University of Tehran Press

Interdisciplinary Journal of Management Studies (IJMS)
Home Page: https://ijms.ut.ac.ir

Online ISSN: 2981-0795

Predicting Student Academic Performance: A Machine Learning Approach and Feature Analysis

Maryam Taher Mazandarani[1] | Zahra Zand[2] | Mohammad Hossein Khodabandelou[3] | Fatemeh Mozaffari[4] | Babak Sohrabi[5*]

1. Department of Information Technology Management, University of Tehran, Tehran, Iran. Email: maryam.taher@ut.ac.ir
2. Department of Information Technology Management, University of Tehran, Tehran, Iran. Email: zahra.zand@ut.ac.ir
3. Department of Industrial Management, Islamic Azad University, Tehran, Iran. Email: mohammadkhodabandelou.official@gmail.com
4. Department of Information Technology Management, University of Tehran, Tehran, Iran. Email: famozaffari@ut.ac.ir
5. Corresponding Author, Department of Information Technology Management, University of Tehran, Tehran, Iran. Email: bsohrabi@ut.ac.ir

ARTICLE INFO

Article type:
Research Article

Article History:
Received 23 July 2023
Revised 25 January 2025
Accepted 15 February 2025
Published Online 01 June 2025

Keywords:
Educational Data Mining (EDM),
Machine learning,
Academic performance,
Intrinsic motivation,
Regression algorithms,
Self-regulation.

ABSTRACT

Predicting student academic performance is a challenging task and, at the same time, has significant implications for educators and policymakers in the field of education. By utilizing machine learning techniques, this article seeks to explore the relationship between various features across six categories: demographic factors, personality traits, skills, favorite activities, relationships with others, out-of-school activities on one hand, and academic performance in terms of Grade Point Average, on the other. The data utilized in this study has been collected through several surveys conducted in one of the schools in Iran over multiple years and educational levels, which form the basis of the analysis. Using CRISP-DM methodology, a predictive model is developed based on CatBoost Regressor. A predictive model with an R-squared value of 0.87 is developed. Moreover, the analysis of feature importance reveals that positive personality traits such as "Interest in studying," "The quality of homework," "Contentment," "Self-regulation," and "Logical thinking and reasoning" skills are among the most predictive features affecting students' academic performance which is rooted in and supported by some of the well-known psychological theories such as Self-Determination Theory. The contribution of the current research includes the development of a highly accurate prediction model based on the machine learning approach to predict student academic performance in terms of their GPA and to extract the most important features that influence it. This study is unique in this field due to the incorporation of various features and data collection across different years and educational stages.

Cite this article: Taher Mazandarani, M.; Zand, Z.; Khodabandelou, M. H.; Mozaffari, F. & Sohrabi, B. (2025). Predicting Student Academic Performance: A Machine Learning Approach and Feature Analysis. *Interdisciplinary Journal of Management Studies* (*IJMS*), 18 (3), 425-440. http//doi.org/10.22059/ijms.2025.362506.676053

 Publisher: University of Tehran Press.
DOI: http//doi.org/10.22059/ijms.2025.362506.676053

1. Introduction

Education plays a vital role in the development of a nation. Therefore, predicting student academic performance has long been considered an important research area, and simultaneously, a challenging task. In fact, academic achievement is a multifaceted phenomenon influenced by various factors, such as demographics, personality traits, socioeconomic, and other environmental factors (Bilal et al., 2022). Some features, such as educational persistence, are theorized from motivational models, while other models of students' engagement with school consider several dimensions including student, family, peer, and school (Moreira et al., 2013). Understanding these factors and their impact on student performance can help in managing their effect (Bilal et al., 2022). From a practical point of view, identifying personality traits (Furnham & Mitchell, 1991) and behavioral skills, such as communication, self-efficacy, and collaboration (Siddiq et al., 2020), is key to predicting academic performance, providing vocational guidance, and achieving other academic goals.

Some studies have explored the role of students' skills and competencies as critical factors in their learning and achievement (Siddiq et al., 2020). However, since ability factors alone are insufficient to account for individual differences, researchers have also sought to identify non-cognitive features, including variables related to personality tendencies (O'Connor & Paunonen, 2007). Terms such as "personal characteristics," "non-cognitive skills," and "soft skills" are often utilized for a wide range of features such as resilience, teamwork skills, and honesty, which are regarded as important factors in several settings, including education and work (Lievens & Sackett, 2012).

Hence, in the educational literature, numerous studies have investigated the relationship between personality factors and academic performance, implying that personality traits help in explaining individual differences in academic achievements. Due to the obvious practical implication of personality in educational psychology, a considerable amount of research has been conducted on the relationship between personality and intelligence (Furnham & Mitchell, 1991). Among these characteristics, conscientiousness and neuroticism have been identified as strong predictors that influence academic performance both positively and negatively (De Feyter et al., 2012).

Demographic factors are another category recognized as important predictors of academic performance. Generally, these factors relate to parents' education level, profession, age, income, and religious affiliations (Farooq et al., 2011).

Moreover, academic achievement is influenced by how students spend their out-of-school hours, which educators recognize as an important opportunity to improve students' performance and engage them in social-oriented activities (Valentine et al., 2002).

Nowadays, educational institutes are generating a large volume of data related to students' features and performance. These data can be processed and analyzed to find insights that may help educators and policymakers make decisions regarding educational matters, especially about the students and their well-being (Bilal et al., 2022). As such, Educational Data Mining (EDM) has emerged as a discipline of data analytics in which machine learning algorithms are utilized for extracting underlying patterns (Ismail et al., 2021). The main goals of EDM include predicting students' learning outcomes, understanding their learning process, and providing a better understanding of education-related phenomena. Achieving these goals can help institutes and organizations understand and improve their educational processes (Czibula et al., 2022). Analyzing educational data also leads to developing an academic failure prevention plan by providing strategies for weak learners to improve their overall performance (Ismail et al., 2021). In recent years, institutions have become more informed about the potential value of analyzing educational data. To this end, machine learning methods are employed to extract useful information from educational databases for a deeper understanding of students' psychological and emotional aspects (Baashar et al., 2022). Machine learning classification and prediction models are also used to predict students' performance based on various features. These models are evaluated using heterogeneous datasets and evaluation metrics, such as accuracy (Ismail et al., 2021).

While a large amount of research has been conducted in EDM, predicting academic performance is still challenging for many educational bodies, such as schools (Baashar et al., 2022). The main challenge originates from the fact that, as mentioned, educational performance is a multidimensional phenomenon. Hence, an integrative approach combines aspects of different schools of thought and considers factors affecting academic performance to understand it holistically. Therefore, this research is unique in terms of the diversity of features investigated to predict academic performance. Some

categories of features are emphasized in the literature, such as demographic features and personality traits, and some are less investigated, such as favorite activities and relations with others. Therefore, in this article, after a literature review of different methods used to predict academic performance and important factors, each predictor's role is investigated by emphasizing the most important ones. A wide range of features has been analyzed in six categories, including demographic, personality traits, skills, favorite activities, relationships with others, and out-of-school activities. The analysis is performed on the data collected from the students at a school in Iran over several years and across different levels of education. This study, to the best of our knowledge, is unique in this field due to its variety of features and data collection across different years and stages.

This research has been designed to answer the following questions regarding the relationship between various feature categories and academic performance:

> **RQ1:** How can a prediction model be developed based on machine learning algorithms to envisage student academic performance in terms of Grade Point Average (GPA)?
>
> **RQ2:** Which category of features (demographic, personality traits, skills, favorite activities, relationships with others, and out-of-school activities) has more influence on student performance?
>
> **RQ3:** Overall, what are the most important features that affect student performance more?

In this study, the above questions will be addressed using different machine learning algorithms, and their accuracy will be compared to determine the best model to predict students' GPAs.

The rest of the paper is organized as follows; Section 2 illustrates the literature review; Section 3 describes the research method based on the CRISP-DM methodology. Section 4 reports the results of the research questions. Section 5 discusses the contributions of the research. Finally, Section 6 summarizes the findings and implications as a conclusion.

2. Literature Review

In recent years, many studies have been conducted about employing machine learning algorithms to get valuable insights from educational data, predict students' performance, and to understand factors influencing academic outcomes. In this regard, algorithms such as Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), along with ensemble methods, have been used. This section provides a review of some similar studies and their findings.

Various prediction algorithms were compared in the research by Hashim et al. (2020) in order to predict students' grades. The algorithms evaluated include DT, NB, KNN, SVM, ANN, LR, and SMO. The results indicated that the LR prediction model achieved the highest accuracy in predicting students' final grades, with 68.7% accuracy for passing and 88.8% for failing. It also demonstrated that demographic features, educational background, and behavioral features are the most crucial factors in predicting students' performance (Hashim et al., 2020).

Ahmad et al. (2021) compared ANN and RF models for predicting student performance based on demographic and evaluation data. They used the Open University Learning Analytics Dataset (OULAD), which includes demographic factors such as age, gender, location, deprivation index of the place of residence, disability status, and final outcome. Moreover, some features of student activities, such as interaction with virtual learning environments and evaluation grades, were also considered. Their results demonstrated the superiority of ANN in achieving 91.08% accuracy. The study's findings emphasized the significance of demographic factors and the suitability of the ANN model for predicting the students' performance. In addition, the authors highlighted the need for large and accurately selected datasets to implement the ANN model successfully (Ahmad et al., 2021). Comparison between various classification models, such as DT, NB, ANN, SVM, and RF, was performed in a study conducted by Ismail et al. (2021) on the Portuguese student performance dataset. It is emphasized that various social, demographic, psychological, and familial factors influence student performance.

Moreover, statistical methods, such as correlation, can also identify important factors. In a study by Prestes et al. (2021), the correlations between school environment issues, professional development, school management, and learning analytics were analyzed. The study utilized Pearson and Spearman

correlation methods and identified important features in education analysis, including gender, ICT usage, creativity, critical thinking, and problem-solving skills. Demographic features were also utilized in the study conducted by Wu (2021), along with other features such as motivational tendencies and students' actual cognitive participation.

Multiple linear regression is another algorithm considered by Dabhade et al. (2021), along with SVM. Moreover, demographic features, hobbies, interests, family income, and students' behavioral characteristics were considered to develop a model. Their findings indicated a significant relationship between students' behavioral features and academic performance. These include time spent on social media and watching movies, attention to academic studies, students' behavioral participation in academic activities, and peer influence (Dabhade et al., 2021).

In another study, Ouatik et al. (2022) considered the challenges of big data analytics while investigating the problem of predicting educational performance. The data were stored in the Hadoop Distributed File System (HDFS), and the prediction algorithm was applied using MapReduce. They also considered various feature selection approaches in their study, which led to the introduction of a model based on the SMO algorithm at the feature selection phase and the SVM algorithm at the classification phase, resulting in the best model with an accuracy of 87.32%. The study also considered the utilization of academic assessment followed by economic status, parent educational level, distance from home, student interest, psychological disorder, and the number of access to virtual classrooms as the most important features in predicting student performance (Ouatik et al., 2022). The importance of demographic features such as gender, residing in affluent and developed regions, and being in Years 3 and 4 on academic success was investigated by Yakubu and Abubakar (2022) using the LR algorithm.

Choosing a major field is another important issue that can be influenced by various factors that predict student behavior, preferences, and performance. In this regard, machine learning algorithms, such as ANN, NB, and SVM, were employed by Veluri et al. (2022) to predict the factors affecting major selection. The ANN outperforms other algorithms with 95% accuracy. Table 1 summarizes the best algorithms from some recent EDM studies. Studies are categorized based on the machine learning algorithms, with each study's highest accuracy classification/prediction algorithm identified. It should be noted that the number of studies in this area is not limited to those presented in Table 1.

Table 1. Literature Review on Classification and Prediction Algorithms to Predict Academic Performance

Study	Classification/Prediction Algorithm	Accuracy
(Hashim et al., 2020)	LR	68.7% for passing 88.8% for failing
(Rahman, Islam, et al., 2021)		99%
(Yakubu & Abubakar, 2022)		83.5%
(Ahmad et al., 2021)	ANN/Multi-Layer Perceptron (MLP)	91.08%
(Veluri et al., 2022)		95%
(Jacob & Henriques, 2023)		85%
(Goga et al., 2015)	DT/RF	99%
(Ismail et al., 2021)		>72%
(Yağcı, 2022)		74%
(Meghji et al., 2023)		93.4%
(Prestes et al., 2021)	Pearson and Spearman correlation	
(Dabhade et al., 2021)	SVM/Linear Support Vector Regression	83.44%
(Bilal et al., 2022)	(SVR)	92%
(Ouatik et al., 2022)		87.32%
(Kananda & Mwangi, 2023)	Apriori Algorithm	Confidences of top three rules: conf:(1) conf:(0.97) conf:(0.91)
(Guang-yu & Geng, 2019)	CatBoost/XGBoost	73%
(Oreshin et al., 2020)		91%
(Ramaswami et al., 2022)		75 ± 2.1 %
(Asselman et al., 2023)		78.75%

As can be observed in Table 1, previous studies, especially in recent years, emphasize the effectiveness of machine learning models in predicting students' performance. These studies utilized

LR, ANN, DT/RF, Correlation analysis, SVM/SVR, Apriori, and ensemble techniques, such as CatBoost/XGBoost, for academic performance classification or prediction. Considering the accuracy achieved by these algorithms, based on Table 1, all the algorithms could achieve high prediction accuracy. LR and DT/RF algorithms could predict academic performance with an accuracy of 99% (Goga et al., 2015; Rahman, Islam, et al., 2021). Though many factors, such as the volume and quality of data and the collected features, could affect the accuracy and outperformer algorithm in each study, Table 1 indicates that all categories of algorithms could effectively predict academic performance.

Previous research also highlights the importance of considering various features, such as demographic factors, psychological and behavioral features, students' skills, as well as out-of-school and favorite activities, to predict their academic performances. Demographics such as parents' educational levels, profession, age, income, religious affiliations, residing in affluent and developed regions, economic status, distance from home, and living location (urban or rural) are among the features considered important in predicting academic performance in many studies (Ahmad et al., 2021; Bilal et al., 2022; Farooq et al., 2011; Goga et al., 2015; Hashim et al., 2020; Ismail et al., 2021; Kananda & Mwangi, 2023; Oreshin et al., 2020; Ouatik et al., 2022; Yakubu & Abubakar, 2022). Personality or behavioral features such as conscientiousness, neuroticism, resilience, honesty, decisiveness, student interests, and psychological disorders were also investigated by several researchers as the key features that can predict academic achievement (Dabhade et al., 2021; De Feyter et al., 2012; Lievens & Sackett, 2012; Meghji et al., 2023; Ouatik et al., 2022). Skills such as interpersonal and mathematical capabilities (Asselman et al., 2023; Lievens & Sackett, 2012), out-of-school activities such as participation in clubs, co-circular and social-oriented activities (Ismail et al., 2021; Rahman, Islam, et al., 2021; Valentine et al., 2002), and favorite activities and hobbies (Dabhade et al., 2021; Meghji et al., 2023) are also among categories identified by researchers as predictive factors using educational data mining.

Reviewing previous studies reveals that demographic features have been studied the most, followed by personality or behavioral traits. While all of these features are considered important in separate studies, having a comprehensive approach and considering all of the categories along with each other to find out which of them has more predictive power in academic performance can lead to insightful results. Some studies have investigated two or more categories of features. For example, Guang-yu and Geng (2019) considered personality/behavioral traits and out-of-school activities. In a study by Dabhade et al. (2021), five categories of features were considered; however, the size of the dataset was small, and consisted only of 85 samples. Meghji et al. (2023) also investigated demographics, personality/behavioral traits, and favorite activities.

The findings indicate that integrating various categories of features can lead to an accurate and comprehensive model for predicting students' performance and understanding the factors influencing their academic outcomes. Therefore, in this research, a more comprehensive approach to the issue has been adopted by using various features of all the categories from 3,701 students. Based on the specific data gathered for this research, another category, not investigated in previous studies, "Relationships," is also considered. These features relate to questions designed to measure different types of relationships students create. Therefore, this research is unique in terms of the variety and comprehensiveness of the features considered.

3. Data and Methodology

This section discusses the main building blocks of the proposed student performance prediction model, which leads to identifying the most important features affecting academic performance. Since the problem is analyzed through the data mining approach, the methodology in the present study is CRoss Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 1999). The CRISP-DM methodology, developed by a consortium of leading data mining users and suppliers, presents a standard framework for performing data mining projects independent of industry and technology. This methodology includes six phases, with each consisting of several tasks. Therefore, the life cycle of a data mining project can be broken into six stages, starting with business understanding, followed by data understanding and preparation, modeling and evaluation, and finally, deployment (Wirth & Hipp, 2000). Figure 1 provides the overall framework of the main steps of this research methodology.

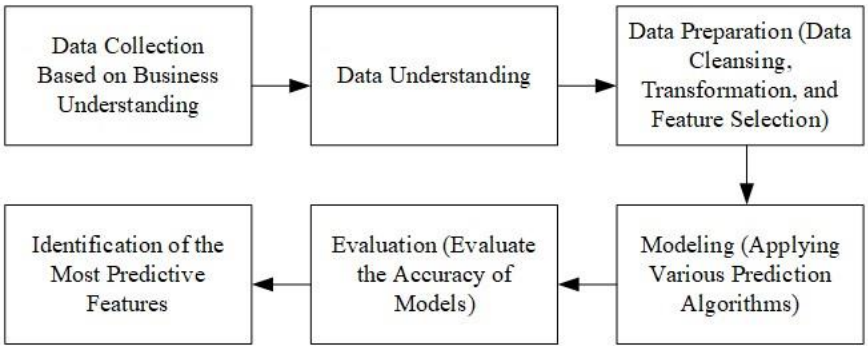

Fig. 1. The Flow Diagram of the Methodology to Predict Student Performance and Identify GPA Determinants

3-1. Data Collection

The dataset utilized in this research is collected from the repeated surveys conducted in different years at one of the prominent schools in Iran. The questionnaires, filled out by students, parents, and teachers, concerned students' features at different educational levels, from elementary to high school. Experts in education and psychology fields designed the survey questions, which are vast enough to be assessed to predict student performance. Therefore, the dataset can be considered as unique in terms of its variety of categories of features. Reviewing the literature and identifying the most important features considered by different studies and using the data gathered for this research have led to the dataset concentrated on six important feature categories: demographics, personality traits, skills, favorite activities, relationships with others, and out-of-school activities.

3-2. Data Understanding

The dataset consists of 3,701 records and 197 features categorized into six groups, as mentioned before. The records pertain to various students at different educational levels and in different years. Table 2 presents these six categories of features along with some samples of the features in each category, which has been extracted from the corresponding survey questions.

Table 2. Understanding Variables (Features) Used in This Research

No.	Feature Category	Features	
1	Demographics	Parents' educational level	
		Parents' profession	
		Parents' field of study	
		Student's level of education	
		Number of children in the family	
2	Personality	Positive Personality Traits	Self-esteem
			Empathy and helping others
			Responsibility
			Interest in studying
			Self-regulation
		Negative Personality Traits	Practical obsession
			ADHD
			Tic disorders
			Seclusion
3	Skills	Interpersonal skills	
		Mathematical and logical skills	
		Linguistic skills	
		Musical and artistic skills	
		Sports skills	
4	Favorite Activities	Mathematics	
		Science	
		Sports	
		Poetry and literature	
		Nature and animal-related activities	
5	Relationships	Relationships with teachers and other students	
		Relationship with parents	
		Ability to find friends	
6	Out-of-School Activities	Interested in extracurricular activities	
		Interest in games and recreational activities	
		Activities related to information technology such as computer, internet, etc.	

It should be noted that each feature corresponds to several questions in the questionnaires. For example, "interpersonal skills" can be assessed by questions about interest in teamwork, discussion skills, social interaction skills, etc. Therefore, the score for each feature can be obtained by assessing the scores of the responses to the related questions, ranging from poor (1) to excellent (5). Hence, to calculate the overall score for each feature, the scores of all questions related to that feature are averaged for each student.

3-3. Data Preparation

After merging the results obtained from different surveys for each Student's Class ID, which is unique at each level of education, several data preparation steps were applied to the dataset. These steps can be divided into three parts. Figure 2 represents the tasks done in each part of the data preparation aimed at finding the ones suitable for modeling with prediction algorithms.

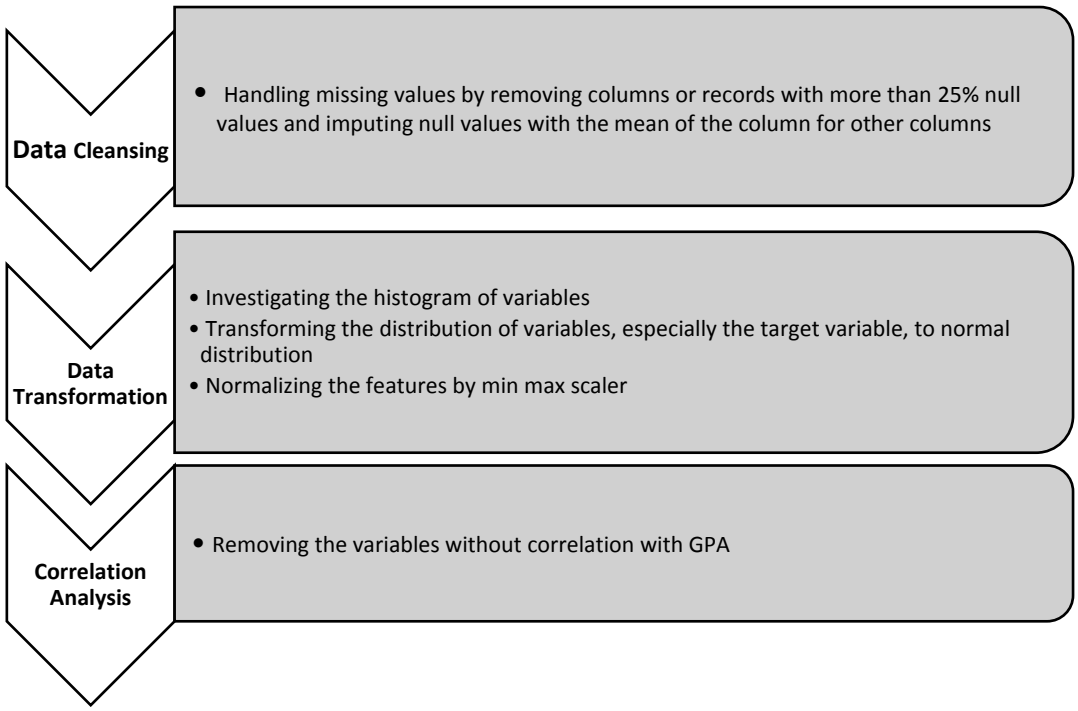

Fig. 2. Data Preparation Step for Developing a Student Performance Prediction Model

As illustrated in Figure 2, the data cleansing process consists of handling missing value problems in the data. First, the records in which more than 25% of the data are null were dropped. Then, for the columns, the ones with more than 1000 null values were removed. Finally, for the remainder, the null values were imputed by the mean of the corresponding column. After data cleansing, the dataset consists of 2,776 records and 73 features, which were split into 5%, which equals 140 records for the test, and 95% equals 2,636 records for the training dataset.

In the data transformation part, histograms of all variables were investigated since there was a requirement for normal distribution for correlation and regression analyses to be performed. The Yeo-Johnson power transformation was used to reduce skewness and approximate normality (Yeo & Johnson, 2000). This step needed to be done as a pre-requisite for regression analysis and played an important role in increasing the accuracy of the model. Another transformation employed was feature scaling via MinMaxScaler since the ranges of variables were different from each other. Then, correlation analysis was conducted to remove the variables lacking any significant correlation with the target variable, i.e., GPA. Eliminating these variables also played an important role in reducing the overfitting of the model and obtaining more accuracy on the test data. All the aforementioned tasks were performed using Python libraries such as "Scikit-learn"(Pedregosa et al., 2011).

3-4. Modeling

Since GPA is a continuous variable, several regression algorithms were used to predict students' GPAs based on the features explained in the previous subsection. Therefore, in this study, some of the prominent and advanced regressors such as CatBoost, RF, DT, K-Neighbors, Polynomial, and SVM regressors were applied to the dataset.

Many previous studies, similar to the current one, have used the regression analysis method Since traditional procedures such as the Ordinary Least Squares (OLS) regression, the Stepwise regression, and the Partial Least Squares regression are highly sensitive to random errors (Farahani et al., 2010). Many improvements have been proposed in the literature over the past few decades, such as the Ridge regression beside other variants (Muthukrishnan & Rohini, 2016). For example, the RF regressor combines the performance of numerous DT algorithms to predict the value of a variable. When RF receives an input vector made up of the value of various features, it builds a number K of regression trees and averages the results. The SVM regression model can also be defined as follows (Rodriguez-Galiano et al., 2015):

$$f\left(\boldsymbol{X}\right)=\boldsymbol{W}^{T}\,\phi\left(\boldsymbol{X}\right)+b$$

where, $\phi:\boldsymbol{X}\rightarrow\phi\left(\boldsymbol{X}\right)\in\mathbb{R}^{H}$ assumed to be a nonlinear function utilized for mapping input data into the high-dimensional feature space. Therefore, the SVM regressor can cope with the non-separable features (Rodriguez-Galiano et al., 2015).

The Categorical Boosting (CatBoost) Regressor is a new gradient boosting algorithm that handles categorical features well. When the input features are categorical, it can improve accuracy and reduce overfitting problems (Panigrahi et al., 2022). All the algorithms were utilized through the Scikit-learn and CatBoost libraries in Python.

3-5. Evaluation

This study has utilized the R-squared metric to evaluate the performance of prediction model. R-squared is defined as "the proportionate reduction in uncertainty, measured by the Kullback-Leibler divergence, due to the inclusion of regressors" (Cameron & Windmeijer, 1997). The R-squared can be formulated as follows (Chicco et al., 2021):

$$R^{2}=1-\frac{\sum_{i=1}^{m}(\hat{Y}_{i}-Y_{i})^{2}}{\sum_{i=1}^{m}(\bar{Y}_{i}-Y_{i})^{2}}$$

where $\hat{Y}_i$ is the predicted ith value, the Y_i is the actual ith value, and the $\bar{Y}_i$ is the mean of true values. The regression model predicts the $\hat{Y}_i$ element for the corresponding Y_i element. Therefore, the R-squared can be interpreted as the proportion of the variance in the dependent variable that can be predicted by independent variables (Chicco et al., 2021).

For the linear regression model, the coefficient of determination, R^2 , is a widely used goodness-of-fit measure whose usefulness and limitations are well understood. Cameron and Windmeijer (1997) proposed the R-squared measure of goodness of fit for some common nonlinear regression models as well.

Hence, in this research, the performances of prediction algorithms, which are various regression models, were compared based on the R-squared measure to select the one with the highest goodness-of-fit.

3-6. Identification of the Most Predictive Features

Finally, the feature importance of the prediction model was determined through the RF Regressor. Feature selection using RF falls under the category of embedded methods, which have their own built-in feature selection techniques. The feature importance of ensemble models, such as RF, represents an aggregation of the feature importance from its base models (Hwang et al., 2023). RFs consist of many DTs, with each being built over a random extraction of the samples and a random extraction of the features from the dataset. At each node, the tree divides the dataset into several classes, with each including samples that are more similar to one another and different from those in other classes. Therefore, the importance of each feature can be extracted, taking into account how pure each of the resulting classes is.

4. Results

This section presents the results based on the research method and the data analysis described in Section 3. Further, the results are used to evaluate the importance of features and identify the top 10 most important features extracted through the RF method. Table 3 provides the performance of each regressor used in this research based on the evaluation metric, which is the R-squared. It should be noted that these results were achieved after hyperparameters tuning. First, a random search was utilized to identify some optimal values for model parameters. A random search sampled a specified number of hyperparameter combinations at random rather than exhaustively searching through all possible combinations. Then, the results were optimized by using a grid search. In this process, the model was trained and evaluated for each combination of hyperparameters in the grid, and the best combination was selected.

Table 3. Performance of Various Prediction Models

Algorithm	MAE	MSE	RMSE	R-Squared
CatBoostRegressor	0.0527	0.0052	0.0725	**0.87**
RF Regressor	0.0568	0.0057	0.0758	**0.85**
Polynomial Regression	0.0754	0.0096	0.0983	**0.76**
SVR Regressor	0.0740	0.0095	0.0976	**0.76**
KNeighbors Regressor	0.0645	0.0071	0.084	**0.82**
DecisionTree Regressor	0.0765	0.0097	0.098	**0.75**

In Table 3, Mean Average Error (MAE) and Mean Squared Error (MSE) are also presented as the ones to assess the quality of fit in terms of the distance of the regressor to the actual training points. The difference between them lies in their evaluating metrics, which are linear or quadratic, respectively. Moreover, Root of Mean Square Error (RMSE) is provided as a natural derivation to standardize the units of measures of MSE (Chicco et al., 2021).

It can be observed that the CatBoost regressor and RF regressor performances outperform the other algorithms, and the CatBoost regressor can predict the students' GPAs with the R-squared of 0.87. It should be noted that the metrics provided in Table 3 are based on the test dataset. Moreover, Table 4 shows the results of the CatBoost regressor for the train and test datasets to compare them with each other.

Table 4. Evaluation of Catboost Regressor for the Train Set and Test Dataset

Dataset	MAE	MSE	RMSE	R-Squared
Train set	0.0445	0.0035	0.0597	0.9121
Test set	0.0527	0.00526	0.0725	0.8706

Based on Table 4, the model has overcome the problem of overfitting since the results of the model evaluation based on train and test datasets are almost similar to each other. This was achieved by removing irrelevant features and increasing the number of samples used for training the model.

As the next step, explained in Section 3, the importance of the features was obtained by the RF feature selection method. RF can be utilized as a feature selection technique due to the random exploration of features. The estimate of feature importance is obtained by permuting the values of the feature along all of the examples and observing the result on the estimate of generalization error. At each node, in creating a random forest, a feature is selected randomly and utilized to split the node and maximize the information gain. This information gain can be used to measure the correlation between the feature and the class (Rogers & Gunn, 2005). Hence, in this study, by identifying the most important features, the determinants of academic performance can be selected from various features gathered about students. Figure 3 illustrates the result.

As illustrated in Figure 3, features were sorted according to their importance based on the RF algorithm. It indicated that the top 10 most important features were among the positive or negative personality traits since "pq" and "nq" represented positive and negative qualities in the personality questionnaire, respectively. There were also some features related to skills among the important features. At the same time, GradeID, which was among the demographic features and represented the level of education for students, was also a determinant factor when predicting the GPA. However, other categories of features showed less importance than personality traits. To identify the description

of features based on Figure 3, Table 5 presents the top 10 important features and their description extracted from the survey questions.

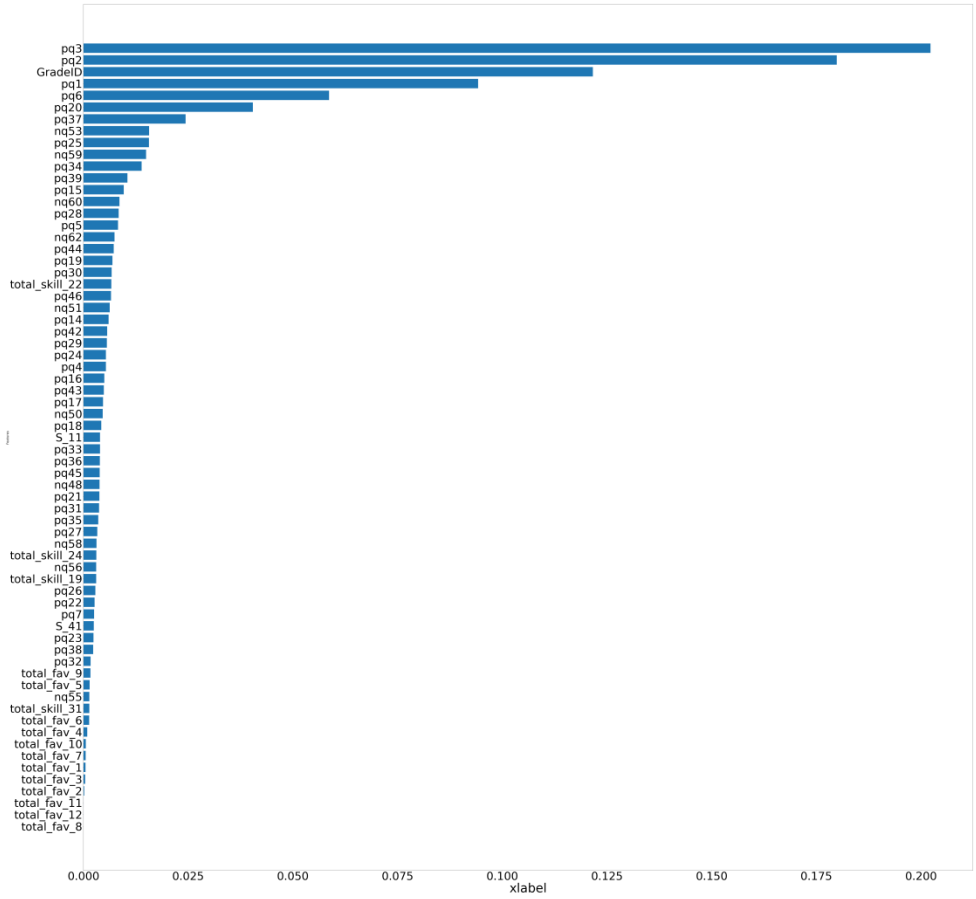

Fig. 3. Feature Importance Derived by RF Feature Selection Method

Table 5. The Ten Most Important Features Identified by the RF Algorithm

No	Feature	Description
1	pq3	Interest in studying
2	pq2	The quality of doing homework
3	GradeID	The student's level of education
4	pq1	Academic level
5	pq6	Accompanying and cooperating with the school
6	pq20	Self-regulation
7	pq37	Contentment
8	nq53	Fraud
9	pq25	Being cautious
10	nq59	Tic disorders

Moreover, the results indicate that among skill features, "logical thinking and reasoning" indicate high importance. Although other features, such as demographic features, are not among the most important ones, some correlate more with GPA. For example, if one of the parents studied mathematics in high school, it positively affects the student's GPA. Meanwhile, the number of children in the family and being the Nineth child of the family has a negative correlation with the GPA.

5. Discussion

This section highlights the contributions to the existing knowledge based on the research questions and the results achieved in this study.

RQ1: How can a prediction model be developed based on machine learning algorithms to envisage student academic performance in terms of Grade Point Average (GPA)?

First, this research aimed to develop a prediction model to predict the students' educational performance in terms of GPA based on various features collected through surveys. As presented in Section 4, several regression algorithms were utilized to predict GPA, among them the CatBoost regressor outperformed the other algorithms, and it could predict GPA with the R-squared of 0.87, indicating a high goodness of fit.

CatBoost regressor, as mentioned, is a new gradient-boosting algorithm that can improve accuracy and reduce overfitting when the input data is categorical. Since the data used in this study is based on the scores of 4 or 5 categories, say, from poor to excellent, this algorithm can be applied adeptly. At the same time, the RF regressor also shows a good performance, which is confirmed by the literature (Ismail et al., 2021). Compared with other research that developed their classification or prediction model based on CatBoost/XGBoost, as presented in Table 1, this study achieved an acceptable accuracy since their accuracy ranges from 73% to 91% (Asselman et al., 2023; Guang-yu & Geng, 2019; Oreshin et al., 2020; Ramaswami et al., 2022).

RQ2: Which category of features (demographics, personality traits, skills, favorite activities, relationships with others, and out-of-school activities) has more influence on student performance?

As presented in Table 5, the most predictive features, the importance of which is higher than the others, are among the positive personality traits. There are also two negative personality characteristics, i.e., skill and demography, in the ten most important features. Besides the fact that several studies emphasized the importance of personality and behavioral traits in academic achievements (Dabhade *et* al., 2021; De Feyter et al., 2012; Furnham & Mitchell, 1991; Hashim et al., 2020; Ismail et al., 2021; Ouatik et al., 2022; Veluri et al., 2022), there are also psychological and educational theories related to the important features identified in this study. Based on Table 5, "interest in studying" is the most important feature which is aligned with theories of intrinsic motivation, such as Self-Determination Theory (SDT) (Ryan & Deci, 2000). Based on this theory, individuals tend to engage in activities and perform better with autonomy. Therefore, interest in studying can be an indicator of a student's choice and their willingness to do it. Some of the features, such as the "quality of doing homework" and "self-regulation," relate to Social Cognitive Theory and Self-Regulated Learning (SRL) (Zimmerman, 2002). Based on SRL, the students' self-regulation is a way to compensate for their individual differences in learning. Self-regulated promoting practices encourage students' metacognition, motivation, and strategic action development. These learning competencies enhance students' academic, social, emotional, and career outcomes. In other words, metacognitive learners are aware of their personal learning strengths and challenges, are aware of learning strategies, and are in agreement with others' needs and interests (Brenner, 2022). Two descriptive features are also among the important features: "Student's level of education" and "Academic level." A student's level of education (Grade ID) can have a negative correlation with the GPA since, at higher stages of education, the courses become more difficult. Piaget's Theory of Cognitive Development (Huitt & Hummel, 2003) also outlines the different stages of intellectual growth as students proceed through their education.

Another important feature obtained by the machine learning approach is "Contentment," which aligns with well-being theories such as Seligman's PERMA model (Seligman, 2011). This model emphasizes key factors contributing to well-being and success, among which positive emotions such as contentment can be observed. "Fraud" and "Tic disorders" are among negative personality traits and are also recognized as important features. Tic disorders are specific psychological conditions that consist of unwanted movements. While these features are not directly connected to a specific theory, they can be considered potential factors influencing the students' well-being and behavior, impacting their academic outcomes. The reasons for the co-occurrence of tics and learning problems are diverse and include common factors such as shared neurodevelopmental and neurotransmitter genes that disrupt cognition and behavior, as well as comorbid conditions (Eapen et al., 2013). According to the literature, children with tics, especially those with associated attention deficit hyperactivity disorder,

have a high frequency of comorbid learning disabilities and are up to 5 times more likely to need special educational services compared with the general population. However, studies revealed that early academic support and modification of environmental characteristics can reduce the negative effect of this feature on poor academic performance (Cubo et al., 2013).

Hence, the results of this research highlight the importance of personality and behavioral features rooted in some theories, such as SDT, SRL, as well as the Theory of Cognitive Development. The findings indicate the critical role of intrinsic motivation and self-regulation in academic success and achievement compared to other factors. Some other studies also emphasized the importance of behavioral or personality features to predict academic performance (Dabhade et al., 2021; De Feyter et al., 2012; Lievens & Sackett, 2012; Meghji et al., 2023; Ouatik et al., 2022).

RQ3: Overall, what are the most important features that affect student performance more?

While most of the important and predictive features can be categorized as personality traits, especially the positive ones, the role of features in other categories is also significant. For example, "logical thinking and reasoning" falls within the skill category and parents' "mathematics field of study" falls within the demographic category. Given that reasoning ability is considered the most central component of analytical intelligence and problem-solving, the relationship between this feature and academic achievements has been investigated in several studies, indicating a strong effect of reasoning ability on predicting GPA and academic success (Freund & Holling, 2008; Valanides, 1997). Notably, according to the findings of this research, both "logical thinking and reasoning" skills and parents' "mathematics field of study" are aligned to highlight the significance of reasoning ability, which can be fostered through curricula and teaching interventions (Valanides, 1997).

As a matter of fact, the scientific contribution of the current research includes developing a prediction model with high accuracy based on the machine learning approach to predict student academic performance in terms of their GPA and extract the most important features that affect it. As presented in Table 1, some previous studies also recognized CatBoost/XGBoost as the algorithm with the highest accuracy. These studies achieved an accuracy between 73% to 91% (Asselman et al., 2023; Guang-yu & Geng, 2019; Oreshin et al., 2020; Ramaswami et al., 2022). Therefore, the accuracy obtained in this study can be considered acceptable. The features extracted as the most predictive features are mostly among personality and behavioral traits as well as other important features, such as logical thinking and reasoning ability. The findings emphasize the importance of intrinsic motivation, self-regulation, cognitive development, and analytical intelligence, which represented higher predictive power than many other features investigated in this research. Moreover, from a practical perspective, the findings of this research would help educators and policymakers develop and foster these features in students for better academic achievements and successful education.

6. Conclusion

In recent years, educational data mining has turned into a subject of data analytics, in which machine learning approaches are applied to extract insights in the field of education. Although a huge number of studies have been conducted in this regard, there is still a necessity for an integrative approach that considers various factors affecting academic performance to obtain a holistic understanding of it. In this study, educational data mining was used as a means to predict the students' academic performance. Since the GPA is still the most popular measure for academic achievements, it is utilized as a target variable to be predicted by various regression models. Different regression algorithms, from polynomial regression to more advanced techniques such as CatBoost, SVR, and RF regression, were employed in this regard.

Moreover, based on 3,701 records collected through surveys in one of the prominent schools in Iran, a significant number of features in six categories, including demographics, personality, skills, favorite activities, relationships, and out-of-school activities, were analyzed to find out predictive features affecting the students' performance. Accordingly, a predictive model based on CatBoost Regressor was proposed, which can predict the students' GPA with 0.87 R-squared. This prediction is also based on important features extracted in this research as the most predictive ones. The analysis of feature importance indicates that positive personality traits play a more significant role in student

educational performance. Among these characteristics, "Interest in studying," "The quality of doing homework," "Student's level of Education," "Academic level," "Accompanying and cooperating with the school," "Self-regulation," "Contentment," "Fraud," "Being cautious," and "Tic disorders" are identified as the most important features in predicting GPA. It can also be observed that most of these important features are rooted in and supported by various psychological and educational theories such as SDT, SRL, the Theory of Cognitive Development, and well-being theories; among them, the intrinsic motivation theories stand out. The importance of "Logical thinking and reasoning" is also emphasized based on the findings of this study.

Therefore, due to the uniqueness of the features investigated and the comprehensive approach—considering all categories of features together—this research provides novel insights into which categories and specific features can play a more significant role in predicting academic performance. It was found that personality and behavioral traits, especially the positive ones, have a greater effect on academic performance. Simultaneously, among skills and demographic features, reasoning and analytical skills, as well as the mathematical background of parents, play the most crucial role. The findings of this research, derived from data mining techniques, align with several psychological theories, demonstrating the effectiveness of Educational Data Mining (EDM) in educational decision-making.

The findings can be noticeable for practitioners, such as educators and policymakers, in their decision-making regarding educational issues or intentions to develop and foster these features in students. Some of the important features identified in this study, such as "Interest in studying," "The quality of doing homework," and "Accompanying and cooperating with the school," are considered school features. These features can be fostered and improved by various methods, such as utilizing personalized approaches by understanding the student's interests and strengths. Employing disruptive technologies, storytelling, gamifying the process, interactive resources, and providing clear feedback to students can be considered as some of the interventions that schools and teachers can conduct to improve these features in students. Other important features, including "self-regulation" and "contentment" as positive personality traits, can also be enhanced by various exercises and practices employed by teachers, parents, and students. For example, teaching goal-setting and planning, developing mindfulness, and encouraging curiosity by providing opportunities for exploration and self-directed learning are among such activities that can foster intrinsic motivation. Moreover, features such as "Fraud," "Being cautious," and "Tic disorders" that are categorized as negative personality or behavioral traits, can be mitigated by various solutions used by teachers and parents. Flexible assessments, including diverse evaluation methods such as projects or open-book tests, focus on learning over grades, building critical thinking skills, supporting behavioral therapy as well as stress reduction techniques, may help. Hence, educators can use the prediction model and the important features achieved in this study to improve academic performance. In this regard, implementing pilot programs in some schools to test the practical applications of the model is suggested to provide real-world feedback and improve the model.

7. Limitations and Future Work

While this study presents a robust predictive model for student academic performance, it also has some limitations. The first limitation is the dataset size, which includes 3,701 records collected from one of the schools in Iran. This relatively small and demographically limited dataset may affect the generalizability of the findings. Moreover, the study relies on self-reported data, especially those concerning behavioral features, which can lead to biases or inaccuracies.

Future research should address these limitations by expanding the dataset to include more samples. This would enhance the validity and applicability of the predictive model. In addition, future work should focus on longitudinal data analysis, given the nature of the dataset, which pertains to different years and levels of education for various students.

Exploring other machine learning algorithms, such as ensemble techniques and deep learning, represents another area for improvement that could enhance the accuracy of the prediction model.

References

Ahmad, M. S., Asad, A. H., & Mohammed, A. (2021). A machine learning based approach for student performance evaluation in educational data mining. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)* (pp. 187–192). http://dx.doi.org/10.1109/MIUCC52538.2021.9447602

Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

Baashar, Y., Hamed, Y., Alkawsi, G., Capretz, L. F., Alhussian, H., Alwadain, A., & Al-amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. *Alexandria Engineering Journal*, *61*(12), 9867–9878. https://doi.org/10.1016/j.aej.2022.03.021

Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports*, *12*(1), 12508. https://doi.org/10.1038/s41598-022-15880-6

Brenner, C. A. (2022). Self-regulated learning, self-determination theory and teacher candidates' development of competency-based teaching practices. *Smart Learning Environments*, *9*(1), 1–14. https://doi.org/10.1186/s40561-021-00184-5

Cameron, A. C., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*(2), 329–342. https://doi.org/10.1016/S0304-4076(96)01818-0

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999, March). The CRISP-DM user guide. In *4th CRISP-DM SIG Workshop in Brussels in March* (Vol. 1999). https://the-modeling-agency.com/crisp-dm.pdf

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Cubo, E., Trejo, J., Ausín, V., Sáez, S., Delgado, V., Macarrón, J., Cordero, J., Louis, E. D., Kompoliti, K., & Benito-León, J. (2013). Association of tic disorders with poor academic performance in central Spain: A population-based study. *The Journal of Pediatrics*, *163*(1), 217–223. https://doi.org/10.1016/j.jpeds.2012.12.030

Czibula, G., Ciubotariu, G., Maier, M.-I., & Lisei, H. (2022). IntelliDaM: A machine learning-based framework for enhancing the performance of decision-making processes. A case study for educational data mining. *IEEE Access*, *10*(2), 80651–80666. http://dx.doi.org/10.1109/ACCESS.2022.3195531

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *47*(8), 5260–5267. http://dx.doi.org/10.1016/j.matpr.2021.05.646

De Feyter, T., Caers, R., Vigna, C., & Berings, D. (2012). Unraveling the impact of the Big Five personality traits on academic performance: The moderating and mediating effects of self-efficacy and academic motivation. *Learning and Individual Differences*, *22*(4), 439–448. https://doi.org/10.1016/j.lindif.2012.03.013

Eapen, V., Črnčec, R., McPherson, S., & Snedden, C. (2013). Tic disorders and learning disability: Clinical characteristics, cognitive performance and comorbidity. *Australasian Journal of Special Education*, *37*(2), 162–172. https://doi.org/10.1017/jse.2013.2

Farahani, H. A., Rahiminezhad, A., & Same, L. (2010). A comparison of partial least squares (PLS) and ordinary least squares (OLS) regressions in predicting of couples mental health based on their communicational patterns. *Procedia-Social and Behavioral Sciences*, *5*, 1459–1463. http://dx.doi.org/10.1016/j.sbspro.2010.07.308

Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of Quality and Technology Management*, *7*(2), 1–14.

Freund, P. A., & Holling, H. (2008). Creativity in the classroom: A multilevel analysis investigating the impact of creativity and reasoning ability on GPA. *Creativity Research Journal*, *20*(3), 309–318. https://doi.org/10.1080/10400410802278776

Furnham, A., & Mitchell, J. (1991). Personality, needs, social skills and academic achievement: A longitudinal study. *Personality and Individual Differences*, *12*(10), 1067–1073. https://doi.org/10.1016/0191-8869(91)90036-B

Goga, M., Kuyoro, S., & Goga, N. (2015). A recommender for improving the student academic performance. *Procedia-Social and Behavioral Sciences*, *180*, 1481–1488. https://doi.org/10.1016/j.sbspro.2015.02.296

Guang-yu, L., & Geng, H. (2019). The behavior analysis and achievement prediction research of college students based on XGBoost gradient lifting decision tree algorithm. In *Proceedings of the 2019 7th International*

Conference on Information and Education Technology (pp. 289–294). https://doi.org/10.1145/3323771.3323803

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing. https://doi.org/10.1088/1757-899X/928/3/032019

Huitt, W., & Hummel, J. (2003). Piaget's theory of cognitive development. *Educational Psychology Interactive*, *3*(2), 1–5.

Hwang, S.-W., Chung, H., Lee, T., Kim, J., Kim, Y., Kim, J.-C., Kwak, H. W., Choi, I.-G., & Yeo, H. (2023). Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar. *Journal of Wood Science*, *69*(1), 1–12. https://doi.org/10.1186/s10086-022-02073-y

Ismail, L., Materwala, H., & Hennebelle, A. (2021). Comparative analysis of machine learning models for students' performance prediction. In *Advances in Digital Science: ICADS 2021* (pp. 149-160). Springer International Publishing. https://doi.org/10.1007/978-3-030-71782-7_14

Jacob, D., & Henriques, R. (2023). Educational data mining to predict bachelors students' success. *Emerging Science Journal*, *7*, 159–171. http://dx.doi.org/10.28991/ESJ-2023-SIED2-013

Kananda, T. N., & Mwangi, H. (2023). Forecasting student academic performance in kenyan secondary schools using data mining. *International Journal of Innovative Science and Research Technology (IJISRT)*, *8*(3), 1626-1629. https://doi.org/10.5281/zenodo.7793063

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, *97*(2), 460-8. http://dx.doi.org/10.1037/a0025741

Meghji, A. F., Shaikh, F. B., Wadho, S. A., Bhatti, S., & Ayyasamy, R. K. (2023). Using educational data mining to predict student academic performance. *VFAST Transactions on Software Engineering*, *11*(2), 43-49. https://doi.org/10.1007/s10639-022-11152-y

Moreira, P. A. S., Dias, P., Vaz, F. M., & Vaz, J. M. (2013). Predictors of academic performance and school engagement—Integrating persistence, motivation and study skills perspectives using person-centered and variable-centered approaches. *Learning and Individual Differences*, *24*, 117–125. http://dx.doi.org/10.1016/j.lindif.2012.10.016

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)* (pp. 18–20). https://doi.org/10.1109/ICACA.2016.7887916

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971–990. https://doi.org/10.1016/j.paid.2007.03.017

Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A., Glukhov, I., Kaliberda, Y., Masalskiy, D., Serdyukov, A., & Kazakovtsev, V. (2020, October). Implementing a machine learning approach to predicting students' academic outcomes. In *Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System* (pp. 78–83). https://doi.org/10.1145/3437802.3437816

Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning (Online)*, *17*(12), 236-251. https://doi.org/10.3991/ijet.v17i12.30259

Panigrahi, R., Patne, N. R., Pemmada, S., & Manchalwar, A. D. (2022). Regression model-based hourly aggregated electricity demand prediction. *Energy Reports*, *8*(4), 16–24. http://dx.doi.org/10.1016/j.egyr.2022.10.004

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Prestes, P. A. N., Silva, T. E. V, & Barroso, G. C. (2021). Correlation analysis using teaching and learning analytics. *Heliyon*, *7*(11), e08435. https://doi.org/10.1016/j.heliyon.2021.e08435

Rahman, S. R., Islam, M. A., Akash, P. P., Parvin, M., Moon, N. N., & Nur, F. N. (2021). Effects of co-curricular activities on student's academic performance by machine learning. *Current Research in Behavioral Sciences*, *2*, 100057. http://dx.doi.org/10.1016/j.crbeha.2021.100057

Rahman, M. M., Watanobe, Y., Kiran, R. U., Thang, T. C., & Paik, I. (2021). Impact of practical skills on academic performance: A data-driven analysis. *IEEE Access*, *9*, 139975–139993. http://dx.doi.org/10.1109/ACCESS.2021.3119145

Ramaswami, G., Susnjak, T., & Mathrani, A. (2022). On developing generic models for predicting student outcomes in educational data mining. *Big Data and Cognitive Computing*, *6*(1), 6. http://dx.doi.org/10.3390/bdcc6010006

Resmi, T. J., Mathews, M. K., & Padmanabhan, S. (2024). Statistical analysis of student data and machine learning models for performance prediction. In *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)* (pp. 1–5). https://doi.org/10.1109/ICDECS59733.2023.10502482

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rogers, J., & Gunn, S. (2005). Identifying feature relevance using a random forest. *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, 173–184. https://doi.org/10.1007/11752790_12

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68-78. http://dx.doi.org/10.1037/0003-066X.55.1.68

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster.

Siddiq, F., Gochyyev, P., & Valls, O. (2020). The role of engagement and academic behavioral skills on young students' academic performance—A validation across four countries. *Studies in Educational Evaluation*, *66*, 100880. https://doi.org/10.1016/j.stueduc.2020.100880

Valanides, N. (1997). Formal reasoning abilities and school achievement. *Studies in Educational Evaluation, 23*(2), 169–185. https://doi.org/10.1016/S0191-491X(97)00011-4

Valentine, J. C., Cooper, H., Bettencourt, B. A., & DuBois, D. L. (2002). Out-of-school activities and academic achievement: The mediating role of self-beliefs. *Educational Psychologist*, *37*(4), 245–256. https://doi.org/10.1207/S15326985EP3704_4

Veluri, R. K., Patra, I., Naved, M., Prasad, V. V., Arcinas, M. M., Beram, S. M., & Raghuvanshi, A. (2022). Learning analytics using deep learning techniques for efficiently managing educational institutes. *Materials Today: Proceedings*, *51*, 2317–2320. https://doi.org/10.1016/j.matpr.2021.11.416

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (Vol. 1, pp. 29-39).

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, *163*, 104066. https://doi.org/10.1016/j.compedu.2020.104066

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. http://dx.doi.org/10.1186/s40561-022-00192-z

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

Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2