electronics

Article

Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach

Zain ul Abideen [1], Tehseen Mazhar [2], Abdul Razzaq [1], Inayatul Haq [3], Inam Ullah [4,*], Hisham Alasmary [5] and Heba G. Mohamed [6,*]

1 Department of Computer Science, MNSUA Multan, Multan 60650, Pakistan
2 Department of Computer Science, Virtual University of Pakistan, Lahore 54000, Pakistan
3 School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
4 BK21 Chungbuk Information Technology Education and Research Center, Chungbuk National University, Cheongju 28644, Republic of Korea
5 Department of Computer Science, College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
6 Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
* Correspondence: inam@chungbuk.ac.kr (I.U.); hegmohamed@pnu.edu.sa (H.G.M.)

Abstract: Out-of-school children (OSC) surveys are conducted annually throughout Pakistan, and the results show that the literacy rate is increasing gradually, but not at the desired speed. Enrollment campaigns and targets system of enrollment given to the schools required a valuable model to analyze the enrollment criteria better. In existing studies, the research community mainly focused on performance evaluation, dropout ratio, and results, rather than student enrollment. There is a great need to develop a model for analyzing student enrollment in schools. In this proposed work, five years of enrollment data from 100 schools in the province of Punjab (Pakistan) have been taken. The significant features have been extracted from data and analyzed through machine learning algorithms (Multiple Linear Regression, Random Forest, and Decision Tree). These algorithms contribute to the future prediction of school enrollment and classify the school's target level. Based on these results, a brief analysis of future registrations and target levels has been carried out. Furthermore, the proposed model also facilitates determining the solution of fewer enrollments in school and improving the literacy rate.

Keywords: enrollment criteria; enrollment predictions; random forest; enrollment analysis; machine learning; AI

Citation: Abideen, Z.u.; Mazhar, T.; Razzaq, A.; Haq, I.; Ullah, I.; Alasmary, H.; Mohamed, H.G. Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach. *Electronics* **2023**, *12*, 694. https://doi.org/10.3390/electronics12030694

Academic Editors: Nikos Petrellis, Nikolaos Voros, Christos P. Antonopoulos and Costas Psychalinos

Received: 6 December 2022
Revised: 20 January 2023
Accepted: 25 January 2023
Published: 30 January 2023

1. Introduction

1.1. Background

In this research, we used this power of the entire field for educational purposes. Low enrollment of students in public schools is the prime challenge in developing countries. Pakistan is a signatory of the sustainable development goal (SDG) 2030, and Article IV of SDG 2030 states that primary education is the right of every individual in the whole world [1]. Article 25(A) of the Islamic Republic of Pakistan constitution says that fundamental education will be given to every citizen of this country [2]. According to the Ministry of Federal Education and Professional Training, Pakistan's current literacy rate is 62.3%, implying that an estimated 60 million people are illiterate in the country [3].

Following the Punjab Education Sector Reform Program (PESRP), the school census report published by the Punjab government in 2020–2021 demonstrates the importance of increasing student enrollment and the reasons for the dropping out of young children. The Institute of Statistics defines the dropout ratio as the proportion of students in the same class who are no longer enrolled in the next school year in the same school [4].

The authors predicted the academic performance of architecture students using data from their earlier research and machine learning models. The researchers used linear regression analysis and a K-nearest neighbor (k-NN) neighbor study. In terms of accuracy, the the K-nearest neighbor (k-NN) model significantly outperformed the linear discriminant analysis model. Additionally, how well architecture students performed on math exams (at the ordinary level) greatly impacted their grades [5].

Information Technology University (ITU), Lahore, Pakistan, used Restricted Boltzmann Machines, Matrix Factorization, and Collaborative Filtering to analyze data from the real world (RBM). The authors looked at the electrical engineering departmental grades of ITU undergraduates. RBM was found to be more reliable in predicting a student's performance in a particular course than other approaches [6].

Researchers discussed a case study showing how machine learning methods can forecast students' academic performance. They created a regression algorithm that can forecast a student's performance based on their performance on a small number of written assignments and their demographics. A software tutoring aid prototype was developed [7].

The primary goal of their paper is to identify the key factors that influence school academic performance and to investigate their relationships using a two-stage analysis of a sample of Tunisian secondary schools. To deal with undesirable outputs in the first stage, we employ the Directional Distance Function Approach (DDF) [8]. According to the author, the past work witnesses the researcher's focus on Student Performance, slow learning, and dropout rate. It is a dire necessity to pay special attention to low enrollment. Low enrollment means a decrease in literacy rate; this is directly related to the progress and development of the country. The current system demands the design of a model that will help the schools to achieve their targets. It is a dire need to maximize enrollment by applying a suitable model that allows the school administration and policymakers. Another solution is to find out the school category in which school enrollment lies, far from the target, Below Target, and on target [9].

Many countries have shown growing interest and concern about the problem of low school enrollment and its primary causes in recent years. This problem is known as the "100-factor problem". A lot of research has been done to identify factors that affect student performance (school failure and dropouts) at different academic levels (primary, secondary and tertiary) [10,11].

In this context, machine learning can contribute to a remarkable achievement in understanding and analyzing the challenges of low enrollment. The authors stated that the use of machine learning algorithms to evaluate the complete/incomplete nature of a thesis—is accurate—is one of the study's innovative results. These assessment models allow for a more balanced match between students and instructors. Machine learning is a cornerstone of artificial intelligence and big data analysis. It includes powerful algorithms capable of recognizing patterns, classifying data, and, in essence, learning to perform a specific task independently. This field has grown in popularity recently, but it is still unknown to most people, including professionals [12]. Machine learning approaches are presented in Figure 1.

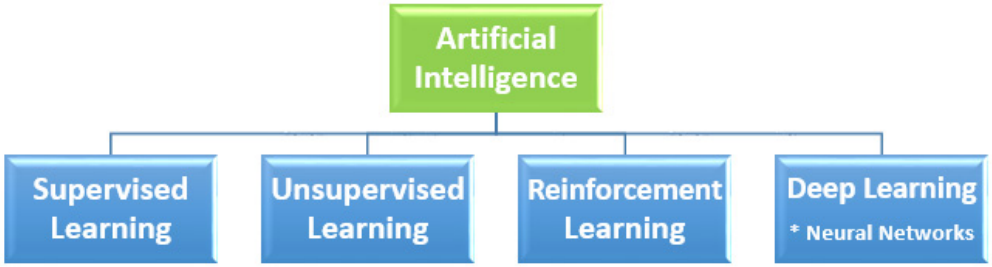

Figure 1. Machine Learning Model A-Z process.

1.2. Research Gaps and Limitations

1.2.1. Gaps in Previous Research

Research has been increasingly interested in issues such as forecasting student performance, preventing failure, and determining what causes kids to drop out of school in recent years. Nagy and Molontay used and evaluated several machine learning methods based on information available at the time of enrollment to identify at-risk individuals and estimate student dropout from university programs (secondary school performance, personal details). They also provided a platform for data-driven decision assistance to the education directorate and other stakeholders. They based their models on data from 15,825 undergraduate students who registered at the Budapest University of Technology and Economics between 2010 and 2017 and either completed or dropped out of their programs [13].

Mengash demonstrated how data mining techniques can be used to anticipate applicants' academic success at universities to aid admissions decisions. The proposed methodology was validated using data from 2039 students enrolled in the Computer Science and Information College of a Saudi state university between 2016 and 2019. According to the findings, candidates' early university performance can be predicted before admission based on specific pre-admission traits (high school grade average, Scholastic Achievement Admission Test score, and General Aptitude Test score). Furthermore, the data show that the Scholastic Achievement Admission Test score is the best predictor of future student achievement. As a result, admissions algorithms should give this score more weight [14].

A multitude of academic and non-academic factors influence a student's academic achievement at a university. While students who previously failed due to familial distractions may be able to focus away from home and thrive at university, students who previously succeeded in secondary school may lose focus due to peer pressure and a social lifestyle. In Nigeria, university admission is heavily based on a student's cognitive entry criteria, which are predominantly intellectual and may not always transfer to excellence if a student enrolls in a university [15].

Learning analytics and educational data mining have improved tremendously in a short amount of time. Baker had a vision for the field's future directions, including increased interpretability, generalizability, transferability, application, and clearer evidence of effectiveness. The keynote talk was gently revised and delivered in 2019 at the Learning Analytics and Knowledge Conference. They offer these future approaches as a set of six competitions, the Baker Learning Analytics Prizes, with particular standards for what would represent forward advancement in each of these routes (BLAP). By addressing these challenges, the field will be able to more effectively use data to benefit students and enhance education [16].

Muralidharan and Prakash investigated the effects of an innovative project in the Indian state of Bihar that provided girls who progressed to secondary school with a bicycle to make it simpler for them to commute to school to bridge the gender gap in secondary enrollment. They analyzed data from a large representative household survey using a triple difference approach, with boys and the neighboring state of Jharkhand serving as comparison groups. They discovered that being part of a cohort that was exposed to the Cycle program increased girls' age-appropriate secondary school enrollment by 32% and reduced the associated gender gap by 40%. Furthermore, they identified an 18% increase in the number of girls taking the important secondary school certificate exam and a 12% increase in the proportion of girls passing it. The triple-difference estimate as a function of distance to the nearest secondary school reveals that enrollment increases tended to occur in villages further from a secondary school, indicating that the ability of the bicycle to reduce the time and safety costs of school attendance was the mechanism of impact [17].

Many educational institutions place a high value on reducing student dropouts. Peréz et al. examined the findings of a case study in educational data analytics aimed at identifying undergraduate Systems Engineering (SE) students who had dropped out after six years of enrollment at a Colombian institution. Original data were enlarged and enriched using a feature engineering technique. The experiment's findings indicated that dropout

predictors can be determined with consistent levels of accuracy using simple algorithms. The findings of Decision Trees, Logistic Regression, Naive Bayes, and Random Forest were compared to recommend the best option. In addition, Watson Analytics is evaluated to see how well it works for non-expert users. The major findings are presented to lower dropout rates by identifying reasonable explanations [18].

However, the current study focuses on the impact of prior academic achievement on the academic performance of architecture students. Several factors affect student academic performance [19].

Academic failure is a serious worry at a time when postsecondary education is becoming increasingly vital to economic success. The authors analyzed student data available at registration, such as school records and environmental circumstances, to predict potential failure early, with the goal of swift and successful remediation and/or study reorientation. Three algorithms for artificial neural networks, logistic regression, and random forest were changed. They developed approaches to improve forecast accuracy when specific classes are of great importance. These strategies are applicable across multiple disciplines and are context-independent [20].

It is critical to pay special attention to low enrollments. Low enrollment leads to a drop in literacy rates, directly related to the country's progress and development.

The objectives of this research are cited as follows [9,21].

1. To find a model that will predict the upcoming enrollment of schools before time.
2. To highlight the categories of schools that require close attention in the future and highlight these schools to enroll the maximum number of students.
3. To improve the school's targets for enhancement of literacy rate.

1.2.2. Limitations of Our Work

- Our work is limited to secondary Schools.
- Our investigation of enrollment criteria is limited to the schools of a particular geographical area.

1.3. List of Abbreviations

The notations used in our study are presented in Table 1.

Table 1. List of abbreviations.

Abbreviations	Full Form
SDG	sustainable development goal
PESRP	Punjab Education Sector Reform Program
k-NN	K-nearest neighbor
ITU	Information Technology University
DDF	Directional Distance Function Approach
ANNs	Artificial Neural Networks
IPL	Iran Pro League
LDM	Learning Data Mining
HSS	Higher Secondary School
RFC	Random Forest Classifier
SVM	support vector machines
DT	Decision tree
NB	Naive Bayes
MLR	Multiple Linear Regression
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

The first part of the paper is related to the introduction of our research. The rest of the paper is arranged as follows. Section 2 presents the related work in our research area. Section 3 is about the research methodology and techniques. Section 4 presents simulation

work, and its results are discussed in detail. The conclusion is presented in Section 5. Finally, future work is discussed in Section 6.

2. Related Work

Uskov et al. performed a recent analysis of three New Jersey schools which shows that current public-school enrollments provide many future enrollments. Three New Jersey school districts were implemented to project enrollments three years into the future. Stochastic forecasting is commonly used in large geographic domains such as provinces or countries; it has not been widely used for congested fields such as the school level. At the elementary level, gender, age, father's income, and literacy rate significantly impact student performance. One of the hot issues is improving the productivity of students entering school. The primary reason for dropping out of school is to observe students and take precautions as early as possible to determine the reasons for dropping out. The model uses educational methods of data mining. The research plan is divided into six stages, i.e., data collection, data integration, data pre-processing (e.g., cleaning, normalization conversion), feature selection, template extraction, model optimization, and evaluation. Comparison results were obtained and discussed [22].

According to Adebayo and Chaubey, educational data analysis in the recent area of research has been welcomed over the past decade due to its ability to monitor student performance and predict future development. Many machine learning methods, significantly controlled learning algorithms, have developed accurate models to predict student characteristics and stimulate their behavior [23].

Uskov et al. examined and evaluated the effectiveness of two packaging methods for supervised learning algorithms to expect student performance at the final exam discussed in this paper. Preliminary numerical experiments show that the advantage of the technique under observation is significantly improving classification accuracy by developing reliable prediction models using markers and a lot of unmarked data [22]. It was stated that educational institutions' primary purpose is to facilitate pupils' productive education. One of the ways to attain the highest quality is to replace the traditional model of teaching in the classroom by opening the predictive knowledge of students participating in the course [23].

According to Zhang, this exploration investigates the expected nearness in the market. It sets up a choice help structure that merchants can use to turn recommended bits of knowledge into future stock, the value bearing with the critical potential for taking that stock. It is fascinating to many people to foresee the outcome of sports matches, from viewers to punters. It is also important as a research issue because of its complexity to a certain extent. The result of a sports match depends on several variables, such as a team's morale (or a player's), abilities, current score, etc. So even for sports experts, the exact results of sports matches are complicated to predict [24].

Research concerns using an approach to machine learning, Artificial Neural Networks (ANNs), to predict the results of one week, specifically applied to the 2013–2014 football matches in the Iran Pro League (IPL). The data from the past matches in the last seven leagues are used to make better predictions for the games to come. Results showed a remarkable ability of neural networks to predict football match results [25].

Adebayo and Chaubeysaid that the markets are unstable, and strategies that create strong expectations on one platform can allow more traders to take that action. In a perfect world, if such "idea floats" can be standard, the broker should store models for use in each new market circumstance (or thought) and afterwards model those models in the coming information that must apply [23].

According to Cortez et al. [26], the future is unpredictable, so the possible concepts are unknown. Keeping up a model with the most forward-thinking cost information is not generally the most advantageous alternative as the market is recouping, and old data are helping later. Therefore, short preparation occasions permit the changed classifier to work with high-recurrence stock information, which can bring about a loss of items because of

an absence of appropriate practice time. The model adopts an alternate strategy to learning by streaming ideas, streams the thought, and creating them as a model.

Similarly, the design acknowledges these adjustments in the market by building a vast number of conventional base scientific classifications (SVMs, Choice Trees, and Neural Systems), covering specific (area) shares, using a random subset of previous data, and the best of these taxonomic models. When the market changes, the base classification of the ensemble is adjusted to be complex and to maintain a high model efficiency level. Policies improve already established algorithms. This study also discusses specific issues related to learning with existing data resources, especially class inequality. Media releases and feature creation due to time of day (such as technical and emotional analysis), dimensionality reduction, and model output. It addresses standard transactional methods, identifying unfair practices used in online exams and identifying outliers in the results, worksheets, student performance forecasts, etc. Knowledge is hidden in educational datasets and can be extracted using data mining techniques [21].

Grading exercises were used to measure student performance, and because of the many methods used to classify data, the decision tree approach was used here. With the help of this assignment, we can extract knowledge that describes the student's academic performance in the final exam. This helps identify dropouts and students who need special attention early and allows teachers to provide relevant advice and counseling [21].

The method is used to predict student performance based on essential features, such as the age of the student, the school in which they study, place of residence, number of households, previous performance scores, and activities, to verify the effectiveness of the model [27].

Compression of the proposed method with other well-known classifiers has been performed. Studies of existing student data show that this method is suitable for evaluating student performance. Research and analysis of educational data, especially student performance, is fundamental. Learning Data Mining (LDM) is a research area that produces educational data to identify interesting patterns and knowledge in educational institutions [21]. The study of Soofi and Awan [27] also focuses on computational education, specifically exploring the factors that theoretically affect student performance in higher education and finding a qualitative model for relevant personal and social factors that classifies based on students' performance.

Iqbal et al. evaluated the effectiveness of various university admissions criteria in Pakistan using a case study of the Information Technology University (ITU) in Lahore. They focused on the applicants' academic standing for ITU's Bachelor of Science in Computer Science program. It was discovered that some of the admissions requirements strongly correlated with the student's overall academic performance. The results of the admissions test and a High School Diploma were the best indicators of academic success (HSSC). Their main finding was that when determining admission to a university, a candidate's performance on the entrance exam and the HSSC should be given a lot of weight [28].

According to Pal and Pal, considering the sum of the above investigations, it very well may be said with sureness that the abilities and experience of the administrators assume a significant job in surveying the presentation of proposition ventures, which from one perspective, affirms past exploration [29].

Results show that the proposed algorithm can predict student dropouts within 4–6 weeks after the course and is trustworthy, and can be used in early warning systems. In the intellectual analysis of educational data, the central problem in discovering knowledge from data are the determination of representative data sets and constructing classification models based on their individual demographic and social foundations. The results show that the accuracy of the classification model generated by the Random Forest algorithm and the J48 algorithm exceeds 71%. [21].

The programmer generates a continuous software response algorithm to adjust and improve the pattern using these input signals. The model is further enhanced with each

new data set provided in the program to clearly distinguish between "humans" and "non-humans" [30].

There is no denying that machine learning can make people work more innovatively and reliably. Finally, machine learning allows us to read, save, and initiate very complex or stressful situations in machine records, such as paper invoices, processing, processing, and editing. Machine learning helps to train the model on the data set before deployment. Some AI models are ceaseless and on the web. This reproducible web-based displaying technique prompts enhancements in the associations between information parts. Those examples and affiliations could undoubtedly have been ignored by human perception because of their unpredictability and size. After a model has been prepared, a method called "gaining from the information" can be used continuously [27].

AI strategies are significant for more exact evaluation models. Depending on the data's form and quality, different methods are used depending on the business problem's nature. We discuss the types of machine learning. Supervised practice usually begins with understanding defined data collection and analysis. The supervised practice aims to identify data patterns applicable within the analytics framework [31].

The data used in Superby et al.'s research include features assigned to explain the data's function. For example, if we want to develop a machine learning application that differentiates millions of species based on features, i.e., images and written descriptions [12].

Unsupervised practice is used when the problem involves significant amounts of unlisted data; life applications such as Twitter, Instagram, and Snapchat all contain much-unlisted information. Understanding the essentialness behind this information requires a philosophy that characterizes information as dependent on personality or group. Unskilled training follows a clear example: testing knowledge without human mediation [32].

This is used for spam location innovation via phone. Official and spam communications have too many variables for an observer to tag unsolicited bulk emails. The emails-learning classification focused on clustering to identify spam emails [33].

Akinode and Bada investigated the impact of various pre-admission factors (WAEC grades, JAMB Scores, etc.). The field survey method was used in their study. A dataset of 560 students enrolled in various courses at a Federal Polytechnic in South-West Nigeria from 2017 to 2018 was used to validate the proposed methodology. Their research used machine learning methods to examine the impact of various factors on student enrollment. The analysis employed the decision tree algorithm (ID3) and support vector machine (SVM) techniques. The Scikit learn tool was used for pre-processing, processing, and experimenting. Results were obtained by comparing the ID3 Decision Algorithm to other ML Algorithms such as Artificial Neural networks, Logistics, etc. Regression analysis reveals that the ID3 algorithm outperforms other ML algorithms. The Decision Tree is the most accurate. This research is beneficial for enhancing the students' enrollment and finding the below targets schools in time so that necessary action can be taken to improve the performance of schools. Prediction and classification techniques were used to analyze students' performance and dropout [34].

Kim and Sunderman looked at student achievement data from six states to see if any demographic differences existed between schools that were identified as requiring improvement and those that achieved the federal standards for adequate annual progress. It is shown that using mean proficiency scores creates a selection bias and that requiring students in schools with a high racial diversity to meet multiple performance goals contributes to these disparities using school-level data from Virginia and California. The authors suggested new approaches to creating accountability systems, including using multiple student achievement indicators, such as student growth in reading and math achievement tests and state accountability ratings of school performance [35].

Machine learning is a subfield of computer science that grew out of the study of pattern recognition and computational learning theory in the field of artificial intelligence [36].

According to the statistics, the most used data mining algorithms are ANN and Random Forest, while WEKA is gaining popularity as a way of forecasting student achievement.

Previous academic success and demographic traits are the most accurate predictors of a student's potential. This study demonstrates that including extraneous features in a dataset reduces prediction accuracy and raises the model's excessive computing cost. This work paves the path for future researchers to employ a wide range of inputs and approaches to obtain remarkably accurate prediction results for a wide range of scenarios. The research also teaches educational institutions how to use data mining techniques to increase their ability to foresee and improve student achievements through the quick provision of additional support services [37].

The authors used an ANN to assess candidates' eligibility for admission to a university based on their O-level scores, CGPAs, departmental rankings, and other information. Positive results from performance analysis using the Confusion Matrix and the AUC ROC suggested effective prediction and provided an overall accuracy of 99% [38].

The authors proposed a brand-new machine learning-based approach to anticipate undergraduate students' final exam grades using midterm exam results as the source data. The performance of the machine learning methods, random forests, nearest neighbor, support vector machines, logistic regression, Naive Bayes, and k-nearest neighbor, were calculated and compared to forecast the students' final exam marks. This research determines the most efficient machine learning algorithms and contributes to the early identification of students who are at high risk of failing [39–42].

Several classification methods, including decision trees, random forests, SVM classifiers, SGD classifiers, Ada Boost classifiers, and LR classifiers, were used to analyze the dataset. The results show that random forest outperforms the other methods (98%). Decision tree, Ada Boost, logistic regression, and SVM receive 90%, 89%, and 88%, respectively, whereas SGD, SVM, and SGD receive 84%. According to the research, technological factors have a significant influence on children's academic achievement. Students who used social media daily performed worse than those who used it just sometimes on the weekends. Additionally, assessments are made on how various factors affect student results [43].

3. Materials and Methods

3.1. Workflow of Research

After defining the problem of school enrollment, convenient features were selected based on related work and the current scenario. The dataset of student information including many important features (age, gender, family Size, and physical health) was collected from the head office of the education department of Punjab for research purposes. This dataset (after preprocessing) is reliable for the proposed research because it was collected from a real-time monitoring survey by the School Education Department. Data pre-processing is the backbone of further analysis and accuracy of algorithms. It involves removing missing values, feature scaling, managing categorical data, and many other useful tasks to make the data consistent. Here is the brief overview of data after preprocessing i.e., missing values is filled with technique of median and feature scaling is done. The data after pre-processing has shown in the Table 2.

Feature reduction methods were analyzed, and we selected the Backward Elimination Method for feature reduction. Now it is time for data pre-processing. Data split in the training and testing phase is done using the Stratified Shuffled split technique.

There are multiple techniques for splitting the data into testing and training data:

(1) Test Train split (Simple divide dataset into purposed ratio)
(2) Stratified Shuffled Split (used for equal division of categorical data)
(3) K-fold (iterative technique for data split).

Stratified shuffled split is used because dataset has a feature known as "GENDER". Based on this, data are divided into the test and train sets. After applying stratified shuffled split. The dataset is perfectly divided. For example, the 100-testing dataset has 24 Female schools and 380 training dataset has 76. The OLS final summary is presented in Table 3.

Table 2. After preprocessing of data.

Sr No.	Feature	Count	Status	Data Type
1	Emis_code	500	non-null	int64
2	GENDER	500	non-null	int64
3	Age	500	non-null	float64
4	P Status	500	non-null	float64
5	Medu	500	non-null	float64
6	FEdU	500	non-null	float64
7	MJOB	500	non-null	float64
8	FJOB	500	non-null	float64
9	GUARDIAN	500	non-null	float64
10	FAMSIZE	500	non-null	float64
11	School	500	non-null	float64
12	Family	500	non-null	float64
13	HEALTH	500	non-null	float64

Table 3. Final OLS summary.

Features	Coef	Std Error	t	$p > \|t\|$
GENDER	4.1097	1.345	3.057	0.002
Age	−7.3404	12.055	−0.609	0.047
Medu	−80.4832	13.186	−6.103	0.041
FEdU	92.4556	12.83	7.206	0.033
MJOB	3.3693	6.216	0.542	0.049
FJOB	16.6516	10.952	1.52	0.028
P Status	−1.8119	6.987	−0.259	0.017
Family edu support	0.0755	0.273	0.277	0.017
HEALTH	−0.0884	0.121	−0.732	0.046

Algorithms Implementation for both classification and regression is the next objective.

After a brief comparative analysis, a conclusion is drawn. Research design (workflow is presented in Figure 2) is as follows.

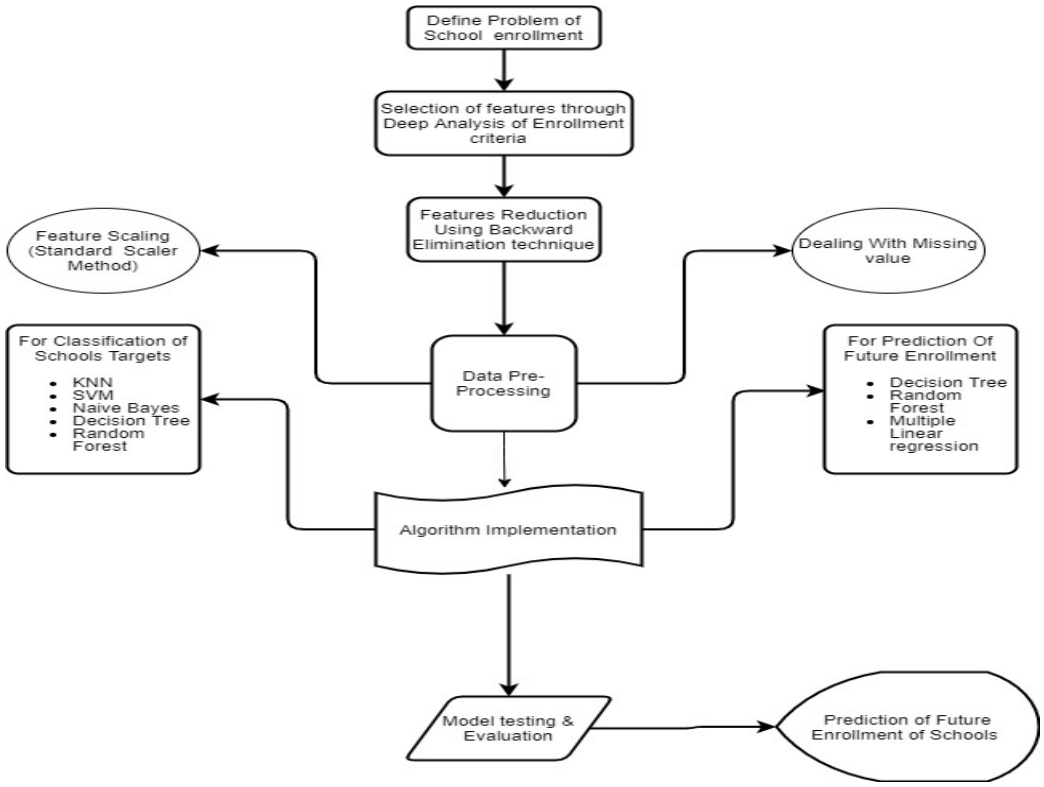

Figure 2. Workflow of research.

3.2. Algorithms Used for Classification

Algorithms used for classification are

- Decision tree Classifier
- Random Forest Classifier
- KNN Classifier
- SVM Classifier
- Naïve Bayes Classifier

And algorithms used for the prediction (Regression) the future enrollment are:

- Multiple Linear Regression
- Random Forest Regression
- Decision Tree Regression

3.3. Data Collection

Five hundred students' data were collected from public schools in the Punjab province of Pakistan, especially schools with low enrollment in early classes. In addition, rural and urban students are represented. The dataset of student information, including many important features (age, gender, family size, physical health), was collected from the head office of the education department of Punjab for research purposes. This dataset (after pre-processing) is reliable for the proposed research because it is compiled from a real-time monitoring survey of the School Education Department [26].

The selection of required features demands special attention, because all the upcoming processes and models' accuracy are correlated with selected features. Twelve features are chosen for this research. The data for genders containing testing and training data are presented in Table 4.

Table 4. The Dataset.

	Testing Dataset	**Training Dataset**	**Total**
Male	76	304	380
Female	24	96	120
Total	100	400	500

This raw data demands to be clean and noise-free on the same scale. Therefore, pre-processing is the only solution to make this data consistent and reliable [30].

3.4. Data Mining Process

Data mining is widely used by businesses, organizations, and governments to find "hidden" patterns and connections in their transactional data. There has been a lot of progress in the last several years. Data mining algorithms can address any problems that come up while analyzing data. Models for Data Mining and Knowledge Discovery reached their peak with the release of SEMMA in 2000. The five-step SEMMA framework is used by the SAS Institute to organize the phases of data mining. SEMMA stands for Sample, Explore, Modify, Model, and Evaluate. An overview of each SEMMA process phase is shown in Figure 3.

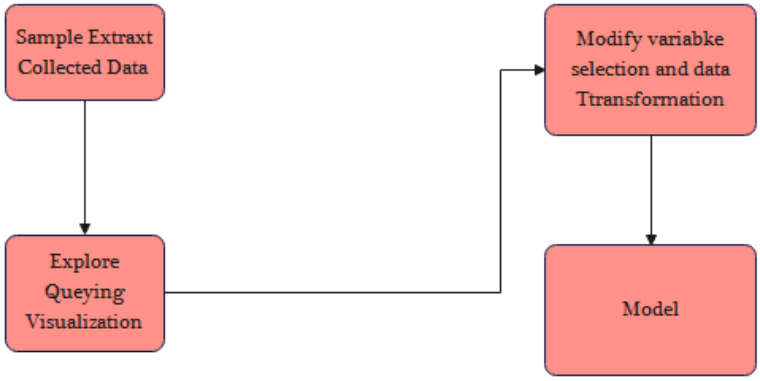

Figure 3. Workflow Process with Data Mining.

3.4.1. Sampling

A relevant subset of data comprising the most important and easily manipulable information must be extracted as the first step in the SEMMA data mining process, which starts with a large dataset. It may be possible to greatly shorten the time required for data mining for extremely large datasets by concentrating on a subset rather than the complete dataset. A statistically significant sample will reveal the existence of a statistically significant trend in the data. Sub-data must be taken into account in the context of datasets.

3.4.2. Explore

The second step is to visually or statistically analyze the data to look for underlying classifications or patterns. Research allows the refinement and modification of the discovery process. This promotes conceptual and cognitive growth. If a visual review of the data does not show any clear patterns, statistical methods, such as factor analysis, correspondence analysis, and clustering can be employed to analyses the data.

If all of the data are processed at once, likely, more intricate patterns will not be found until further exploration is done.

3.4.3. Model

The data will be cleaned and organized, and statistical models will be created to show how the patterns behave. Data mining uses a variety of statistical models, including time series analysis, memory-based reasoning, and principal component models, among others, as well as modeling approaches including artificial neural networks, decision trees, rough set analysis, support vector machines, and logistic models. Each model has distinctive qualities and attributes that make it the best choice for various data mining tasks.

3.4.4. Evaluation of the Model's Performance Is the Last SEMMA Stage

The user's assessment of the model is used to evaluate how accurately it predicts the criteria. It is typical to divide the data into training and testing sets when evaluating a model. While the second is used to assess the model, the first is used to instruct it. This step's goal is to evaluate how closely the model resembles the training data set's outcomes. Both the training sample and the reserved sample can make use of the model if it is accurate.

3.4.5. Tools and Data Analysis Techniques

Python is used for data pre-processing and machine learning algorithm implementation. We also have a significant variety of software related to our domain. Still, the python tool is user-friendly and easy to apply to Artificial Intelligence algorithms, especially Machine learning and deep learning algorithms. It is suitable for large and small datasets [44].

Scikit Learn performs early data processing using Scikit learn (Sklearn), the most well-liked and dependable machine learning library for Python. Clustering, regression classification, and dimension reduction are a few of the powerful machine learning and

data mining modeling tools accessible through a Python-compatible interface [26]. Python code for this library is created by combining NumPy, SciPy, and Matplotlib. Outliers are eliminated, missing values are filled in, and features are scaled during the pre-processing data stage [45].

We will go into detail on three different categories in what follows:

Based on a set of data, estimators are computer programs that can make educated assumptions about specific parameters. The fit and transform methods are available. The fit procedure performs a proper operation and calculates internal parameters, to obtain useful data-related information [46].

The transform method takes in data and outputs results as close as possible to the input data. Another function that carries out both fitting and modifying is fit transform [44].

Models such as the linear regression model and predictors are similar. Fit and prediction are very similar when compared. It also has a scoring feature that can assess the forecasts' accuracy [47].

A total of five hundred entries are present in the data set with the help of the GENDER count function. One-hundred twenty Female Schools' and 380 Male Schools' data are present in the data set, and a significant part of the data needs scaling [48,49].

Feature scaling is used to make all the attributes on one scale. Primarily, two types of feature scaling methods:

1. Min-max scaling (Normalization) (value—min)/ (max-min) Sklearn provides a class called Min Max Scaler for this.
2. Standardization (value—mean)/std Sklearn provides a class called Standard Scaler.

Backward elimination is a technique for selecting significant features when constructing a machine learning model. This technique eliminates elements that do not affect the dependent variable [50]. There are multiple ways to construct a Machine Learning model; some are given below:

- Backward Elimination
- All-in
- Forward Selection
- Score Comparison
- Bidirectional Elimination

Already there are potential methods for training the model in Machine Learning, but we used the Backward Elimination approach as it is the quickest way. A list of proposed features is presented in Table 5, and the Final OLS summary in Table 6.

Table 5. List of proposed features.

Sr #	Attributes	Description (domain)
1	GENDER	Male = 1 Female = 0
2	Age	students Age
3	Parents STATUS	parents' contribution (living together or not)
4	M.Edu	Mother's education
5	F.Edu	Father's education
6	M_JOB	Mother's job
7	F_JOB	father's job
8	GUARDIAN (Parents %)	std guardian (parents or other)
9	FARM SIZE	Family size
10	School extra edu support	extra educational school support
11	Family edu support	family educational school support
12	HEALTH (good health %)	Health Status of Students

Table 6. Final OLS Summary.

Features	Coef	Std Error	t	$p > \|t\|$
GENDER	4.1097	1.345	3.057	0.002
Age	−7.3404	12.055	−0.609	0.047
Medu	−80.4832	13.186	−6.103	0.041
FedU	92.4556	12.83	7.206	0.033
MJOB	3.3693	6.216	0.542	0.049
FJOB	16.6516	10.952	1.52	0.028
P Status	−1.8119	6.987	−0.259	0.017
Family edu support	0.0755	0.273	0.277	0.017
HEALTH	−0.0884	0.121	−0.732	0.046

Below are some primary steps which are used to apply the backward elimination process:

1. First, we will decide on the model's level of significance. (SL = 0.05)
2. The second step involves fitting the entire model, which includes all independent variables and potential predictors.
3. Third, as shown in the image below, pick the predictor with the highest *p*-value.
4. Proceed to Step 4 if the *p*-value is less than the SL. However, when you are done, our model will be too.
5. Step 4 is to remove the predictor.
6. Fifth, rebuild and re-fit the model after removing some of the variables.

Three things we discovered could be eliminated from our model to improve its accuracy. Our work will become worse if we try to eliminate even more. The alfa value increased in the eliminated traits (family Size, School Extra Education, Parent's Status). There is further analysis done on the remaining features.

There are several ways to split datasets into training and test sets, including:

- Test Train split (Simple divide dataset into purposed ratio)
- Stratified Shuffled Split (used for equal division of categorical data)
- k-fold (an iterative approach for data split.) [47].

Stratified shuffled split is used because the dataset has a feature named "GENDER." After applying a stratified shuffled split, this data are divided into test and train sets. The dataset is perfectly divided. Such as the 100-testing dataset has 24 female and 76 male schools, and the training dataset has 120 female and 380 male schools [46].

Manual enrollment targets are assigned to all public schools, and approximately 70% of marks are achievable. The rest of the targets always remain unachieved. This research significantly contributes to increasing enrollments in public schools and helps to achieve the given targets.

This research aims to predict and Identify low enrollment schools and classify them according to targets.

4. Results and Discussion

4.1. Classification

Classification is the technique of machine learning used for predicting group composition in data instances. Several classification methods are available and can be used for classification purposes. In this section, we discussed Linear and Sigmoid kernels, the basic classification techniques, and some major classification method types, including the k-nearest neighbor, decision tree, classifier and support vector machines (SVM), Naive Bayes, and Random Forest [51].

Class-wise categories are mentioned below:

- Class 1: Below Target
- Class 2: Far from Target
- Class 3: On target

4.1.1. Random Forest

Random forests are popularly used for data science competitions and practical problems. They are accurate, do not require scaling or categorical encoding of features, and need little parameter tuning. It consists of several random decision trees [52]. Inside the trees are built two types of randomness. First, each tree is constructed from the original data on a random sample. Second, a subset of functions is randomly selected at each tree node to produce the best split. Random Forest Classifier (RFC) is the most common ensemble learning classifier proven to be a very popular and effective cognition and machine learning technique for high-dimensional classification and skewed problems. RF classifier performance with truth data and results are presented in Figure 4.

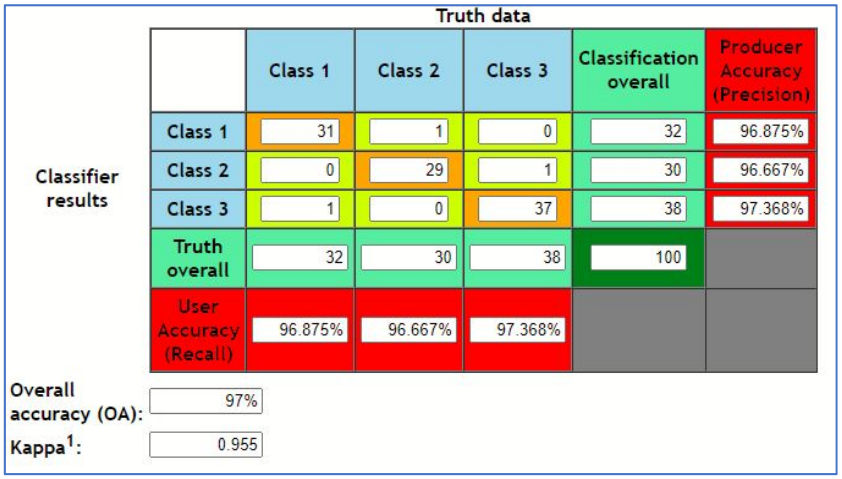

Figure 4. RF Classifier Performance. The accuracy of Random Forest is 97% as shown in the truth table (Figure 4).

4.1.2. Decision Tree

Growing machine learning algorithms have their benefits and motivations for implementation. The decision tree algorithm is one of the most widely used algorithms. A decision tree is an upside-down tree that makes decisions based on the conditions of the data [10]. Decision tree performance with truth data classifier results and accuracy is presented in Figure 5.

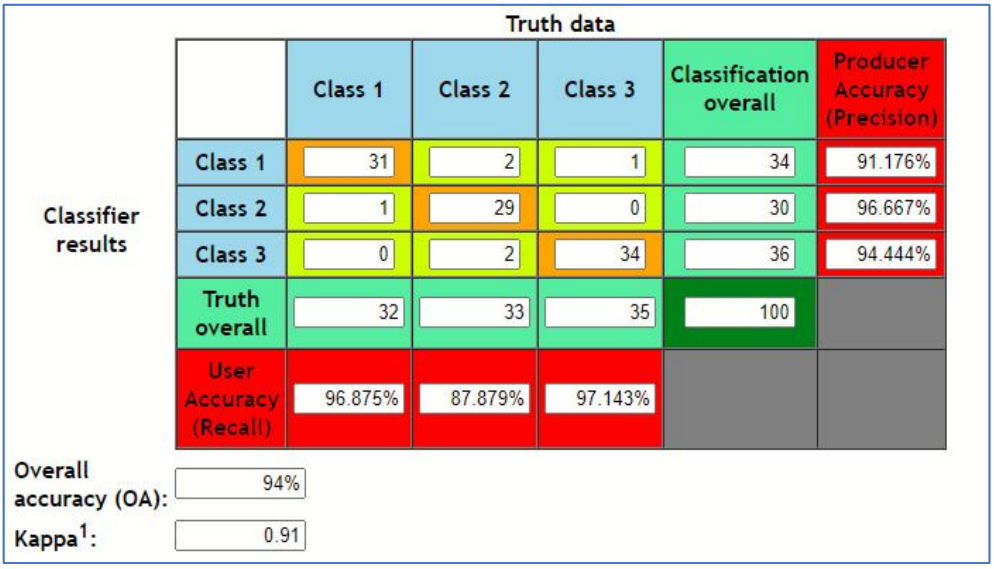

Figure 5. DT Performance. The accuracy of Decision Tree is 94% as shown in the truth table (Figure 5).

4.1.3. SVM (Linear)

A support vector machine (SVM) is a supervised machine learning model that employs classification algorithms to solve problems involving classification into two classes. After providing an SVM model with sets of labeled training data for each category, they can categorize new text. As a result, we are attempting to solve the enrollment target classification issue. We are working to improve our training results and may have even experimented with Naive Bayes. However, now that we can look forward to enjoying the dataset, we want to go a step further. Enter Support Vector Machines (SVM) is a straightforward and trustworthy classification algorithm with sparse data performing admirably [53]. SVM performance with truth data classifier results and accuracy is presented in Figure 6.

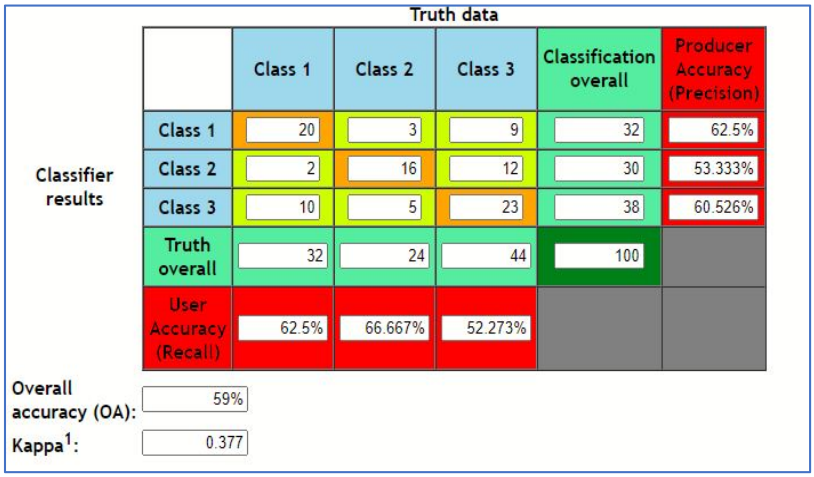

Figure 6. SVM Performance. The accuracy of SVM (linear) is 59% as shown in the truth table (Figure 6).

Having searched a little deeper and bumped into concepts such as linearly separable, kernel trick, and kernel functions [53]. The idea behind the SVM algorithm is simple and does not require much of the complicated stuff to apply to the classification of targets.

4.1.4. SVM (Sigmoid Karnal)

The Accuracy of this Model is not up to the mark. A detailed Confusion table and essential parameters are given in Figure 7.

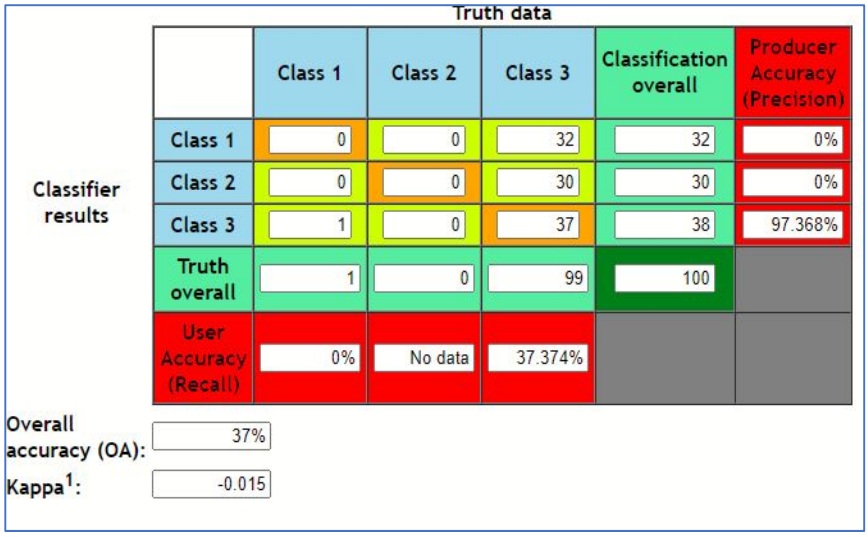

Figure 7. SVM Sigmoid Confusion Matrix. The accuracy of SVM (Sigmoid Karnal) is 37% as shown in the truth table (Figure 7).

4.1.5. KNN

The classification technique k-Nearest-Neighbors (kNN) is one of the common ways in machine learning; it is essentially classification by identifying the most similar pieces of information in the training data and making an informed guess based on their classification. As it is very easy to understand and implement, this approach has seen wide use in many fields, for example, in recommendation systems, semantic searching, and the identification of anomalies [53]. KNN performance, with truth data, classifier results, and accuracy, is presented in Figure 8.

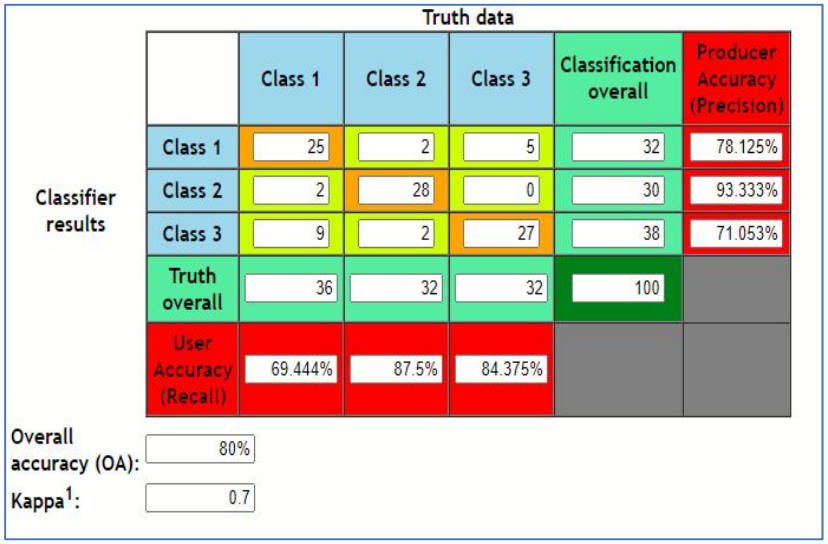

Figure 8. KNN Performance. The accuracy of KNN is 80% as shown in the truth table (Figure 8).

4.1.6. Naive Bayes

Naive Bayes is a classifier for machine learning that is easy but efficient and commonly used. It is a probabilistic classifier making classifications in a Bayesian setting, using the Maximum A Posteriori decision law. It can also be represented using a Bayesian network which is very basic. Naive Bayes classifiers are particularly popular for classification and are a traditional solution to problems such as spam detection [54–56]. NB performance, with truth data, classifier results, and accuracy, is presented in Figure 9.

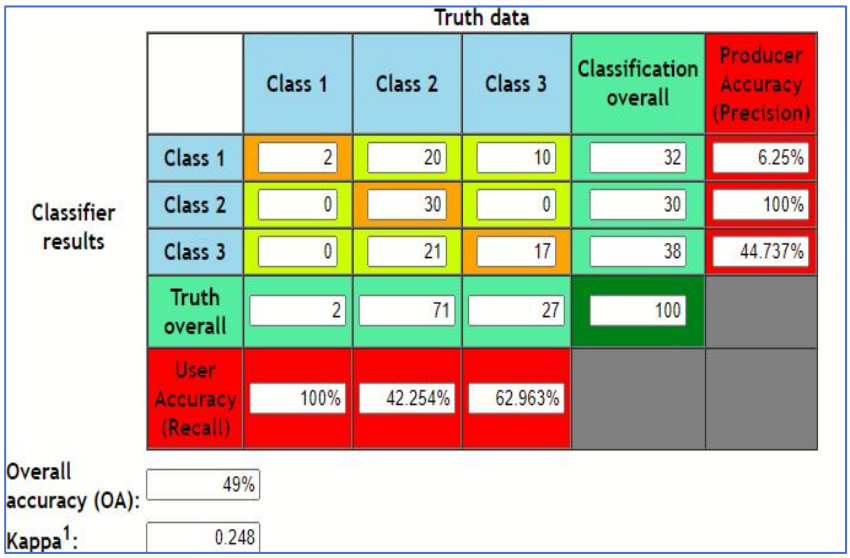

Figure 9. NB Performance. The accuracy of Naive Bayes is 49% as shown in the truth table (Figure 9).

4.1.7. Models Summary

A summary of all models containing precision, Recall, fl-score, and accuracy is presented in Table 7. The accuracy of the random forest model is high, among others.

Table 7. Classification Summary.

		Precision	Recall	f1-Score	Accuracy
SVM Linear	Below Target	0.69	0.78	0.74	59
	Far From Target	0.88	0.93	0.9	
	On Target	0.84	0.71	0.77	
	Macro avg	0.8	0.81	0.8	
	Weighted avg	0.81	0.8	0.8	
SVM Sigmoid	Below Target	0	0	0	37
	Far From Target	0	0	0	
	On Target	0.37	0.97	0.54	
	Macro avg	0.12	0.32	0.18	
	Weighted avg	0.14	0.37	0.21	
KNN	Below Target	0.69	0.78	0.74	80
	Far From Target	0.88	0.93	0.9	
	On Target	0.84	0.71	0.77	
	Macro avg	0.8	0.81	0.8	
	Weighted avg	0.81	0.8	0.8	
Random Forest	Below Target	0.94	0.91	0.92	95
	Far From Target	1	0.97	0.98	
	On Target	0.93	0.97	0.95	
	Macro avg	0.95	0.95	0.95	
	Weighted avg	0.95	0.95	0.95	
Decision tree	Below Target	0.97	0.91	0.94	94
	Far From Target	0.88	0.97	0.92	
	On Target	0.97	0.94	0.96	
	Macro avg	0.94	0.94	0.94	
	Weighted avg	0.94	0.94	0.94	
Naïve Bayes	Below Target	1	0.06	0.12	49
	Far From Target	0.42	1	0.59	
	On Target	0.63	0.45	0.52	
	Macro avg	0.68	0.5	0.41	
	Weighted avg	0.69	0.49	0.41	

The accuracy of all models in which Random Forest has high accuracy is shown in Figure 10.

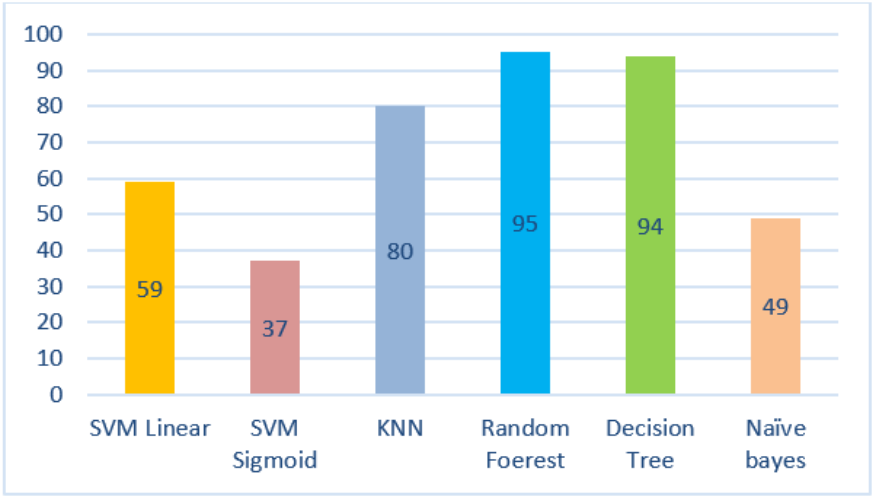

Figure 10. Classification Models Summary.

4.2. Prediction

The second part of this research consists of prediction. Prediction of future enrollment is the prime objective, and we implemented three major algorithms for prediction. Results and comparative analysis are discussed in the below section. Parametric Charts and compression tables highlighted the comparisons of parameters of regression models. In addition, the highest performance in terms of R2 and RMSE is the Random Forest Regressor.

4.2.1. Multiple Linear Regression

Multiple regression analysis methods are used to look into relationships that run in a straight line between two or more variables. The independent variables (IVs), also known as stand-alone variables, are the Xs. Y is the dependent variable (DV) [53]. The subscript j stands for the observation's number (row). The regression coefficients are in the form of the. They base their projections on b's.

In contrast, b is close to the original unknown parameter (population). Although $n = 9$ is used in this study, a general equation with independent variables is provided. The features coefficient is presented in Figure 11.

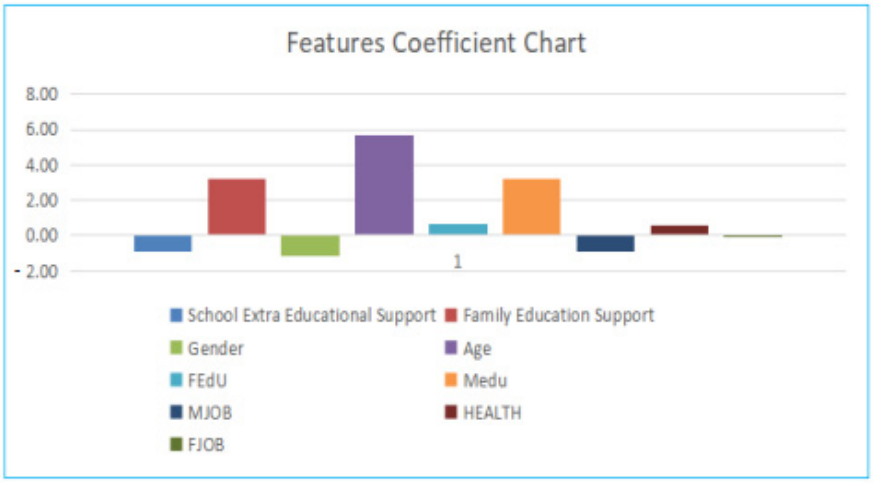

Figure 11. Coefficient of Features.

Multiple Linear Regression Evaluation Parameters are presented in Table 8, and Intercept and Coefficients are presented in Table 9.

Table 8. Multiple Linear Regression Evaluation Parameters.

Explained variance	0.1746
Mean squared log error	0.0668
r2	0.1739
MAE	14.1085
MSE	292.1716
RMSE	17.093

Table 9. Multiple Linear Regression Intercept and Coefficients.

Sr No.	Features	Coefficients
1	School Extra Educational Support	−0.91
2	Family Education Support	3.21
3	Gender	−1.16
4	Age	5.67
5	FEdU	0.66
6	Medu	3.17
7	MJOB	−0.9
8	HEALTH	0.61
9	FJOB	−0.01

Though various techniques can solve the regression problem, the most commonly used approach is the least squares. The bs are chosen to minimize the sum of the squared residuals in the least square regression analysis. This set of bs is not necessarily the set we want because outliers—points that are not representative of the data can distort them.

Intercept is 65.43

- Mean Absolute Error (MAE) is the mean of the absolute value of the errors;
- Mean Squared Error (MSE) is the mean of the squared errors;
- Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors;
- MSLE Is the loss being the mean over the seen data of the squared differences between the log-transformed true and predicted values, or writing it as a formula;
- Where ŷ is the predicted value, this loss can be interpreted as a ratio between the true and predicted values. Because MSE "punishes" larger errors, it is more popular than MAE. However, RMSE is more popular than MSE because it can be expressed in the "y" unit. E. Model Evaluation parameters are presented in Figure 12.

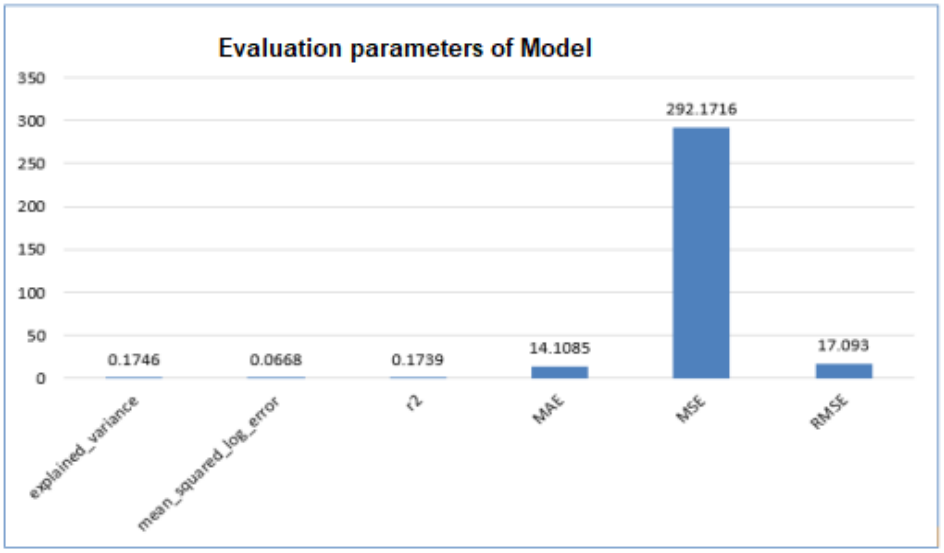

Figure 12. Model Evaluation parameters.

4.2.2. Random Forest Regression

The Radom Forest regressor's performance is comparatively high compared to other models used in this research. Detailed analyses are given below in Figure 13.

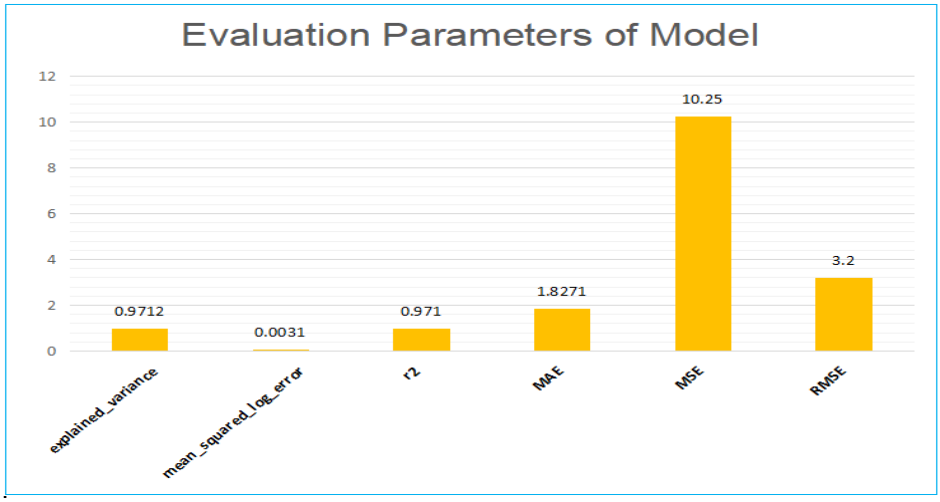

Figure 13. Random Forest Regressor Parameters.

4.2.3. Decision Tree

The evaluation parameters of the decision tree model are presented in Figure 14.

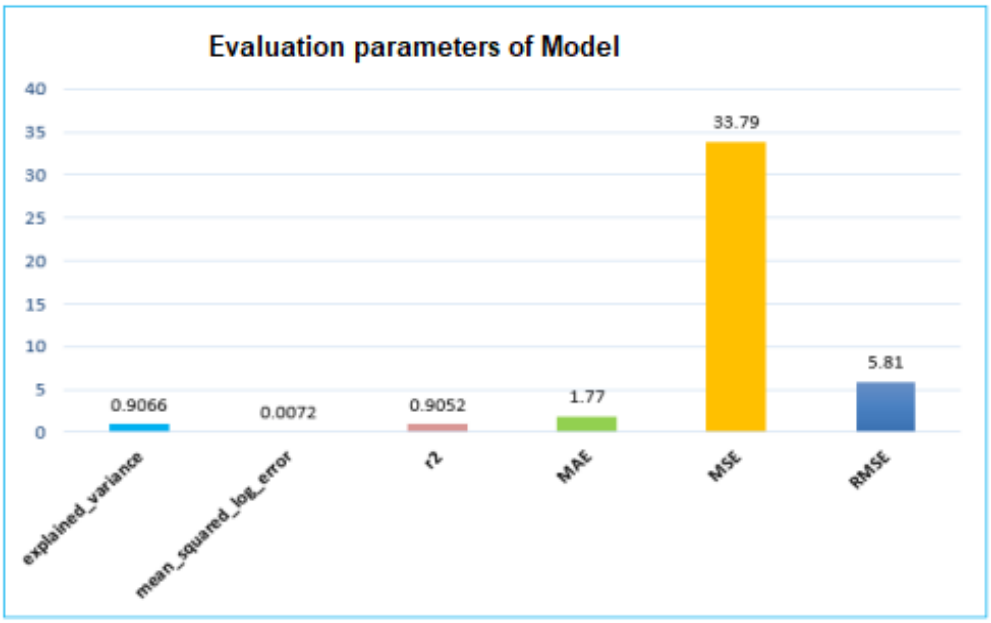

Figure 14. DT Performance.

4.2.4. Models Summary

A comparative Summary of Models MLR, RF, and DT are presented in Figure 15.

Figure 15. Comparison of Regression Models.

Now going to the conclusion of the Regression part, the best model for predicting School Enrollment is a Random Forest Regressor. Model Performance is above 97%. The parametric analysis of all subject models is presented in Figure 16.

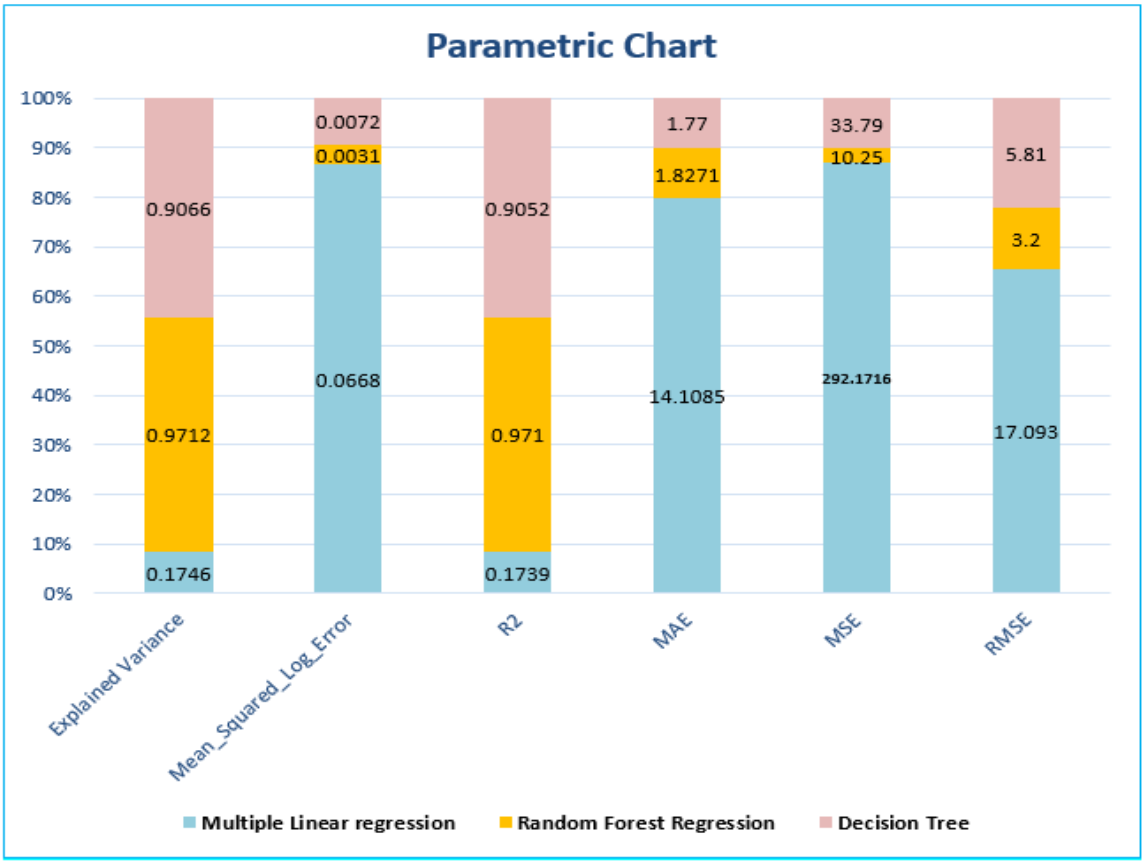

Figure 16. Parametric Analysis.

Parametric Charts highlighted the comparisons of parameters of regression models. The highest performance in terms of R2 and RMSE is the Random Forest Regressor. In addition, a comparison of regressors models is presented in Table 10.

Table 10. Comparison of Regressors Models.

Algorithms	Explained Variance	Mean Squared Log Error	R2	MAE	MSE	RMSE
Multiple Linear Regression	0.1746	0.0668	0.1739	14.1085	292.1716	17.093
Random Forest Regression	0.9712	0.0031	0.971	1.8271	10.25	3.2
Decision Tree	0.9066	0.0072	0.9052	1.77	33.79	5.81

4.3. Testing of Regression Model

Under the supervised learning system, different model evaluation techniques can be used, which helps us determine how well our model is doing. A straightforward approach for testing a model is to find the variance that is the difference between the expected and actual values, but it is not the best solution and can lead to poor decision making. Furthermore, we need more measures to evaluate the different models, and choosing the appropriate evaluation measure is crucial in choosing and separating the suitable model from other models.

Unlike classification, In Regression (where we can count the number of results we have classified correctly), we are often incorrect since our estimates are either greater or smaller than the original value (rarely the same as the original value). Therefore, we are also not concerned with how many times we have been incorrect but rather with the quantity of variance between the actual and expected value

The determination coefficient is the most critical method of evaluating a Regression model and is much more common than SSE/MSE/RMSE. The coefficient of determination in statistics is the ratio of the variance in the dependent variable that is predictable from the independent variable(s).

Based on the total outcome variance explained by the model, R-Square measures how well the model replicates well-observed outcomes. The coefficient of determination ranges from 0 to 1. R-Square is a statistic that provides information about the goodness of fit of a model. In Regression, the R-Square determination coefficient is a statistical indicator of how well the regression line approximates the actual data points. An R-Square of one indicates that the data fit well with the regression line. R2 informs us, after training our model, we create a job lib file of the model. This is simply the trained model and it is ready for applications. We performed multiple tests on this model, and the results are below. Take five rows from the test data for testing this model. The output of five random samples is presented in Table 11.

Table 11. Final testing Summary.

Sr No.	Inputs		Observed Enrollment	Actual Enrollment
1	Gender	0.56195	47.96	47
	Family Education Support	0.46871		
	School Extra Educational Support	1.9199		
	Age	1.60163		
	FEdU	1.06295		
	Medu	1.59448		
	MJOB	0.36765		
	HEALTH	−0.1086		
	FJOB	−0.0724		
2	Gender	0.56195	71.62	72
	Family Education Support	0.63557		
	School Extra Educational Support	0.48892		
	Age	0.2189		
	FEdU	0.18339		
	Medu	0.6387		
	MJOB	0.17332		
	HEALTH	−0.017		
	FJOB	−0.0631		
3	Gender	0.56195	47.51	47
	Family Education Support	−1.1999		
	School Extra Educational Support	−1.9914		
	Age	−1.6957		
	FEdU	−0.8721		
	Medu	−1.0817		
	MJOB	0.17332		
	HEALTH	−0.1452		
	FJOB	−0.0668		
4	Gender	0.56195	93.02	95
	Family Education Support	−1.3667		
	School Extra Educational Support	−1.7052		
	Age	−1.3766		
	FEdU	−1.2239		
	Medu	−1.5596		
	MJOB	−0.021		
	HEALTH	−0.028		
	FJOB	−0.05		
5	Gender	0.56195	87.03	91
	Family Education Support	0.96929		
	School Extra Educational Support	0.48892		
	Age	1.17618		
	FEdU	0.71113		
	Medu	1.59448		
	MJOB	0.65915		
	HEALTH	−0.0389		
	FJOB	−0.0537		

5. Conclusions

This study is divided into two sections. The first section includes different models for predicting school enrollment, such as Random Forest Regression, Decision Tree Regression, and Multiple Linear Regression. Enrollment is based on essential student characteristics. Pre-processing techniques extracted useful information from collected raw data. For feature reduction, backward elimination is used. Stratified shuffled split is a data splitting technique that uses cross-validation. Train the model on multiple machine learning algorithms and analyze the best model to predict future enrollment. The highest Regressor model is the Random Forest Regressor having with an accuracy of 97.1%. The second part contains the classification of schools into three categories. The Random Forest classifier has the highest accuracy (94%), whether the school is on target, below target, or far from the target. The goal is to determine which schools will be on track or not in the coming academic year. This research significantly improves literacy and allows for the implementation of specific initiatives for low-performing schools ahead of schedule.

6. Future Work and Directions

- The study can be expanded to include an examination of student retention. Twelve features were chosen based on the scope of the research. However, this research can be taken to a higher level by analyzing the level of study of enrolled students using Machine Learning.
- We can implement deep learning and Federated learning techniques to improve the accuracy of models.
- We can also extend our study to higher secondary schools and colleges, Universities for the analysis of enrollment criteria.

Author Contributions: Conceptualization, Z.u.A., T.M. and A.R.; methodology, T.M., I.H. and Z.u.A.; software, T.M. and Z.u.A.; validation, A.R. and I.U.; formal analysis, T.M., I.H.; investigation, T.M., I.U. and I.H.; resources, I.U. and H.A.; data curation, I.U., Z.u.A.; writing—original draft preparation, T.M., I.H. and I.U.; writing—review and editing, I.H. and I.U.; visualization, A.R. and H.A.; Project administration, I.U., H.A. and H.G.M.; Funding acquisition, I.U., H.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023TR140), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Acknowledgments: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023TR140), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. SDGs. N.I.F.S.D.G. Article IV of SDG 2030. 2022. Available online: https://www.sdgpakistan.pk/ (accessed on 25 December 2022).
2. National Assembly of Pakistan. Article 25(A) of the Constitution of the Islamic Republic of Pakistan. Available online: https://na.gov.pk/en/downloads.php (accessed on 25 August 2022).
3. Ministry of Federal Education and Professional Training Pakistan. Literacy Rate. 2022. Available online: http://mofept.gov.pk/Detail/NDM1NDI0ZTQtZmFjMy00ZTVlLWE5M2YtYjgxOTE4YTkyYWNi (accessed on 25 December 2022).
4. PESRP. The School Census Report 2020–2021. 2022. Available online: https://www.pesrp.edu.pk/ (accessed on 25 December 2022).
5. Aluko, R.O.; Adenuga, O.A.; Kukoyi, P.O.; Soyingbe, A.A.; Oyedeji, J.O. Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. *Constr. Econ. Build.* **2016**, *16*, 86–98. [CrossRef]
6. Iqbal, Z.; Qadir, J.; Mian, A.N.; Kamiran, F. Machine learning based student grade prediction: A case study. *arXiv* **2017**, arXiv:1708.08744.
7. Kotsiantis, S.B. Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artif. Intell. Rev.* **2012**, *37*, 331–344. [CrossRef]
8. Rebai, S.; Yahia, F.B.; Essid, H. A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Econ. Plan. Sci.* **2020**, *70*, 100724. [CrossRef]
9. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* **2009**, *53*, 950–965. [CrossRef]

10. Ciolacu, M.; Tehrani, A.F.; Beer, R.; Popp, H. Education 4.0—Fostering student's performance with machine learning methods. In Proceedings of the 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME), Constanta, Romania, 26–29 October 2017.
11. Doan, T.; Kalita, J. Selecting machine learning algorithms using regression models. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015.
12. Superby, J.-F.; Vandamme, J.; Meskens, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proceedings of the Workshop on Educational Data Mining*; Citeseer: University Park, PA, USA, 2006.
13. Nagy, M.; Molontay, R. Predicting dropout in higher education based on secondary school performance. In Proceedings of the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, 21–23 June 2018.
14. Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access* **2020**, *8*, 55462–55470. [CrossRef]
15. Adekitan, A.I.; Noma-Osaghae, E. Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Educ. Inf. Technol.* **2019**, *24*, 1527–1543. [CrossRef]
16. Baker, R.S. Challenges for the future of educational data mining: The Baker learning analytics prizes. *J. Educ. Data Min.* **2019**, *11*, 1–17.
17. Muralidharan, K.; Prakash, N. Cycling to school: Increasing secondary school enrollment for girls in India. *Am. Econ. J. Appl. Econ.* **2017**, *9*, 321–350. [CrossRef]
18. Pérez, B.; Castellanos, C.; Correal, D. Predicting student drop-out rates using data mining techniques: A case study. In *IEEE Colombian Conference on Applications in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2018.
19. Aluko, R.O.; Daniel, E.I.; Oshodi, O.S.; Aigbavboa, C.O.; Abisuga, A.O. Towards reliable prediction of academic performance of architecture students using data mining techniques. *J. Eng. Des. Technol.* **2018**, *16*, 385–397. [CrossRef]
20. Hoffait, A.-S.; Schyns, M. Early detection of university students with potential difficulties. *Decis. Support Syst.* **2017**, *101*, 1–11. [CrossRef]
21. Azar, A.T.; Elshazly, H.I.; Hassanien, A.E.; Elkorany, A.M. A random forest classifier for lymph diseases. *Comput. Methods Programs Biomed.* **2014**, *113*, 465–473. [CrossRef] [PubMed]
22. Uskov, V.L.; Bakken, J.P.; Byerly, A.; Shah, A. Machine learning-based predictive analytics of student academic performance in STEM education. In Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 8–11 April 2019.
23. Adebayo, A.O.; Chaubey, M.S. Data mining classification techniques on the analysis of student's performance. *GSJ* **2019**, *7*, 45–52.
24. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 27386492. [CrossRef] [PubMed]
25. Abdelhamid, N.; Thabtah, F. Associative classification approaches: Review and comparison. *J. Inf. Knowl. Manag.* **2014**, *13*, 1450027. [CrossRef]
26. Cortez, P.; Silva, A.M.G. *Using Data Mining to Predict Secondary School Student Performance*; EUROSIS-ETI: Oostende, Belgia, 2008.
27. Soofi, A.A.; Awan, A. Classification techniques in machine learning: Applications and issues. *J. Basic Appl. Sci.* **2017**, *13*, 459–465. [CrossRef]
28. Iqbal, Z.; Qadir, J.; Mian, A.N. Admission criteria in pakistani universities: A case study. In Proceedings of the 2016 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 19–21 December 2016.
29. Pal, A.K.; Pal, S. Classification model of prediction for placement of students. *Int. J. Mod. Educ. Comput. Sci.* **2013**, *5*, 49.
30. Yadav, S.K.; Bharadwaj, B.; Pal, S. Data mining applications: A comparative study for predicting student's performance. *arXiv* **2012**, arXiv:1202.4815.
31. Boswell, D. *Introduction to Support Vector Machines*; Departement of Computer Science and Engineering University of California San Diego: La Jolla, CA, USA, 2002.
32. Taheri, S.; Mammadov, M. Learning the naive Bayes classifier with optimization models. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 787–795. [CrossRef]
33. Cortez, P.; Morais, A.D.J.R. *A Data Mining Approach to Predict Forest Fires Using Meteorological Data*; Associação Portuguesa Para a Inteligência Artificial: Braga, Portugal, 2007.
34. Akinode, J.; Bada, O. Student Enrollment Prediction using Machine Learning Techniques. Presented at the 5th National Conference of the School of Pure & Applied Sciences Federal Polytechnic, Ilaro, Nigeria, 29–30 September 2021.
35. Kim, J.S.; Sunderman, G.L. Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educ. Res.* **2005**, *34*, 3–13. [CrossRef]
36. Huang, X.-L.; Ma, X.; Hu, F. Machine learning and intelligent communications. *Mob. Netw. Appl.* **2018**, *23*, 68–70. [CrossRef]
37. Batool, S.; Rashid, J.; Nisar, M.W.; Kim, J.; Kwon, J.-Y.; Hussain, A. Educational data mining to predict students' academic performance: A survey study. *Educ. Inf. Technol.* **2022**, *7*, 1–67. [CrossRef]
38. Ekong, A.; Silas, A.; Inyang, S. A Machine Learning Approach for Prediction of Students' Admissibility for Post-Secondary Education using Artificial Neural Network. *Int. J. Comput. Appl.* **2022**, *184*, 44–49. [CrossRef]
39. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. [CrossRef]

40. Yousafzai, B.K.; Khan, S.A.; Rahman, T.; Khan, I.; Ullah, I.; Ur Rehman, A.; Baz, M.; Hamam, H.; Cheikhrouhou, O. Student-performulator: Student academic performance using hybrid deep neural network. *Sustainability* **2021**, *13*, 9775. [CrossRef]
41. Ahmad, I.; Ullah, I.; Khan, W.U.; Ur Rehman, A.; Adrees, M.S.; Saleem, M.Q.; Cheikhrouhou, O.; Hamam, H.; Shafiq, M. Efficient algorithms for E-healthcare to solve multiobject fuse detection problem. *J. Healthc. Eng.* **2021**, *2021*, 9500304. [CrossRef]
42. Tufail, A.B.; Ullah, K.; Khan, R.A.; Shakir, M.; Khan, M.A.; Ullah, I.; Ma, Y.K.; Ali, M. On Improved 3D-CNN-Based Binary and Multiclass Classification of Alzheimer's Disease Using Neuroimaging Modalities and Data Augmentation Methods. *J. Healthc. Eng.* **2022**, *2022*, 1302170. [CrossRef]
43. Amjad, S.; Younas, M.; Anwar, M.; Shaheen, Q.; Shiraz, M.; Gani, A. Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of High School Students. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 9299115. [CrossRef]
44. Raschka, S. *Python Machine Learning*; Packt Publishing Ltd.: Birmingham, UK, 2015.
45. Saa, A.A. Educational data mining & students' performance prediction. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*.
46. Macintyre, J. Engineering Applications of Neural Networks. In Proceedings of the 20th International Conference, EANN 2019, Xersonisos, Greece, 24–26 May 2019; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1000.
47. Marquez-Vera, C.; Romero, C.; Ventura, S. Predicting school failure using data mining. In *Educational Data Mining*; CiteSeer: University Park, PA, USA, 2011.
48. Al-Obeidat, F.; Tubaishat, A.; Dillon, A.; Shah, B. Analyzing students' performance using multi-criteria classification. *Clust. Comput.* **2018**, *21*, 623–632. [CrossRef]
49. Tahir, M.E.; Abbas, N.; Sayat, M.F.; Nasir, M. Statistical analysis of crowd behaviour in catastrophic situation. *Mehran Univ. Res. J. Eng. Technol.* **2022**, *41*, 104–112. [CrossRef]
50. Livieris, I.E.; Drakopoulou, K.; Tampakas, V.T.; Mikropoulos, T.A.; Pintelas, P. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *J. Educ. Comput. Res.* **2019**, *57*, 448–470. [CrossRef]
51. Nair, P.B.; Choudhury, A.; Keane, A.J. Some greedy learning algorithms for sparse regression and classification with mercer kernels. *J. Mach. Learn. Res.* **2002**, *3*, 781–801.
52. Trawiński, B.; Smętek, M.; Telec, Z.; Lasota, T. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *Int. J. Appl. Math. Comput. Sci.* **2012**, *22*, 867–881. [CrossRef]
53. Goetz, J.; Brenning, A.; Petschko, H.; Leopold, P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **2015**, *81*, 1–11. [CrossRef]
54. Khalil, H.; Rahman, S.U.; Ullah, I.; Khan, I.; Alghadhban, A.J.; Al-Adhaileh, M.H.; Ali, G.; ElAffendi, M. A UAV-Swarm-Communication Model Using a Machine-Learning Approach for Search-and-Rescue Applications. *Drones* **2022**, *6*, 372. [CrossRef]
55. Haq, I.; Mazhar, T.; Malik, M.A.; Kamal, M.M.; Ullah, I.; Kim, T.; Hamdi, M.; Hamam, H. Lung Nodules Localization and Report Analysis from Computerized Tomography (CT) Scan Using a Novel Machine Learning Approach. *Appl. Sci.* **2022**, *12*, 12614. [CrossRef]
56. Tufail, A.B.; Ullah, I.; Rehman, A.U.; Khan, R.A.; Khan, M.A.; Ma, Y.K.; Hussain Khokhar, N.; Sadiq, M.T.; Khan, R.; Shafiq, M.; et al. On Disharmony in Batch Normalization and Dropout Methods for Early Categorization of Alzheimer's Disease. *Sustainability* **2022**, *14*, 14695. [CrossRef]