



Organisation of Islamic Cooperation

**Prediction of academic achievement of undergraduate students in engineering program**

**NJOYA PEFENSIE MOHAMED**

**201031407**

**M.Sc. T.E. (Specialization in Computer Science and Engineering)**

**THE ORGANISATION OF THE ISLAMIC COOPERATION (OIC)**

**ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)**

**Department of Technical and Vocational Education (TVE)**

**DHAKA, BANGLADESH**

**May 2023**



**Prediction of academic achievement of undergraduate students in engineering program**

*A thesis submitted for the partial fulfillment of the degree of MScTE in Technical Education  
at the Islamic University of Technology in 2023*

NJOYA PEFENSIE MOHAMED

201031407

Submitted to:

Department of Technical and Vocational Education (TVE)

Islamic University of Technology (IUT)

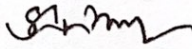
A subsidiary organ of OIC

Board Bazar, Gazipur 1704, Dhaka

## RECOMMENDATION OF THE BOARD OF EXAMINERS

The Thesis title is “**Prediction of academic Achievement of undergraduate students in Engineering program**” Submitted by **Njoya Pefensie Mohamed**, Master of Science in Technical Education with specialization in **CSE**, **Student ID: 201031407** of the AY 2021-2022 has been found satisfactory and accepted as partial fulfillment of the requirement of the degree of **Master of Science in Technical Education (M.Sc.TE) in May 2023**.

### MEMBERS OF THE EXAMINATION BOARD



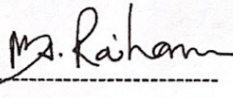
Asso. Prof. Md. Rashedul Huq Shamim

Supervisor (Chairman)



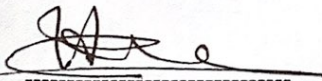
Prof. Dr. Md. Shahadat Hossain Khan  
Head TVE Dep., IUT, TVE

(Ex-Officio)



Prof. Dr. Md. Abu Raihan

Member



Prof. Dr. M. Tariq Ahsan

External Member

**Dedication**

I dedicated my work to The Almighty Allah without his support, strength and mercy I would be nowhere and to my beloved family have made it possible because of their countless prayers, love, support, guidance, motivation, cooperation and wisdom that held me firm and strong through my program. May Allah S.W.T render their countless support, keep them safe and away from fitnah.

## **Acknowledgement**

First and foremost, praises and thanks to the Allah, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor, Mr. Md. Rashedul Huq Shamim, for giving me the opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honour to work and study under his guidance. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, empathy, and great sense of humour.

I am fortunate that, I had the kind association as well as supervision of for his patience, effort, suggestion, criticism and encouragement throughout the research. I would like to thank him for his care and support. I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

A deep appreciation goes my mother Mrs. Ngajine Awawou, my late father Mr. Njoya Ahamadou, Words cannot express how grateful I am to my mother, and father for all the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far.

Also, I express my thanks to my friends Njayou Youssouf for his moral and technical support for the successful completion of this journey. I can't thank you enough for encouraging me throughout the experience.

Special thanks and gratitude go to the Organization of Islamic Cooperation for giving me a scholarship which provided me the opportunity to further my studies and later participate in the socio-economic development of my country.

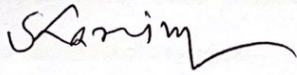
I am thankful to the Ministry of Higher Education of Cameroon (MINSUP) for providing me with the air-ticket which enable me to travel to Bangladesh.

It is my pleasure to be indebted to various people, who directly or indirectly contributed to the development of this research and my friends with whom we shared our day-to-day experiences and received lots of suggestions that influenced our thinking, behaviour and acts during the course of my study and improved my quality of work.

Finally, I thank Allah, for letting me through all the difficulties. I have experienced Your guidance day by day. You are the one who let me finish my degree. I will keep on trusting You for my future. Thank you, Allah.

**Declaration of the Author**

This is to certify that the work presented in this thesis is my original work. This thesis has neither been submitted nor previously been accepted for the award of any other degree in this university or elsewhere. I also declare that the sources used in this thesis were explicitly acknowledged with proper citation and references.



**Mr. Muhammad Rashedul Huq Shamim**

Supervisor and Associate professor

Department of TVE

Islamic University of Technology Gazipur-1704, Bangladesh

Year: 2021/2022



**Njoya Pefensie Mohamed**

StudentNo: 201031407

## **Abstract**

The purpose of this study was to predict academic achievement among engineering students and to pinpoint the elements or characteristics that influence this performance. To predict academic achievement, the study employed several machine learning models such as Linear Regression, Random Forest, Xgbooster, Artificial neural network and an ensemble of 3, and then compared their performance to find the best model. The study also looked at the important variables that affect academic accomplishment, such as demography, socioeconomic position, high school academic performance, and other pertinent variables. The study's conclusions could enhance academic support, counselling for engineering students, and instructional strategies.

An ensemble model surpassed any individual machine learning model, according to the study, which assessed the accuracy and precision of several machine learning models. Furthermore, the study found that past academic success in particular disciplines, such as Biology, English Language, Critical Reading, Citizen Competencies, and Mathematics, significantly influenced the academic performance of engineering students. However, while high schools and institutions had a positive or negative impact, the socioeconomic background of the students had no discernible impact on their academic achievement. While having no influence on female students, the demographic factor of gender had a beneficial effect on the academic performance of male students.

The lack of information on the students' academic achievement across various subject areas throughout their time at university also presented problems for the study. Making more precise prediction in the future will depend on gathering more detailed data on students' academic performance across various university-level courses. This knowledge might aid in developing better strategies for increasing academic outcomes in certain areas and help us better understand how certain disciplines impact academic achievement.



<b>Dedication.....</b>	<b>iii</b>
<b>Acknowledgement.....</b>	<b>iv</b>
<b>Declaration of the Author .....</b>	<b>vi</b>
<b>Abstract .....</b>	<b>vii</b>
<b>Contents.....</b>	<b>viii</b>
<b>List of tables .....</b>	<b>ix</b>
<b>Chapter 1. Introduction.....</b>	<b>1</b>
1.1. Background study and present state of the problem .....	1
1.2. Machine Learning Techniques .....	1
1.3. Factors Considered in Machine Learning Models.....	2
1.4. Potential Benefits of Predicting Academic Achievement .....	3
1.5. Motivation of the Research .....	4
1.6. Research Objectives .....	5
<b>Chapter 2. Literature review.....</b>	<b>6</b>
<b>Chapter 3. Methodology .....</b>	<b>18</b>
3.1. Machine learning algorithm .....	18
3.1.1 <i>Linear Regression</i> .....	19
3.1.2 <i>Random Forest</i> .....	21
3.1.3 <i>Naïve Bayes</i> .....	23
3.1.4 <i>XGBoost</i> .....	24
3.1.5 <i>Artificial Neural Network</i> .....	25
3.2. The performance evaluation metrics .....	26
3.2.1. <i>Accuracy</i> .....	26
3.2.2. <i>Recall</i> .....	27
3.2.3. <i>F1 Score</i> .....	27
3.2.4. <i>Mean Absolute Error (MAE)</i> .....	27
3.2.5. <i>Mean Squared Error (MSE)</i> .....	27
3.3. DATA COLLECTION .....	27
3.4. DATA PREPROCESSING .....	33
3.4.1. Data cleaning .....	34
3.4.2. Data Encoding .....	36
3.4.3. Data Normalization .....	36

<b>Chapter 4. RESULTS AND DISCUSSION.....</b>	<b>38</b>
4.1 Linear Regression.....	38
4.1.1 Comparison between actual and predicted values.....	38
4.1.2 Top important attributes.....	40
4.2 Random Forest.....	42
4.2.1 Comparison between actual and predicted values.....	42
4.2.2 Top important attributes.....	45
4.2.3 Partial Dependence Plots.....	48
4.3 Xbooster.....	51
4.3.1 Comparison between actual and predicted values.....	51
4.3.2 Top important attributes.....	53
4.3.3 Partial Dependence Plots.....	55
4.4 Artificial Neural Network.....	59
4.4.1 Comparison between actual and predicted values.....	59
4.5 Ensemble model.....	61
4.5.1 Comparison between actual and predicted values.....	62
<b>Chapter 5.....</b>	<b>64</b>
<b>CONCLUSION.....</b>	<b>64</b>
<b>Further research.....</b>	<b>66</b>
<b>REFERENCES.....</b>	<b>67</b>

#### List of tables

Table 1: Academic variables.....	29
Table 2 Socioeconomic variables.....	30
Table 3 Engineering programs.....	31
Table 4 Actual and predicted values LR.....	38
Table 5 Top important attributes LR.....	40
Table 6 actual and predicted values RF.....	42
Table 7 Top important attributes RF.....	45
Table 8 actual and predicted values XGB.....	51
Table 9 Top important attributes XGB.....	53
Table 10 actual and predicted values ANN.....	59
Table 11 actual and predicted values ENSEM.....	62

## List of figures

Figure 1 Data pre-processing and machine learning process .....	37
Figure 2 Actual and predicted values LR scatter plot .....	39
Figure 3 Top important attributes LR .....	41
Figure 4 Actual and predicted values RF scatter plot .....	44
Figure 5 Top important attributes RF .....	46
Figure 6 PDP BIO_S11 RF .....	48
Figure 7 PDP ENG_S11 RF .....	49
Figure 8 PDP CR_S11 RF .....	50
Figure 9 actual and predicted scatter plot XGB .....	52
Figure 10 Top important attributes XGB .....	54
Figure 11 PDP ENG_S11 XGB .....	55
Figure 12 PDP BIO_S11 XGB .....	57
Figure 13 PDP CR_S11 XGB .....	58
Figure 14 actual and predicted scatter plot ANN .....	61
Figure 15 actual and predicted scatter plot ENSEM .....	63

## **Chapter 1. Introduction**

### **1.1. Background study and present state of the problem**

For educators, institutions, and students equally, the academic success of undergraduate engineering students is of utmost significance. The future careers of students, institution rankings, and the Caliber of the engineering personnel are all significantly influenced by it. It enables educators to identify students who may require additional support, resources, or interventions. Furthermore, accurate predictions can aid in the evaluation and improvement of academic programs and curricula. Machine learning (ML) has become a potent instrument for scholastic performance prediction and the identification of success-related factors in recent years. This preliminary research seeks to present a summary of the state-of-the-art ML methods used in college engineering programs to forecast scholastic success, the variables considered in such models, and the possible advantages of these predictions.

### **1.2. Machine Learning Techniques**

The ability of machine learning methods to evaluate large and complicated datasets, find invisible patterns, and produce precise forecasts has increased their appeal in educational research (Kotsiantis et al., 2003). Several supervised ML techniques, such as regression-based approaches, decision trees, support vector machines (SVMs), and artificial neural networks (ANNs), have been used to forecast academic success in engineering schools (Kelleher et al., 2015).

To determine relationships between students' academic success and various variables, such as demographic data, previous academic performance, and participation in learning activities, regression-based techniques, such as linear regression and logistic regression, are frequently used (Marbouti et al., 2016). To find combinations of variables that result in various degrees of success and visualize them, decision trees and random forests have been used by (Romero & Ventura, 2010).

To find at-risk students and provide them with targeted assistance, support vector machines have been used to divide students into different groups based on their expected scholastic success. Due to their capacity to recognize intricate and nonlinear connections between input characteristics and target variables, artificial neural networks, in particular deep learning models, have demonstrated encouraging outcomes in the prediction of scholastic success (Dien et al., 2020).

### **1.3. Factors Considered in Machine Learning Models**

To forecast academic success in engineering schools, various variables have been considered in machine learning models. These elements can be roughly divided into three categories: non-academic, academic, and socioeconomic (Abu Saa et al., 2019).

- Socioeconomic: Age, gender, race, and financial position are examples of socioeconomic variables that has an impact on students' scholastic success. These elements are frequently included in machine learning algorithms to account for possible biases in the data.
- Academic variables: Prior scholastic accomplishment is thought to be a powerful indicator of students' success in engineering schools, including high school GPA, standardized test results, and marks in required classes.

- Non-academic factors: psychological, social, and environmental factors like self-efficacy, time management, and social support can affect students' scholastic success. These elements can be included in machine learning models to give a more complete knowledge of the variables influencing students' success in engineering programs.

#### **1.4. Potential Benefits of Predicting Academic Achievement**

Using ML to predict academic achievement can help a variety of stakeholders in several ways:

- ❖ For educators: Early detection of at-risk students enables educators to offer focused assistance and solutions, enhancing student retention and total success rates (Adnan et al., 2021).
- ❖ For institutions: Accurate forecasts of students' academic achievement can guide resource allocation and program design, creating more productive teaching and learning settings. Institutions can also use these forecasts to assess the success of their initiatives and come to data-driven decisions for enhancement (Costa et al., 2017)
- ❖ For Students: predictive models can help them make educated choices about their study strategies and seek the proper assistance, when necessary, by using predictive models to better understand their strengths and weaknesses (Davidson et al., 2012).
- ❖ For policymakers: they can use the knowledge obtained from ML models to develop tailored policies that will advance equality and inclusion in engineering education and remove structural obstacles that prevent minority student groups from succeeding.

The ability of machine learning techniques to forecast academic success in undergraduate engineering schools has shown tremendous promise. When these strategies are combined with demographic, academic, and non-academic variables, they can offer insightful

information about the variables influencing students' performance and help guide focused initiatives to enhance student results.

### **1.5. Motivation of the Research**

There are several reasons why predicting academic achievement in undergraduate engineering programs using machine learning techniques is important. For starters, forecasting academic success can assist institutions in identifying students who are at danger of failing or dropping out of their course. When at-risk students are identified early, universities may give tailored assistance and interventions to help them succeed in their studies.

Second, forecasting academic accomplishment can assist educational institutions in determining factors that influence student success. Universities may establish programs and interventions that are personalized to the requirements of their students by evaluating data and determining the elements that are most closely connected with academic performance. This has the potential to enhance overall student results and guarantee that graduates are well-prepared for future professions.

Third, forecasting academic accomplishment can assist employers in identifying graduates who are likely to succeed in their chosen sector. Employers may discover graduates who have exhibited academic distinction and have the potential to become high-performing workers by employing machine learning algorithms to examine academic performance data. This can assist to improve workforce quality and guarantee that graduates are well-suited to their desired career pathways.

Finally, utilizing machine learning approaches to predict academic success can assist to stimulate educational innovation. Researchers can obtain a better understanding of the elements that contribute to student performance by constructing and refining machine learning models to predict academic accomplishment. This can help to inform the development of new teaching methods and strategies that are more effective at improving student outcomes.

### **1.6. Research Objectives**

The main objective of this research is to create machine learning models that reliably predict student performance at the conclusion of their undergraduate engineering studies.

Specific objectives:

- To develop a machine learning models that reliably predict students' performance and identify features that impact students' academic achievement.
- To compare machine learning models based on their accuracy in forecasting student achievement.



## Chapter 2. Literature review

In 1970 there was a need for a pictorial method for predicting student achievement in engineering technology programs, based on Composite American College Test Score (ACT scores) and high school grades, and motivated by the need to assist students in determining their odds of success, assisting students in defining their objectives, and indicating necessary institutional actions. Hazard (1974) used Regression analysis was used to evaluate the relationship between a composite of high school grades, ACT scores and college success.

Due to the low degree of accuracy of the methodology, employing merely regression analysis or probability to forecast academic accomplishment was ineffective. Later researchers began utilizing machine learning methods to determine the academic achievement. Yakubu & Abubakar (2022) used logistic regression and found an accuracy of 84.7%, both logistic regression and vector machines were used and discovered that the first is less accurate than the latter by (Iraqi et al., 2020).

Doleck et al. (2020) compared the performance of machine learning and deep learning algorithms and present the results as a comparative evaluation. For the MOOC dataset, the predictive accuracy of different machine learning algorithms ranged from 63.04% to 69.31%. For the CEGEP dataset, the predictive accuracy of different machine learning algorithms ranged from 84.16% to 90.60%. They overall findings suggest that machine learning algorithms achieve prediction performance like deep learning algorithms.

This study aimed to predict the fifth year and cumulative grade point averages (CGPA) of engineering students in a Nigerian university using data mining techniques. The Konstanz

Information Miner (KNIME) based data mining Tree Ensemble achieved an accuracy of 87.884%, while the Decision Tree predictor had the third-best accuracy (87.85%), and the Random Forest predictor had the fourth-best accuracy (87.70%). The Naive Bayes and PNN predictors achieved accuracies of 86.438% and 85.89%, respectively. Regression models yielded R2 values of 0.955 and 0.957 for linear and pure quadratic models, respectively, indicating that students' CGPA can be predicted based on their GPA performance in the first three years of study. The Logistic Regression algorithm achieved a maximum accuracy of 89.15%, while the PNN algorithm had the least accuracy of 85.895% (Adekitan et al., 2019). The performance of various machine learning algorithms based on their accuracy criteria. The algorithms compared in this study were Random Forest, Naïve Bayes, Multilayer Perceptron, Support Vector Machine, and DT-J48. The results showed that Multilayer Perceptron had the highest accuracy (76.07%), followed by Support Vector Machine (75.40%), DT-J48 (73.60%), Random Forest (67.40%), and Naïve Bayes (64.40%). This indicated that Multilayer Perceptron and Support Vector Machine are the most effective algorithms for the given task, while Naïve Bayes and Random Forest performed relatively poorly. DT-J48 felled in the middle with an accuracy of 73.60% (Jalota, &Agrawal ,2019).

This study displays the outcomes of utilizing the top classifiers from distinct families in boosting and bagging methods either separately or as functions. The classifiers' effectiveness is assessed based on accuracy, F-measure, and ROC metrics, and the results demonstrated that the boosting method performed better than the bagging method. The J48 classifier exhibited consistent performance when used as a function in the Adaboost\_J48 boosting method, increased its accuracy by 4.1% from 0.943 to 0.983. Thus, the J48 classifier was selected as the first function to be combined with the best overall classifier in the EMT model (Almasri et al., 2019).

Livieris et al. (2019) explored the efficacy of two wrapper methods, self-training and YATSI, for semi-supervised learning to predict the academic performance of high-school students in final exams. The performance of these methods was evaluated against classic supervised techniques and two popular semi-supervised algorithms. The paper used various assessment criteria, such as written assignments, oral exams, short tests, and exams during the academic year, to evaluate the final grade using semi-supervised learning methods, which yielded high accuracy based on the experimental results. The study concluded that semi-supervised algorithms could enhance classification accuracy by utilizing a limited number of labelled and numerous unlabelled data to build dependable prediction models.

Academic institutions are concerned about student achievement, and data generated by learning management systems can be used to improve academic performance. A hybrid algorithm that combines clustering and classification was proposed and applied to student data, revealing a strong relationship between student behaviour and academic performance. The proposed model achieved an accuracy of 0.7547 when applied to academic, behaviour, and extra features of the student data, outperforming existing algorithms. This model can help educators identify weak learners and improve the learning process, while administrators can better manage the learning system. The model can be extended to support a wider range of student dataset features in the future (Francis & Babu, 2019).

Waheed et al. (2020) utilized deep artificial neural network to predict which students were at risk by analysing unique handcrafted features extracted from clickstream data in virtual learning environments. This information can be used to intervene early and provide support to at-risk students. The results indicated that the proposed model achieved a high

classification accuracy of 84%-93%. Additionally, the study showed that the deep artificial neural network performed better than the logistic regression and support vector machine models, which achieved classification accuracies of 79.82%-85.60% and 79.95%-89.14%, respectively.

Hasan et al. (2020) proposed a data classification model to predict student academic performance using data from Moodle and edify. The dataset contained 772 samples from one academic year, with 18 features and one meta-attribute used to form the dataset. The Tree-based classification model, specifically Random Forest, outperformed the other techniques with an accuracy of 88.3%, using equal width data transformation and information gain ratio selection technique. Feature reduction using genetic algorithm and PCA was inconclusive, but multivariate analysis identified nine variables that successfully predicted academic performance. The CN2 Rule Inducer algorithm achieved 87.4% accuracy and was easier to interpret for non-expert users such as faculty.

The studies aimed to predict student academic performance and pass/fail outcomes using classic data mining algorithms and Auto-WEKA. The experiments were conducted on datasets from three different courses, with each course split into chronological sets. The results showed that Bagging, Random Forest, and SMOreg were among the best performers in predicting student grades and pass/fail outcomes, with tree-based methods being primarily suggested by Auto-WEKA. This research also demonstrated that using Auto-WEKA enhanced prediction accuracy and mean absolute error considerably. Overall, the experiments demonstrated the ability of data mining and Auto-WEKA to predict and improve student academic achievements (Tsiakmaki et al., 2020).

Mengash (2020) intended to enhance university admission choices by employing data mining to forecast applicants' academic performance. The methodology was validated using data from 2,039 students enrolled in a Saudi public university. Results showed that applicants' early university performance can be predicted based on certain pre-admission criteria, with the Scholastic Achievement Admission Test score being the most accurate predictor. Artificial Neural Network technique had an accuracy rate above 79%, making it the most superior classification technique compared to Decision Trees, Support Vector Machines, and Naïve Bayes. The study recommended assigning more weight to the Scholastic Achievement Admission Test score in admission systems.

Hooshyar et al. (2020), in his study a new algorithm called PPP was developed to predict the performance of students with learning difficulties based on their procrastination behaviour, considering pre-due date behaviour as well as late or non-submissions. The algorithm uses feature vectors to label students as procrastinators, procrastination candidates, or non-procrastinators and applies classification methods to predict performance. Results from a course with 242 students showed that PPP accurately predicted performance with 96% accuracy, and linear support vector machine was the best classifier for continuous features, while neural network performs better for categorical features.

This paper presented a two-phase machine learning approach that combined unsupervised and supervised learning techniques to predict outcomes for students in higher education programs. The approach was tested on a case study of undergraduate computer science students at the University of Thessaly in Greece. The students were initially clustered based on education-related factors and metrics using the K-Means algorithm, resulting in three coherent clusters. Two machine learning models were then trained for each cluster to predict

time to degree completion and student enrolment in the educational programs. The paper suggests that the clustering-aided approach could be useful for learning analytics in higher education. The developed models were found to produce predictions with relatively high accuracy (Iatrellis et al., 2021).

In this paper YILDIZ & BÖREKÇİ (2020), educational data from ninth-grade students was analysed using data mining methods to develop insight into demographic information, studying routines, attending learning activities, and epistemological beliefs about science. The aim was to solve a classification problem and estimate the success of students in the exam. The supervised classification algorithms were compared, and the Neural Network algorithm was found to have the highest accuracy rate (98.6%). The study revealed that demographic variables of the family, scientific epistemological beliefs of the student, study routines, and attitudes towards some courses affected the classification. These findings can help support students' academic success by understanding the relationship between these variables and academic success.

Academic institutions and educators consider it crucial to analyse students' academic performance to identify ways to enhance individual student performance. This project (Oyedeji et al., 2020) involved examining past performance records of students, including their age, demographic distribution, family background, and study attitudes. Machine learning tools were employed to analyse this data, and three models were tested, including Linear regression for supervised learning, linear regression with deep learning, and neural network. The test and train data were used to evaluate the models, and the results showed that Linear regression for supervised learning had the lowest mean average error (MAE) of 3.26

Alsaman et al., (2019) examined the use of data mining techniques to predict the academic performance of Jordanian university students. They used Decision Tree and Artificial Neural Network classification techniques to build a model that predicts students' expected GPA. They collect data through an online questionnaire and select relevant attributes to test their correlation with academic performance. The study finds that MLP classifier in ANN has the highest accuracy of up to 97%, indicating the potential of ANN in predicting student academic performance.

The study examined various ML algorithms for predicting student academic performance in STEM courses, including Linear Regression, Logistic Regression, k-Nearest Neighbor Classification, Naïve Bayes, Artificial Neural Network, Decision Tree, Random Forest, and Support Vector Machine. The outcomes of ML-based analytics depended on input dataset size, type of data, selected ML algorithm focus, and algorithm setup. Linear regression had the best accuracy with an average error of 3.70% for predicting student performance based on individual assignment scores. SVM and Random Forest had the second-best accuracy, with an error range of 6-7%, and were recommended for ML-based predictive analytics in education (shmawy et al., 2019)

Supportive learning has been found to enhance educational quality, with school and family tutoring offering personalized help and positive feedback to improve students' understanding. Predicting student performance is important for building a strong foundation for post-secondary studies and career success. This paper proposed an improved algorithm, ICGAN-DSVM, based on deep support vector machines, to predict student performance under supportive learning through school and family tutoring. With low sample sizes in students' academic datasets, ICGAN-DSVM increases data volume and enhances prediction accuracy.

Results showed that the proposed algorithm outperforms related works by 8-29% in terms of specificity, sensitivity, and AUC (Chui et al., 2020).

Educational data generated from various platforms can be analysed using educational data mining techniques to gain insights into student performance. Predicting student performance is a desirable application of educational data mining, and there is a need for automated techniques. Previous studies primarily use conventional feature representation schemes, but recent advancements in deep learning allow for automatic extraction of high-level features. In this work, the attention based Bidirectional Long Short-Term Memory network was used to predict student performance from historical data, achieving a 90.16% prediction accuracy. This technique has important applications for universities and government departments in early performance prediction. The proposed method outperforms existing state-of-the-art techniques (Alshantiti & Namoun, 2020).

The growing amount of data from institutional technology, e-learning resources, and virtual courses can be used by educators to understand students' learning behaviours. Educational Data Mining (EDM) can extract hidden information from raw data to predict students' academic performance with high accuracy. A hybrid 2D CNN architecture was developed to predict whether students would pass or fail a class, achieving 88% accuracy, outperforming previous models. Future research includes exploring the impact of different performance metrics and features on academic performance and employing explainable AI for smaller datasets. Poudyal et al., (2022) demonstrated the potential for using CNN architecture to anticipate student academic performance and assist them accordingly.



Yağcı (2022) presented a new machine learning model that uses educational data mining to predict the final exam grades of undergraduate students based on their midterm grades. The model compares the performance of various machine learning algorithms, such as random forests and logistic regression, to achieve an accuracy of 70-75%. The dataset included academic achievement grades of 1854 students taking the Turkish Language-I course at a state university in Turkey during the fall semester of 2019-2020. This study contributes to early identification of students at high risk of failure and determining the most effective machine learning methods in higher education decision-making.

The goal of Alamri et al., (2020) was to use classification algorithms, specifically Support Vector Machines (SVM) and Random Forest (RF), to predict the academic performance of students and improve the results of educational organizations. They used binary classification and regression techniques to predict the final grades of mathematics and Portuguese language courses using datasets of 369 and 649 records, respectively. The experimental results indicated that both SVM and RF algorithms achieve high levels of accuracy, with a superior accurate prediction of up to 93% for binary classification and the lowest RMSE of 1.13 in the case of RF for regression.

Data mining is widely used in various fields, including education, where it is known as Educational Data Mining (EDM). Educational institutions can utilize the data contained in higher education to analyse student performance and anticipate problems that may cause delays in the study period. In this study, two algorithm models, to predict student performance, K-Nearest Neighbor and Decision Tree C4.5 were utilized. The best accuracy rate was 59.32% for the K-Nearest Neighbor algorithm model, and 54.80% for the Decision

Tree C4.5 model. The study highlights the possibility for employing EDM to contribute to educational progress through data mining (Yulianto et al., 2020).

Zhang et al. (2021) planned to create a deep learning-based model dubbed Sparse Attention Convolutional Neural Networks (SACNN) to predict student grades in Chinese universities. Sparse attention layers, convolutional neural layers, and a fully connected layer comprised the model. The sparse attention layers considered the varying contributions of courses to the grade prediction, the convolutional neural layers captured the temporal features of the courses, and the fully connected layer classified the achieved features. The model was evaluated using a dataset of 54k grade records from 1307 students and 137 courses, achieving 81% prediction precision and 85% accuracy on failure prediction. The model also provided an explanation for the predicted results.

There has been considerable development in the application of machine learning techniques to the field of educational data mining during the last decade. Many machine learning models have been used to evaluate datasets from educational institutions all around the world, with excellent accuracy in classification and prediction tasks. Notably, various machine learning algorithms have emerged as strong tools in this sector, including Linear Regression, Xgboost, Random Forest, and Artificial Neural Networks.

Linear regression, a basic machine learning method, is commonly employed in educational data mining. It uses statistical approaches to construct correlations between variables, which makes it useful for forecasting student performance, assessing teaching methods, and discovering factors that impact educational results. Its ease of use and interpretability make it an appealing option for evaluating educational statistics.

Because of its stability and capacity to handle complicated datasets, Xgboost, an advanced gradient boosting technique, has gained favour. To improve accuracy, it merges many

decision trees and improves their predictions. Xgboost has been effectively used in educational data mining to predict student outcomes such as graduation rates or academic performance by considering characteristics such as demographics, historical achievements, and engagement measures.

Random Forest, another ensemble learning approach, is frequently used in educational data mining because of its capacity to handle high-dimensional data and capture complicated correlations between factors. Random Forest excels at classification tasks such as identifying at-risk pupils and recognizing learning trends in huge educational datasets by creating a slew of decision trees and aggregating their predictions.

Artificial Neural Networks (ANNs), which are modelled after the structure of the human brain, have revolutionized machine learning and found widespread use in educational data mining. Artificial neural networks (ANNs) are made up of linked layers of artificial neurons that analysed and learn from input. Their capacity to capture nonlinear correlations and adapt to complicated patterns has proven useful in a variety of educational activities such as student performance prediction, recommendation systems, and anomaly identification.

We want to use the aforementioned machine learning models, notably Linear Regression, Xgboost, Random Forest, and Artificial Neural Networks, in this study to assess and forecast our dataset. We anticipate getting significant insights into the elements impacting educational results, recognizing patterns, and generating accurate forecasts in our individual case by utilizing these strong algorithms.

It is vital to highlight that the selection and application of these machine learning models will be determined by the dataset's particular characteristics, and study objectives. We think that by using the potential of these sophisticated methodologies, we may add to the increasing

body of knowledge in educational data mining and improve our understanding of the elements that influence educational performance.

## Chapter 3. Methodology

### 3.1. Machine learning algorithm

Machine learning algorithms are a subset of artificial intelligence that include creating models or programs that can learn from data and anticipate or make judgments without being explicitly programmed. These algorithms evaluate and learn from data using statistical approaches, and their performance improves with time. Image and audio recognition, natural language processing, predictive analytics, and recommendation systems are all examples of how machine learning algorithms are applied.

One of the most important properties of machine learning algorithms is that they are data-driven, which means that they need a huge quantity of data to train and increase their accuracy. There are three types of algorithms: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning entails training a model on a labelled dataset in which the outcome variable for each input sample is known. The system learns to relate inputs to outputs and may then anticipate new, previously unknown data. Supervised learning is often used for classification and regression problems, such as forecasting a house's price based on its attributes.

The ImageNet Large Scale Visual Recognition Challenge, an annual competition that attempted to advance the science of computer vision by pushing academics to construct models that could classify images into 1,000 separate categories, is an example of supervised learning (Russakovsky et al., 2015).

Unsupervised learning, on the other hand, entails training a model on an unlabelled dataset with an unknown output variable. The algorithm learns to recognize patterns and structures in data, such as grouping together similar data points. Unsupervised learning is often used for exploratory data analysis and dimensionality reduction.

Anomaly detection in complex systems, such as cybersecurity, is an intriguing use of unsupervised learning. Unsupervised learning algorithms may be used in this application to learn patterns of typical behaviour in a system and detect variations from those patterns that may signal a security breach (Vikram & Mohana, 2020).

Reinforcement learning entails teaching a model to do actions in an environment to maximize a reward signal. The algorithm learns by trial and error, experimenting with various behaviours and gaining knowledge from the input it gets. Reinforcement learning is widely employed in robotics and video games.

A study undertaken by NVIDIA researchers built a reinforcement learning algorithm to operate a virtual automobile in a racing game as an example of this sort of application case. After trial and error, the algorithm learnt to navigate the course and improve its lap timings (Bojarski et al., 2016).

### **3.1.1 Linear Regression**

Linear regression is an effective machine learning approach for predicting the connection between one or more independent variables and a dependent variable. It is frequently utilized in a variety of sectors such as finance, economics, social sciences, and marketing. Linear

regression, according to Hastie et al (2017), is a basic yet effective approach that has been frequently utilized for over a century.

Machine learning linear regression is like classic statistical linear regression in that it predicts the value of the dependent variable based on the values of the independent variables. Linear regression is used in machine learning for both regression and classification issues. The purpose of regression is to predict a continuous variable, whereas the goal of classification is to predict a categorical variable.

In machine learning, the linear regression equation is expressed as  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \epsilon$ , where  $y$  represents the dependent variable,  $x_1, x_2, \dots, x_n$  identifies the independent variables,  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  represent the coefficients, and  $\epsilon$  is the error term. The coefficients are calculated using a training set of data, and the model that results may be used to forecast fresh data.

Ordinary least squares (OLS), gradient descent, and stochastic gradient descent are all techniques used to determine the coefficients in linear regression. OLS is a popular approach for minimizing the sum of squared errors between anticipated and actual values. The iterative optimization procedures gradient descent and stochastic gradient descent alter the coefficients to minimize the cost function.

The interpretability of linear regression is one of its benefits. The coefficients represent the change in the dependent variable resulting from a one-unit change in the independent variable while maintaining all other variables constant. This simplifies understanding the connections between variables and making predictions based on the model.

Linear regression, on the other hand, has certain restrictions. It assumes a linear connection between the dependent and independent variables, which may or may not be correct in all

circumstances. It also presumes that the error terms are normally distributed with constant variance, which may not be the case for all datasets. Montgomery et al. (2012) argues that violating any of these assumptions might result in biased and inefficient coefficient estimations.

Finally, linear regression is a strong machine learning approach that may be used to describe the connection between a dependent variable and one or more independent variables. Linear regression may be used for both regression and classification problems, and the coefficients can be estimated in a variety of ways. While linear regression has significant drawbacks, it is an effective tool for analysing variable connections and generating predictions based on the model.

### **3.1.2 Random Forest**

Random Forest is a machine learning technique used for classification, regression, and feature selection. It is a form of ensemble learning method that makes predictions by combining numerous decision trees. A random portion of the training data and a random subset of the input characteristics are used to create each tree in the forest. Because of its excellent accuracy, resilience, and capacity to handle high-dimensional data, Random Forest has grown in prominence.

Random Forest's central idea is to minimize the variation of decision trees by incorporating randomization into the tree-building process. By selecting a random subset of the data and attributes, the approach provides a diverse group of trees that are less likely to overfit the data. Each tree in the forest produces a unique prediction during prediction, and the final prediction is the mode (for classification) or mean (for regression) of all the trees' forecasts.



Random Forest has been applied in a wide range of fields, including bioinformatics, finance, image classification, and educational data mining. Random Forest has been used in bioinformatics to predict protein structure and gene expression (P. Wang et al., 2021). It has been used in finance to detect fraud and to provide credit ratings (Shiyang et al, 2018). Random Forest has been used in image classification to recognize objects and categorize scenes (Xia et al., 2018), and Vijayalakshmi & Venkatachalapathy, (2019) used it to predict student academic progress in educational data mining

Recent research has concentrated on improving the performance and interpretability of Random Forest. To reduce tree bias, one method is to modify the tree-building process. For example, the Extremely Randomized Trees (ERT) approach constructs trees by randomly selecting the splitting point for each feature rather than searching for the ideal split point (Geurts et al., 2006). This reduces tree bias and may improve Random Forest performance. Another possibility is to incorporate feature importance or domain knowledge into the tree-building process. The Random Forest method with Dominant Feature Selection (RF-DFS). This can increase model interpretability while decreasing computational expense.

Recent research has concentrated on expanding Random Forest to new activities and places, as well as improving the algorithm itself. For example, Random Forest has been used to anticipate the toxicity of chemicals and drugs (Wu & Wang, 2018), as well as the outcome of Surgery (Merali et al., 2019).

Random Forest is a powerful machine learning approach that has been widely applied in a variety of applications and fields. Its excellent accuracy and ability to handle high-dimensional data make it a popular choice for a wide range of applications. The current focus

of research is on improving the algorithm's performance and interpretability, as well as applying it to new applications and domains.

### **3.1.3 Naïve Bayes**

Naive Bayes is a machine learning algorithm used for text classification, spam filtering, sentiment analysis, and educational data mining. It is a probabilistic method that uses Bayes' theorem to classify new instances based on their properties. The approach is simple and efficient, using only a little amount of training data to get correct results.

The Naive Bayes approach is founded on Bayes' theorem, which states that the probability of a hypothesis given some observed evidence is proportional to the likelihood of the evidence given the hypothesis multiplied by the prior probability of the hypothesis. In other words, it computes the probability of a certain occurrence based on previous knowledge of factors that may be related with the event. The method implies that the data characteristics are independent of one another, thus the term "naive."

The algorithm presumes that Naive Bayes algorithms are classified into three types: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Gaussian Naive Bayes assumes that the features are normally distributed, whereas Multinomial Naive Bayes is used for discrete data, such as text classification, and Bernoulli Naive Bayes is used for binary classification problems.

The simplicity and effectiveness of the Naive Bayes algorithm are two of its main features. The system can be trained quickly on large datasets and provide predictions in real time. It is also resistant to irrelevant properties, making it suitable for large datasets. Another benefit is that it requires little training data to get trustworthy results, making it appropriate for applications with little data.

Recent research has demonstrated that Naive Bayes is a trustworthy algorithm for text classification applications, notably sentiment analysis. In research done by Wang (2020), Naive Bayes was compared with various machine learning algorithms for sentiment categorization of Chinese microblogs. According to the findings, Naive Bayes outperforms other algorithms in terms of accuracy and efficiency.

Kontsewaya et al. (2021) did another research that compared Naive Bayes to different machine learning methods for spam filtering. The study discovered that Naive Bayes outperformed other algorithms in terms of accuracy and efficiency, especially for datasets with many characteristics.

Finally, Nahar et al. (2021) employed naive bayes among other machine learning algorithms to predict student accomplishment, with naive bayes achieving the highest accuracy among all methods.

Hence, Naive Bayes is a simple and effective algorithm that is popular option for many machine learning applications tasks due to its tolerance to irrelevant variables and ability to generate correct results with a minimal quantity of training data.

### **3.1.4 XGBoost**

XGBoost (Extreme Gradient Boosting) is a well-known machine learning technique that has had a lot of success in data science contests and real-world applications. It is an ensemble approach that makes predictions using many decision trees and combines them to increase

accuracy. Because of its speed, scalability, and accuracy, XGBoost has become a go-to method for many machine learning problems. According to Chen and Guestrin (2016), XGBoost beat other prominent algorithms like Random Forest and Neural Networks on a variety of datasets.

Chen and Guestrin initially announced XGBoost in 2016, and it has subsequently attracted substantial interest from the machine learning field. XGBoost is based on gradient boosting, which is a technique that adds decision trees to the model in a sequential manner, with each new tree correcting the errors caused by the preceding ones. XGBoost improves on gradient boosting by using a more regularized model formalization for better control over fitting and a complex parallelizing tree construction approach for greater computing efficiency.

One of the most important characteristics of XGBoost is its versatility. It provides a diverse set of loss functions and assessment measures and may be utilized for regression and classification applications. XGBoost can also handle numerical and categorical data and has support for missing values built in.

Another feature of XGBoost is its interpretability. XGBoost provides feature relevance ratings to help users determine which features are most important in predicting the target variable. XGBoost additionally includes visualization tools like tree plots and feature interaction plots to help users understand the model's decision-making process.

### **3.1.5 Artificial Neural Network**

Nielsen (2015) Artificial Neural Networks (ANNs) are defined as a subset of machine learning approaches inspired by the structure and operation of the human brain. ANNs are composed of connected nodes known as neurons that can analyse input and generate

predictions based on it. They are widely used in fields like as computer vision, natural language processing, and predictive modelling.

ANNs have been around since the 1940s, when Warren McCulloch and Walter Pitts created the first neuron model, which consisted of a simple on/off switch. ANNs, on the other hand, were not become widely used until the 1980s, due to the development of backpropagation, a method for training neural networks (Lecun et al., 2015). Since then, ANNs have risen in popularity due to their ability to learn from massive amounts of data, their versatility in processing many types of data, and their ability to generalize to new data.

One advantage of ANNs is their adaptability to diverse types of data. ANNs can handle a wide range of data types, including numerical, category, and text data, making them adaptable to a wide range of applications. For example, in finance, ANNs may be used for predictive modelling to forecast stock values based on past data. In healthcare, ANNs may be used to predict illness outcomes based on patient data and it can be used in education to predict the student academic achievement based of the student's data.

## **3.2. The performance evaluation metrics**

### **3.2.1. Accuracy**

This metric measures the percentage of correctly classified instances by the model. It is calculated as  $(TP + TN) / (TP + TN + FP + FN)$ . Accuracy is useful when the number of positive and negative instances in the dataset is roughly balanced.

### **3.2.2. Recall**

This metric measures the proportion of true positive predictions out of all actual positive instances in the dataset. It is calculated as  $TP / (TP + FN)$ . Recall is useful when the cost of false negatives is high

### **3.2.3. F1 Score**

This metric is the harmonic mean of precision and recall. It is useful when the dataset is imbalanced, and one metric alone cannot effectively evaluate the model's performance.

### **3.2.4. Mean Absolute Error (MAE)**

This metric measures the average absolute difference between predicted and actual values. It is useful when the target variable has a linear relationship with the features.

### **3.2.5. Mean Squared Error (MSE)**

This metric measures the average squared difference between predicted and actual values. It is useful when the target variable has a non-linear relationship with the features.

## **3.3. DATA COLLECTION**

The data will be using here presents the results of national assessments for engineering students in secondary and university education in Colombia. The dataset includes academic, social, and economic information for 12,411 students. The data were obtained by merging databases from the Colombian Institute for the Evaluation of Education (ICFES) (Delahoz-Dominguez et al., 2020). The observations reflect results from two educational stages:

secondary and professional evaluations, along with social context variables of the students' living environment.

The first moment of evaluation corresponds to the secondary evaluation of engineering students in Colombia. In this evaluation, students are tested on a variety of subjects, including mathematics, science, and social studies. The findings of this review are an important indicator of the secondary school system's success in preparing students for further education in engineering.

The second evaluation point refers to the professional evaluation of engineering students in Colombia. This assessment measures students' mastery of technical skills and information necessary for professional activity. The examination includes civil, industrial, and mechanical engineering, as well as electrical engineering and telecommunications, Mechatronics engineering, textile engineering, topographic engineering, and aeronautical engineering. The findings of this review are critical in determining the efficacy of university-level engineering programs in Colombia.

The dataset includes social and economic information about the students in addition to academic information. This information gives insight into the students' social background and can aid in identifying potential hurdles to success in engineering school. The dataset contains information on the students' socioeconomic status, family background, such as the parents' education level and occupation, as well as the range of their incomes, the number of people living in the house, and access to resources such as computers, the internet, a washing machine, a car, and a microwave oven.

**Table 1: Academic variables**

<b>Variable</b>	<b>Full name</b>
MAT_S11	Mathematics
CR_S11	Critical Reading
CC_S11	Citizen Competencies S11
BIO_S11	Biology
ENG_S11	English
ENG_PRO	English
WC_PRO	Written Communication
FEP_PRO	Formulation of Engineering Projects
QR_PRO	Quantitative Reasoning
CR_PRO	Critical Reading
G_SC	Global Score
PERCENTILE	Percentile
2ND_DECILE	Second Decile
QUARTILE	Quartile
CC_PRO	Citizen Competencies SPRO
SEL	Socioeconomic Level
SEL_IHE	Socioeconomic Level of The Institution of Higher Education

The table contains a collection of variables and their complete names from national examinations given to engineering students in secondary and university education. The first set of factors comprises academic topics assessed in secondary school, such as Mathematics,



Critical Reading, Citizen Competencies, Biology, and English. The second set of factors comprises academic disciplines assessed throughout the professional education stage, such as English, Written Communication, Engineering Project Formulation, and a Global Score. Variables such as Percentile, Second Decile, Quartile, Citizen Competencies in the Professional Stage, and Socioeconomic Level are also included in the table.

**Table 2 Socioeconomic variables**

Variable	Full Name	Levels	Variable	Full Name	Levels
GENDER	Gender	2	DVD	DVD	2
EDU_FATHER	Father's education	12	FRESH	Fresh	2
EDU_MOTHER	Mother's education	12	PHONE	Phone	2
OCC_FATHER	Father's occupation	13	MOBILE	Mobile	2
OCC_MOTHER	Mother's occupation	13	REVENUE	Revenue	3
STRATUM	Stratum	7	JOB	Job	8
SISBEN	Sisben	6	SCHOOL_NAME	School name	3,735
PEOPLE_HOUSE	People in the house	13	SCHOOL_NAT	Nature of School	2
INTERNET	Internet	2	SCHOOL_TYPE	Type of School	4

The table displays a collection of variables, their complete names, and the levels to which they relate. The variables in the datasets are connected to numerous demographic and socioeconomic aspects. The first set of variables includes gender, fathers and mother's education level, fathers and mother's occupation, stratum, and Sisben score. The second set of variables includes whether the household has access to the internet, and whether they have a DVD player, fresh produce, a mobile phone, and revenue level. The third set of variables includes job type, school name, the nature of the school, the number of people in the house, and the type of school. The levels for each variable vary, for example, gender only has two levels (male and female), while job type has eight levels.

**Table 3 Engineering programs**

<b>Academic Program</b>	<b>% Women</b>	<b>% Men</b>	<b>% Public School</b>	<b>% Private School</b>	<b>FEP_PRO</b>	<b>G_SC</b>
Civil constructions	42.86%	57.14%	85.71%	14.29%	154.36	151.86
Aeronautical Engineering	27.27%	72.73%	43.18%	56.82%	138.52	155.80
Cadastral Engineering and Geodesy	58.97%	41.03%	48.72%	51.28%	78.08	174.60
Civil Engineering	35.87%	64.13%	49.94%	50.06%	144.46	161.11
Control Engineering	41.67%	58.33%	50.00%	50.00%	163.42	177.08
Production Engineering	51.67%	48.33%	46.67%	53.33%	135.32	172.90
Productivity and	55.17%	44.83%	68.97%	31.03%	62.55	162.10

quality Engineering						
Transportation and road Engineering	48.15%	51.85%	96.30%	3.70%	172.19	167.74
Electric Engineering	21.94%	78.06%	51.44%	48.56%	139.98	173.99
Electromechanical Engineering	14.71%	85.29%	73.53%	26.47%	141.32	148.62
Electronic Engineering	19.55%	80.45%	55.95%	44.05%	145.87	166.87
Electric Engineering and telecommunications	19.15%	80.85%	42.55%	57.45%	149.53	160.43
Industrial Automation Engineering	36.36%	63.64%	68.18%	31.82%	160.41	166.09
Automation Engineering	30.00%	70.00%	20.00%	80.00%	160.30	165.50
Control Engineering	0.00%	100.00%	75.00%	25.00%	65.38	164.75
Control Engineering and industrial automation	0.00%	100.00%	0.00%	100.00%	94.00	113.00

The table shows the percentages of women and men enrolled in various academic programs, along with the percentage of students who attended public and private schools. It also displays the Global Score (G\_SC) and Formulation of Engineering Projects (FEP\_PRO)

values for each program. The academic programs range from Civil Constructions to Topographic Engineering.

The percentage of women enrolled in these programs ranges from 7.41% for Mechatronics Engineering to 100% for Textile Engineering. The percentage of men enrolled ranges from 0% for Textile Engineering to 92.59% for Mechatronics Engineering. Most of the programs have a higher percentage of men than women.

The percentage of students who attended public schools ranges from 20% for Automation Engineering to 96.30% for Transportation and Road Engineering. Most programs have a relatively balanced mix of students from public and private schools.

The G\_SC values for the programs range from 113.00 for Control Engineering and Industrial Automation to 177.08 for Cadastral Engineering and Geodesy. The FEP\_PRO values range from 62.55 for Productivity and Quality Engineering to 172.19 for Transportation and Road Engineering.

### **3.4. DATA PREPROCESSING**

In this research, we will use the Python programming language for data preparation, which has become a popular choice for data analysis and manipulation due to its wide libraries and strong capabilities. Python has several libraries, such as Pandas, NumPy, and SciPy, that provide comprehensive data preparation, cleaning, and transformation capabilities.

Pandas, a powerful data manipulation toolkit, will be essential in managing the dataset. It has efficient data structures and operations for loading, exploring, cleaning, and transforming

data. Pandas' broad range of data manipulation capabilities allows us to preprocess the dataset by managing missing values, reducing outliers, and doing feature engineering as needed.

Furthermore, we will use Google Colab to expedite the data preparation process and enable collaborative collaboration. Google Colab is a web-based tool that integrates a Jupyter Notebook environment with Google Drive. Google Colab provides free computing resources, including GPU support, which helps speed up computationally intensive operations like training machine learning models on huge datasets.

We can easily handle, clean, and convert the dataset for further analysis by utilizing the Python programming language and tools such as Pandas, NumPy, and SciPy. And, by using Google Colab, we can take advantage of its computer capabilities to speed up the data preparation and analysis process.

As a whole, the combination of Python and its powerful libraries, will provide us with a robust and efficient framework for data preparation in this study, allowing us to derive meaningful insights and draw accurate conclusions from the dataset.

#### **3.4.1. Data cleaning**

During the data preparation step, we concentrated on assuring the dataset's quality and relevance for our research. One critical aspect was dealing with missing data to avoid biases or errors in our results. Depending on the conditions, we used several ways to resolve missing values, such as imputation or elimination.

For example, we noticed columns in the dataset that were not required for our study, such as the student ID in high school and university. These columns provide no useful insights into the elements that influence educational results or forecasts. As a result, we decided to

eliminate these columns from the dataset to simplify our analysis and decrease superfluous noise.

In addition, we discovered a "unnamed: 9" column in the dataset that had no values. This column appears to be a data gathering artifact or a mistake during data recording. We opted to omit this column from our study since it supplied no relevant information.

Also, for our unique study aims, we selected percentile columns that were redundant. While percentiles can be useful in some studies, we discovered that they were not directly contributing to the specific predictions or insights we hoped to glean from the dataset. As a result, we chose to delete these percentile columns from our study to simplify it and focus on the most important attributes.

We verified that the dataset was simplified and optimized for our study by deleting these superfluous columns and eliminating missing data using appropriate procedures. We were able to work with a cleaner and more relevant dataset because of this strategy, which improved the accuracy and reliability of our following analyses and forecasts.

It is crucial to emphasize that the choice to eliminate columns or manage missing data was taken after thorough consideration of the unique study objectives, data type, and potential influence on analysis results. We hoped to achieve a more robust and concentrated dataset that would produce important insights and accurate predictions in our study by employing these data preparation strategies.

### **3.4.2. Data Encoding**

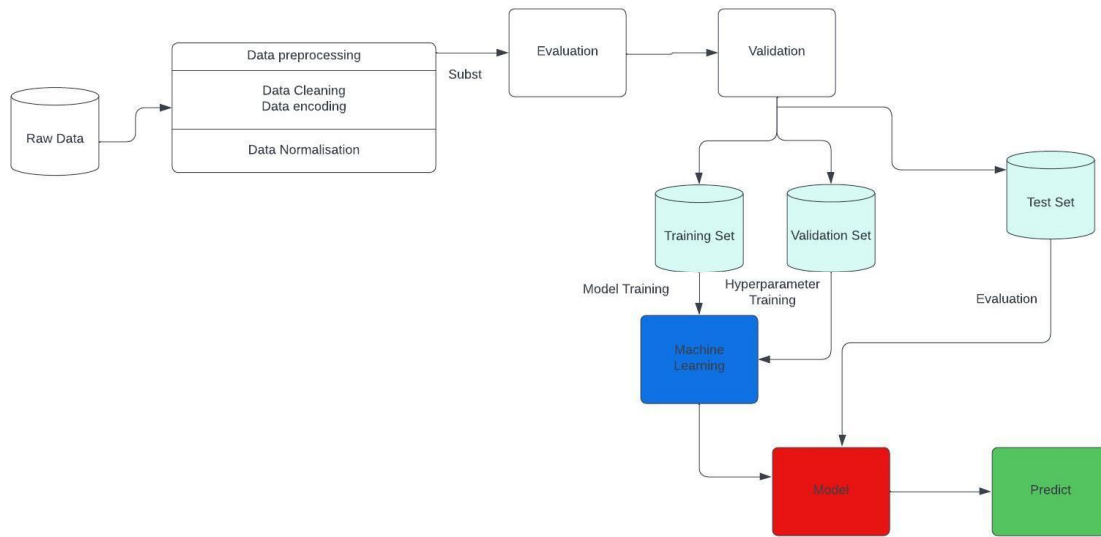
In our study, we utilize a regression model and other models to predict students' scores at the end of their engineering degree programs, therefore it is critical that all data used for modeling is in numerical format. However, categorical or discrete variables are popular, especially when addressing demographic and socioeconomic aspects.

We used strategies to translate categorical or discrete data into numerical representations to address this issue. This procedure, known as encoding or feature transformation, enables us to successfully include these variables into our regression model.

One-hot encoding is a popular approach for transforming category variables to numerical values. This method entails separating binary columns for each category inside a categorical variable. For example, if we had a categorical variable called "gender" with the categories "male" and "female," we would generate two binary columns, one for "male" and one for "female." These columns' values would be 1 or 0, signifying the existence or absence of each category for a given data item. This allows us to use the complete spectrum of available data to reliably forecast students' results at the end of their engineering degree programs.

### **3.4.3. Data Normalization**

The attribute columns that we are attempting to predict possess values that span a significant range, ranging from 0 to 300. This broad range of values can pose a challenge in achieving accurate predictions. Thus, to improve the precision of our predictions, it is imperative that we narrow down the range of values of the attribute that we wish to predict. To achieve this, we shall apply a data normalization technique, which involves scaling the range of values down to a more manageable range, say from 0 to 10. By doing so, we can obtain a more accurate and reliable prediction model that is better suited to handle the given data.



**Figure 1 Data pre-processing and machine learning process**



## Chapter 4. RESULTS AND DISCUSSION

### 4.1 Linear Regression

Our MSE using the linear regression model is 8.23. Mean Square Error (MSE) a frequently used metric in statistical modelling and machine learning to evaluate the average squared difference between a given dataset's predicted and actual values.

#### 4.1.1 Comparison between actual and predicted values

**Table 4 Actual and predicted values LR**

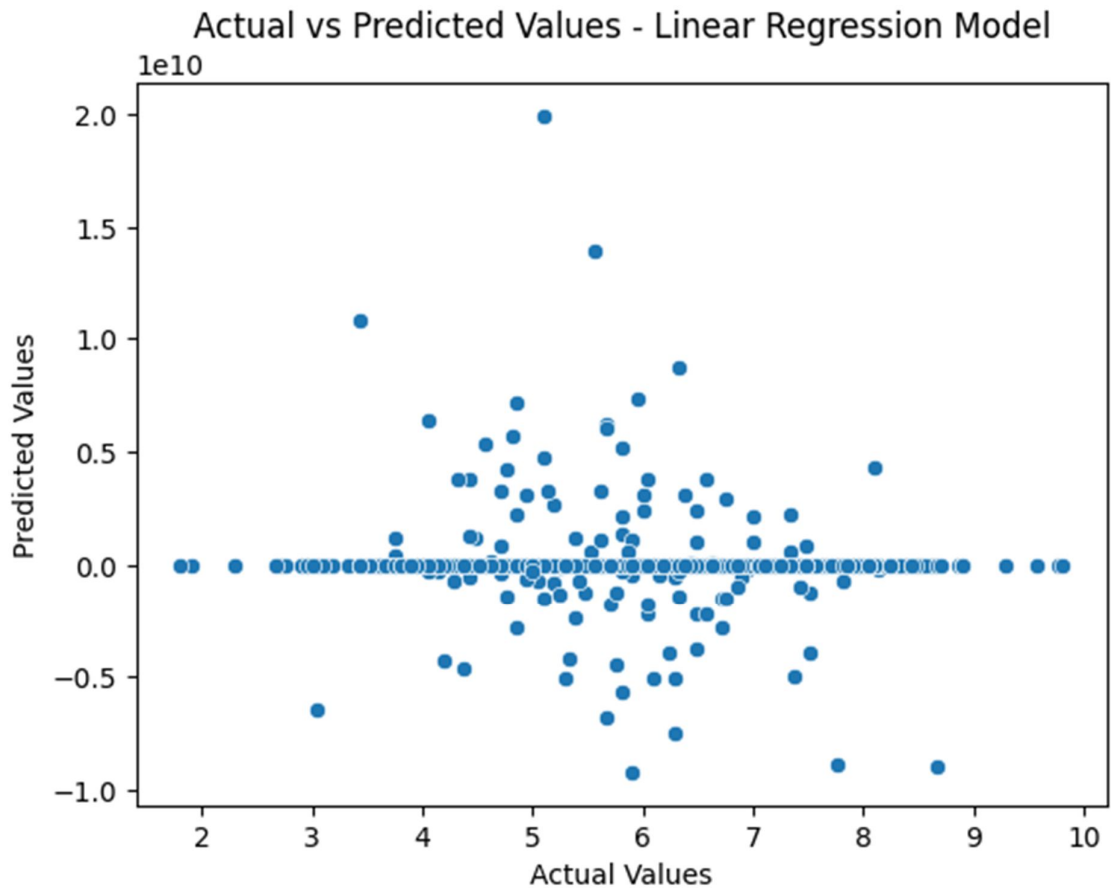
<b>no</b>	<b>actual</b>	<b>predicted</b>	<b>Diff</b>
<b>0</b>	<b>7.24</b>	<b>7.40</b>	<b>-0.16</b>
<b>1</b>	<b>6.90</b>	<b>6.50</b>	<b>0.41</b>
<b>2</b>	<b>7.00</b>	<b>6.75</b>	<b>0.25</b>
<b>3</b>	<b>6.57</b>	<b>5.92</b>	<b>0.65</b>
<b>4</b>	<b>4.10</b>	<b>5.02</b>	<b>-0.92</b>
<b>5</b>	<b>5.67</b>	<b>4.99</b>	<b>0.67</b>
<b>6</b>	<b>7.43</b>	<b>6.87</b>	<b>0.56</b>
<b>7</b>	<b>5.52</b>	<b>6.05</b>	<b>-0.53</b>
<b>8</b>	<b>7.43</b>	<b>7.10</b>	<b>0.32</b>
<b>9</b>	<b>5.62</b>	<b>3.2</b>	<b>-3.26</b>
<b>10</b>	<b>5.24</b>	<b>5.18</b>	<b>0.06</b>

Table 1 is a comparison between the actual and predicted values of a variable in the International Curriculum for Educational Evaluation (ICFES) in the following table. The "no" column represents the index of each observation in the database.

Actual values are provided in the "actual" column, whereas predicted values are provided in the "predicted" column.

"Diff" displays the difference between actual and anticipated values. Negative values in the "Diff" column indicate that the predicted value was greater than the actual value, while positive values indicate that the anticipated value was lower.

The discrepancy between the actual and predicted values is quite large for certain results, while it is relatively minor for others. Furthermore, observation 9 shows a very significant negative difference, suggesting that what was predicted was much greater than the actual value.



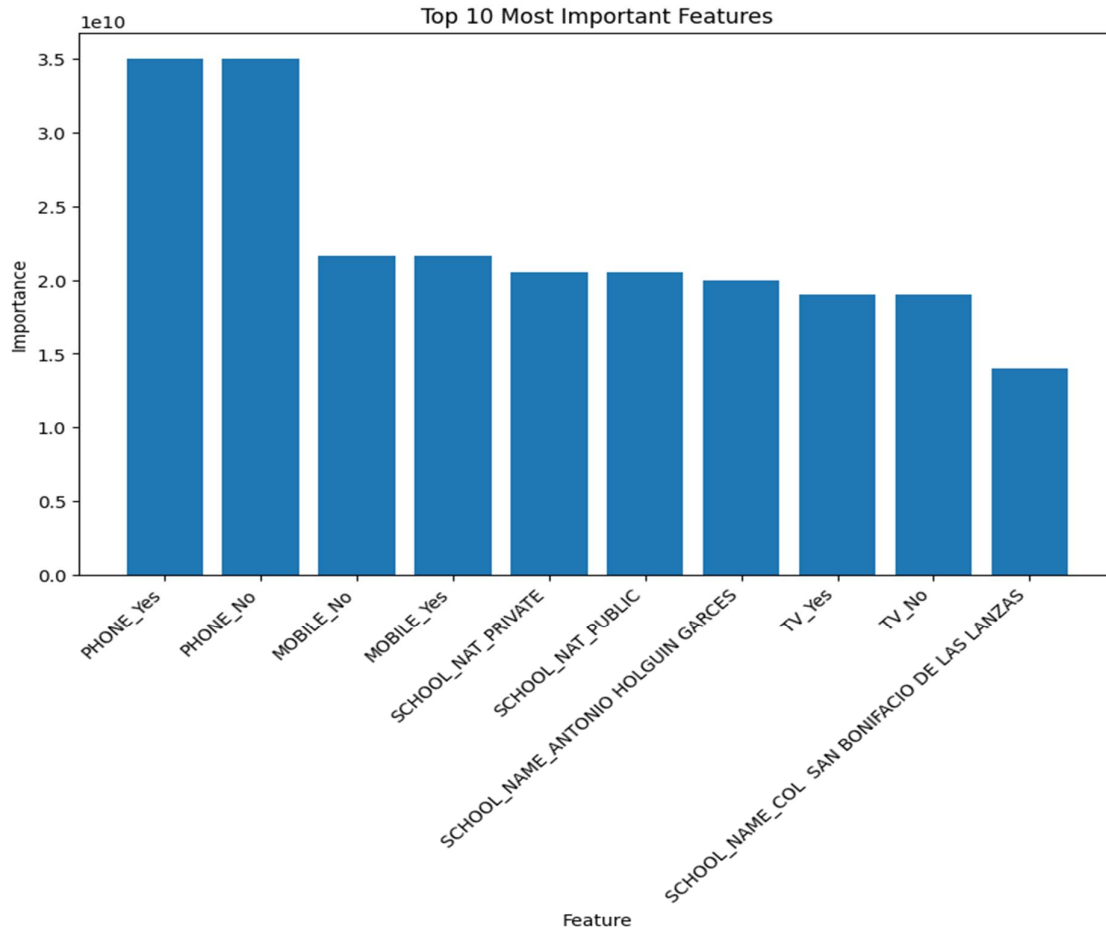
**Figure 2 Actual and predicted values LR scatter plot**

With the mean square error of this model being quite large in comparison to others, the high discrepancy between predicted and actual values is visible in the scatter plot above.

#### 4.1.2 Top important attributes

**Table 5 Top important attributes LR**

<b>Index</b>	<b>Feature</b>	<b>Value</b>
<b>1</b>	<b>PHONE_Yes</b>	<b>3.5</b>
<b>2</b>	<b>PHONE_No</b>	<b>3.5</b>
<b>3</b>	<b>MOBILE_No</b>	<b>2.16</b>
<b>4</b>	<b>MOBILE_Yes</b>	<b>2.16</b>
<b>5</b>	<b>SCHOOL_NAT_PRIVATE</b>	<b>2.05</b>
<b>6</b>	<b>SCHOOL_NAT_PUBLIC</b>	<b>2.05</b>
<b>7</b>	<b>SCHOOL_NAME_ANTONIO HOLGUIN GARCES</b>	<b>2</b>
<b>8</b>	<b>TV_Yes</b>	<b>1.9</b>
<b>9</b>	<b>TV_No</b>	<b>1.9</b>
<b>10</b>	<b>SCHOOL_NAME_COL SAN BONIFACIO DE LAS LANZAS</b>	<b>1.4</b>



**Figure 3 Top important attributes LR**

This table and figure display the relative relevance of several characteristics in a prediction model. The index or number of each feature in the model is represented by the "Index" column.

Each feature's name or description is specified in the "Feature" column, while the "Value" column displays the importance value assigned to each feature. It is worth noting that the values in the "Value" column may be relative rather than absolute.

The variables given in the table are thought to be predictive of the outcome in question, with more relevant characteristics awarded greater significance levels.

For example, phone ownership (PHONE\_Yes and PHONE\_No) and school type (SCHOOL\_NAT\_PRIVATE and SCHOOL\_NAT\_PUBLIC) all have significance ratings of

2 or above, indicating that they are highly effective in predicting the result of interest. Specific school name characteristics (SCHOOL\_NAME\_ANTONIO HOLGUIN GARCES and SCHOOL\_NAME\_COL SAN BONIFACIO DE LAS LANZAS) have lower significance ratings, indicating that they have less predictive.

## 4.2 Random Forest

the random forest model has an MSE value of 0.48, indicating that the model's projected values are on average 0.48 units off from the actual values. This suggests that the model is working properly and can generate accurate predictions.

### 4.2.1 Comparison between actual and predicted values

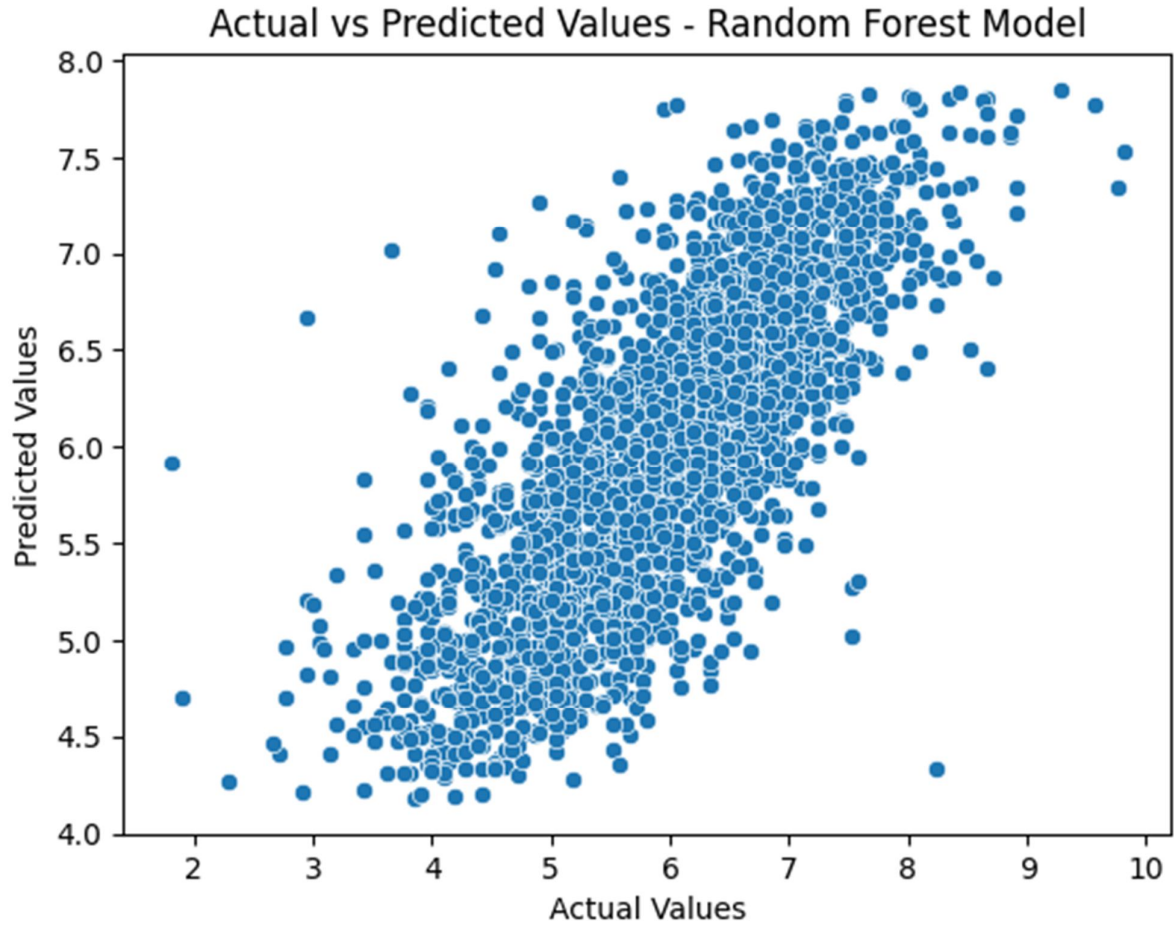
**Table 6 actual and predicted values RF**

<b>Index</b>	<b>Actual</b>	<b>Predicted</b>	<b>Diff</b>
<b>0</b>	<b>7.24</b>	<b>6.92</b>	<b>0.32</b>
<b>1</b>	<b>6.90</b>	<b>6.03</b>	<b>0.87</b>
<b>2</b>	<b>7.00</b>	<b>6.10</b>	<b>0.90</b>
<b>3</b>	<b>6.57</b>	<b>6.52</b>	<b>0.06</b>
<b>4</b>	<b>4.10</b>	<b>4.99</b>	<b>-0.89</b>
<b>5</b>	<b>5.67</b>	<b>5.17</b>	<b>0.50</b>
<b>6</b>	<b>7.43</b>	<b>6.37</b>	<b>1.06</b>
<b>7</b>	<b>5.52</b>	<b>5.88</b>	<b>-0.36</b>
<b>8</b>	<b>7.43</b>	<b>7.02</b>	<b>0.41</b>
<b>9</b>	<b>5.62</b>	<b>5.68</b>	<b>-0.06</b>
<b>10</b>	<b>5.24</b>	<b>5.17</b>	<b>0.07</b>

The table is divided into four columns: "actual", "predicted", "Diff", and an index column labelled from 0 to 10. The "actual" column provides real score of the test, the "predicted" column has predicted values determined by the machine learning model, and the "Diff" column contains the difference between the two. The "actual" column values vary from 4.09 to 7.42, whereas the "predicted" column values range from 4.1 to 7.01.

For each row, the "Diff" column shows the difference between the "actual" and "predicted" values. A positive number in the "Diff" column indicates that the predicted value is more than the actual value, whereas a negative value in the "Diff" column indicates that the projected value is less than the actual value. The values in the "Diff" column range from -0.9 to 1.06.

For reference, the index column simply labels each row from 0 to 10. Overall, the table compares actual and expected numerical values, as well as the difference between them.



**Figure 4 Actual and predicted values RF scatter plot**

With point forming a diagonal line from top to bottom, translate that the accuracy in the prediction is good.

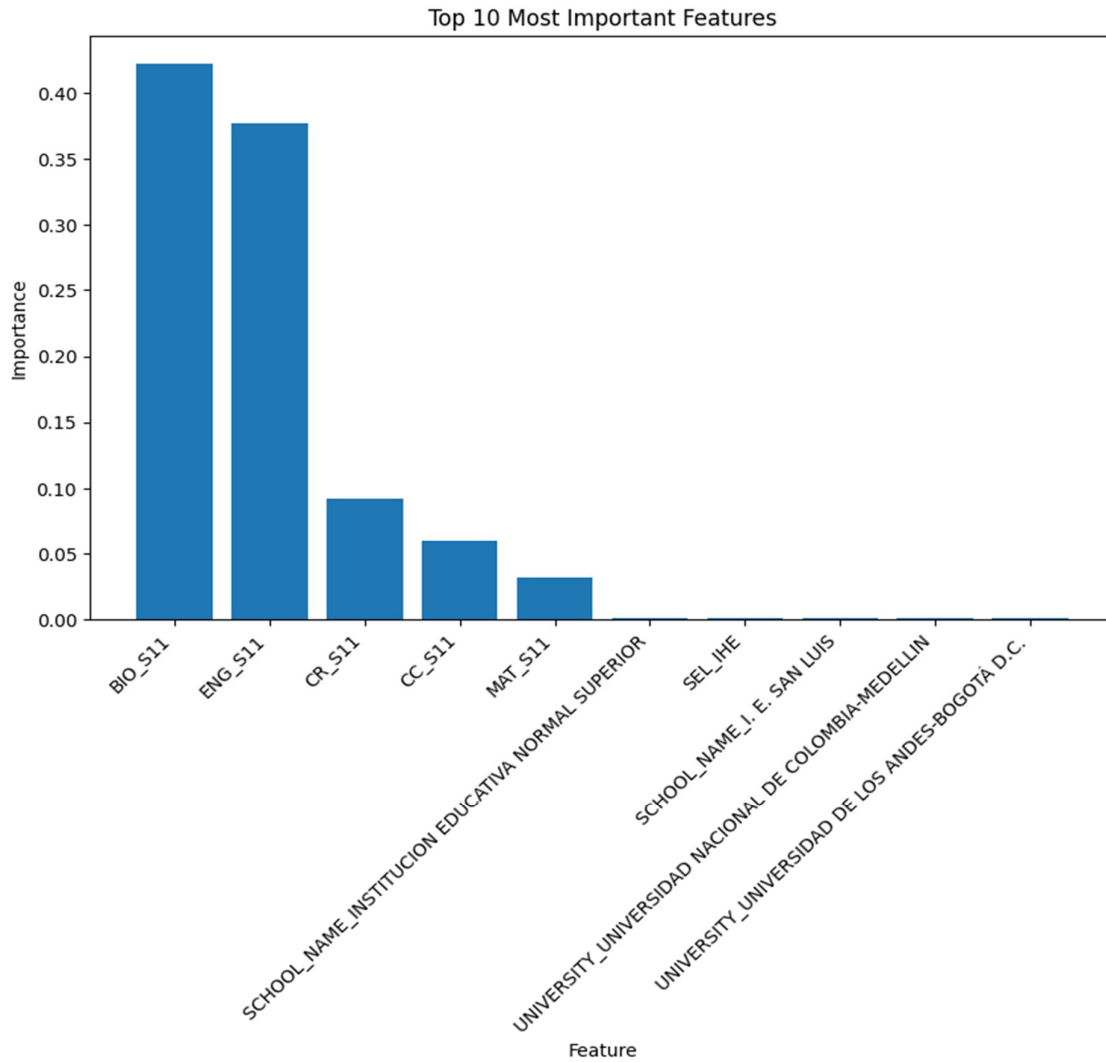
### 4.2.2 Top important attributes

We determined the top ten characteristics that had the greatest effect on prediction using their correlation coefficients, as shown in the table and chart below.

**Table 7 Top important attributes RF**

<b>Index</b>	<b>Feature</b>	<b>Importance</b>
<b>1</b>	<b>BIO_S11</b>	<b>0.421994</b>
<b>2</b>	<b>ENG_S11</b>	<b>0.377319</b>
<b>3</b>	<b>CR_S11</b>	<b>0.092560</b>
<b>4</b>	<b>CC_S11</b>	<b>0.059743</b>
<b>5</b>	<b>MAT_S11</b>	<b>0.031781</b>
<b>6</b>	<b>SCHOOL_NAME_INSTITUCION EDUCATIVA NORMAL SUPERIOR</b>	<b>0.001247</b>
<b>7</b>	<b>SEL_IHE</b>	<b>0.000991</b>
<b>8</b>	<b>SCHOOL_NAME_I. E. SAN LUIS</b>	<b>0.000858</b>
<b>9</b>	<b>UNIVERSITY_UNIVERSIDAD NACIONAL DE COLOMBIA-ME...</b>	<b>0.000712</b>
<b>10</b>	<b>UNIVERSITY_UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.</b>	<b>0.000596</b>





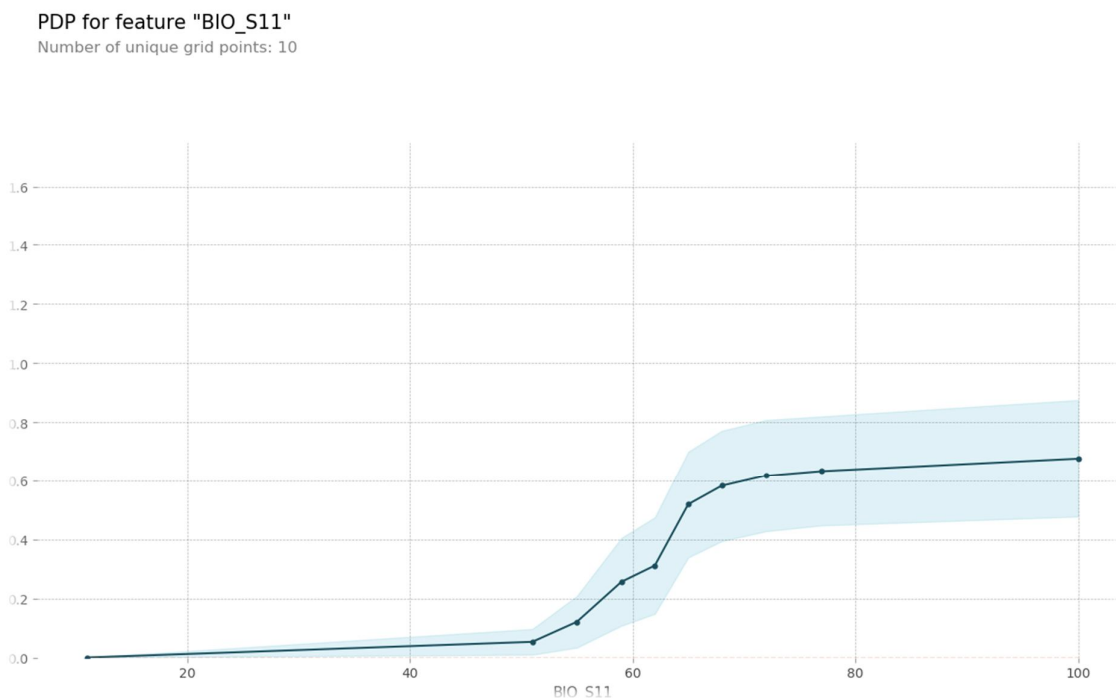
**Figure 5 Top important attributes RF**

The table and the figure above demonstrate the significance of numerous characteristics in predicting the results. Subject-specific scores such as BIO\_S11 (Biology), ENG\_S11 (English), CR\_S11 (Critical Reading), CC\_S11 (Social Sciences), and MAT\_S11 (Mathematics) are included, as well as school-related information such as the name of the school (in the case of both high schools and universities) and a binary variable SEL\_IHE indicating whether the student was selected to attend higher education (university or technical training).

The relevance column displays each feature's relative relevance in predicting exam outcomes, with higher numbers suggesting more importance. BIO\_S11 is awarded the greatest significance value, followed by ENG\_S11, CR\_S11, CC\_S11, and MAT\_S11, showing that Biology and English scores are the most significant determinants in predicting ICFES exam outcomes. The relevance scores for school-related variables, such as school names and SEL\_IHE, are very low, indicating that they may be less relevant predictors of test outcomes when compared to subject-specific scores.

### 4.2.3 Partial Dependence Plots

We will investigate how and if some of the features have a negative or positive impact on the model by using Partial Dependence Plots, which show the marginal effect of a variable on the predicted result while controlling for all other features. It will aid in seeing how the influence of a feature on the prediction model varies as its value changes. PDPs will be used to determine if important features have a positive or negative influence on the model's performance.

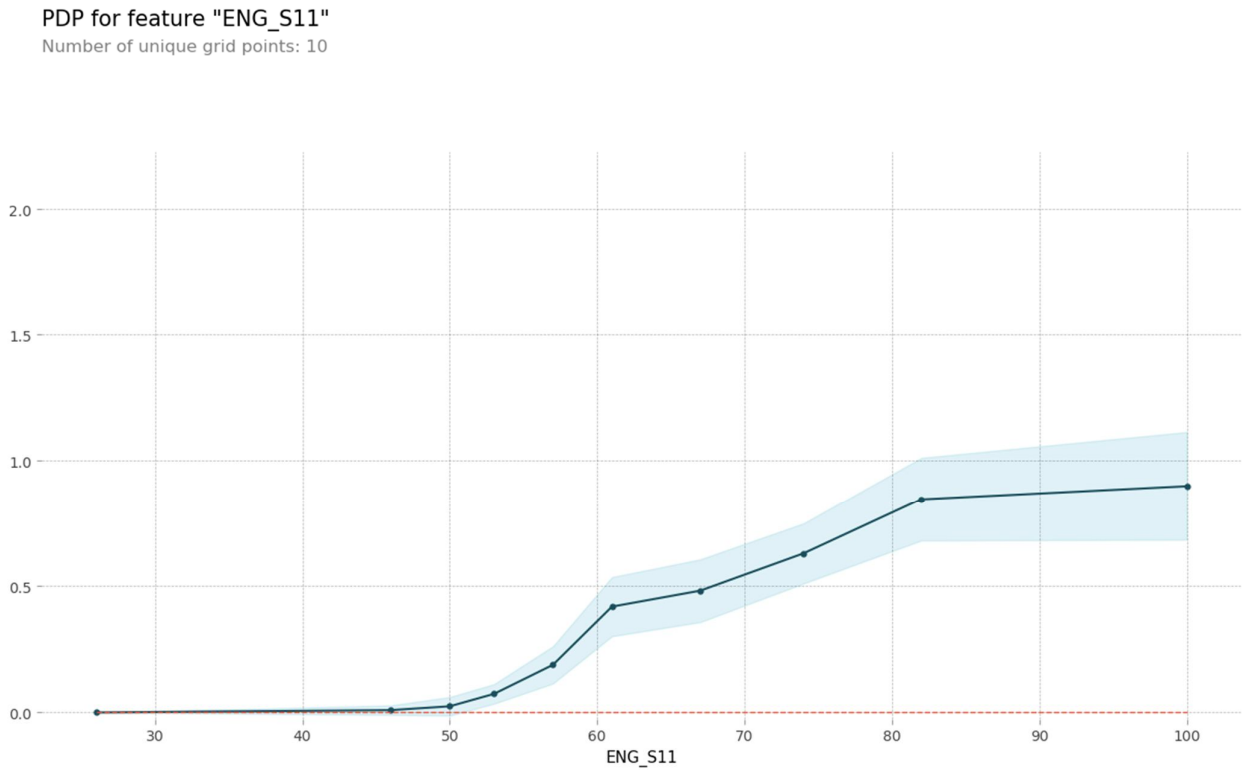


**Figure 6 PDP BIO\_S11 RF**

The graph depicts the link between students' BIO\_S11 scores and the influence they have on the model for predicting the overall score. The statistics in the graphic clearly show that there is a high link between the two variables.

The image specifically shows that the higher the students' BIO\_S11 score, the bigger the favorable influence it has on the model for predicting the overall result. This indicates that

when the BIO\_S11 score rises, the projected global score improves in accuracy and dependability.

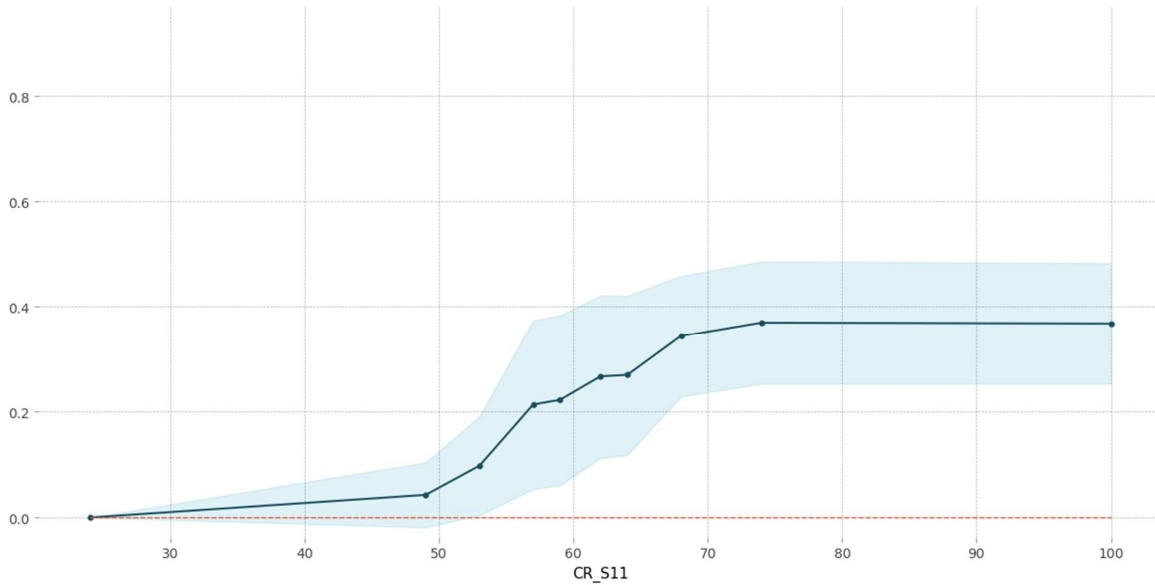


**Figure 7 PDP ENG\_S11 RF**

The second most influential variable in this prediction model has been discovered as ENG\_S11. Unlike BIO\_S11, ENG\_S11 has a positive influence on the model when the score surpasses a particular threshold.

When the ENG\_S11 score hits 40.5, the model for predicting the global score begins to favorably impact it. This indicates that raising the ENG\_S11 score over this level will result in a more accurate forecast of the overall score.

PDP for feature "CR\_S11"  
Number of unique grid points: 10



**Figure 8 PDP CR\_S11 RF**

CR\_S11 has been found as the third most significant variable in predicting the result in this prediction model. The influence of CR\_S11 on prediction is not as strong as that of ENG\_S11, but it has a substantial impact on the ultimate result.

The influence of CR\_S11 on the prediction model is not consistent over the whole range of scores. When the CR\_S11 score is between 0 and 50, the influence is limited, and the expected result does not alter much. However, when the CR\_S11 score rises, the influence on the expected outcome is more significant and exponentially.

This means that when the CR\_S11 score surpasses 50, the effect on the expected outcome becomes more significant, and even minor changes in the score might result in a significant variation in the outcome. As a result, it is crucial to pay attention to the CR\_S11 score when using this model to generate predictions.

### 4.3 Xbooster

The xbooster machine learning model has an MSE of 0.48, which means that the average squared difference between predicted and actual values is 0.48. This implies that the model may make quite accurate predictions, with an error of roughly 0.69 units (since the square root of 0.48 is approximately 0.69).

#### 4.3.1 Comparison between actual and predicted values

The following table displays the actual, predicted, and difference between the two.

**Table 8 actual and predicted values XGB**

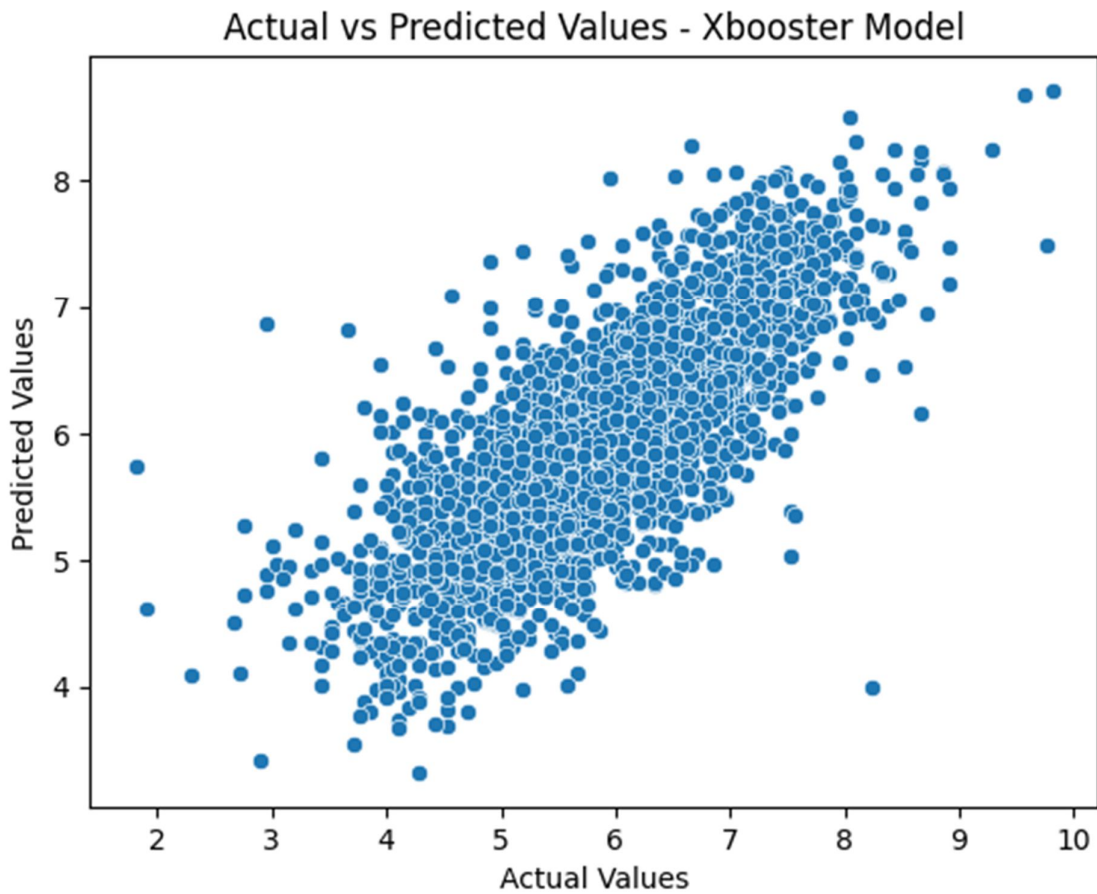
<b>Index</b>	<b>actual</b>	<b>predicted</b>	<b>Diff</b>
<b>0</b>	<b>7.24</b>	<b>7.03</b>	<b>0.20</b>
<b>1</b>	<b>6.90</b>	<b>6.30</b>	<b>0.60</b>
<b>2</b>	<b>7.00</b>	<b>6.37</b>	<b>0.63</b>
<b>3</b>	<b>6.57</b>	<b>6.67</b>	<b>-0.10</b>
<b>4</b>	<b>4.10</b>	<b>4.93</b>	<b>-0.84</b>
<b>5</b>	<b>5.67</b>	<b>5.14</b>	<b>0.53</b>
<b>6</b>	<b>7.43</b>	<b>6.84</b>	<b>0.59</b>
<b>7</b>	<b>5.52</b>	<b>5.97</b>	<b>-0.45</b>
<b>8</b>	<b>7.43</b>	<b>7.17</b>	<b>0.26</b>
<b>9</b>	<b>5.62</b>	<b>5.58</b>	<b>0.04</b>
<b>10</b>	<b>5.24</b>	<b>5.05</b>	<b>0.19</b>

The table has three columns: actual, predicted, and difference. The actual column reveals the test result's real values, and predicted column displays the values predicted using an Xbooster

model. The Difference column represents the difference between the actual and predicted values, with positive values indicating that the predicted value is greater than the actual value and negative values indicating the opposite.

The table's rows each represent a distinct observation, with the top row being the first observation, the second row representing the second observation, and so on.

For example, in the first row, the actual value is 7.23, the predicted value is 7.03, and the difference is 0.2. Similarly, in the second row, the actual value is 6.9, the predicted value is 6.303529, and the difference is 0.6.



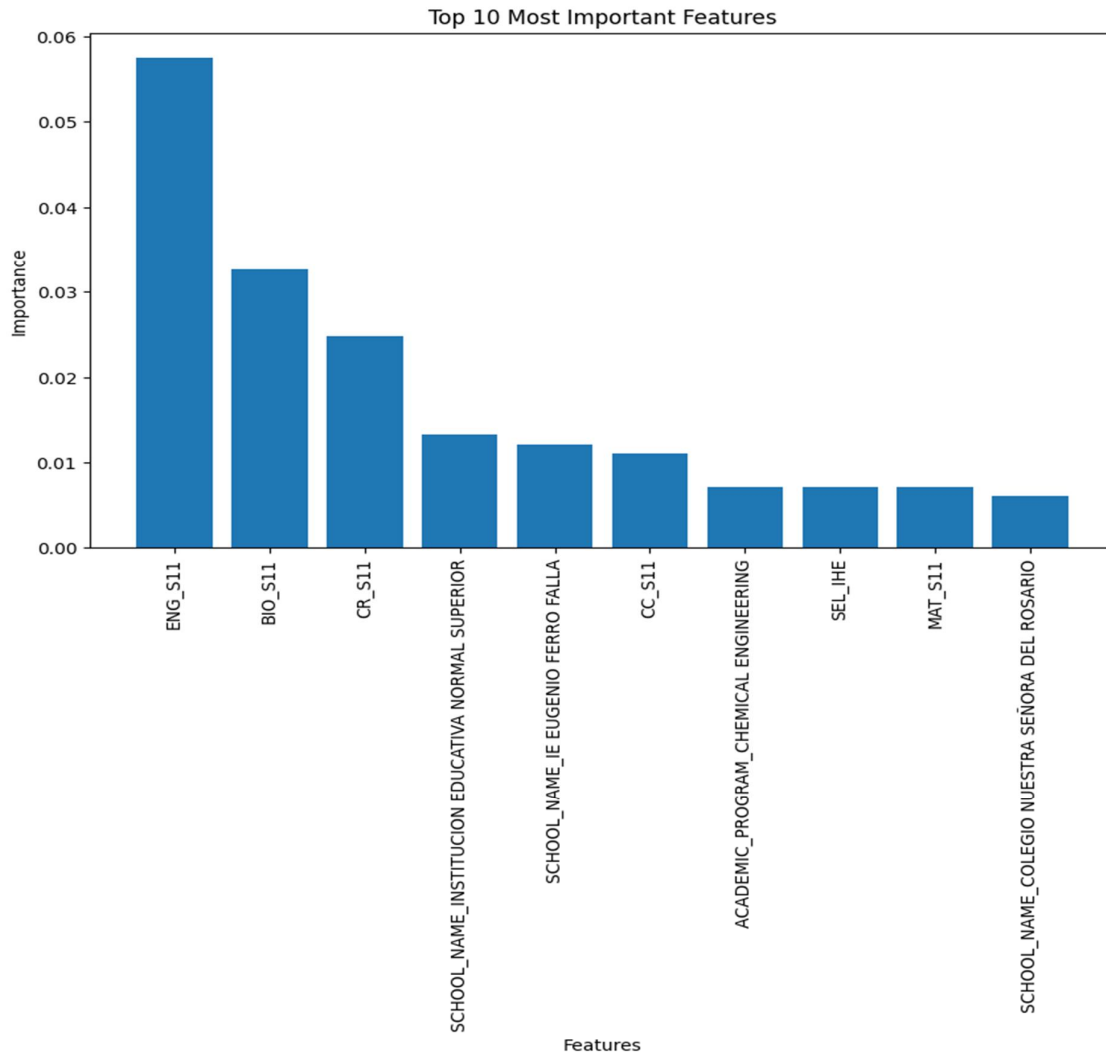
**Figure 9** actual and predicted scatter plot XGB

### 4.3.2 Top important attributes

**Table 9 Top important attributes XGB**

<b>Index</b>	<b>Features</b>	<b>Importance</b>
0	ENG_S11	0.058
1	BIO_S11	0.033
2	CR_S11	0.024
3	SCHOOL_NAME_INSTITUCION EDUCATIVA NORMAL SUPERIOR	0.013
4	SCHOOL_NAME_IE EUGENIO FERRO FALLA	0.012
5	CC_S11	0.011
6	ACADEMIC_PROGRAM_CHEMICAL ENGINEERING	0.007
7	SEL_IHE	0.007
8	MAT_S11	0.007
9	SCHOOL_NAME_COLEGIO NUESTRA SEÑORA DEL ROSARIO	0.006





**Figure 10 Top important attributes XGB**

The table and chart reveal the top 10 criteria that influenced the forecast, along with their significance values.

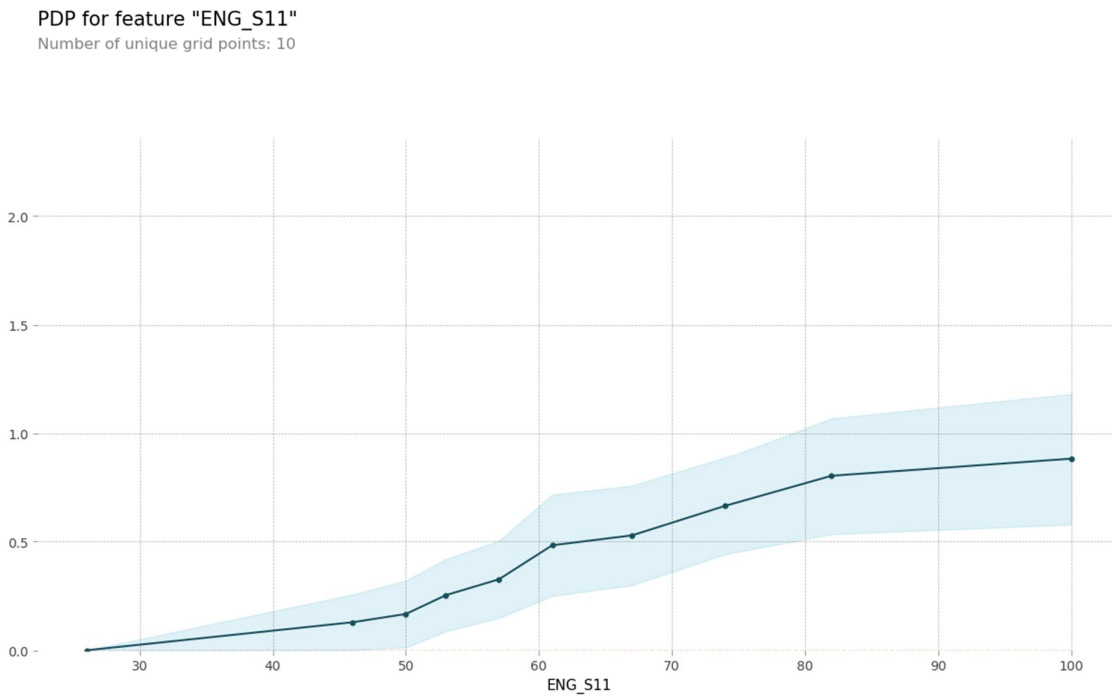
The table includes data on numerous aspects of education, including subject-specific scores for English, Biology, and Critical Reading, as well as the overall score for the College Entrance Exam (CC\_S11) and Mathematics (MAT\_S11). The significance of school names and academic programs is also included in the table.

The feature significance values indicate how much each feature adds to the prediction model's overall performance. Consequently, higher significance ratings indicate that a characteristic has a greater impact on the model's output.

As a result, while using this model to create predictions, it is crucial to consider the ENG\_S11 score. The model's accuracy and reliability may be greatly improved by thoroughly assessing and comprehending the impact of ENG\_S11 on the expected output.

### 4.3.3 Partial Dependence Plots

We'll check if any of the top traits have an influence on the Xbooster prediction model, either negatively or positively.



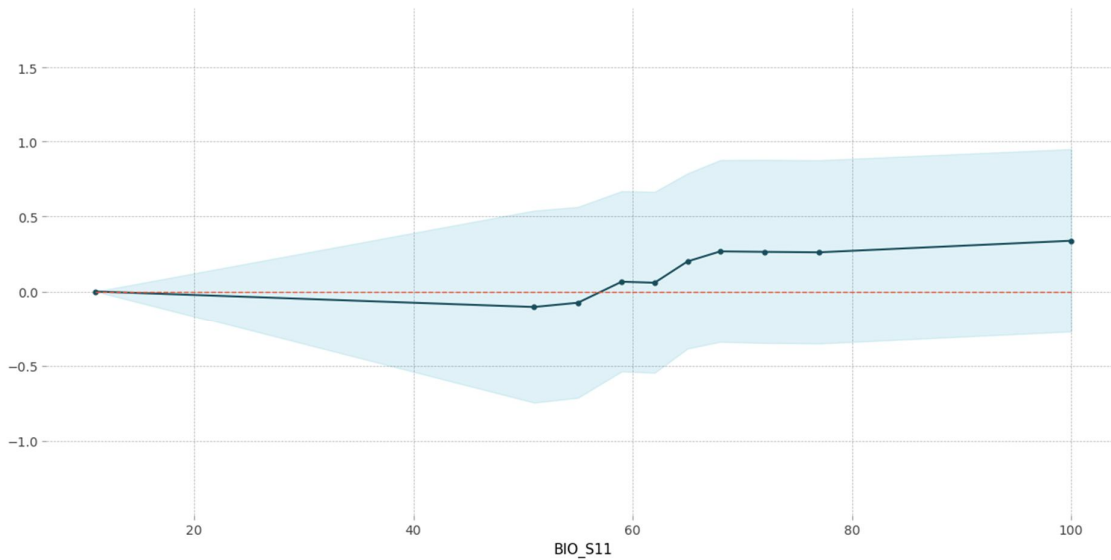
**Figure 11 PDP ENG\_S11 XGB**

The feature known as ENG\_S11 was discovered to be the most significant component determining the result, with a positive influence, in this prediction model. This implies that changes in the ENG\_S11 score can have a significant influence on the predicted outcome.

However, the magnitude of this effect varies within the ENG\_S11 score range. When the ENG\_S11 score is less than 50, the prediction model has minimal influence and the difference in outcome is insignificant. The influence is substantial when the ENG\_S11 score hits 60, resulting in a major shift in the expected outcome.

As a result, while using this model to make predictions, it is critical to take the ENG\_S11 score into account. By properly examining and appreciating the influence of ENG\_S11 on the projected output, the model's accuracy and dependability may be considerably enhanced.

PDP for feature "BIO\_S11"  
Number of unique grid points: 10



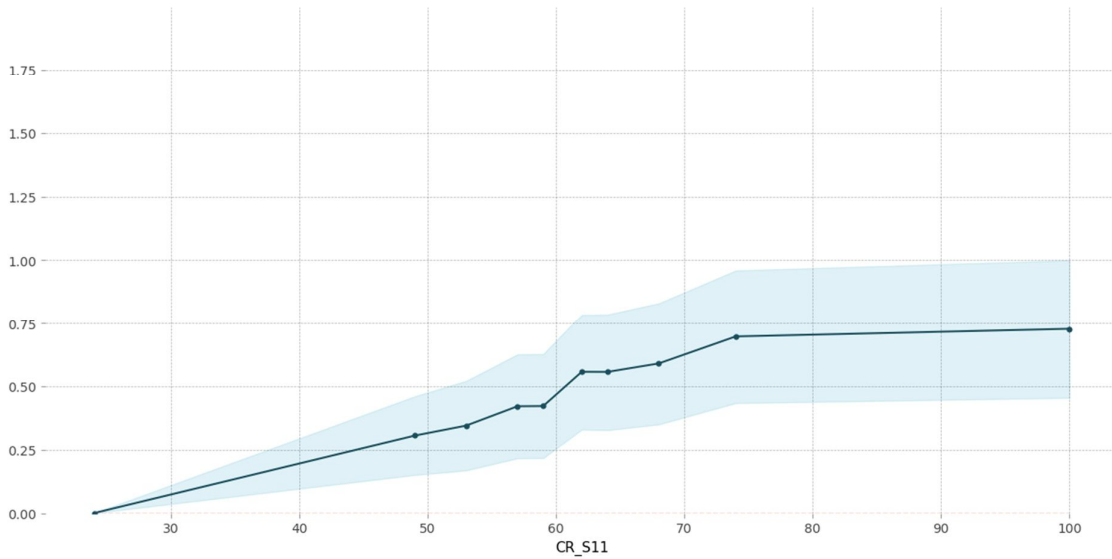
**Figure 12 PDP BIO\_S11 XGB**

BIO\_S11 is a prediction model variable that has been identified to have a significant impact on the predicted outcome. When the value of BIO\_S11 is less than 48, it has a detrimental influence on prediction in this model. This means that when the BIO\_S11 score goes below 48, the predicted outcome is less likely to be true, and the model's accuracy declines.

When the value of BIO\_S11 exceeds 48, the prediction model's influence becomes positive. In fact, the beneficial effect of BIO\_S11 on the model develops as the score increases. This indicates that when the BIO\_S11 score grows, the predicted outcome becomes more accurate and reliable.

The negative effect of BIO\_S11 when its value is less than 48 is most likely due to the variable's complex connection with other variables in the prediction model. However, when the BIO\_S11 score rises beyond this threshold, it appears to play a larger role in forecasting the result and contributes favourably to the model's overall accuracy.

PDP for feature "CR\_S11"  
Number of unique grid points: 10



**Figure 13 PDP CR\_S11 XGB**

The prediction model variable CR\_S11 was discovered to be the third most important factor influencing the expected result. The statistics show that it has a positive influence on the prediction model, and that this influence develops as its score increases.

While the beneficial influence of CR\_S11 on the prediction model is not as great as the top two factors, it is still an essential factor to consider when making predictions. By taking the CR\_S11 score into account, one may improve the accuracy of the predicted outcome and make more informed decisions.

#### 4.4 Artificial Neural Network

The mean square error for this model is 0.57. It signifies that the projected values generated by the model depart from the actual values seen in the data by a squared difference of 0.57 on average.

Overall, a mean square error of 0.57 indicates that the model's predictions are generally accurate, with only a modest amount of error between anticipated and observed values in the data.

##### 4.4.1 Comparison between actual and predicted values

**Table 10 actual and predicted values ANN**

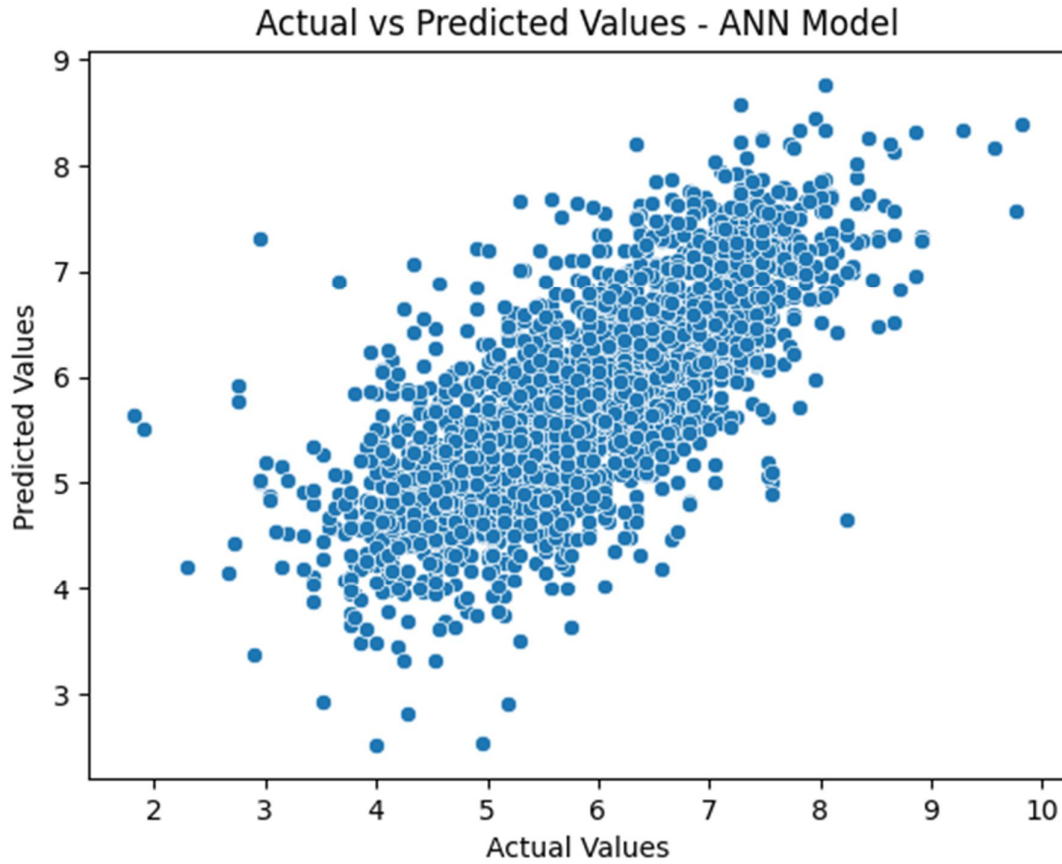
<b>actual</b>	<b>predicted</b>	<b>Diff</b>
0	7.24	7.43
1	6.90	6.49
2	7.00	6.79
3	6.57	6.01
4	4.10	5.07
5	5.67	5.01
6	7.43	6.89
7	5.52	6.04
8	7.43	7.17

9	5.62	5.70
10	5.24	5.21

For 11 observations, this table compares the actual values, predicted values, and the difference (or error) between them. The first column contains the actual values, the second predicts the values, and the third the difference between the actual and predicted values.

For example, the actual value for the first observation is 7.24, the projected value is 7.43, and the difference is -0.19. This suggests that the anticipated value is somewhat greater than the actual value, with a -0.188999 error.

The table gives an overview of how well the prediction model is functioning. Ideally, predicted values should be as near to the actual values as feasible, with the discrepancy being as little as possible. In this scenario, some of the predicted values are close to the actual values (for example, observation 2), while others diverge significantly (for example, observation 4).



**Figure 14 actual and predicted scatter plot ANN**

#### **4.5 Ensemble model**

I chose to merge the best three models to develop a more accurate prediction model. Rather of depending on a single model, I combined the three models by averaging their predictions. The resultant model is a combination of the strengths of each of the original models, and any flaws may be minimized.

The mean square error in the linear regression model is the highest. This suggests that it does not adequately match the data and may not be the best forecast for the situation. The other three models, however, outperform the linear regression model and are regarded the top three models.



The mean square error of the three models is 0.46, indicating that the average is more accurate than any of the individual models. This method of mixing models is widely used in machine learning and statistics to increase prediction accuracy.

#### 4.5.1 Comparison between actual and predicted values

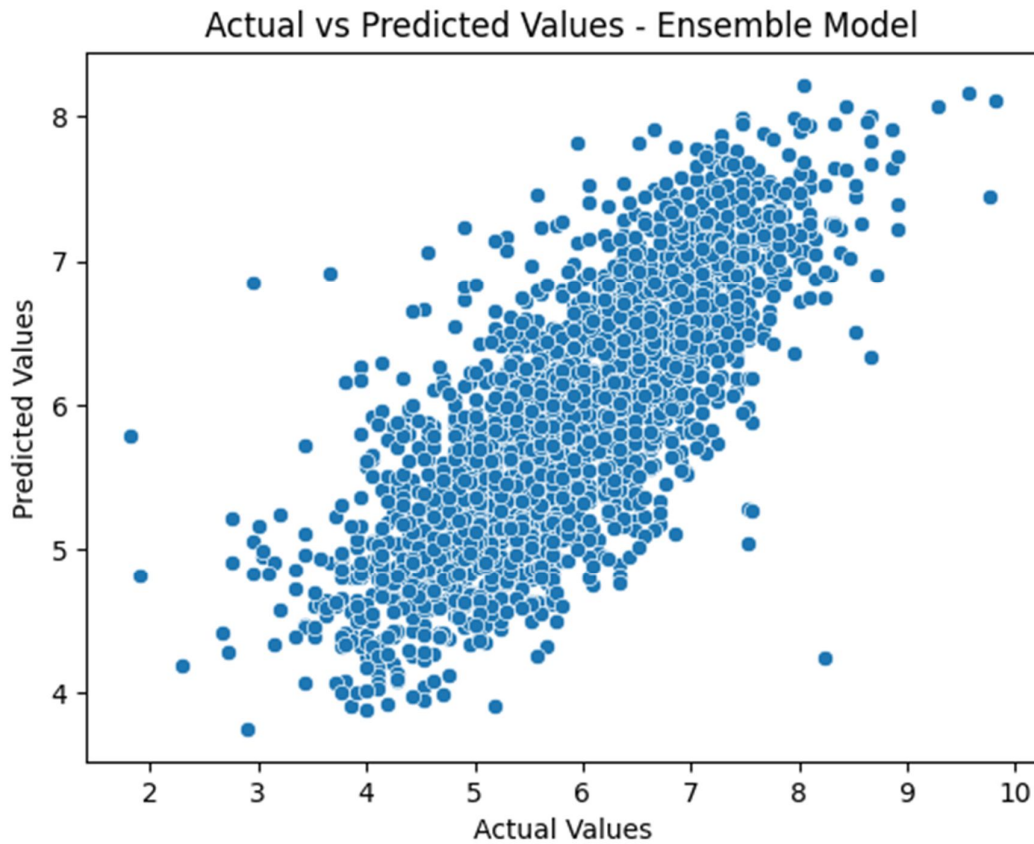
**Table 11 actual and predicted values ENSEM**

<b>Index</b>	<b>Actual</b>	<b>Predicted</b>	<b>Diff</b>
0	7.24	6.98	0.26
1	6.90	6.17	0.74
2	7.00	6.23	0.77
3	6.57	6.59	-0.02
4	4.10	4.96	-0.87
5	5.67	5.15	0.51
6	7.43	6.60	0.83
7	5.52	5.93	-0.40
8	7.43	7.09	0.34
9	5.62	5.63	-0.01
10	5.24	5.11	0.13

For 11 observations, the table displays the actual values, anticipated values, and the difference between them (Diff). Each row is a single observation.

In the first row, for example, the actual value is 7.24, the projected value is 6.98, and the difference is 0.26.

Similarly, the actual value for the second row is 6.904762, the projected value is 6.17, and the difference is 0.74.



**Figure 15 actual and predicted scatter plot ENSEM**

This clearly shows that this model has more accurate prediction than other models.

## **Chapter 5.**

### **CONCLUSION**

The main objective of this study was to predict students' academic progress after completing their engineering degree, which is an important component in determining their total academic success. To do this, multiple machine learning models were used to predict students' academic progress, and the performance of these models was compared to determine the most effective model.

The study sought to identify the aspects or characteristics that influence academic success among engineering students, as well as to predict academic achievement. The study took into account student demographics, socioeconomic status, high school academic success, and other relevant factors.

By identifying these aspects, the study hopes to get insight into the key causes of academic accomplishment among engineering students. This information might then be used to improve teaching techniques, academic support, and engineering student counseling.

Finally, the study looked at the accuracy and precision of several machine learning models. Among the models investigated were Random Forest, xgbooster, articiel neural network, and linear regression. By analyzing the performance of several models, the study tried to identify the most effective model for predicting academic accomplishment among engineering students.

We were able to develop multiple machine learning models using data from the Colombian Institute for the Evaluation of Education (ICFES). These models were used to forecast a student's total score during their ICFES exam. After evaluating each machine learning model, it was determined that linear regression performed the worst, with a mean square error of 8. However, random forest and xgbooster performed the best, with mean square errors close to 0.48. The artificial neural network performed third best, with a mean square error of 0.56. Following that, we integrated the three best-performing models to build an ensemble model, which outperformed any individual machine learning model with a mean score error of 0.46. We were also able to determine the most significant elements that influenced the predictions by using these machine learning models. These determinants included past academic successes in topics such as Biology, English Language, Critical Reading, Citizen Competencies, and Mathematics, which have an impact on engineering students' academic performance. The pupils' socioeconomic background had no meaningful effect on the forecasts. However, certain high schools and institutions, such as "INSTITUCION EDUCATIVA NORMAL SUPERIOR, IE EUGENIO FERRO FALLA, COLEGIO NUESTRA SEORA DEL ROSARIO, UNIVERSIDAD NACIONAL DE COLOMBIA-ME..., UNIVERSIDAD DE LOS ANDES-BOGOT D.C." had a favourable or negative impact. Furthermore, demographic characteristics such as gender had a beneficial effect on male students' forecasts while having no influence on female students.

We encountered limitation with our analysis that prevented us from making predictions that were more accurate and closer to the actual numbers. The restriction was the absence of data on the students' academic performance throughout their time at university in several subject areas. Without this vital information, we were unable to identify the precise university courses that had the greatest impact on the forecasts. We think that having access to this data

would have allowed us to improve our machine learning models and forecast outcomes more accurately. Therefore, gathering more thorough information on students' academic performance across a range of university-level disciplines would be crucial to making future projections that are more accurate. This information might improve our comprehension of how specific subjects affect a student's overall academic performance and assist us in creating better methods for enhancing academic outcomes in those areas.

Further research will look into the impact of feature selection techniques on predictive models in educational data mining.

We will investigate the influence of feature selection approaches on the performance of predictive models in educational data mining in this follow-up study. Identifying the most relevant and informative characteristics from a given dataset is critical to improve the efficiency and performance of prediction models.

Further research may also identify at-risk students, create effective early warning systems, and recommend ways to help high-risk students in higher education institutions.

## REFERENCES

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. *Technology, Knowledge and Learning*, 24(4), 567–598. <https://doi.org/10.1007/s10758-019-09408-7>

Adekitan, A. I., Salau, O., Ng, A. I., & Adekitan, A. I. (2019). *The impact of engineering students' performance in the first three years on their graduation result using educational data mining*. <https://doi.org/10.1016/j.heliyon.2019>

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, 9, 7519–7539. <https://doi.org/10.1109/ACCESS.2021.3049446>

Alamri, L. H., Almuslim, R. S., Alotibi, M. S., Alkadi, D. K., Ullah Khan, I., & Aslam, N. (2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest. *ACM International Conference Proceeding Series, Part F168981*, 100–107. <https://doi.org/10.1145/3446590.3446607>

Almasri, A., Celebi, E., & Alkhaldeh, R. S. (2019). EMT: Ensemble meta-based tree model for predicting student performance. *Scientific Programming*, 2019. <https://doi.org/10.1155/2019/3610248>

Alshantqi, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. *IEEE Access*, 8, 203827–203844. <https://doi.org/10.1109/ACCESS.2020.3036572>

Ashmawy, A. K., Schreiter, S., IEEE Education Society., & Institute of Electrical and Electronics Engineers. (2019). *Proceedings of 2019 IEEE Global Engineering Education Conference (EDUCON) : date and venue, 9-11 April, 2019, Dubai, UAE.*

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). *End to End Learning for Self-Driving Cars.* <http://arxiv.org/abs/1604.07316>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Chui, K. T., Liu, R. W., Zhao, M., & Ordóñez de Pablos, P. (2020). Predicting Students' Performance with School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine. *IEEE Access*, 8, 86745–86752. <https://doi.org/10.1109/ACCESS.2020.2992869>

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247–256. <https://doi.org/10.1016/j.chb.2017.01.047>

Davidson, O. B., Feldman, D. B., &Margalit, M. (2012). A focused intervention for 1st-year college students: Promoting hope, sense of coherence, and self-efficacy. *Journal of Psychology: Interdisciplinary and Applied*, 146(3), 333–352.

<https://doi.org/10.1080/00223980.2011.634862>

Delahoz-Dominguez, E., Zuluaga, R., &Fontalvo-Herrera, T. (2020). Dataset of academic performance evolution for engineering students. *Data in Brief*, 30.

<https://doi.org/10.1016/j.dib.2020.105537>

Dien, T. T., Luu, S. H., Thanh-Hai, N., & Thai-Nghe, N. (2020). Deep learning with data transformation and factor analysis for student performance prediction. *International Journal of Advanced Computer Science and Applications*, 11(8), 711–721.

<https://doi.org/10.14569/IJACSA.2020.0110886>

Doleck, T., Lemay, D. J., Basnet, R. B., &Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, 25(3), 1951–1963. <https://doi.org/10.1007/s10639-019-10068-4>

Francis, B. K., &Babu, S. S. (2019). Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *Journal of Medical Systems*, 43(6).

<https://doi.org/10.1007/s10916-019-1295-4>

Geurts, P., Ernst, D., &Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>



Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences (Switzerland)*, 10(11). <https://doi.org/10.3390/app10113894>

Hazard, F. E. (1974). *.4. . 1 1 PREDICTING STUDENT ACHIEVEMENT IN TWO-YEAR ENGDINEERING TECHNOLOGY PROGRAMS.*

Hooshyar, D., Pedaste, M., & Yang, Y. (2020). Mining educational data to predict students' performance through procrastination behavior. *Entropy*, 22(1), 12. <https://doi.org/10.3390/e22010012>

Iatrellis, O., Savvas, I., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, 26(1), 69–88. <https://doi.org/10.1007/s10639-020-10260-x>

IEEE Systems, M., & Institute of Electrical and Electronics Engineers. (2018). *ICNSC 2018 : the 15th IEEE International Conference on Networking, Sensing and Control : March 27-29, 2018,*

*Zhuhai, China.* Institute of Electrical and Electronics Engineers, & ManavRachna International Institute of Research and Studies. (2019). *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing : trends, prespectives and prospects: COMITCON-2019 : 14th-16th February, 2019.*

Institute of Electrical and Electronics Engineers, & PPG Institute of Technology. (n.d.). *Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES 2020): 10-12, June 2020.*

Iraqi, K. M., IEEE Computer Society, University of Karachi. Department of Computer Science, & Institute of Electrical and Electronics Engineers. (2020). *ICISCT'20 : 2nd International Conference on Information Science and Communication Technology : 8th-9th February 2020.*

Jordan University of Science & Technology, Institute of Electrical and Electronics Engineers. Jordan Section, & Institute of Electrical and Electronics Engineers. (2019). *2019 10th International Conference on Information and Communication Systems (ICICS) : 11-13 June, 2019, Jordan University of Science and Technology, Irbid, Jordan.*

Kelleher, J. D., Mac, B., Aoife, N., & Arcy, D.'. (n.d.). *FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS Algorithms, Worked Examples, and Case Studies.* Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the Effectiveness of Machine Learning Methods for Spam Detection. *Procedia Computer Science*, 190, 479–486. <https://doi.org/10.1016/j.procs.2021.06.056>

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2774 PART 2, 267–274. [https://doi.org/10.1007/978-3-540-45226-3\\_37](https://doi.org/10.1007/978-3-540-45226-3_37)

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. <https://doi.org/10.1038/nature14539>

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. *Journal of Educational Computing Research*, 57(2), 448–470. <https://doi.org/10.1177/0735633117752614>

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*, 103, 1–15. <https://doi.org/10.1016/j.compedu.2016.09.005>

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8, 55462–55470. <https://doi.org/10.1109/ACCESS.2020.2981905>

Merali, Z. G., Witiw, C. D., Badhiwala, J. H., Wilson, J. R., & Fehlings, M. G. (2019). Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS ONE*, 14(4). <https://doi.org/10.1371/journal.pone.0215133>

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. H. M. S. (2021). Mining educational data to predict students performance: A comparative study of data mining techniques. *Education and Information Technologies*, 26(5), 6051–6067. <https://doi.org/10.1007/s10639-021-10575-3>

Nielsen, M. (2020). *Neural Networks and Deep Learning*.  
<http://neuralnetworksanddeeplearning.com>

Oyedeki, A. O., Salami, A. M., Folorunsho, O., & Abolade, O. R. (2020). Analysis and Prediction of Student Academic Performance Using Machine Learning. *JITCE (Journal of Information Technology and Computer Engineering)*, 4(01), 10–15.  
<https://doi.org/10.25077/jitce.4.01.10-15.2020>

Porter, D. R. (1993). Introduction to Linear Regression Analysis. *Technometrics*, 35(2), 224–225. <https://doi.org/10.1080/00401706.1993.10485050>

Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. *Electronics (Switzerland)*, 11(7).  
<https://doi.org/10.3390/electronics11071005>

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. In *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* (Vol. 40, Issue 6, pp. 601–618). <https://doi.org/10.1109/TSMCC.2010.2053532>

Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the American Statistical Association*, 99(466), 567–567.  
<https://doi.org/10.1198/jasa.2004.s339>

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual

Recognition Challenge. *International Journal of Computer Vision*, 115(3), 211–252.  
<https://doi.org/10.1007/s11263-015-0816-y>

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., &Ragos, O. (2020). Implementing autoML in educational data mining for prediction tasks. *Applied Sciences (Switzerland)*, 10(1).  
<https://doi.org/10.3390/app10010090>

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

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104. <https://doi.org/10.1016/j.chb.2019.106189>

Wang, P., Zhang, G., Yu, Z. G., & Huang, G. (2021). A Deep Learning and XGBoost-Based Method for Predicting Protein-Protein Interaction Sites. *Frontiers in Genetics*, 12. <https://doi.org/10.3389/fgene.2021.752732>

Wang, Y. (2020). Iteration-based naive Bayes sentiment classification of microblog multimedia posts considering emoticon attributes. *Multimedia Tools and Applications*, 79(27–28), 19151–19166. <https://doi.org/10.1007/s11042-020-08797-7>

Wu, Y., & Wang, G. (2018). Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis. *International Journal of Molecular Sciences*, 19(8). <https://doi.org/10.3390/ijms19082358>

Xia, J., Ghamisi, P., Yokoya, N., & Iwasaki, A. (2018). Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1), 202–216. <https://doi.org/10.1109/TGRS.2017.2744662>

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1). <https://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>

YILDIZ, M., & BÖREKÇİ, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. *Journal of Educational Technology and Online Learning*. <https://doi.org/10.31681/jetol.773206>

Yulianto, L. D., Triayudi, A., & Sholihati, I. D. (2020). Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5. In *JurnalMantik* (Vol. 4, Issue 1). <https://iocscience.org/ejournal/index.php/mantik/index>

Zhang, Y., An, R., Cui, J., & Shang, X. (2021). Undergraduate grade prediction in Chinese higher education using convolutional neural networks. *ACM International Conference Proceeding Series*, 462–468. <https://doi.org/10.1145/3448139.3448184>