-category and in table 2.7, 5-category classification for regular students while in the 3-category and 5-category classification for diploma students, Random Forest and Bagging performed the best. In addition, Boosting also performed the best in 5-category classification for diploma students.

2.12 STUDENT PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS

The research is focused on classifying students as either weak if the student average university score is less than 4.5 or strong if the score is equal to or higher than 4.5 (Kabakchieva, 2012). The dataset was obtained from the University of National and World Economy Sofia, Bulgaria.

At first, it contained 10330 instances but after data preprocessing, it contained 10067 instances and 14 attributes which include: gender, age, birth year, PlacePrevEdu, ProfilePrevEdu, ScorePrevEdu, Admission Year, Admission Exam, Admission Exam Score, Admission Score, UnivSpecialityName, CurrentSemester, NumFailures and Student Class (weak or strong). The data mining algorithms used are Rule learner (OneR), Decision Tree (J48), Neural Network and K-nearest neighbour. These algorithms were implemented using the WEKA toolkit. The metrics used to evaluate the algorithms are percentage of correctly/incorrectly classified instances, Kappa Statistic, True Positive (TP) and False Positive (FP) Rates, Precision, Recall, F-Measure and ROC Area. The results are shown in table 2.8.

Table 2.8: Achieved Results for the data mining algorithms

Data mining algorithms	Overall Accuracy
Rule learner (OneR)	67.4554%
Decision Tree (J48)	72.7432%
Neural Network	**73.5904%**
K nearest Neighbour	70.47%

Table 2.8 shows that Neural Network has the highest among all the models with an accuracy of 73.5904%.

2.13 SUMMARY

In this section, the knowledge discovery of databases and its steps were discussed. The literature review shows different many regression and classification algorithms used to predict student's performance using different features. Some of the regression and classification algorithms will be discussed in the next chapter to understand how they work.

CHAPTER THREE

3.0 INTRODUCTION

The system is a web-based application designed to predict the final score of the student and store information about students, teachers and the school administrator. In order to predict the final score of the student, certain attributes are required such as the student's gender, mother's job, father's job etc. In this chapter, the research methodology used in this project is described with an in-depth look at the proposed system and the processes undertaken in the development of the system. The approach adopted is structured and systematic enabling the study of how the system will perform. This study involves a series of choices, which include the choices about what information and/or data to gather and how to analyze the data gathered. The chapter starts by roughly describing the data selection and pre-processing.

3.1 DATA SELECTION AND PRE-PROCESSING

Data can be defined as set of facts and statistics collected together for reference and analysis. In computing, data refers to information that has been translated into a form that is efficient for processing. Data usually needs to be processed into reliable information for utilization. In order to develop user friendly software, data needs to be collected and analyzed.

Previous academic result plays a major role in predicting the student's future academic outcome. The dataset used include student grades, demographic, social and school related features collected from two Portuguese secondary schools in two distinct subjects: Mathematics and Portuguese Language. It was collected using school reports and questionnaires. The dataset comprises of 649 instances and 33 attributes (Cortez and Silva, 2008). These attributes are either numerical data which are attributes composed of only numbers such as integers and floatingpoint values or categorical data which are attributes that contain label values. There are 14 numerical attributes and 17 categorical attributes. Table 3.1 below show the different attributes.

Table 3.1 Attributes in the dataset

Attribute	Information
School	Student's school (categorical (binary)): 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira).
Sex	Student's sex (categorical (binary)): 'F' - female or 'M' - male)
Age	Student's age (numeric: from 15 to 22)
Address	Student's home address type (categorical (binary)): 'U' - urban or 'R' - rural)
famsize	family size (categorical (binary): 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (categorical (binary): 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Reason	Reason to choose this school (categorical (nominal): close to 'home', school 'reputation', 'course' preference or 'other')
Guardian	Student's guardian (categorical (nominal): 'mother', 'father' or 'other')
Traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
Schoolsup	extra educational support (categorical (binary): yes or no)
Famsup	family educational support (categorical (binary): yes or no)
Paid	extra paid classes within the course subject (Math or Portuguese) (categorical(binary): yes or no)
Activities	extra-curricular activities (categorical (binary): yes or no)
Nursery	attended nursery school (categorical(binary): yes or no)
Higher	wants to take higher education (categorical (binary): yes or no)
Internet	Internet access at home (categorical (binary0: yes or no)
Romantic	with a romantic relationship (categorical (binary): yes or no)
Famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Freetime	free time after school (numeric: from 1 - very low to 5 - very high)
Gout	going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Health	current health status (numeric: from 1 - very bad to 5 - very good)
Absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

3.2 DATA CLEANING AND NORMINALIZATION

Data cleaning refers to the process of detecting and correcting errors in the dataset that may negatively impact the prediction model (Brownlee, 2016). The dataset used did not contain any missing values for any of the attributes. The mother's education (Medu) and father's education (Fedu) attributes where converted from numeric data to categorical (nominal) data. The nominal attributes were changed to dummy variables. A dummy variable takes 0 or 1, where the values indicate the presence or absence of an attribute. It is used when categorical attribute has more than two categories. This allows categorical to fit well into regression models (Bock T. , 2018).

3.3 SOFTWARE DEVELOPMENT MODEL

The Incremental Model is a method of software development where the model is designed, implemented and tested repeatedly. Each repeated model comes with more functionality to the previous model (Ghahrai, 2016). In incremental model, the whole requirement is divided into various builds. There are multiple development cycles, each cycle similar to the waterfall life cycle. These cycles are smaller and easier to manage. The first phase of this process involved developing the Student Information System where data about the students, teachers, admin, subjects and scores on those subjects can be stored, retrieved and modified. The second phase of the process involved the implementation of the machine learning model used for predicting the final grade of the students.

3.4 SOFTWARE REQUIREMENT SPECIFICATION

The software requirement specification describes in detail, the functional and non-functional requirements required by the system. This defines how the system will interact with hardware, internal modules and communicate with other programs (Admin, 2015).

3.4.1 FUNCTIONAL REQUIREMENTS

Functional requirement is defined as the means of entailing the services that the software offer. It is also a means of specifying what the system should do (Ulf, 2012). In some cases, the functional requirements may also explicitly state what the system should not do. The users and the functional requirements of the system include the following:

1. School Administrator: This is the individual that manages routine activities and provide instructional leadership in educational institutions. The functional requirements of this user in the system include the following:

    1. The school administrator can login to the system.
    2. The school administrator can change the password of their account.
    3. The school administrator can view their profile.
    4. The school administrator can create, edit and delete student account.

5. The school administrator can create, edit and delete teacher account.
6. The school administrator can add subjects and assign them to teachers.
7. The school administrator can view the total number of pupils.
8. The school administrator can view the total number of subjects.
9. The school administrator can view the total number of school administrator.

2. Teacher: This is an individual who imparts knowledge to or instructs a person or people as to how to do something. The functional requirements of this user in the system include the following:

1. The teacher can login to the system.
2. The teacher can change the password of their account.
3. The teacher can grade the student's g1, g2 and g3 scores.
4. The teacher can view their profile.
5. The teacher can view the total number of subjects they are lecturing.
6. The teacher can view the number of pupils predicted to fail the subjects they are lecturing.
7. The teacher can view the predicted scores of all their pupils.

3. Pupil – This is a young individual who is in an educational institution for the purpose of learning or acquiring a skill. The functional requirements of this user in the system include the following:

1. The pupil can register the subjects required.
2. The pupil can view their grades in the subjects registered.
3. The pupil can change their password.
4. The pupil can view their profile.

3.4.2 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements describe how a system must behave and how its functionality must be constrained. The non-functional requirement defines the system operational capabilities which are not specifically requested by the customer but are expected to be implemented in the system (Altexsoft, 2019). The non-functional requirements for the system include the following:

1. Security: this ensures that the system and storage data should be secure from unwanted access or malware attack.
2. Availability: this ensures that the system should be accessible at all times.
3. Reliability: this defines how likely failure of the system will occur under predefined conditions.
4. Performance: this requirement illustrates the system responsiveness to different user interactions.

3.5 UNIFIED MODIFIED LANGUAGE DIAGRAMS

1. Class diagram – This is a graphical notation used to construct and visualize object-oriented systems. A class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods) and the relationships among objects.

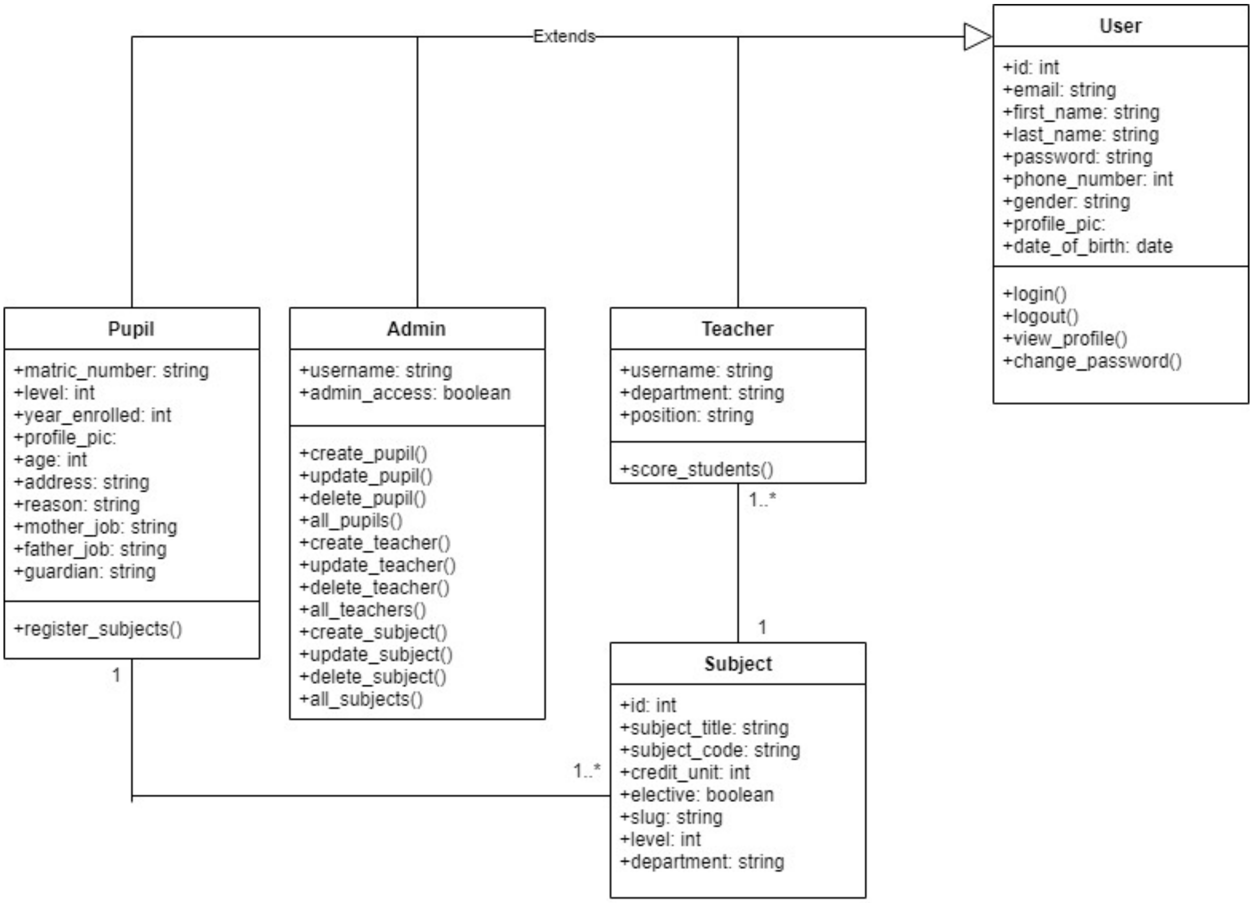


Fig 3.1 Class Diagram for the School Management System

2. Use case diagram – this is used to describe a set of actions that can be performed by the users of the system. The users are called actors and are represented by stick figures and functions by oval. Actors are associated with functions they can perform.

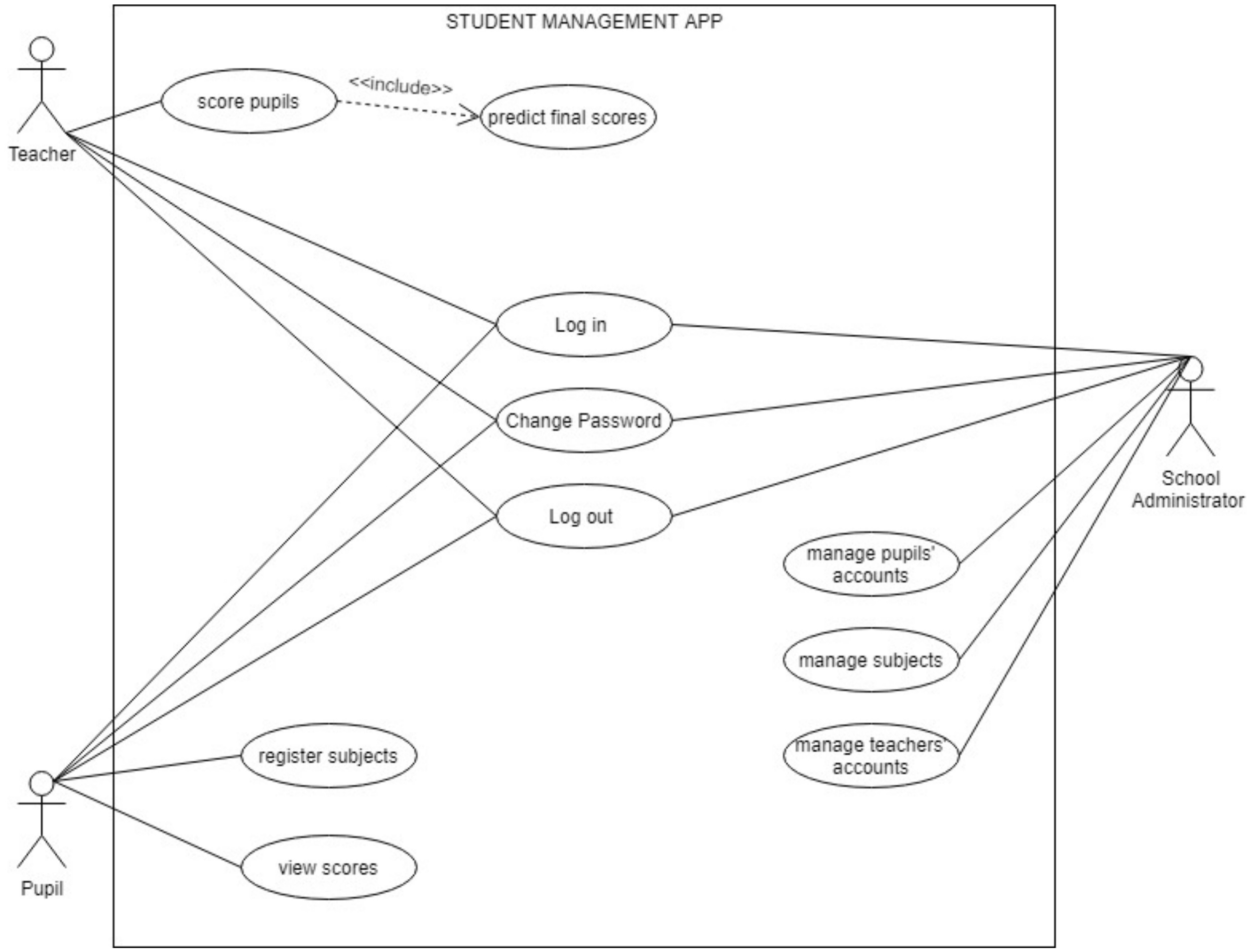


Fig 3.2 Use Case Diagram for the School Management System

In this system, there are three actors namely the school administrator, the teacher and the pupil. Each sphere represents a particular function that the user of the system can perform.

3.6 DEVELOPMENT

The programming languages used in development of the system includes:

1. Hyper Text Markup Language (HTML) – HTML helps the developers to create web pages and applications that contain sections, paragraphs, headings, links and blockquotes (G, 2019).

2. Cascading Style Sheets (CSS) – CSS is a simple design language intended to specify the HTML document style which includes the colors of text and background, page layouts, fonts etc. (Morris, n.d.).
3. JavaScript (JS) - JS is a scripting language used to create and control dynamic website content (Morris, n.d.).
4. Python - Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting.
5. Django - Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design.

3.7 MACHINE LEARNING

Machine learning is an application of machine learning that teaches computers how to learn and act without being explicitly instructed. It involves the process of building models or algorithms that improve from experience which is the data being fed into the model (Team, 2020).

3.7.1 CATEGORIES OF MACHINE LEARNING

1. Supervised learning: This involves building models with labelled data to predict values or classes.
2. Unsupervised learning: This involves building models which identify patterns in unlabeled data (Heath, 2018).

3. Semi-supervised learning: this approach mixes supervised and unsupervised learning. It relies on a small amount of labelled data and large amount of unlabeled data to train the model (Team, 2020).
4. Reinforcement learning: This is a learning method where the agent that interacts with its environment by producing actions and discovers errors. The agent is rewarded if the action performed yielded positive results and punished if the action performed yielded negative results. The goal of the agent would be to find the sets of actions that would produce the maximum reward (Team, 2020)..

3.7.2 MACHINE LEARNING ALGORITHMS USED

1. Linear Regression- It is one of the machine learning techniques that fall under supervised learning. Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable (Bock T. , 2018). Linear regression is used to predict values and classes. The common formula for a linear regression is:

$$y = \beta X + \alpha \text{ ....1}$$

In equation 1, y represents the vector of the response values, X symbol represents the matrix of features/predictor variables used to predict the y vector, $\alpha(alpha)$ symbol represents the bias which represents the prediction baseline when all the features/predictor variables have values of zero. $\beta(beta)$ symbol represents the vector of coefficients that a linear regression model uses with the bias to create the prediction. The goal of linear regression is to find the best set of beta coefficients and alpha to minimize a cost function given the squared difference between the predictions and the real output values, the cost function is represented in equation 2 and 3 (Mueller & Massaron, 2016).

.

$$minimize\ \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

2. Logistic Regression – Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable based on the concept of probability (Pant, 2019). Logistic regression uses a more complex cost function called the sigmoid function that returns a value between zero and one.

$$f(x) = \frac{1}{1+e^{-(x)}}\text{c} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4$$

In the function 4, e also known Euler number represents the base of the natural logarithms (Brownlee, Logistic Regression for Machine Learning, 2016).

3. Decision Tree – the decision tree is a supervised learning algorithm that is used to predict the class or value of the target variable by learning simple decision rules inferred from prior data (Chauhan, 2020). A decision tree is like a tree-like graph in which each internal node represents a condition on an attribute/features/inputs, (Brid, 2018). Decision trees is an easy and intuitive algorithm to use and can be used in making effective solutions without performing some preprocessing practices e.g. changing categorial data format, transforming the data etc. the algorithms works by looking into a sample of observations, create the rules that generated the output classes or values by dividing the input matrix in smaller and smaller parts until a rule for stopping is triggered. One of the disadvantages of decision tree is that its outputs are influenced by noise and data errors easily (Mueller & Massaron, 2016).

4. Random Forest - It is a classification and regression algorithm that uses multiple decision trees. The idea behind it being that when many of the decision trees predict a value or class and the results are averaged, better results are produced. These decision trees have an improved performance when they are uncorrelated with each other. Thus, the decision tress is built on different sets of bootstrapped examples which is a method used to estimate the sampling distribution of statistics and subsampled features. Its creators Adele Cutler and Leo Brieman wanted to create something that was easy to use and understandable to non-experts (Mueller & Massaron, 2016).
5. Gradient Boosting – Opposed to random forest, boosting is an approach whose strategy is to create interrelated ensembles of simple machine learning algorithms to solve complex target functions. The simple models can predict different parts of the target functions well and models used here are stumps which are single split branch, linear models, perceptron etc. The summation of the functions helps in providing a more accurate solution. Gradient boosting is one of the most powerful tools used in machine learning and uses gradient descent optimization to determine the right weights for learning (Mueller & Massaron, 2016).
6. Support Vector Machine (SVM) – This is a machine learning algorithm that by used for regression, binary and multiclass classification and detection of anomalous data. It has robust handling of overfitting, noisy data and outliers. It also can detect nonlinearity in data with the use of kernel functions which map the original feature space into a new feature space reconstructed to achieve better classification or regression results. The strategy of the Support Vector Machine is to look for a straight line with the largest separating margin between boundary of the points belonging to different classes. The

middle of the margin is known as the maximum margin and optimal margin hyperplane. The points that are used as reference on judging where the straight line will be located is known as support vectors (Mueller & Massaron, 2016).

7. K-Nearest Neighbors (KNN) – This is a machine learning algorithm that works by finding the most similar observations or characteristics to the one to be predicted. In the case, there are two or more observations similar to the predicted one, averaging the values or by picking the most frequent answer class among them. KNN is fast at training because it involves data recording but slow at predicting because it involves searching for similar observations. It is also memory intensive as data recording involves storing the dataset in memory. KNN is also sensitive to outliers. These factors limit its use in practical terms especially when dealing with big data. It is ideally used on classification problems involving hundreds of labels. KNN works out its closest neighbors using a measure of distance such as Euclidean distance which is most commonly used or Manhattan distance which works better when redundant features are present in the data. Its only hyper parameter is the k parameter which represents the number of closest neighbors that the algorithm has to consider to choose the best class or value (Mueller & Massaron, 2016).

3.7.3 EVALUATION METRICS

Evaluation metrics are used to measure the quality and performance of a machine learning model (Tavish, 2019). This is an integral component of the project and helps in selecting what machine learning model works best on future unseen or unlabelled data. The type of predictive model is crucial in order to determine which evaluation metric to use. There are two types of predictive models which are:

1. Regression model – this is a form of predictive modelling technique which investigates the relationship between a dependent (target) and one or more independent variable. The dependent variable is usually a continuous value.
2. Classification model – this is a form of predictive modelling technique which features or attributes as input and return a value in the form of a class. A class is a set of enumerated target values for a label.

The different evaluations metrics are:

1. Confusion matrix – This is the N X N matrix representation of the prediction results of the performance of the classification model on a set of test data for which the true values are known. N represents the number of classes being predicted. There are terms needed to understand the confusion matrix:
    1. True Positives (TP)– is the total number of outcome where the model correctly predicts the positive class.
    2. True Negatives (TN) – is the total number of outcome where the model correctly predicts the negative class.
    3. False Positives (FP)– is the total number of outcome where the model incorrectly predicts the positive class.
    4. False Negatives (FN) - is the total number of outcome where the model incorrectly predicts the negative class.
2. Precision and Recall – Precision for a class is the number of true positives i.e. the number of items correctly labeled as belonging to the positive class divided by the total number of elements labeled as belonging to the positive class i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class (Singh,

2019). High precision means that an algorithm returned substantially more relevant results than irrelevant ones. The precisions' formula is located at equation 5.

$$\text{Precision} = \frac{TP}{(TP+FP)} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 5$$

Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been. High recall means that an algorithm returned most of the relevant results. The recalls' formula can be seen at equation 6.

$$\text{Recall} = \frac{TP}{(TP+FN)} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 6$$

Precision and Recall are in conflict meaning that as Precision increases, Recall decreases. The domain use cases and costs associated with them determine whether high precision or high recall is most desirable

3. F1-Score – is the harmonic mean of precision and recall values for a classification problem. The advantage of using the harmonic mean over the arithmetic mean is that harmonic mean punishes extreme values more and this value tells us how useful a model can be. The formula for f1-score can be seen at equation 7.

$$\text{F1-score} = 2\,.\ \frac{precision\,.recall}{precision+recall} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 7$$

4. Area Under the ROC Curve - ROC stands for Receiver Operating Characteristics curve. ROC is a plot of the False Positive Rate (x-axis) versus the True Positive Rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0 (Jason, 2018). The True Positive Rate is calculated as the number of True Positives divided by the

addition of the number of True Positives and the number of False Negatives. It describes how good the model is at predicting the positive class when the actual outcome is positive. The False Positive Rate is calculated as the number of False Positives divided by the sum of the number of False Positives and the number of True Negative. It summarizes how often a positive class is predicted when the actual outcome is negative.

5. Root Mean Squared Error (RMSE)– RMSE is the most popular evaluation metric used in regression problems. In RMSE, the errors are squared before they are averaged. This basically implies that RMSE assigns a higher weight to larger errors. This indicates that RMSE is much more useful when large errors are present and they drastically affect the model's performance. The formula for RMSE can be seen at equation 8.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \dots\dots\dots\dots\dots\dots\dots 8$$

3.8 SUMMARY

In this section, the dataset was examined and the necessary steps taken for data pre-processing were given. The class diagram and use case diagram of the student management application were shown to illustrate how the application works and who the users of the application would be.

CHAPTER FOUR

DESIGN AND IMPLEMENTATION

4.0 INTRODUCTION

This chapter focuses of the implementation of the machine learning model and the web application using the model for predicting students' score. By definition, Implementation is simply the execution of a method, plan or any idea, design, model, specification, standard or policy for doing something. This chapter also highlights the various evaluation metrics used for evaluating the performance of our machine learning model.

4.1 DATA ANALYSIS

Data analysis is a process that involves using methods and techniques to take raw data and gain insight that is relevant to the goal of the individual or business. Several attributes of the dataset were inspected to see if there are factors that can determine the chances of success or failure of the pupil. Correlation coefficient is used to measure the strength of the relationship between two variables, it measures the linear relationship between attributes assuming that all the attributes are linearly related being compared are linearly related. Table 4.1 shows the linear correlation between the label G3 and other numeric attributes.

Table 4.1 Correlation coefficient of numeric attributes to G3

Attributes	Correlation coefficient to G3
Age	-0.161579
Traveltime	-0.117142
Studytime	0.097820
Failures	-0.360415

Famrel	0.051363
Freetime	0.011307
Gout	-0.132791
Dalc	-0.054660
Walc	-0.051939
Health	-0.061335
Absences	0.034247
G1	0.801468
G2	0.904868

From Table 4.1, G1 and G2 have very high positive correlation to G3. This means that the higher a student score in G! or G2, the higher the score in G3. Also, the feature “failures” has a low negative correlation to G3 which could mean that the higher the G3 score, the lower the number of failures. A Scatterplot is a type of data display that shows the relationship between two numerical variables. As seen in Figure 4.1, it shows that as the G3 score increases, the G1 score increases as well, also in Figure 4.2, it shows the weak negative correlation between failures and G3.

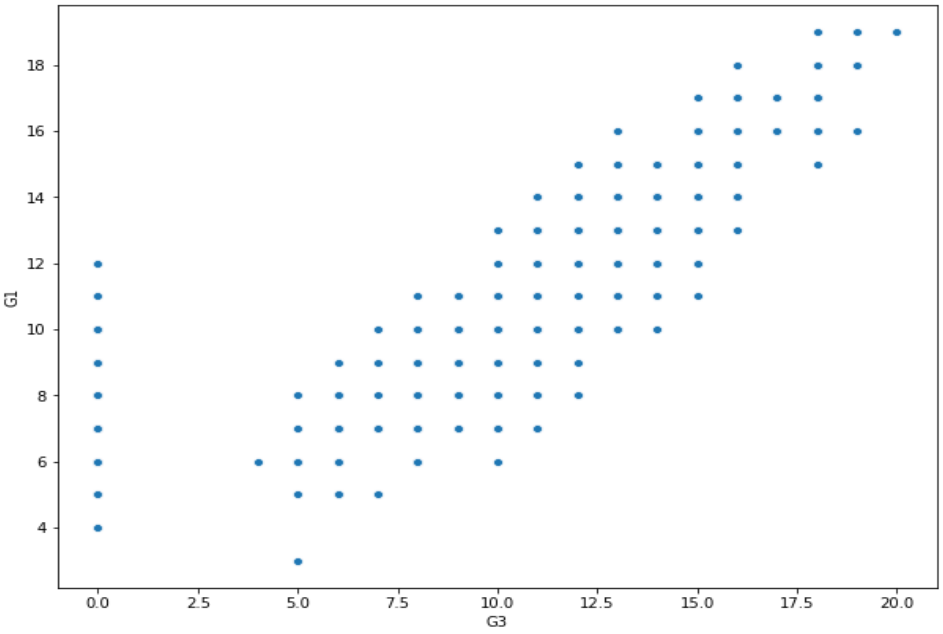


Fig 4.1 Scatterplot of G1 against G3

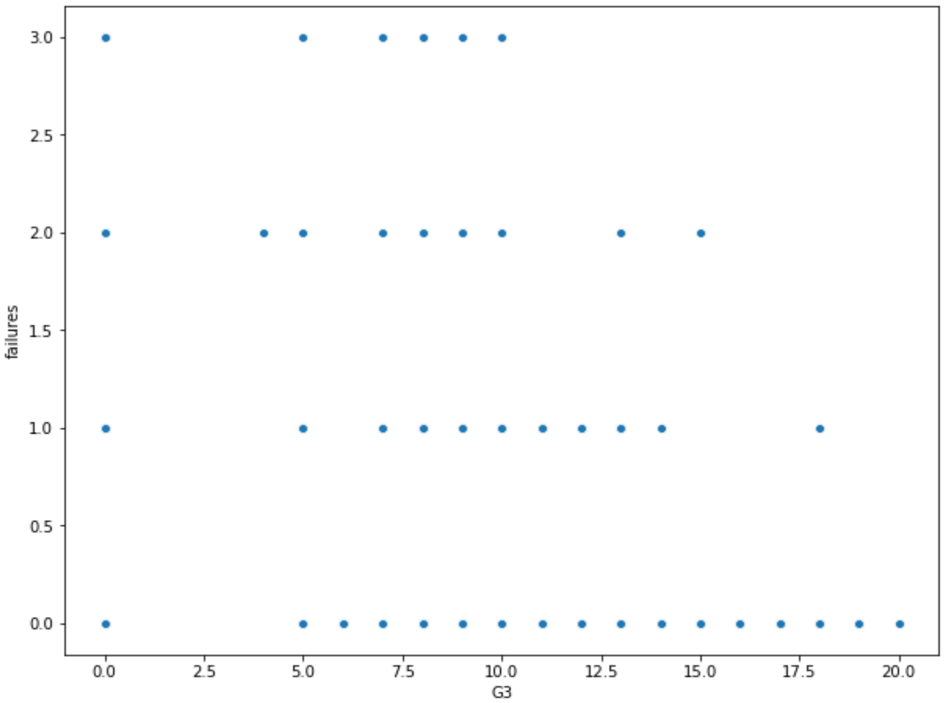


Fig 4.2 Scatterplot of G3 against failures

There were two scenarios considered for creating the predictive models, these scenarios are:

1. Classification and Regression using all the attributes in the dataset

   The dataset was split into two parts for the classification model, the training part which consists of 70% of the dataset which trains the classification and regression model and the test part which contains 30% of the dataset which is used to test how good the model can predict values or classes.

Table 4.2 Evaluation metrics for classification models designed with all features

ML model	Accuracy (%)	Categories	Precision	Recall	F1-score
Logistic Regression	82	Fail Pass	0.76 0.86	0.82 0.81	0.79 0.84
KNN	84	Fail Pass	0.77 0.90	0.88 0.81	0.82 0.85
Decision Tree	82	Fail Pass	0.80 0.84	0.78 0.86	0.79 0.85
Random Forest	84	Fail Pass	0.75 0.95	0.92 0.78	0.83 0.85
Support Vector Classifier	85	Fail Pass	0.79 0.90	0.88 0.83	0.83 0.86
Gradient Boosting	86	Fail Pass	0.80 0.91	0.88 0.84	0.84 0.87

From the results indicated in table 4.2, all the three classification prediction models performed reasonably well in predicting the G3 score of the students. Among all the models, Gradient Boosting produced the most accurate predictions with an accuracy of

86%. The least accurate result was obtained from Logistic Regression and Decision Tree with an accuracy of 82%.

Table 4.3 Evaluation metrics for regression models designed with all features

ML model	Root Mean Squared Error (RMSE)
Linear Regression	2.035
KNN	1.683
Decision Tree	2.531
Random Forest	1.798
Support Vector Regressor	1.972
Gradient Boosting	1.776

From the results indicated in table 4.3, all the three regression prediction models performed reasonably well in predicting the G3 score of the students. The better models are those in which their RMSE is closer to zero. Among all the models, KNN produced the most accurate predictions with an RMSE of 1.683. The least accurate result was obtained from Decision Tree with an RMSE of 2.531.

2. Classification and Regression using some of the attributes for the web application

   Its goal is to find features in the dataset that are relevant in the dataset and can be correctly given as false information can affect the performance of the model. The chosen features include sex, age, address, reason, mother's job, father's job, guardian.

   In order for the web application to provide predictions when only the G1 score is available, two models were made, one of the models uses all the features listed with the

G1 score and the other model uses the all the features listed with both the G1 and G2 score.

Table 4.4 Evaluation metrics for classification models using only G1 score

ML model	Accuracy (%)
Logistic Regression	88.00
KNN	86.00
Decision Tree	84.00
Random Forest	85.00
Support Vector Classifier	87.00
Gradient Boosting	87.00

From the results indicated in table 4.4, all the three classification prediction models performed reasonably well in predicting the G3 score of the students. Among all the models, Logistic Regression produced the most accurate predictions with an accuracy of 88%. The least accurate result was obtained from Decision Tree with an accuracy of 84%.

Table 4.5 Evaluation metrics for classification models designed using G1 and G2 score

ML model	Accuracy (%)
Logistic Regression	91.00
KNN	74..00
Decision Tree	88.00

Random Forest	92.00
Support Vector Regressor	88.00
Gradient Boosting	91

From the results indicated in table 4.5, all the three classification prediction models performed reasonably well in predicting the G3 score of the students. Among all the models, Random Forest produced the most accurate predictions with an accuracy of 92%. The least accurate result was obtained from KNN with an accuracy of 74%.

Random Forest and Gradient Boosting performed well in all the scenarios. Gradient Boosting model was chosen as the model to be used in predicting the pupil's G3 score in the web application.

4.3 THE PROGRAM INTERFACE

As previously stated, there exist three users in the system and this section shows the interface the users can interact with. These include:

1. PUPIL LOG IN PAGE

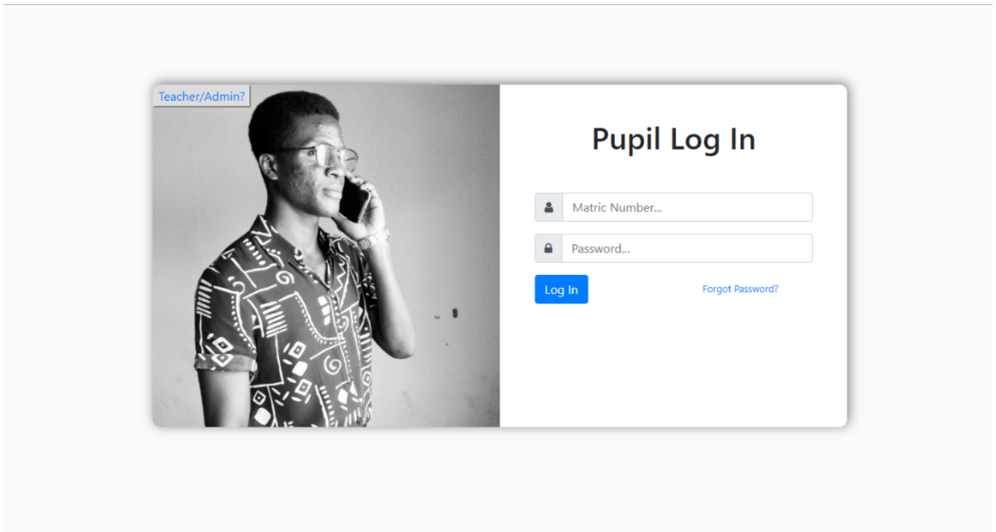
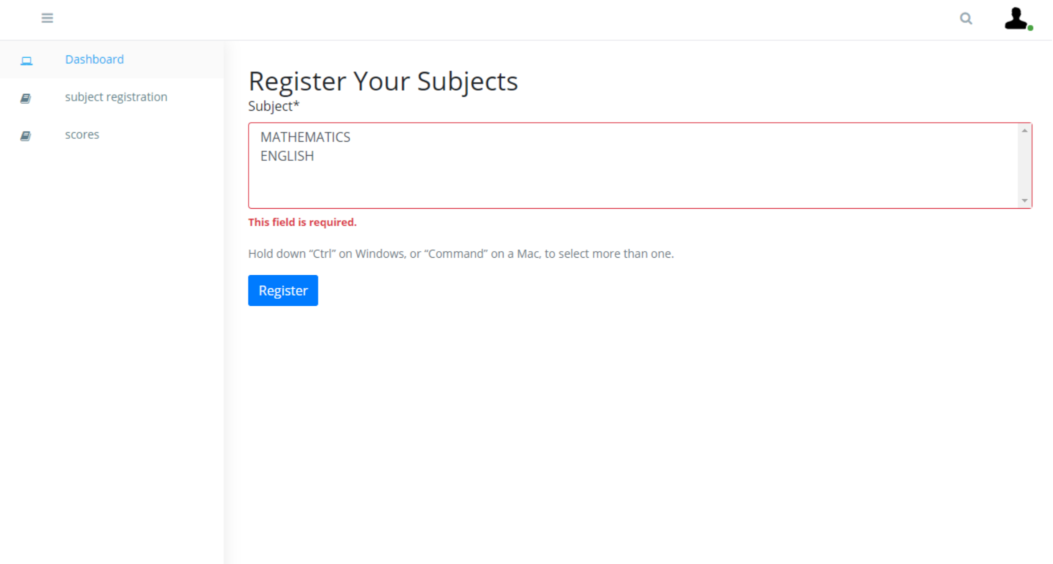


Fig 4.3 Pupil Login Page

This interface allows the pupil, with the appropriate login details access to their account, if invalid access is denied and pop up shows up indicating that either the unique number assigned to the pupil is wrong or the password is wrong.

2. PUPIL REGISTRATION PAGE



Register Your Subjects

Subject*

MATHEMATICS
ENGLISH

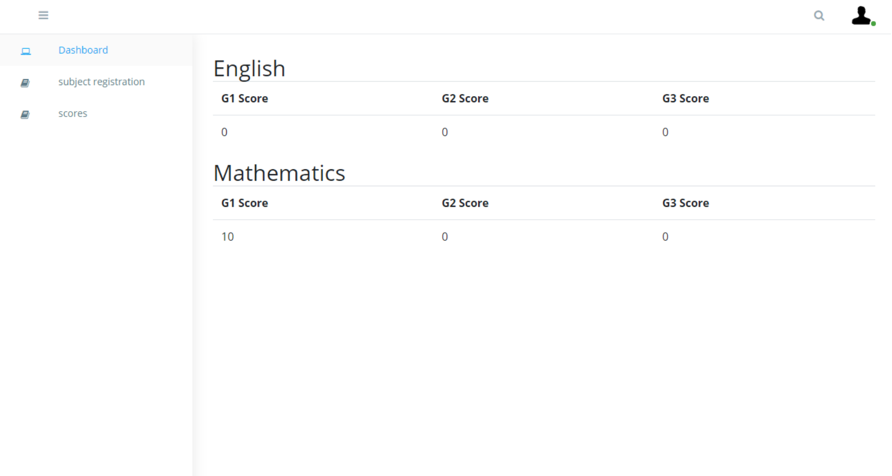
This field is required.

Hold down “Ctrl” on Windows, or “Command” on a Mac, to select more than one.

Fig 4.4 Pupil registration page

This interface is shown after the successful login of the pupil. The pupil is able to register the subjects required for their education.

3. PUPIL VIEW SCORE PAGE



English

G1 Score	G2 Score	G3 Score
0	0	0

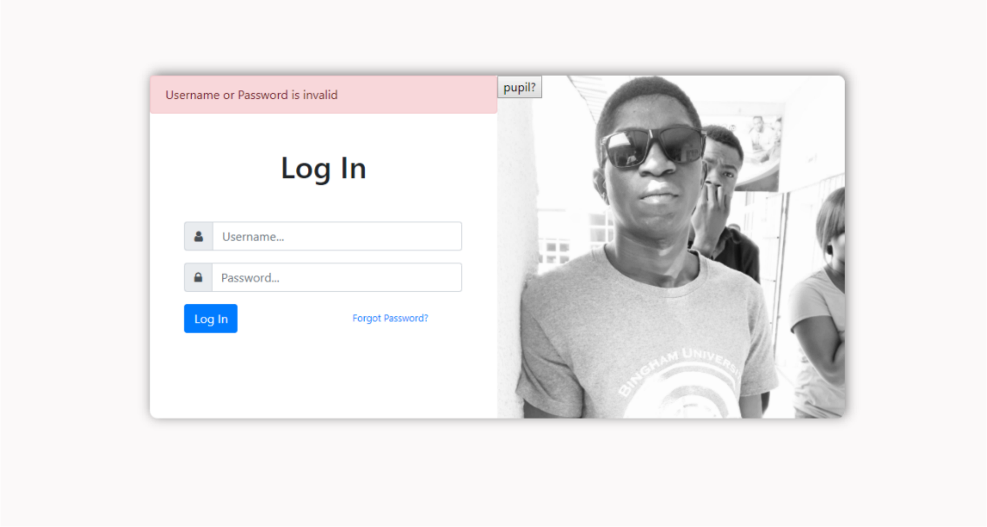
Mathematics

G1 Score	G2 Score	G3 Score
10	0	0

Fig 4.5 Pupil view score page

This interface allows the pupil to see the scores given to them by their various teachers on the subjects that was registered.

4. TEACHER/ADMINISTRATOR LOGIN PAGE

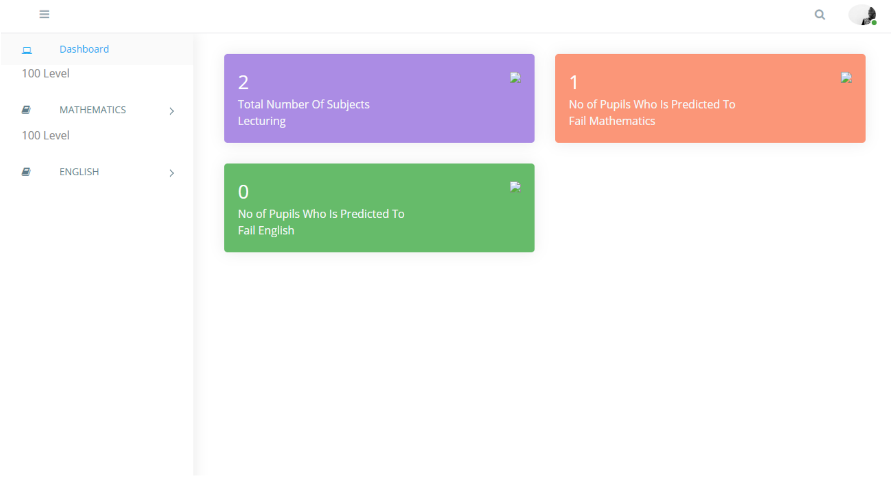


Username or Password is invalid

Fig 4.6 Teacher/Administrator login page

This interface allows the teacher or school administrator, with the appropriate login details access to their account, if invalid access is denied and pop up shows up indicating that either the username assigned to the teacher or school administrator is wrong or the password is wrong.

5. TEACHER DASHBOARD



2
Total Number Of Subjects Lecturing

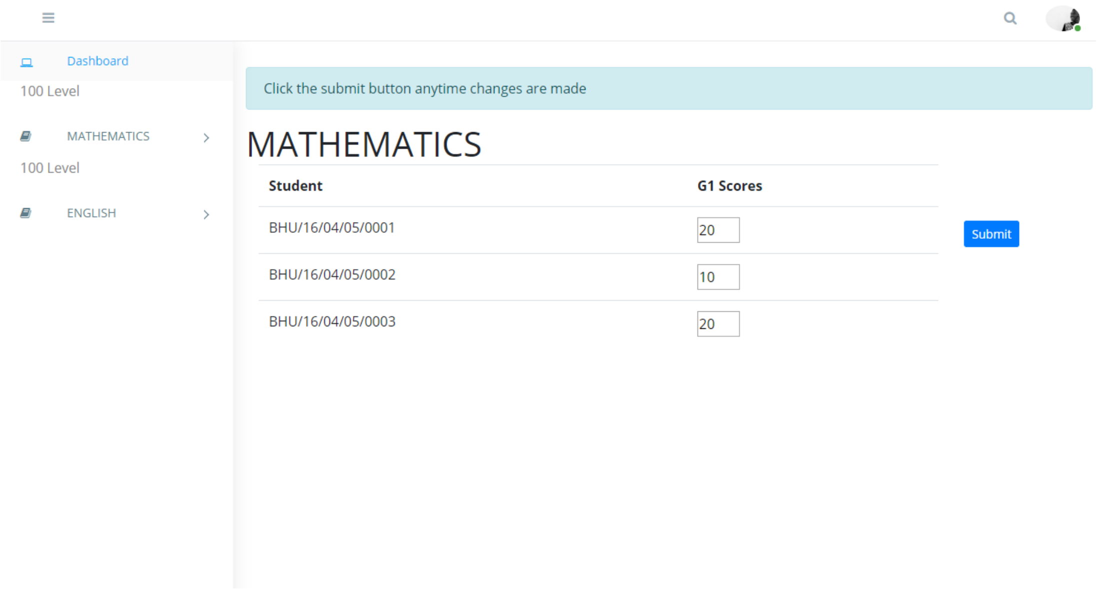
1
No of Pupils Who Is Predicted To Fail Mathematics

0
No of Pupils Who Is Predicted To Fail English

Fig 4.7 Teacher dashboard

This interface shows the teacher information about their students such as the number of students predicted to fail the subjects taught by the teacher. The dashboard also shows the subjects taught by the teacher and each one contains the g1 scores, g2 scores, g3 scores and the predicted g3 score.

6. SCORE PAGE



Click the submit button anytime changes are made

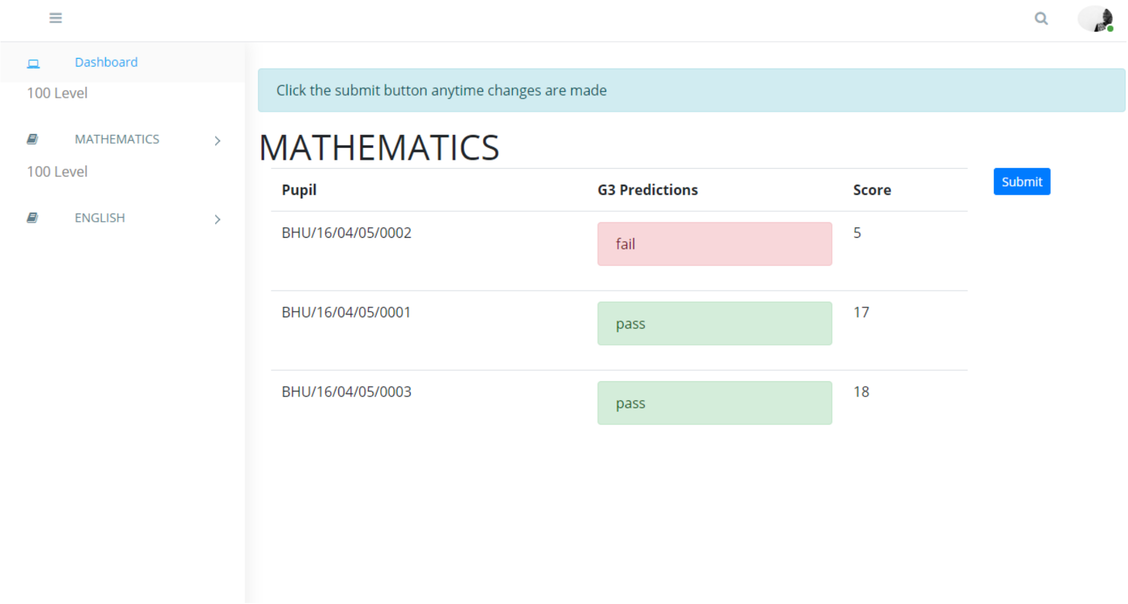
MATHEMATICS

Student	G1 Scores
BHU/16/04/05/0001	20
BHU/16/04/05/0002	10
BHU/16/04/05/0003	20

Fig 4.8 Score page

This interface allows the teacher to give the students registered their g1 score for the subject. The interface for the g2 and g3 scores are the same allowing for grading of student's performance in the subjects.

7. G3 PREDICTED SCORE PAGE



Click the submit button anytime changes are made

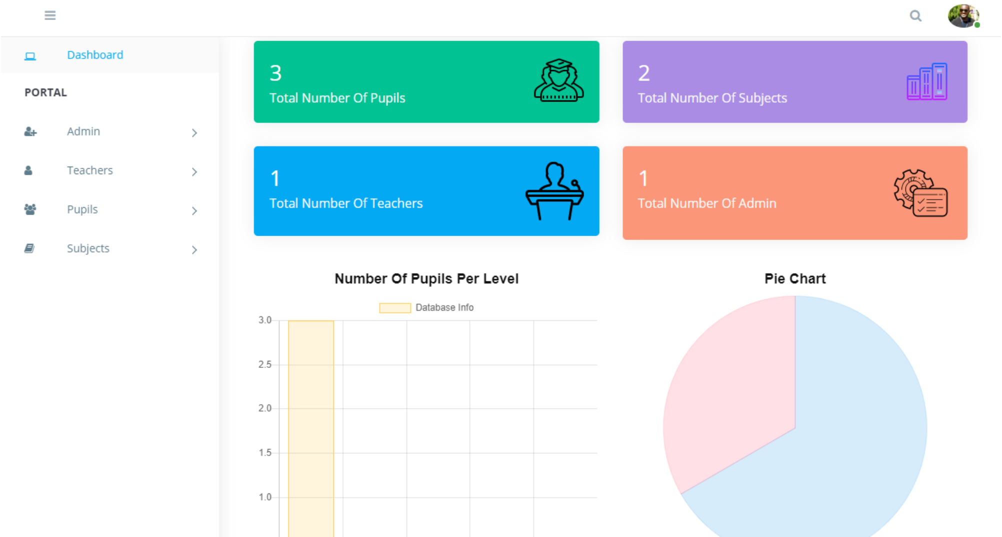
MATHEMATICS

Pupil	G3 Predictions	Score
BHU/16/04/05/0002	fail	5
BHU/16/04/05/0001	pass	17
BHU/16/04/05/0003	pass	18

Fig 4.9 G3 predicted score page

This interface allows the teacher to see each student's predicted score which is generated with machine learning model at the backend. This enables the teacher to make decisions on what actions to take to prevent failure of the pupils.

8. ADMINISTRATOR DASHBOARD



3
Total Number Of Pupils

2
Total Number Of Subjects

1
Total Number Of Teachers

1
Total Number Of Admin

Fig 4.10 Administrator dashboard

This interface allows shows the administrator perform certain functionalities such as creating, deleting and updating of student accounts. The dashboard also shows the administrator information such as the total number of subjects, the total number of teachers etc.

9. ALL TEACHERS PAGE

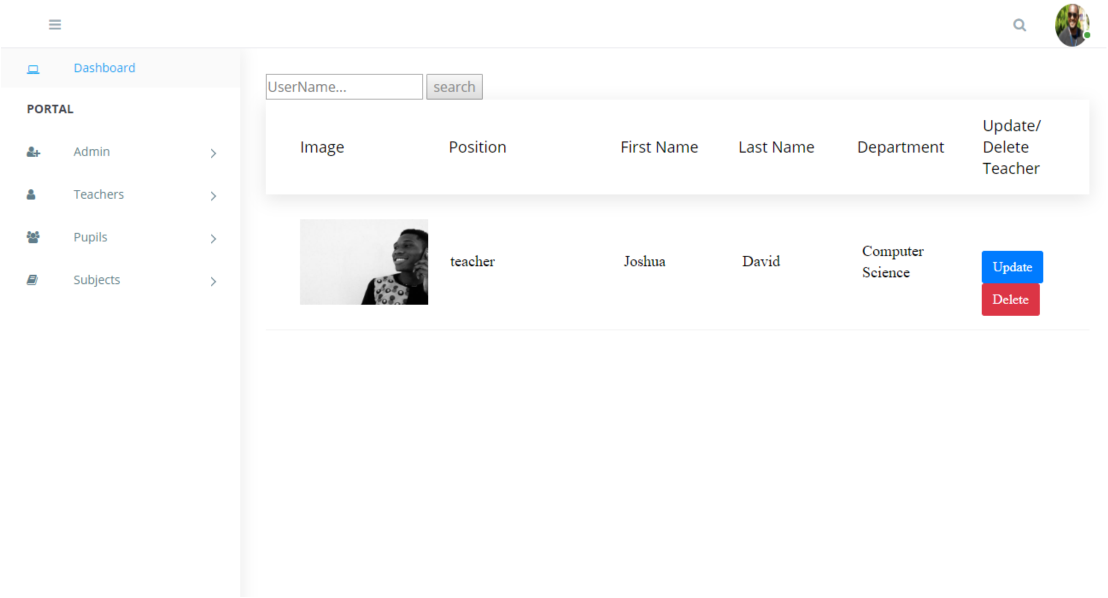
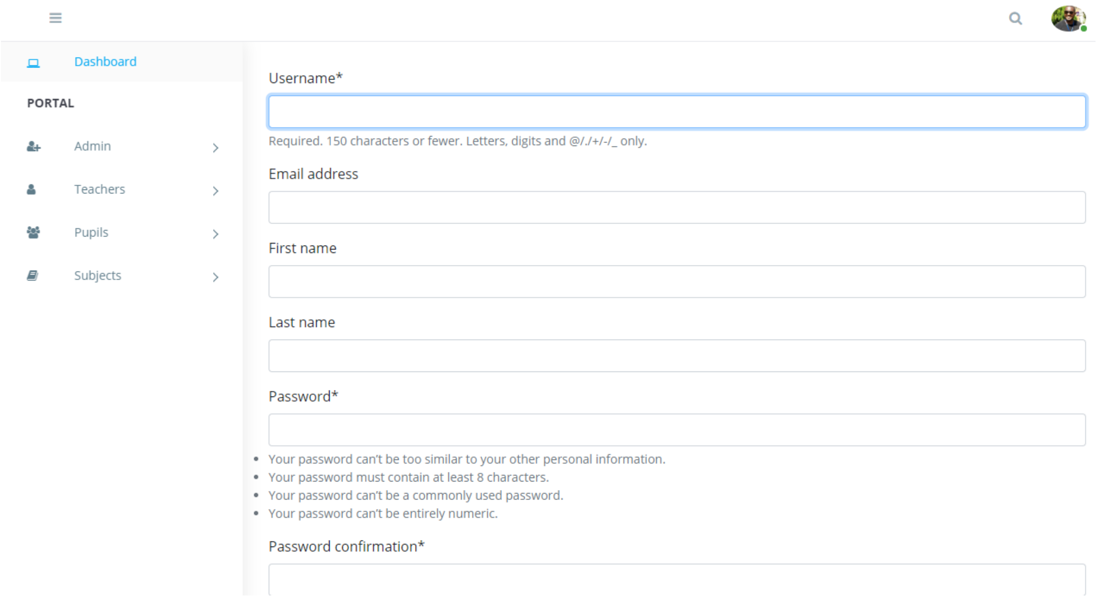


Image	Position	First Name	Last Name	Department	Update/ Delete Teacher
	teacher	Joshua	David	Computer Science	Update Delete

Fig 4.11 All Teachers page

This interface allows the school administrator to see all the teachers who have accounts in the system. It also takes the administrator to different web pages for deleting or updating the teacher account. This interface also allows for the search of teacher account based on the username.

10. CREATE TEACHER PAGE



Username*

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

Email address

First name

Last name

Password*

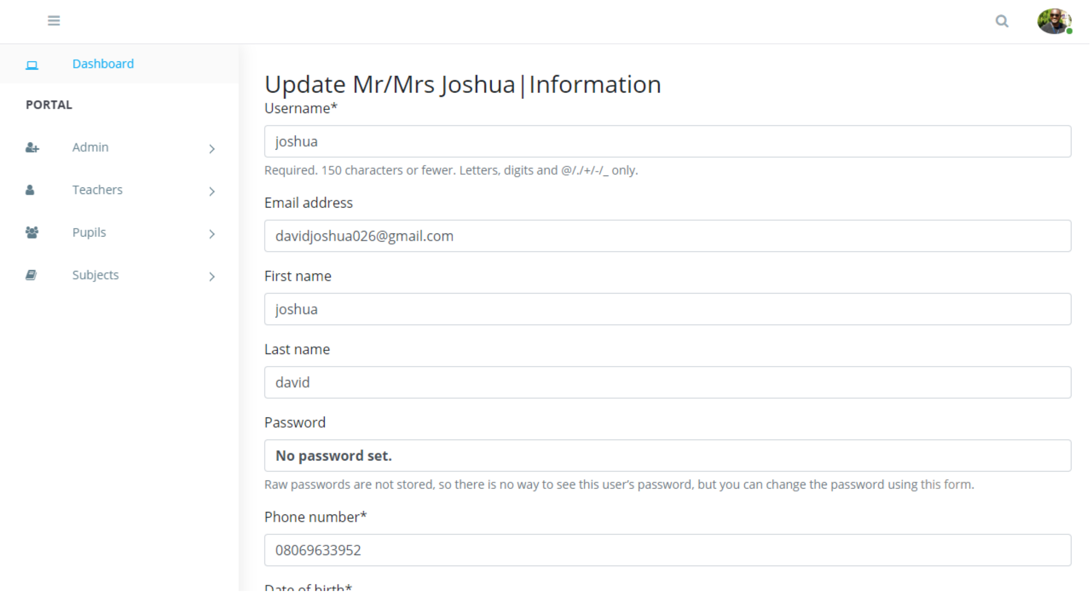
- Your password can't be too similar to your other personal information.
- Your password must contain at least 8 characters.
- Your password can't be a commonly used password.
- Your password can't be entirely numeric.

Password confirmation*

Fig 4.12 Create Teacher Page

This interface allows the school administrator to create teacher account for the school with all the necessary information such as username, email address, first name, last name etc.

11. UPDATE TEACHER PAGE



Update Mr/Mrs Joshua | Information

Username*

joshua

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

Email address

davidjoshua026@gmail.com

First name

joshua

Last name

david

Password

No password set.

Raw passwords are not stored, so there is no way to see this user's password, but you can change the password using this form.

Phone number*

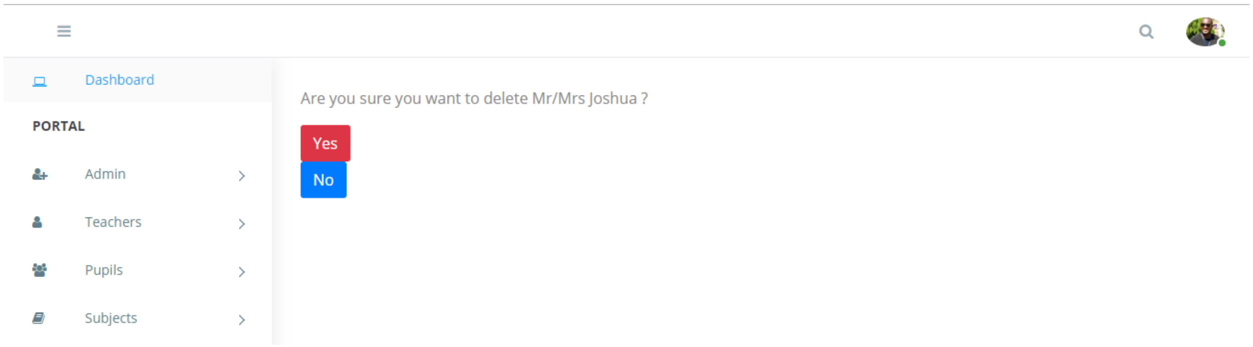
08069633952

Date of birth*

Fig 4.13 Update teacher page

This interface allows the school administrator to update information about the teacher

12. DELETE TEACHER PAGE



Are you sure you want to delete Mr/Mrs Joshua ?

Fig 4.14 Delete teacher page

This interface allows the school administrator to remove the teacher account permanently from the system.

CHAPTER FIVE

5.0 SUMMARY

It has been of great pleasure working on this project. Developing the system has proven to be both challenging and exciting as it has demonstrated the great depths to which a painstaking research into any challenge facing the human race can be effectively tackled and resolved using technology. This project has shown the important role machine learning can play in reducing the number of students who fail by predicting the student scores. This enables the management to make decisions on how to help these students.

This research work begins with an introductory section centered on an overview of what the project entails: the background of study, project limitations with project aim and objectives. It then follows with a second chapter focused on the project literature review encompassing the statement of fundamental concepts and review of related studies conducted by different researchers in the past. The third chapter focuses on the proposed solution with an in-depth look at the research methodology employed in carrying out the project. The fourth chapter incorporates the specific requirements for the design and implementation of the system.

The concept of a school management system with the functionality of predicting student's final score has been thoroughly examined in this research work. It is hopeful that this research would be useful to the readers, students of information management, lecturers and school management boards.

5.1 RECOMMENDATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

After a careful and comprehensive study of the prediction of student's academic performance with its various functions and benefits to institutions and the other end users, the following recommendations were taken into consideration.

The utilization of other machine learning techniques which could include unsupervised learning or semi supervised learning. educational research shows that some socioeconomic, psychological factors, such as learning style, self-efficacy, motivation and interest, and teaching and learning environment, also play a role in student learning and thus affect student achievement. Therefore, future studies should include those above-mentioned variables in the models so as to increase the prediction accuracy.

5.2 LIMITATIONS OF THE STUDY

In machine learning, the size of data is a very important factor in increasing the accuracy at which the model predicts a student's score. There is also issue of reliability if there exist false data in the dataset, this can affect the accuracy of the prediction model.

5.3 CONCLUSION

The main purpose of this project was to show how information stored in institutional databases can be used to predict the academic performance of the students. This project also shows that by picking essential features in the dataset, highly accurate predictive models could be created. Essentially with large amounts of data at our disposal, useful information can be learnt which enable teachers, learners and institutions make decisions that can improve student's performance.

REFERENCES

1. Admin, S. (2015, April 26). *Software Requirement Specification (SRS)*. Retrieved from softwaretestingclass.com: https://www.softwaretestingclass.com/software-requirement-specification-srs/

2. Altexsoft. (2019, November 21). *Non-functional Requirements: Examples, Types, How to Approach*. Retrieved from https://www.altexsoft.com: https://www.altexsoft.com/blog/non-functional-requirements/

3. Berland, M., Baker, R., & Blikstein, P. (2014). Education Data Mining and Learning Analytics: Applications to Constructionist Research. *Technology, Knowledge and Learning*, 205.

4. Bock, T. (2018, August 29). *What are Dummy Variables?* Retrieved from DisplayR: https://www.displayr.com/what-are-dummy-variables/

5. Bock, T. (2018, April 6). *What is Linear Regression?* Retrieved from displayr: https://www.displayr.com/what-is-linear-regression/

6. Brid, R. S. (2018, October 26). Retrieved from A Medium Corporation [US]: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

7. Brownlee, J. (2016, April 1). *Logistic Regression for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

8. Brownlee, J. (2020, March 20). *How to Perform Data Cleaning for Machine Learning with 8Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/

9. Chauhan, N. S. (2020, January 15). *Decision Tree Algorithm Explained*. Retrieved from KDnuggets: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

9. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *FUture BUsiness TEChnology Conference* (pp. 5 -12). Porto: EUROSIS.

10. G, D. (2019, November 25). *What is HTML? The Basics of Hypertext Markup Language Explained*. Retrieved from Hostinger: https://www.hostinger.com/tutorials/what-is-html

11. Ghahrai, A. (2016, September 3). *Software Development Methodologies*. Retrieved from DevQA.io: https://devqa.io/software-development-methodologies/

12. Heath, N. (2018, September 14). *What is machine learning? Everything you need to know*. Retrieved from ZDNet: https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/

13. Jason, B. (2018, August 31). *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

14. Kabakchieva, D. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms. *International Journal of Computer Science and Management Research*, 686-690.

15. Morris, S. (n.d.). *TECH 101: THE ULTIMATE GUIDE TO CSS*. Retrieved from skillcrush: https://skillcrush.com/blog/css/

16. Morris, S. (n.d.). *TECH 101: WHAT IS JAVASCRIPT?* Retrieved from Skillcrush: https://skillcrush.com/blog/javascript/

17. Mueller, J. P., & Massaron, L. (2016). *Machine Learning for dummies.* New Jersey: John Wiley & Sons, Inc.

18. Obsie, E. Y., & Adem, S. A. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*.

19. Oyerinde, O., & Chia, P. (2017). Predicting Students' Academic Performances - A Learning Analytics Approach using Multiple Linear Regression. *International Journal of Computer Applications*, 37-44.

20. Pant, A. (2019, January 22). *Introduction to Logistic Regression*. Retrieved from Towards data science: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

21. Rajni, J., & Malaya, D. B. (2013). A survey on educational data mining and research trends. *Internation Journal of Database Management Systems*, 53-72.

22. Rob, P. (2018, October 1). *6 essential steps to the data mining process*. Retrieved from BarnRaisers: https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/

23. Shetty, D., Shetty, D., & RoundHal, S. (2019). Student Performance Prediction. *International Journal of Computer Applications Technology and Research*, 157-160.

24. Shetty, I. D., Shetty, D., & Roundhal, S. (2019). Student Performance Prediction. *International Journal of Computer Applications Technology and Research*, 157 - 160.

25. Singh, B. (2019, October 25). *Evaluation Metrics for Machine Learning Models*. Retrieved from Medium: https://heartbeat.fritz.ai/evaluation-metrics-for-machine-learning-models-d42138496366

26. Srivastava, J., & Sirvastava., K. (2015). Data Mining in Education Sector: A Review. *Special Conference Issue: National Conference on Cloud Computing & Big Data*, 184-190.

27. Sultana, J., Rani, M. U., & Farquad, M. (2019). Student's Performance Prediction using Deep Learning and Data Mining Methods . *International Journal of Recent Technology and Engineering (IJRTE)* , 1018-1021.

28. Sumit. (2020, June 2). *What is Data Mining in 2020?* Retrieved from Henry Harvin: https://www.henryharvin.com/blog/what-is-data-mining/

29. Tavish, S. (2019, August 6). *11 Important Model Evaluation Metrics for Machine Learning Everyone should know*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

30. Team, E. S. (2020, May 6). *What is Machine Learning? A definition*. Retrieved from Expert System: https://expertsystem.com/machine-learning-definition/

31. Ulf, E. (2012, April 5). *Why is the difference between functional and Non-functional requirements important?* Retrieved from reqtest: https://reqtest.com/requirements-blog/functional-vs-non-functional-requirements/

32. Vanessa, S., & Paska, M. H. (2020). Application of data mining method using association rules apriori to shopping cart analysis on sale transactions (Case study alfamidi burnt stone): . *Journal of Computer Networks, Architecure and High Performance Computing* , 222-226.