Educational Administration: Theory and Practice
2024, 30(6), 1755-1767
ISSN: 2148-2403
https://kuey.net/

Research Article

Enhancing Academic Success: A Novel Approach to Predict Learning Performance with an Advanced Blended Learning Performance Predictor

Veena M[1*], Manjunath Kotari[2]

[1*]Research Scholar, Alva's Institute of Engineering and Technology, Mangalore, Karnataka, India
[2]Department of Computer Science and Engineering., Alva's Institute of Engineering and Technology, Mangalore, Karnataka, India

Citation:, Veena M ,(2024), Enhancing Academic Success: A Novel Approach to Predict Learning Performance with an Advanced Blended Learning Performance Predictor , *Educational Administration: Theory and Practice*, *30*(6), 1755-1767, Doi:. 10.53555/kuey.v30i6.5583

ARTICLE INFO **ABSTRACT**

The concept of educational data mining focuses on the development of methods for investigating and analyzing the enormous amounts of information that are produced by the academic environment. It is in a position to capitalize on the large number of studies produced by the data mining industry and connect that knowledge to academic issues pertaining to teaching, reasoning, and evaluation. Recently, this sector has become predominantly effective at solving various difficulties with scholastic analytics. This achievement can be attributed to the vast processing capacity and information retrieval methods that are used. The primary goal of educational establishments at the university level is to furnish the students who attend those institutes with a curriculum of sufficient calibre. The pursuit of information that can accurately anticipate students' performance is one strategy for elevating the standard of the educational structure at the postsecondary level to the greatest possible standard. In this study, a superior Blended Learning Performance Predictor Toolkit (BLPPT) is presented that can be used to predict how students would perform on their semester examinations utilizing the data collected through surveys. The BLPPT model reaches an MAE score of 2.94×10^{11} and an MSE value of 9.18×10^{23}.

Keywords: *Regression, Performance prediction, Semester result forecast, Voting based regression*

INTRODUCTION

The internet's capacity to store and provide access to data and statistics is growing exponentially, making conventional information management methods obsolete. Further, the right management of globally accessible data can unlock previously unimaginable opportunities in commerce, science, and the classroom [1]. The challenge, though, is figuring out how to make the most of this information without becoming overwhelmed by the volume of numbers itself. The vast field of analytics, also known as Big data, holds the key. Several businesses and academic institutions already use machine learning as part of their operations and initiatives, but this discipline is by no means limited to those with deep pockets [2]. An emerging field, Academic Data Analysis seeks to help explain both learners and the academic environments in which they operate by establishing tools for understanding the distinctive forms of data generated in academic contexts [3]. The goal of data mining is to discover intriguing, latent, largely undiscovered, and potentially beneficial patterns or information in large datasets [4]. Since academic databases typically contain a great deal of information, several data mining techniques have been created and applied to extract the information needed and reveal any hidden relationships [5]. Regression, outlier detection, and prediction are just a few of the many common data mining tasks used in the classroom. Data mining has many potential applications in the field of education, including the prediction of dropout rates, the identification of correlations between pupils' standardized test scores and their subsequent achievement in college, the forecasting of students' grades, the identification of closely relevant threads in curriculum content, the development of new knowledge about students' academic credentials, the categorization of students' training course results by cognitive strategies, and the investment of similarities and differences between data sets [6-8].

When it comes to making accurate predictions, the forecasters in use are crucial. So, before attempting to forecast student outcomes, it is necessary to identify the components affecting the teaching & learning

procedure [9]. However, many institutes today are most concerned about the declining success rate of pupils and the increasing frequency of withdrawals. Institutions could benefit from looking back at students' achievement data.

In modern education, achieving academic success for students is ever more instrumental, propelling the emergence of novel methodologies for predicting and boosting student performance. Nevertheless, most current prediction models tend to overlook the sophisticated dynamical interplay among numerous factors within the learning environment that can be taken to another level with the advent of blended learning. This gap accentuates the necessity for a new unified approach to predict performance appropriately, unlocking the full potential of blended learning and mitigating its unfavorable sides. This paper is designed to fill precisely this gap with the help of providing blended learning performance prediction toolkit of next-generation, ensuring a comprehensive approach to performance prediction reliable enough to yield optimal educational results

The contributions made in this paper are as follows:

- Prepared quetionaries to collect the data from final year engineering college students and collected data.
- Conducted exploratory data analysis (EDA) to understand the insights of the students performance data.
- Explored the categorical data and numerical data through EDA. Handled inconsistency in the data, missing data, analysed the features effectively. Performed feature elimination by filtering technique.
- Converted categorical columns to numerical and merged with numerical columns. Performed normalization of all the columns in the features.
- Proposed and evaluated a novel blended learning performance predictor approach for blending the multiple regression models and applied voting regressor mechanism to the blended model to create a superior model for student performance prediction.

This paper presents a prediction model of student performance in semester seven based on developed models. There are several different models utilized for this prediction. After comparing the model performances with the experimental results, we found that our novel Blended Learning Performance Predictor Toolkit (BLPPT) is superior in predicting an upcoming semester result for the students.

RELATED WORKS

Student modelling is one of the goals that can be accomplished in the educational data mining industry. When it comes to student modelling, there are two primary tasks that are distinguished from one another: structure discovery and prediction. Especially in the case of prediction, we discriminate between two applications: forecasting the unwanted actions of students and predicting the features of students such as their learning methods and performances [10-12]. Classification and regression, have been utilized to construct a model that can correctly forecast students' performance [13]. When the resultant variables are categorical, one can use the classification method, but the regression technique is used when the resultant features are of numerical type [14]. The performance of students can be predicted using a number of different algorithms that fall under the regression technique [15]. Some examples of these algorithms include Linear Regression, Decision Tree Regression, and Random Forest Regression.

Regression is widely used in the field of statistics and has recently gained popularity in the field of data mining. This approach is often utilized in students' performance prediction problems. Regression analysis is used when the target factors considered are numbers or continuous values. It is the simplest approach to examining the causal link between variables. In regression-based prediction, the connection is defined by correlation analysis measures using the explanatory factors and one or more target columns [11]. The most important byproduct of this situation is the equation. The significance of individual predictors can be assessed, and the correlation between variables can be understood in a prediction situation using this method [16]. Sravani *et al.* explored the significant impact that machine learning (ML) applications have on the courses and how they might be used to enhance the educational experience at the university level. Sites like Course Era, Udemy, etc., sprang to prominence as a result of a rise in students' need for digital and online education. They used an ML approach called linear regression to make predictions about the student's academic achievement [17]. Tran et al. suggested combining the two approaches to improve forecast accuracy. The encouraging findings with an RMSE of 1.668 can give students quick feedback regarding their expected performance on coursework, saving time for both students and teachers as they figure out which courses are most suited to each individual's skill level [18]. Stearns et al. employed data mining techniques to analyze 8 million student records from Brazil's National High School Test to determine whether or not socio-economic background was a reliable predictor of exam achievement in mathematics. Gradient boosting was found to be effective in predicting students' performance [19]. Blended learning, which combines traditional classroom instruction with an online component, was investigated by Xu et al., who based their lessons on the flipped classroom and the concept of the individual online course. They examined how different online learning habits affected students' outcomes [20].

METHODOLOGY AND DATASET MANAGEMENT

A. Problem Statement

Prediction is the process of making an educated guess about a dependent factor's future value from the other variables' existing values. Marks and other quantitative data can serve as outcome variables in academic sectors to represent students' performance. In this particular experimental study, we aim to predict the students' sixth-semester results using the collected data. This is a regression task as the target column is numerical. There are two types of linear regression analysis: the former uses a single independent variable to determine the connection between the dependent and the explanatory variable, while the latter uses several independent variables to do the same. In this experimental study, we opted for the multiple linear regression problem.

Usually, we take y as the target variable and X_i as the various dependent variables for this type of problem. Equation 1 represents the mathematical expression for multiple linear regression.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \varepsilon \qquad (1)$$

Here, β_0 is the y-intercept value, $\beta_i X_i$ is the regression coefficient of the i^{th} factor, and ε represents the regression model error.

B. Methodology

Data mining is the practice of discovering meaningful patterns in large datasets. The data are collected through google forms and surveys in this procedure. Then, these data are preprocessed, and correlated features are either processed or generated according to the requirement for the prediction. After this step, the dataset is split into two sections for training and subsequent evaluation of the predictive model. Then, six distinct regression models are applied to the training and evaluation datasets to find the best-fitting model. Figure 1 presents the simple workflow chart for the experimental study for sixth-semester results prediction.

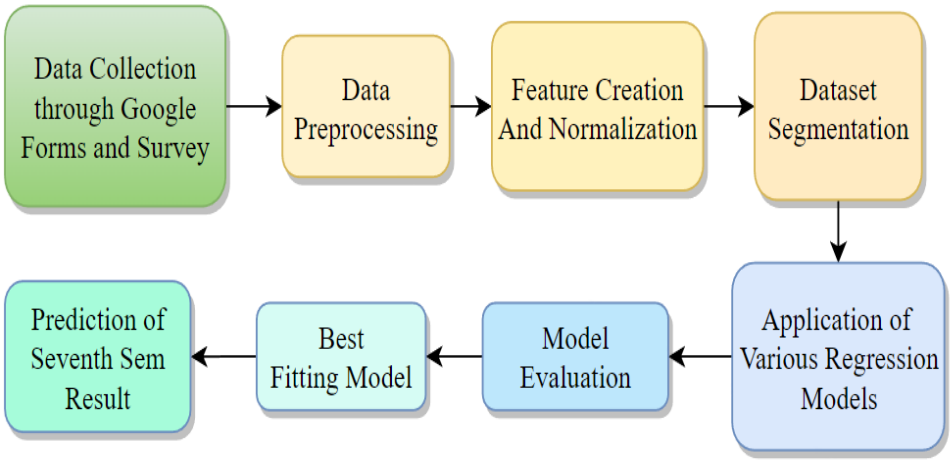

Fig. 1. The General Workflow of the Methodology.

We used Numpy and Pandas libraries to work on mathematical operations and data manipulation stages. Matplotlib's pyplot was used for the graph generation. We utilized sci-kit-learn for model designing in this study.

C. Data Description

Data is collected via google Forms and surveys. Different parameters are considered for students' performance. Various parameters are considered like living habits, time to travel to college, relationships, family relationships, smoking etc. The dataset contains 439 sets of data and 50 feature columns. Thirty-three feature columns contain categorical features, and 17 of them are numerical ones. The 50th feature column belongs to the 'timestamp'. This particular data column is in 'Datetime' format. Tables 1 and 2 present various categorical and numerical feature columns in the dataset used here in this study.

TABLE 1. Categorical feature columns

Sl. No.	Feature name	Details
1	Email Address	Email address of the students.
2	Name	Official Name of the students.
3	USN	Universal Student Number
4	Current Sem of the student	Semester Currently studying in
5	Gender	F(Female), M (Male)

6	KCET / DCET Ranking	KCET Ranking (for PU Students) and DCET Ranking (for Diploma students)
7	Type of Engineering Seat	Seat type for the Engineering course
8	Attendance	How regular the student is in attendance.
9	Failures	Number of past class failures
10	Other technical/Certification Courses	If taken other technical/Certification Courses
11	Internship	Has the student done any internships?
12	Placement offer in hand (7th Sem.)	Number of placements offers in hand [if current 7th Sem students]
13	Placement offer in hand (Pass out)	Number of placements offers You had [if current Passed out students]
14	Area	Belonging to Urban/Rural area
15	State	State the student belongs to
16	Stay in/with	Accommodation type
17	Travelling Time to College	Time taken to go to college
18	Annual Income	Annual Income of the student's family/ income source
19	Family Size	Family size is 'LE3' - less or equal to 3 or 'GT3' - greater than 3
20	Mother Education	Highest education level of the mother
21	Father Education	Highest education level of the father
22	Mother's job	Current occupation of the mother
23	Father's job	Current occupation of the father
24	Study time	Weekly study time
25	Educational Support	Type of educational support student receives
26	Extra-Curricular Activities	Participated in Extra Curricular Activities or not.
27	Interested in Higher Education	If or not
28	Internet access at home/PG/hostel	Internet access at the place of accommodation
29	With romantic relationship (Optional)	If the student is in any relationship or has any partner
30	Leisure Time(weekly)	Schedule free time
31	Going Out with friends	Hangout time
32	Reason to join engineering	The reasoning for which the student joined the engineering
33	Immediate Future plan	Future plan, immediate after course completion

TABLE 2. Numerical feature columns

Sl. No.	Feature names	Sl. No.	Feature names
1	Year of degree completion	10	IV Sem SGPA/ Percentage
2	SSLC Marks/Grade 10 Marks (percentage)	11	V Sem SGPA/ Percentage
3	PU II Marks/Grade 12 Marks (Percentage)	12	VI Sem SGPA/ Percentage
4	JEE Ranking	13	VII Sem SGPA/ Percentage
5	Overall CGPA/Percentage	14	VIII Sem SGPA/ Percentage
6	Year of degree completion	15	Weekend Alcohol Consumption
7	I Sem SGPA/ Percentage	16	Quality of Family Relationship
8	II Sem SGPA/ Percentage	17	Health Conditions
9	III Sem SGPA/ Percentage		

D. *Exploratory Data Analysis*

Exploratory Data Analysis, or EDA for short, is a data analysis technique that focuses on using visualization aids. With the assistance of statistical summaries and graphical visualizations, it can be utilized to identify patterns and trends and validate assumptions.

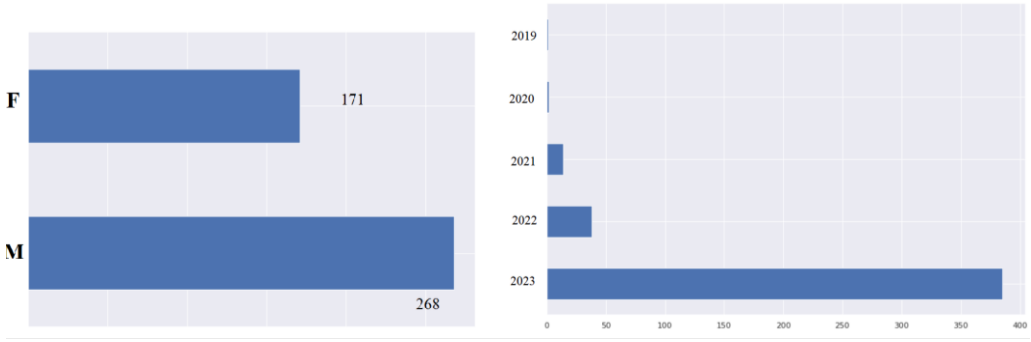

Fig. 2. (a) Gender Column Distinct Value counts, (b) Year of the Degree Completion who responded to the survey

From EDA, we came to know about the details present in each data column. Figure 2(a) shows that there are two categories in the 'Gender' column, consisting of 171 females and 268 males. The responding students of this survey can be divided into five categories based on their 'year of degree completion' (from 2019 to 2023), which is presented in figure 2(b).

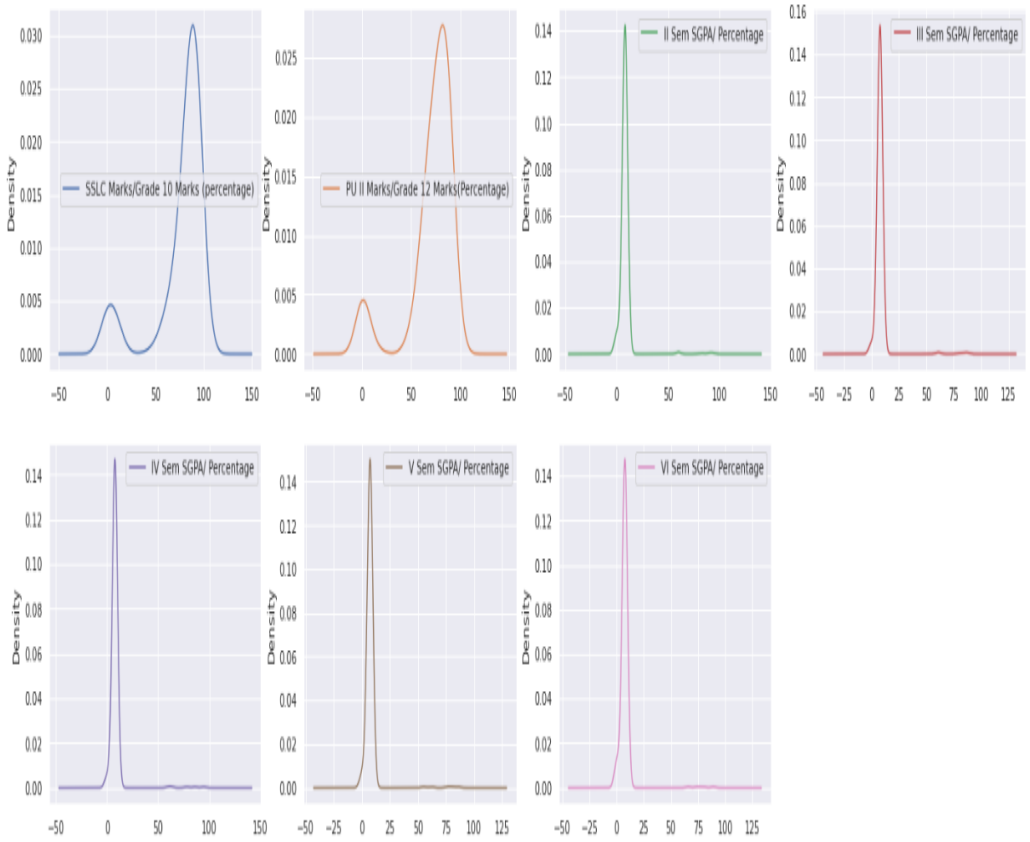

Fig. 3. Density Plot for Grade 10, Grade 12, and Previous Six Semester Results.

Figure 3 presents the density plots for the feature columns for 'SSLC Marks/Grade 10 Marks (percentage)', 'PU II Marks/Grade 12 Marks (Percentage)', and the previous six-semester results in engineering for each student. Grade 10 and grade 12 result data columns are skewed slightly towards the left, indicating that they are negatively skewed. For the case of the semester results, these columns are all right-skewed data columns. There are seven distinct categories in the 'Type of Engineering Seat' column, consisting of the following: 'SC/ST/OBC/CAT-1', 'Ex-Central Armed Police Force', 'Defense', 'SNQ', 'Management', and 'GM'.

The column containing data related to attendance is described in figure 3. The 'number of past class failures' data column is divided into four categories: 'none', 'Greater than three subjects', 'Less than three subjects', and 'None'. The 'none' category is distributed in two categories due to spelling mistakes. The feature column containing data related to the 'Other technical/Certification courses' comprises 43 different types of courses. Due to spelling mistakes, the 'Internship' related column categories are distributed into four classes instead of two. The 'Area' feature is distributed into to classes 'Urban' and 'Rural'. The 'Stay in/with' feature contains data

about students' accommodation arrangements depending upon if they live in a hostel, with a guardian, in a rented

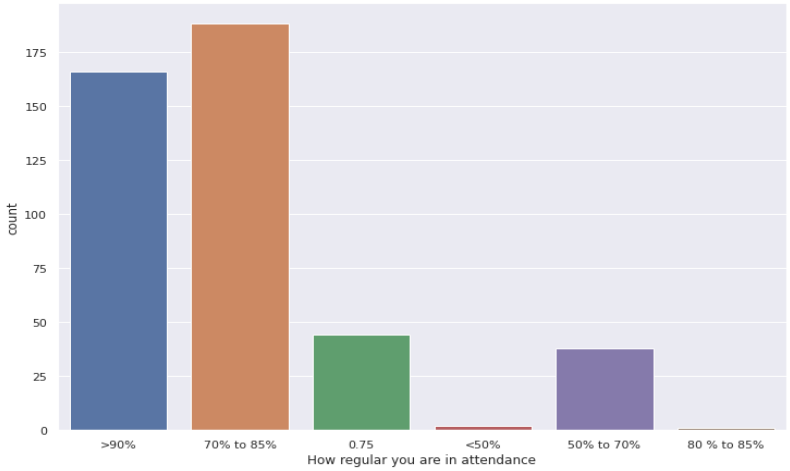

Fig. 4. The Categories in the Attendance Column**.**

The 'Travelling Time to College' feature is distributed into six different categories. The categories are: '<15 min', '15-30 mins', '31 min to 1 hour', '1 hour', and '>1 hour'. Similarly, there are seven distinct classes for the 'Annual Income' feature. Most of the subjects belong to two types among these seven: ' <2.5 lakhs' and '2.5 to 5 Lakhs'. The next feature is the family size of the students, which is divided into three classes: ' less or equal to 3' and 'Greater than 3'. The features, namely 'Father Education' and 'Mother Education', are distributed in 7 and 6 classes, respectively. Likewise, 'Mother's job' and 'Father's Job' is divided into 7 and 5 distinct sections, respectively. The attribute named 'Educational Support' is distributed into three categories instead of two because of a spelling mistake.

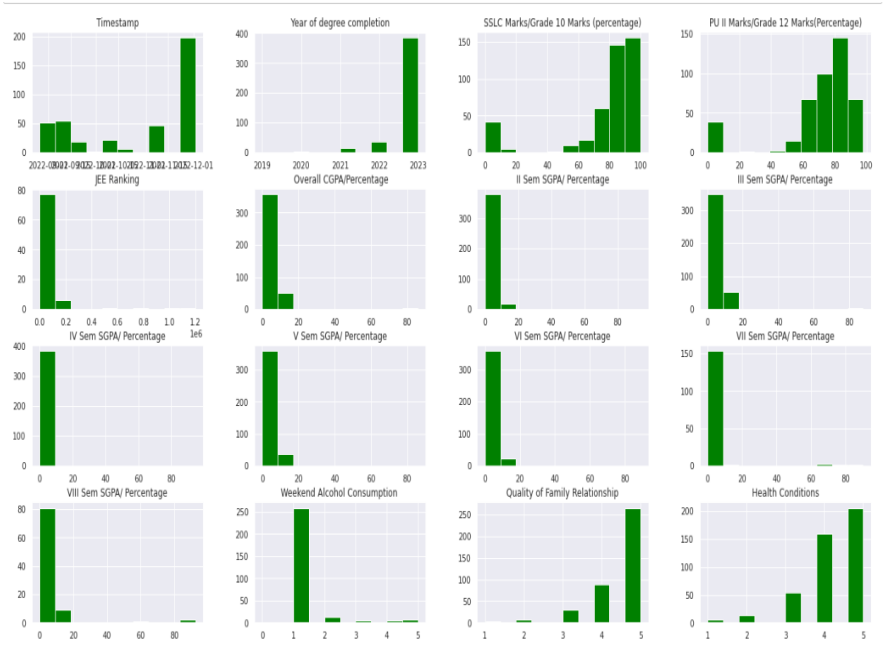

Fig. 5. Feature Column Histogram

Fig. 5 represents the graphical distribution of the provided dataset. It shows the underlying frequency distribution or the probability distribution of the feature columns present in the dataset. The Fig.6. statistical analysis of the "Travelling Time to College" data provides valuable information about students' commuting. The vast majority of 56% of students constitute the group of 147 individuals who reported a commuting time of less than 15 minutes. It indicates that a considerable proportion of students live in the proximity of the college. Another 37% of respondents, accounting for 94 people, take 15 to 30 minutes to travel, which is also a large portion of those who live at a relatively short distance to the college. However, the largest group of students, with 196 people, reports a traveling time of 30 minutes or more.

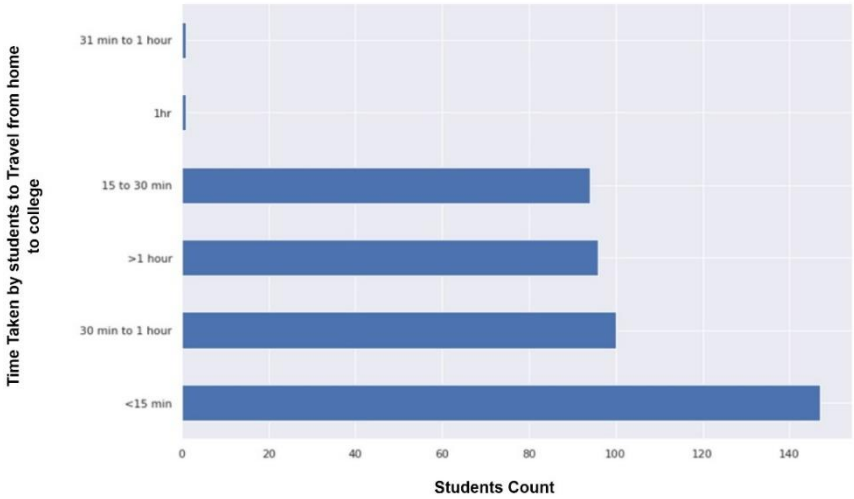

Fig. 6. Statistical analysis on Time taken for students in Travelling

The group includes 51%, or 100 individuals, who reported 30 minutes to 1 hour and 49% or 96 people who reported more than 1 hour. Therefore, the geographic location of the college and students' places of living would have an impact on the transport infrastructure as well as the accessibility and possible distance-related influence on the students' well-being and academic performance . Furthermore, two students reported an exact traveling time of 1 hour and 31 minutes to 1 hour, indicating outliers. In addition, the analysis shown in fig. 7, on the "Father Education" data shows that there are various educational backgrounds represented in the fathers of the students who participated in the survey. The above Fig. 6, shows that 244 fathers have reported Higher Education, which can be considered a relatively large but not the dominant figure, while another 103 fathers have secondary education. Further, one can see that almost 50 fathers have Higher Primary or Primary Education , whereas the 7 fathers did not receive even the basic 4 th standard level of education. Although 17 fathers report knowing nothing about the education, it is curious that there are also fathers of the students who represent the entire spectrum of education within these categories. Furthermore, one could notice that there is a discrepancy between the fathers reporting higher education or ITI modalities, which reflects the importance of standardization in the survey questions and processes. ServiceExceptionally, the analysis provides an important background for considering the factors concerning the fathers' educational influence on the academic intentions and outcomes, as will be discussed further.

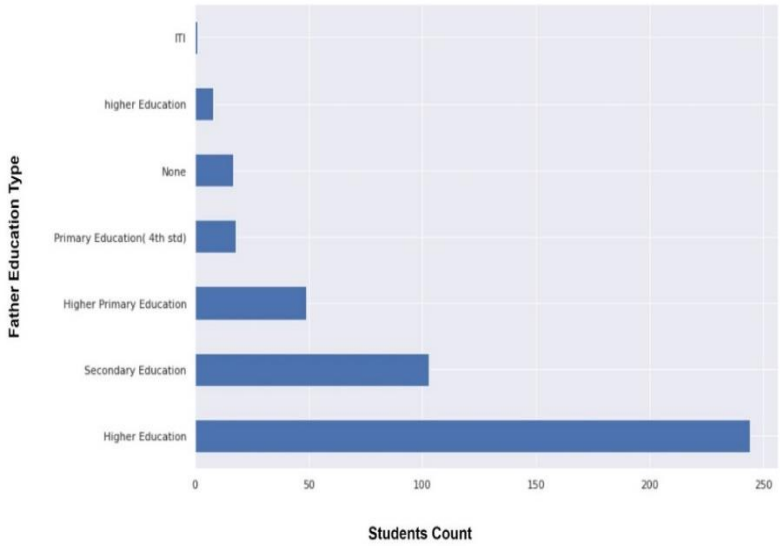

Fig. 7. Statistical analysis on students father education type

On the observation of the "Mother Education" in Fig 8, the data allowed to identify the levels of the completed educational programs by student mothers. The obtained results indicated a wide range of educational backgrounds across the students' mother demographic. The most common category of education is "Higher Education", with 218 responses that describe mothers who have received the highest education. In addition, "Secondary Education" was also quite prevalent , with this option present in 102 responses. Importantly, a large proportion of mothers has "Higher Primary Education" , totaling 69 daughters, while 28 students' mothers completed "Primary Education" at the 4 th standard level. Futhermore, there was a small proportion of mothers with no education at al , totaling 15 students. Finally, the presence of such response variations as "higher Education" , totaling 8 students, indicates the importance of uniformity when collecting and analyzing data.

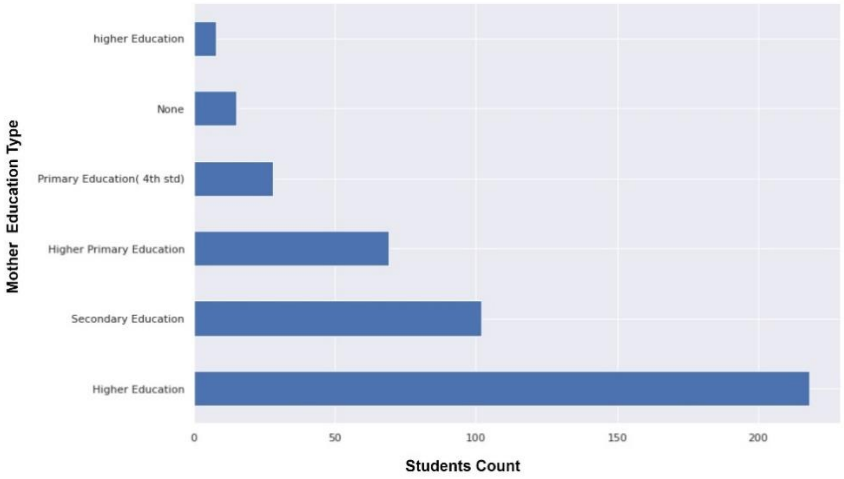

Fig. 8. Statistical analysis on students mother education type

The analysis of the "Weekly Study Time" is shown in Fig.9, the data leads to several conclusions on the surveyed students' studying habits and behaviors.
Namely, the study time's distribution patterns indicate a differentiated mode of academic preparation among the respondents.
The first inclusive group included 180 students who spent 2 to 5 hours per week studying, and, thus, the largest proportion of students invested a moderate amount of time in the study.
The second group included 90 students who spent from 5 up to 10 hours per week studying, and, correspondingly, a considerable proportion of students focused on academic pursuits invest a larger amount of time. Finally, 55 of the respondents shared that they spend more than 10 hours per week studying, and this group indicates the subgroup of learners who invest a substantive amount of time in studying.
However, the data also illustrate that a significant proportion of students spent less than two hours per week studying, which was 110 students, while 10 reported from 3 up to 5 hours . Therefore, the distribution presence study patterns prove multiple and, as a result, the need to explore individual study habits and behaviors.

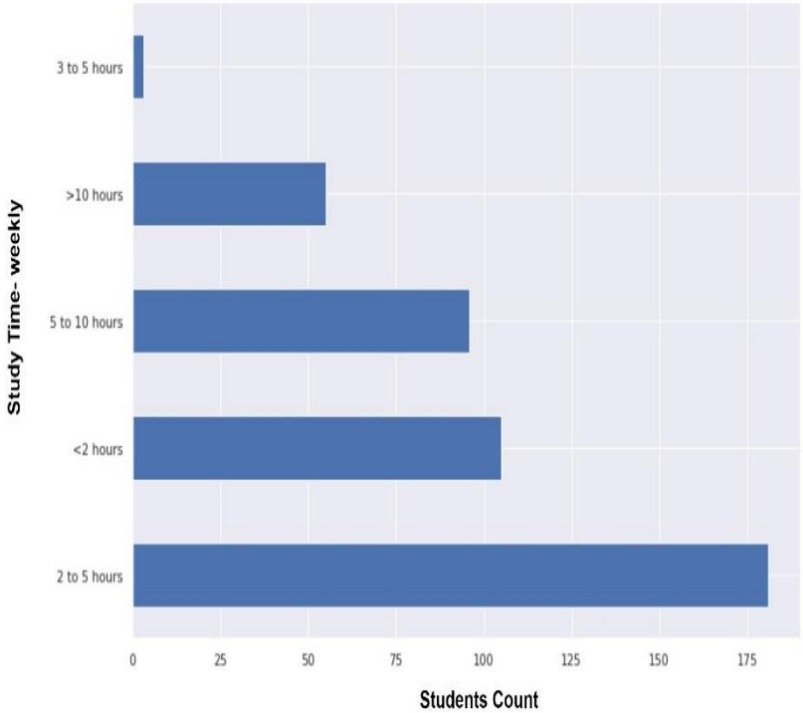

Fig. 9. Statistical analysis on study time of students.

The degree of association that exists between two numerical variables can be graphically represented using a plotting technique known as a correlation heatmap. The heatmap for feature correlation for this particular dataset is presented in Fig. 10.

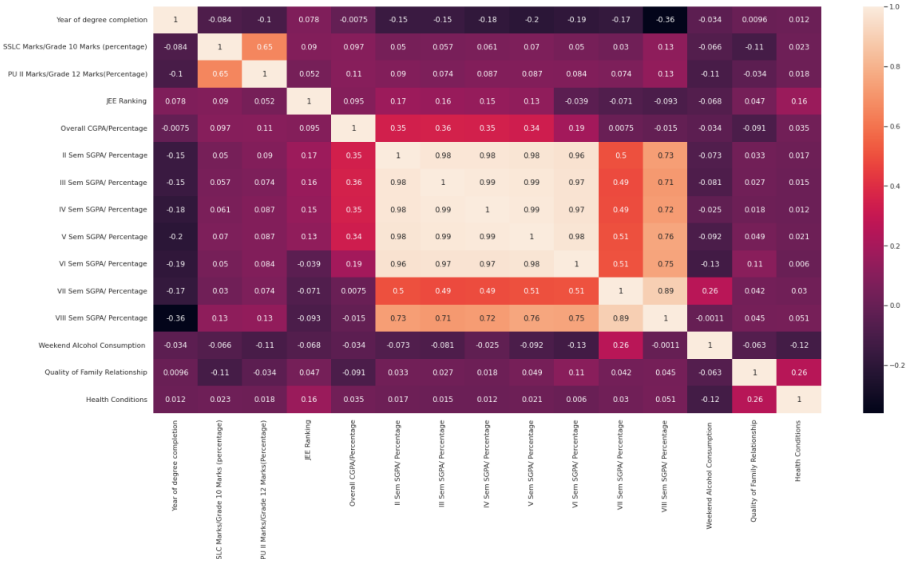

Fig 10. Heatmap for Feature Correlation

E. Data Preprocessing

We found many inconsistencies in various feature columns due to spelling mistakes. We changed and reconstructed the data details to manage this dataset inconsistency. The changes made in the dataset are described below in table 3.

TABLE 3. Data Inconsistency Handling

Feature Column	Inconsistency	Changes/Replaces Made
Type of Engineering Seat	'SC/ST/OBC/CAT-2'	"SC/ST/OBC/CAT-1"
Attendance	'80 % to 85%', 0.75	"70% to 85%"
Failures	'none'	"None"
Internship	'YES', 'no'	"Yes", "No"
Area	'urban'	'Urban'
Stay in/with	'hostel'	'Hostel'
Travelling Time to College	'31 min to 1 hour', '1hr'	"30 min to 1 hour", "30 min to 1 hour"
Annual Income	'<2.5 lakhs', '2.5 to 5 lakhs', '6 Lakhs to 10 Lakhs'	'<2.5 Lakhs', '2.5 to 5 Lakhs', '5 Lakhs to 10 Lakhs.'
Family size	'LE3', 'GT3'	'GT4', 'GT3'
Father Education	'Higher Education', 'ITI'	'Higher Education', 'Secondary Education'
Mother Education	'Higher Education	'Higher Education'
Mother's job	'House Wife'	'Home Maker/House Wife'
Study time	'yes', 'no'	'Yes', 'No'
Interested in Higher Education	'yes'	'Yes'
Internet access	'yes'	'Yes'
With romantic relationship (Optional)	'yes', 'no'	'Yes', 'No'
Leisure Time (weekly)	'6hr to 10 hrs', '10 to 16 hrs', '>16 hrs'	'6 hrs to 10 hrs', '10 to 16hrs', '>16hrs'

There are a lot of missing data present in this dataset. The feature columns which are topping the list of the columns with nonexistent data issues are: 'VIII Sem SGPA/ Percentage', 'JEE Ranking', 'Email Address', 'VII Sem SGPA/ Percentage', 'Weekend Alcohol Consumption', 'Name', 'With romantic relationship' etc. The numerical columns with the missing data are filled with median and mode values to fill the empty spaces. The columns associated with result-based data are replaced with median values of the data columns in place of those empty cells. The categorical data empty spaces are substituted using the term value 'Not provided'.

F. Feature Creation and Elimination

Feature creation refers to the steps used to generate additional features from available information with the goal of developing an ML model. Since an ML model can only adapt from the input we provide, the process of constructing attributes that are essential to a problem is necessary. Due to this reason, we created a new feature for this dataset, the 'Average SGPA till 5th Semester'. Each student's mean value of the previous 5-semester results is this particular feature. After the data processing and the feature creation steps are done, the unnecessary feature columns are removed from the dataset. These feature columns are as follows: "Timestamp", 'Email Address', 'Name', 'USN', '/Semester Currently studying in', ' Overall CGPA/ Percentage', 'VII Sem SGPA/ Percentage', 'VIII Sem SGPA/ Percentage', 'Number of placements offers in hand [if current

7th Sem students]', 'Number of placements offers You had [if current Passed out students]'. There were 41 data columns left after all the data preprocessing and feature column elimination.

G. Handling Categorical Data

There are 28 categorical-type feature columns left after the data preprocessing and non-essential feature column removal steps are done. These categorical data columns should be handled in such a way so that the dataset can be employed in the designed predictive models. We used the 'get dummies()' function from Pandas to manipulate these types of data columns. This method transforms discrete categories into continuous indicators. Dummy variables are used to encode classes in this manner.

H. Feature Normalization

Data normalization refers to the method of rearranging information already present in a dataset to make it more accessible for subsequent inquiries and analyses. It is a method used to provide trustworthy information. This involves cleaning up the data by getting rid of duplicates and organizing the information so that it looks the same in every entry. We utilized 'MinMaxScaler' to scale each feature to a given range. Here the range is [0,1].

I. Dataset Handling

After all these processes are done, the divided datasets are concatenated. Then, the finalized processed dataset is split into two subsets in a 3:1 ratio to create a training and a validation dataset for model training and evaluation processes. The training subset contains 75% of the whole data.

I.MODEL DESIGN

We have employed six different regression models to the dataset to predict sixth-semester results. The models were then evaluated to find the best-fitting one for this case study. The models are described below.

A. Linear Regression

Values of one variable can be predicted from knowledge of the values of other variables using linear regression analysis. The term 'dependent variable' refers to the sixth-semester result. Independent variables are those that can be used to make a prediction about a target variable. Here, we have 418 independent variables for predicting the target variable. When comparing projected and observed output values, linear regression seeks to minimize the distance between the two using a straight line or another smooth surface.

B. Decision Tree Regression

With a decision tree, we may create tree-like regression models. It gradually develops a decision tree in tandem, subletting a dataset into ever-tinier pieces. Consequently, we have a tree structure with leaf and decision nodes. A decision node may have two or more "sections" or "options" for the feature being evaluated. Each leaf node represents a choice on the quantitative goal. The best predictor is described as the top-level decision node in a tree, termed the root.

C. Random Forest Regression

Using a combination of numerous decision trees and a method termed bagging, Random Forest is an aggregation methodology that can solve regression problems. The objective here is to use a combination of decision trees instead of just one to arrive at a conclusion. Several decision trees serve as the foundational ML algorithms in Random Forest. Row and attribute sampling are both arbitrarily performed from the dataset to create test datasets for each model. The Bootstrap framework is responsible for this.

D. XGBoost Regression

A subset of ensemble ML methods, gradient boosting can be applied to both classification and regression predictive modelling issues. Decision-tree approaches are the building blocks of an ensemble. The prediction errors of earlier models are reduced by fitting additional trees to the ensemble. Boosting is an ensemble ML strategy that uses many models to improve performance. Short for 'Extreme Gradient Boosting,' XGBoost is a powerful, freely available program that implements the gradient boosting technique. The method is directly applicable to regression predictive modelling.

E. SVM Regression

The flexibility provided by support vector regression (SVR) enables us to determine the maximum amount of error that can be tolerated in our models, and it can also locate a correct line to fit the information. Instead, then focusing on minimizing the squared error, the role of SVR is to reduce the coefficients. The error is rather managed in the constraints, wherein we specify the absolute error to be lower or equivalent to a defined boundary, which is referred to as the maximum deviation.

F. Blended Education Performance Predictor Toolkit

We employed 'VotingRegressor' for designing the BLPPT to predict sixth-semester results. A voting ensemble is a type of ML ensemble technique that, rather than relying on one framework, makes use of a variety of different approaches in order to improve the overall efficiency of the system. The outcomes of a number of different techniques can be combined using this strategy and can be used to solve regression problems. The 'VotingRegressor' was designed to combine regressors for ML that are clearly distinguishable from one another and return the mean of the projected values. A regressor of this type can be helpful for a group of models with comparable levels of performance by helping to compensate for the shortcomings of the models individually. The VotingRegressor algorithm is at the core of the student performance prediction. It is based on ensemble learning – a popular approach in artificial intelligence that combines the predictions made by several base regressor models. These models include linear regression , decision tree regressor , random forest , XGBoost , and support vector regression . The essential idea behind ensemble learning is that by aggregating the predictions made by the diverse and numerous models, each model with its unique individual bias and source of uncertainty, the combined voting regressor model can "pick the brains" of the whole ensemble. This allows for minimizing prediction accuracy errors while also shedding light on the complicated functional relationships between various indicators of student performance in the age of blended learning. Therefore, the VotingRegressor algorithm is central to the proposed toolkit, promoting further advancement in the field of academic success prediction and enabling more informed decision-making among educational stakeholders. The propsed blended student performance predictor is shown below:

Algorithm: Blended Student Performance Predictor

```
1   Input ← raw student data collected from questioners
2   Analyse the data through exploratory data analysis
3   Find the corelation of the input features
4   Handle the data inconsistency
5   Handle the missing data using mean imputation method
6   Convert the multiple categorical data into numerical data by using get dummies method
7   Perform feature scaling using min max scaler method
8   Merge the converted multiple categorical data and numerical data
9   Split the data into train and test in 80:20 to provide the input for training
10  Create a base model such as linear regressor, decision tree regressor, random forest
    regressor, XGBoost regressor and support vector regressor
11      Blend base models
        Create Voting Regressor ← [('lr', linear_standard_model),
                                   ('dt', dt_standard_model),
                                   ('rf', rf_standard_model),
                                   ('xgb', xgb_standard_model),
                                   ('svr', supportvector_poly_model)]
12      Train the blended model using training data
13      Save the blended model for predicting the new student performance
14  Predict the student performance using testing data
15  Evaluate the prediction performance using MAE, MSE and r2
16  Test the prediction of the blended model by providing the new student data as input
    and save the predicted result in .CSV file.
```

II. Experimental Results & Discussion

In this article, we applied six different regression models to our dataset and analyzed the factors that were generated as a result. We have determined their overall level of precision in predicting via employing them on our evaluation dataset and analyzing the results. We decided on a voting-based regression to use while constructing the best model for additional analysis. When developing the final model, we had to take into account both the precision and three types of error values. These error values are: mean absolute error, mean squared error, and R2-score.

- Mean Absolute Error (MAE)
  The MAE is the mean of the absolute deviation that exists between the readings that were actually observed and those that were anticipated for the dataset. It takes the mean of all the residuals in the dataset into consideration. Equation 2 shows the mathematical representation of MAE, where y indicates the predicted value of the target feature and y_i is the average value of the target feature.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y\right| \qquad (2)$$

- Mean squared Error (MSE)
  The MSE is the mean of the squared variation between the actual data points and the readings that were predicted for those values. It determines how variable the residuals are by measuring their variance. Equation 3 presents the mathematical representation of the evaluation metric.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y)^2 \qquad (3)$$

- R2-Score
  The R2 score is an essential statistic that is utilized in the process of evaluating the effectiveness of a Regression ML model. It is referred to as the determination coefficient or the R-squared value. It does this by calculating the degree of variability in the projections that can be attributed to the dataset.

$$R2 - Score = 1 - \frac{\sum(y_i - y)^2}{\sum(y_i - \bar{y})^2} \qquad (4)$$

TABLE 3. Comparison table of model evaluation metrics of different predictive models

Sl. No.	Model	MAE	MSE	R2 Score
1	Linear Regression	1.47×10^{12}	2.29×10^{25}	4.9×10^{24}
2	Decision Tree Regression	0.759	2.043	0.563
3	Random Forest Regression	0.758	1.966	0.580
4	XGBoost Regression	0.765	1.944	0.584
5	SVM Regression	1.214	3.977	0.151
6	Blended Education Performance Predictor Toolkit	2.94×10^{11}	9.18×10^{23}	-1.96×10^{23}

Table 3 is created with the three different evaluation metrics values collected from the six distinct model outputs and their observed errors. Decision Tree Regression, Random Forest Regression, and XGBoost Regression models showed almost similar MAE, MSE, and R2-scores when implemented to this dataset. The SVM regression model showed an increased value of MAE (1.214) and an increased MSE score of 3.977. The R2-score is reduced to 0.151. The BLPPT model shows better performance than the other state-of-the-art regression model. This particular model is created to combine the individual models into a better one to get an effective output. It showed an MAE value of 2.94×10^{11} and an MSE score of 9.18×10^{23}. R2-score for this blended model reached a value of -1.96×10^{23}. The amount of data present in this dataset is not very much. It only comprises 439 sets of data with 50 feature columns. It means that the dataset is not a large one. This type of regression-based prediction problem requires huge data input to predict the performance of the students efficiently. Large data input can remove the bias issues in prediction. In this performance prediction study, we used the BLPPT model, which uses a voting-based regressor for prediction purposes. This type of regressor helps reduce the weakness present in the predictive model by combining different models. The outcome is the average prediction from different models. In this way, the blended architecture used here reaches a better model efficiency than the conventional machine learning approaches.

III. CONCLUSION

We have provided an analysis of the various contributing factors at play in today's engineering education systems. Course designers and professors benefit from a reliable estimation of student performance because it allows them to teach their students better. Students also can benefit from this type of prediction as they can gain insight into how they will likely perform in a given course, which can help them choose a curriculum that best suits their needs and interests. These forecasted outcomes also provide them with early feedback, allowing us to keep students from dropping out and enhance the course layout annually. The objective of this study was to propose a superior BLPPT that uses the information obtained from surveys to make predictions about how well students would perform on the exams that are administered at the end of a semester. Our aim was to forecast the sixth-semester results of the students for this particular experimental study. BLPPT showed better performance efficiency than the conventional machine learning predictive models for this regression-based prediction. In the future, we want to expand the input dataset by promoting the surveys more to collect more data related to this study. A larger dataset can lead us to a more trained model which can predict better.

REFERENCES

1. Fleming, ""Never Let a Good Crisis Go to Waste": How Consulting Firms Are Using COVID-19 as a Pretext to Transform Universities and Business School Education," *Academy of Management Learning Education,* no. ja, p. AMLE_20220217, 2022.
2. A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

3. K. Kitto and S. Knight, "Practical ethics for building learning analytics," *British Journal of Educational Technology,* vol. 50, no. 6, pp. 2855-2870, 2019.
4. M. Mach-Król and B. Hadasik, "On a certain research gap in big data mining for customer insights," *Applied Sciences,* vol. 11, no. 15, p. 6993, 2021.
5. C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery,* vol. 10, no. 3, p. e1355, 2020.
6. K. Nahar, B. I. Shova, T. Ria, H. B. Rashid, and A. S. Islam, "Mining educational data to predict students performance: A comparative study of data mining techniques," *Education Information Technologies,* vol. 26, no. 5, pp. 6051-6067, 2021.
7. A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques," *Technology, Knowledge Learning,* vol. 24, pp. 567-598, 2019.
8. A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon,* vol. 5, no. 2, p. e01250, 2019.
9. S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, and D. J. Murray, "Identifying key factors of student academic performance by subgroup discovery," *International Journal of Data Science Analytics,* vol. 7, pp. 227-245, 2019.
10. B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Education Information Technologies,* vol. 23, pp. 537-553, 2018.
11. O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A multiple linear regression-based approach to predict student performance," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 1-Advanced Intelligent Systems for Education and Intelligent Learning System*: Springer, 2020, pp. 9-23.
12. O. El Aissaoui, Y. E. M. El Alami, L. Oughdir, and Y. El Allioui, "Integrating web usage mining for an automatic learner profile detection: A learning styles-based approach," in *2018 international conference on intelligent systems and computer vision (ISCV)*, 2018, pp. 1-6: IEEE.
13. A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science,* vol. 72, pp. 414-422, 2015.
14. Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2*, 2019, pp. 737-758: Springer.
15. O. T. Omolewa, A. T. Oladele, A. A. Adeyinka, and O. R. Oluwaseun, "Prediction of student's academic performance using k-means clustering and multiple linear regressions," *Journal of Engineering Applied Sciences,* vol. 14, no. 22, pp. 8254-8260, 2019.
16. A. S. Hadi and S. Chatterjee, *Regression analysis by example*. John Wiley & Sons, 20B. Sravani and M. M. Bala, "Prediction of student performance using linear regression," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-5: IEEE.
17. T.-O. Tran, H.-T. Dang, V.-T. Dinh, and X.-H. Phan, "Performance prediction for students: A multi-strategy approach," *Cybernetics Information Technologies,* vol. 17, no. 2, pp. 164-182, 2017.
18. B. Stearns, F. M. Rangel, F. Rangel, F. F. de Faria, J. Oliveira, and A. A. d. S. Ramos, "Scholar Performance Prediction using Boosted Regression Trees Techniques," in *ESANN*, 2017.
19. Z. Xu, H. Yuan, and Q. Liu, "Student performance prediction based on blended learning," *IEEE Transactions on Education,* vol. 64, no. 1, pp. 66-73, 2020.