ytrts

by Kenil Ghetia

Submission date: 17-May-2024 01:46AM (UTC+0530)
Submission ID: 2381299158
File name: erar.pdf (586.89K)
Word count: 6600
Character count: 38865

CHAPTER 1

INTRODUCTION

The volume of data and improvements in data analysis methods are causing a revolutionary change in the educational sector. The use of educational data mining (EDM) and learning analytics to assess student performance, pinpoint at-risk individuals, and develop customized learning interventions has become increasingly common. Using a large dataset from Colombia, this study explores the variables affecting engineering students' academic success.

1.1 Background

Predicting academic performance in engineering education has become a significant research focus, with numerous studies utilizing data mining and machine learning to pinpoint factors contributing to student success. Early research focused on traditional machine learning models, such as decision trees, random forests, support vector machines, and Naïve Bayes, to predict student outcomes based on demographic, academic, and social factors. More recently, deep learning models have been explored for their improved predictive accuracy.

1.2 Objectives

By using a variety of machine learning algorithms to forecast student performance based on academic, social, and economic characteristics, this study attempts to analyse the factors impacting academic achievement in engineering programs. The specific objectives are:

- To identify the key academic competencies at both secondary and university levels that influence performance in engineering programs.
- To assess the impact of social and economic factors on academic achievement in engineering.
- To evaluate how well various machine learning models predict the performance quartiles of students.
- To offer guidance to educators and legislators to create focused interventions and assistance programs that will improve the academic performance of engineering students.

1.3 Research Questions

This study is guided by the following research questions:

1. How do academic competencies at the secondary level influence performance in engineering programs?

2. What is the impact of social and economic factors on academic achievement in engineering?
3. Can specific factors be identified that significantly contribute to student success in different engineering disciplines?

1.4 Importance of the Research

The study's findings will provide valuable information on the relationships between various factors and academic success in engineering degrees. By keeping these connections in mind, targeted interventions and support systems can be created to enhance educational achievements. This study has implications for enhancing student support services in postsecondary education and contributes to the growing body of knowledge in the field of educational data mining.

CHAPTER 2

LITERATURE REVIEW & PROBLEM IDENTIFICATION

Predicting academic performance in engineering education has become a significant research focus, with numerous studies utilizing data mining and machine learning to pinpoint factors contributing to student success.

This field has seen a surge in using learning analytics (LA) and educational data mining (EDM) to understand and predict student performance.

2.1 Literature Review

Predicting academic performance in engineering education has become a significant research focus, with numerous studies utilizing data mining and machine learning to pinpoint factors contributing to student success. This field has seen a surge in using learning analytics (LA) and educational data mining (EDM) to understand and predict student performance.

Early research focused on traditional machine learning models. Delahoz-Dominguez et al. (2020) [1] and Soto-Acevedo et al. (2023) [2] used standardized test scores and socioeconomic data to analyze and predict academic performance in Colombia, emphasizing the importance of a strong academic foundation established in secondary education. Similar approaches have been explored by other researchers, employing models like decision trees [4], random forests [5], support vector machines [6], and Naïve Bayes [7] to predict student outcomes based on demographic, academic, and social factors.

The use of Learning Management Systems (LMS) has provided new avenues for performance prediction. Khan et al. (2023) [3] investigated the predictive power of LMS activity logs, identifying factors like resource views, activity gaps, and previous academic performance as strong predictors of student success. This highlights the value of student engagement and historical data in assessing learning progress.

More recently, researchers have explored the application of deep learning models for improved predictive accuracy. Alhazmi and Sheneamer (2023) [8] used dimensionality reduction techniques and machine learning models to predict student performance at early stages, using admission scores and first-year course grades. Sun et al. (2023) [9] proposed a model based on multifeature fusion and attention mechanisms to analyze historical academic data from multiple dimensions, demonstrating the potential of this approach for accurate performance prediction. Convolutional

Neural Networks (CNNs), often used for image recognition, have also been adapted for student performance prediction. Chau et al. (2021) [18] used a 2D CNN to analyze temporal educational data by transforming it into 2D images, showcasing the potential of CNNs to uncover patterns in non-image data. Poudyal et al. (2022) [17] further explored this approach by developing a hybrid 2D CNN model trained on 1D numerical data converted into 2D grayscale images, achieving promising results in predicting academic performance. Like our exploration of advanced models, Nabil et al. (2021) [20] investigated the effectiveness of Deep Neural Networks (DNNs) for predicting student performance. Their study focused on predicting student success in a Data Structures course based on grades from previous courses. They found that DNNs outperformed other machine learning models, achieving an accuracy of 89%. This highlights the potential of DNNs in handling complex academic performance data and identifying students at risk of failure. Our findings align with previous research [1], [2], [10]–[12], highlighting the significant influence of both secondary and university-level academic competencies on student performance in engineering programs. As emphasized by Kabakchieva (2013) [13], a strong foundation in secondary education, particularly in mathematics and science as noted by Ramesh, Parkavi & Ramar (2013) [11], plays a crucial role in predicting success at the university level. Furthermore, our analysis, along with the work of Bydzovská (2020) [14], reveals that university-level academic competencies, especially critical reading (CR PRO) and citizen competencies (CC PRO), are strong predictors of student performance. These findings underscore the importance of fostering these skills throughout the educational journey of engineering students. Kanani et al. (2023) [19] demonstrated the effectiveness of LSTM models in time-series predictions, achieving high accuracy in rainfall forecasting, which is comparable to our use of advanced models for predicting student performance.

Like the research by Ramesh, Parkavi & Ramar (2013) [11], our study found that parental occupation, among other socioeconomic factors, has a notable impact on student performance. This aligns with the broader discussion on the complex interplay between socioeconomic background and academic achievement within engineering education [1], [10], [15]. Daud et al. (2017) [15] further emphasize the importance of considering family expenditures and student personal information as potential predictors of academic success. These findings suggest the need for a more nuanced understanding of how various socioeconomic and personal factors interact and influence student outcomes in engineering programs.

While our research primarily employed traditional machine learning models, studies such as Bydzovská (2020) [14] and Thai-Nghe et al. (2010) [16] demonstrate the potential of utilizing social behavior data and collaborative filtering techniques for performance prediction. These approaches could offer valuable insights into student performance by leveraging similarities among students and courses, potentially leading to more personalized interventions and support systems. Future research could explore the integration of collaborative filtering and recommender systems within the context of engineering education to enhance predictive models and develop targeted strategies for student success.

The current research expands upon these efforts by leveraging a hybrid approach. We combine the feature extraction capabilities of CNNs with the sequential processing strength of Long Short-Term Memory (LSTM) networks, aiming to capture both complex relationships within data and temporal dependencies in student learning behaviour. Furthermore, we delve into feature importance analysis to provide insights into the factors most significantly contributing to student success in engineering programs.

Table 2.1 Table of literature survey

Paper Title	Authors	Dataset	Algorithm(s)	Results	Key Findings
Classification and prediction of student performance data using various machine learning algorithms [22]	Pallathadka et al. (2021)	UCI Machine Learning Repository: Student Performance	Naïve Bayes, ID3, C4.5, SVM	SVM achieved the highest accuracy	SVM outperforms other algorithms for classifying student performance data.
Next-Term Student Performance Prediction: A Recommender Systems Approach [23]	Sweeney et al. (2015)	George Mason University	Factorization Machines (FM), Random Forests (RF), Personalized Multi-Linear Regression (PMLR)	FM, RF, and PMLR achieved the lowest prediction errors. Hybrid FM-RF method outperformed individual methods.	Instructor characteristics and differences between transfer and non-transfer students significantly impact student performance.
GritNet: Student Performance Prediction with Deep Learning [24]	Kim et al. (2018)	Udacity Nanodegree programs	GritNet (based on Bidirectional Long Short-Term Memory (BLSTM))	GritNet consistently outperformed the baseline logistic regression model, especially in the early weeks.	Deep learning can effectively predict student graduation, particularly during the initial stages of the course.

Student Performance Prediction Using Dynamic Neural Models [25]	Delianidi et al. (2016)	ASSISTments, FSAI-F1toF3	Recurrent Neural Networks (RNN), Time Delay Neural Networks (TDNN)	RNN outperformed TDNN in all datasets. RNN model surpassed state-of-the-art models in most cases.	RNNs effectively track student knowledge states and predict performance. Initializing skill embeddings with pre-trained vectors did not provide a significant advantage.
Exercise-Enhanced Sequential Modeling for Student Performance Prediction [26]	Su et al. (2018)	iFLYTEK Co., Ltd. (Zhixue online learning system)	Exercise-Enhanced Recurrent Neural Network (EERNN) with two variants: EERNNM (Markov property) and EERNNA (Attention mechanism)	Both EERNNM and EERNNA outperformed baselines, with EERNNA showing better performance due to its attention mechanism.	Incorporating exercise texts alongside student exercise records improves prediction accuracy. EERNN effectively handles cold-start problems for new students and exercises.
A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction [27]	Pandey and Sharma (2013)	Manav Rachna College of Engineering	J48, NBTree, RepTree, Simple Cart	J48 decision tree algorithm found to be the most suitable for model construction, achieving accuracy over 80%.	Decision trees can effectively predict student grades in engineering programs.
A Machine Learning Model to Predict Standardized Tests in Engineering Programs in Colombia [2]	Soto-Acevedo et al. (2023)	Colombian Institute for the Evaluation of Education Quality (ICFES)	K Nearest Neighbours, Generalized Linear Network Model (GLMNET), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Trees (DT), Boosting	GLMNET achieved the highest performance with accuracy and AUC of over 80%.	University accreditation is a crucial factor in predicting student performance on standardized tests.
Study on Student Performance Estimation, Student Progress Analysis, and Student Pot	Yang and Li (2018)	Academic performance data of 60 hig	Back Propagation Neural Network (BP-NN) for classifica	BPNN based models accurately estimated student attributes, performance, and	BPNN based tools can analyze student progress and potential for improvement by considering

ential Prediction base d on Data Mining [28]	h school stud ents	tion and predictio n	causal relationships between attributes.	performance and nonp erformance related attributes.

2.2 Problem Identification

Despite the advances in predictive modelling, several challenges and gaps remain in the literature. This section identifies key problems addressed by the current research.

2.2.1 Diverse Influencing Factors

Previous studies have highlighted the importance of academic competencies, socioeconomic factors, and demographic variables in predicting student performance. However, there is a need for a more comprehensive understanding of how these factors interact and influence outcomes in different contexts, particularly in engineering education.

2.2.2 Model Performance and Comparability

While various models have been employed to predict academic performance, comparisons of model performance across different studies are often limited by variations in datasets, target variables, and evaluation metrics. There is a need for standardized methodologies to evaluate and compare the effectiveness of different models in predicting student outcomes.

2.2.3 Interpretability and Practical Application

The interpretability of machine learning models is critical for their practical application in educational settings. While advanced models such as CNNs and LSTMs offer high predictive accuracy, their complexity can hinder understanding and trust among educators and policymakers. Developing models that balance accuracy and interpretability is essential for practical implementation.

2.2.4 Targeted Interventions

Identifying key predictors of student success is crucial for developing targeted interventions. Academic and socioeconomic factors have an impact on students' performance, as demonstrated by prior studies; nonetheless, actionable insights are needed to guide the creation of support systems and policies to improve educational results.

2.3 Research Contribution

This research aims to address these gaps by:

1. Employing a comprehensive dataset that includes academic, social, and economic information to predict student performance in engineering programs.

2. Comparing the performance of various machine learning models, including traditional, advanced, and hybrid approaches, to identify the most effective predictors.
3. Utilizing interpretability techniques to provide insights into the factors influencing predictions and to develop practical recommendations for educators and policymakers.
4. Providing practical insights to guide the creation of focused interventions and assistance programs to improve engineering students' academic performance.

By addressing these challenges, this study contributes to the growing field of educational data mining and provides valuable implications for improving student support systems and educational outcomes in higher education.

CHAPTER 3

PROPOSED METHODOLOGY AND IMPLEMENTATION.

3.1 Data Description

The dataset used in this study was obtained from the Colombian Institute for the Evaluation of Education (ICFES) and compiled by Delahoz-Dominguez et al. (2020) [1]. It contains information on 12,411 engineering students, including academic performance data from standardized tests at both secondary (Saber 11) and university (Saber Pro) levels, as well as socioeconomic variables such as parental education, household appliances, and type of school attended.

3.1.1 Variables

The dataset includes the following key variables:

- **Academic Performance:**
  - **Secondary Level:** Results on standardized tests (Saber 11) for English (ENG_S11), biology (BIO_S11), citizenship competencies (CC_S11), critical reading (CR_S11), and mathematics (MAT_S11).
  - **University Level:** Results on the Quantitative Reasoning (QR_PRO), Critical Reading (CR_PRO), Written Communication (WC_PRO), English (ENG_PRO), Citizen Competencies (CC_PRO), and Formulation of Engineering Projects (FEP_PRO) standardized examinations (Saber Pro). Additionally, a Global Score (G_SC) is given.
- **Demographic Information:** Gender, socioeconomic level, and educational background of parents.
- **Social and Economic Factors:** Access to resources like internet, television, computers, and household appliances.
- **Educational Background:** Type of school attended (public or private), academic program enrolled in.

3.2 Data Preprocessing

Effective data preprocessing is crucial for ensuring the quality and reliability of the predictive models. The following steps were taken to preprocess the data:

3.2.1 Categorical Encoding

Categorical variables such as gender, parents' education, and occupation were converted into numerical representations using Label Encoding. This step is essential for preparing the data for machine learning algorithms that require numerical input.

3.2.2 Target Variable Transformation

The global score (G_SC) was binned into quartiles (Q1, Q2, Q3, Q4) to create a categorical target variable for classification models. This transformation allows the prediction of students' performance levels rather than exact scores, making it easier to identify students in need of support.

3.3 Model Selection

A range of machine learning models was selected to predict student performance based on the pre-processed dataset. The chosen models include traditional, advanced, and hybrid approaches to provide a comprehensive comparison.

3.3.1 Random Forest Classifier

Several decision trees are used in this ensemble learning technique to decrease overfitting and increase prediction accuracy. It works especially well with complicated datasets that have a lot of variables.

3.3.2 K-Nearest Neighbors (KNN)

This non-parametric algorithm classifies data points based on the k nearest neighbors in the training data, considering the similarity between data points. KNN is simple yet effective for various classification tasks.

3.3.3 Gradient Boosting Classifier

By concentrating on fixing mistakes from earlier models, this ensemble learning technique successively joins weak learners to produce a powerful predictive model. Gradient Boosting is renowned for its strong durability and great accuracy.

3.3.4 Logistic Regression with Cross-Validation (GLMNet)

This statistical model is suitable for binary and multiclass classification problems and is particularly useful for interpreting feature importance. Cross-validation ensures the model's robustness and generalizability.

3.3.5 Support Vector Machine (SVM)

This algorithm finds the hyperplane maximizing the margin between classes, making it effective for high-dimensional data and complex relationships. SVM is known for its precision in classification tasks.

3.3.6 Decision Tree Classifier

This easily interpretable, tree-like model makes decisions by applying a set of rules, offering insights into the variables affecting forecasts. Decision trees are helpful tools for comprehending the process of making decisions.

3.3.7 Convolutional Neural Network (CNN)

This deep learning architecture is particularly well-suited for processing data with grid-like structures. CNNs were adapted to handle tabular data to explore their potential for performance prediction.

3.4 Model Training and Evaluation

The selected models were trained on the pre-processed dataset, and their performance was evaluated using various metrics.

3.4.1 Train-Test Split

The data was split into training and testing sets with an 80/20 ratio. This ensures model evaluation on unseen data and prevents overfitting, providing a reliable measure of model performance.

3.4.2 Evaluation Metrics

The following metrics were used to evaluate the performance of each model:

- **Accuracy:** the percentage of cases that were correctly classified, which gives an overall idea of how good the model is.
- **F1 Score:** the harmonic mean of recall and precision, which provides an impartial assessment for unbalanced datasets.
- **Confusion Matrix:** By displaying the counts of true positives, true negatives, false positives, and false negatives for each class, this table illustrates the effectiveness of the model.
- **Classification Report:** Provides a detailed breakdown of precision, recall, F1-score, and support for each class (quartile of global score), allowing for a comprehensive understanding of the model's behaviour across different performance levels.

- **ROC Curve and AUC:** Shows how true positive rate and false positive rate are traded off, and the area under the ROC curve (AUC) provides an overview of how well the model performs overall in class distinction.

3.5 Interpretability and Feature Importance

Understanding the factors influencing model predictions is crucial for practical application in educational settings. The following techniques were used to interpret the models and identify key features:

3.5.1 Permutation Importance

This technique assesses the importance of each feature by measuring the decrease in model performance when the feature's values are randomly shuffled. It helps in identifying the most critical features influencing predictions.

3.5.2 SHAP (SHapley Additive exPlanations)

This game-theoretic method sheds light on how each attribute contributes to the predictions, explaining the output of any machine learning model. Understanding how each variable affects the model's output is made easier with the use of SHAP values.

3.5.3 LIME (Local Interpretable Model-agnostic Explanations)

This method focuses on explaining individual predictions by creating a locally faithful interpretable model around the prediction. It helps understand how specific features contribute to a particular prediction, offering insights into the model's behaviour on a case-by-case basis.

Using these approaches, the research hopes to offer a thorough grasp of the variables affecting academic achievement in engineering education as well as practical advice for enhancing student outcomes.

CHAPTER 4

RESULT ANALYSIS

The results of the data analysis are presented in this chapter, with an emphasis on how well various machine learning models predict the academic and socioeconomic quartiles of student performance. The study compares the effectiveness of the models, assesses the significance of the features, and draws conclusions from interpretability methods.

4.1 Model Performance Comparison

The performance of various machine learning models was evaluated using metrics such as accuracy, F1 score, confusion matrix, and ROC curve with AUC. The models compared include Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Logistic Regression (GLMNet), Support Vector Machine (SVM), Decision Tree, and Convolutional Neural Networks (CNN).

4.1.1 Accuracy and F1 Score

The accuracy and F1 scores for each model are presented in the table below. These metrics provide an overall measure of each model's effectiveness in predicting student performance quartiles.

Table 4.1 Precision, Recall, F1 Score, Support and Accuracy of the models tested.

Model	Class	Precision	Recall	F1 Score	Support	Accuracy
RandomForest	Q1	0.93	0.88	0.91	657	
	Q2	0.79	0.81	0.80	589	
	Q3	0.80	0.85	0.82	625	**0.86**
	Q4	0.94	0.90	0.92	612	
	Overall			0.86	2483	
KNN	Q1	0.84	0.88	0.86	657	
	Q2	0.69	0.67	0.68	589	
	Q3	0.67	0.71	0.69	625	**0.76**
	Q4	0.87	0.80	0.84	612	
	Overall			0.77	2483	
GMLBoost	Q1	0.97	0.93	0.95	657	
	Q2	0.88	0.90	0.89	589	
	Q3	0.89	0.93	0.91	625	**0.92**
	Q4	0.98	0.95	0.96	612	
	Overall			0.93	2483	
GLMNet	Q1	0.96	0.95	0.95	657	**0.91**

	Q2	0.89	0.88	0.89	589	
	Q3	0.88	0.89	0.88	625	
	Q4	0.95	0.94	0.95	612	
	Overall			0.92	2483	
SVM	Q1	0.96	0.80	0.87	657	
	Q2	0.69	0.78	0.73	589	
	Q3	0.64	0.72	0.68	625	**0.77**
	Q4	0.84	0.79	0.81	612	
	Overall			0.77	2483	
DecisionTree	Q1	0.88	0.89	0.89	657	
	Q2	0.74	0.71	0.73	589	
	Q3	0.76	0.76	0.76	625	**0.82**
	Q4	0.89	0.90	0.90	612	
	Overall			0.82	2483	
CNN	Q1	0.98	0.93	0.95	657	
	Q2	0.90	0.81	0.85	589	
	Q3	0.81	0.94	0.87	625	**0.90**
	Q4	0.94	0.94	0.94	612	
	Overall			0.90	2483	

- **Random Forest:** Obtained an F1 score of 0.86 and an accuracy of 86%, demonstrating a remarkable capacity for prediction.
- **KNN:** Demonstrated a moderate level of competence, scoring an F1 of 0.76 and 77% accuracy.
- **Gradient Boosting:** Emerged as the top performer with an accuracy of 93% and an F1 score of 0.92.
- **GLMNet:** Demonstrated high accuracy 92% and F1 score 0.91, reflecting balanced performance across quartiles.
- **SVM:** Achieved an accuracy of 77% and an F1 score of 0.77, with notable drop in performance for Q3.
- **Decision Tree:** Exhibited good performance in Q1 and Q4, displaying an accuracy of 82% and an F1 score of 0.82.
- **CNN:** 90% accuracy and an F1 score of 0.90 were attained, indicating good predictive power.

4.1.2 Confusion Matrices

Confusion matrices for each model provide a visual depiction of the models' performance by showing the counts of true positives, true negatives, false positives, and false negatives for each class.

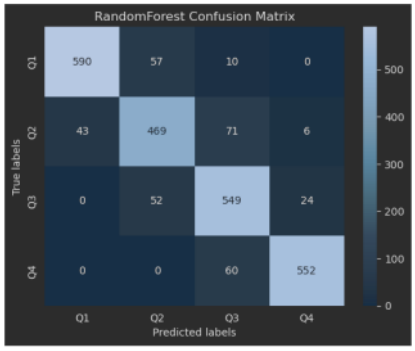

Fig. 4.1 Random Forest

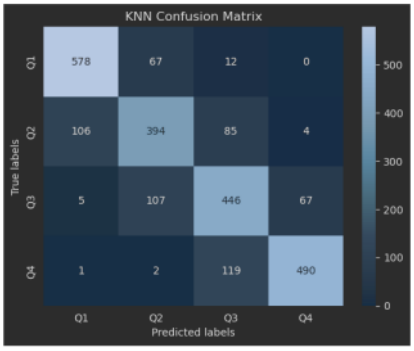

Fig. 4.2 KNN

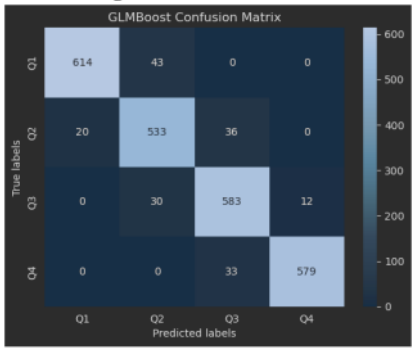

Fig. 4.3 GLMBOOST

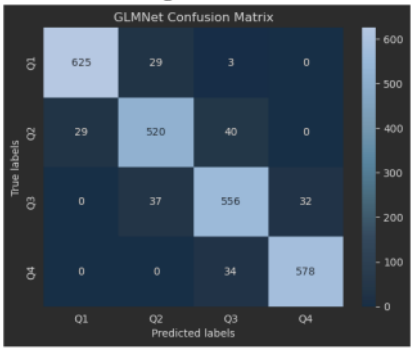

Fig. 4.4 GLMNet

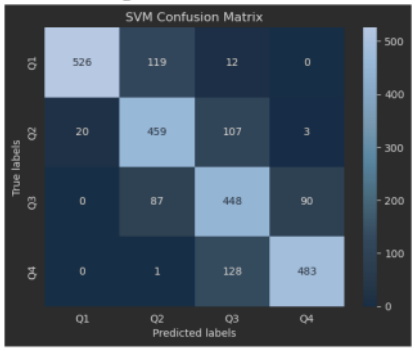

Fig. 4.5 SVM

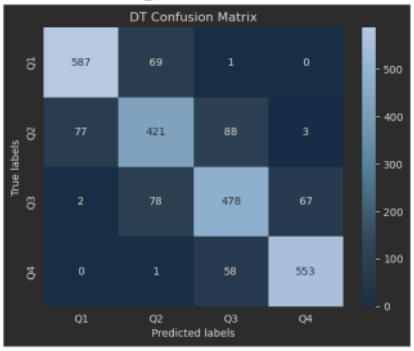

Fig. 4.6 Decision Tree

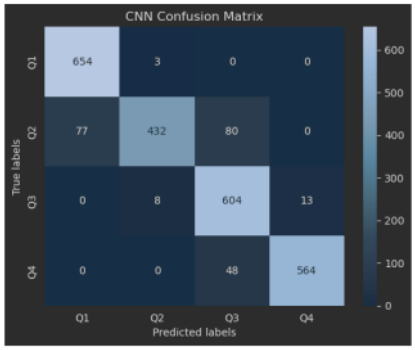

Fig. 4.7 CNN

Fig. 4.1 Random Forest

The confusion matrix for the Random Forest model shows that the model performed quite well in classifying the different categories. Most of the true labels are correctly predicted. However, there is some confusion between Q2 and Q3, as evidenced by the 52 instances where Q3 was predicted as Q2 and 60 instances where Q4 was predicted as Q3. Overall, the Random Forest model demonstrates strong predictive capabilities.

Fig. 4.2 K-Nearest Neighbors (KNN)

The KNN confusion matrix indicates a reasonable performance but shows notable misclassifications. For example, Q2 has significant misclassifications into Q1 (106 instances) and Q3 (107 instances). Similarly, Q4 is frequently misclassified as Q3 (119 instances). This suggests that the KNN model struggles with distinguishing between these categories more than other models.

Fig. 4.3 GLMBoost

The GLMBoost confusion matrix shows that this model performs well in predicting the correct categories, particularly for Q1 and Q4 with minimal misclassifications. The primary confusion occurs between Q2 and Q3, but the overall number of misclassifications is lower compared to the KNN model. This indicates that GLMBoost is a reliable model for this classification task.

Fig. 4.4 GLMNet

The GLMNet confusion matrix reveals a high accuracy, especially in predicting Q1 and Q4, with very few misclassifications. The model shows some confusion between Q2 and Q3, with 37 instances of Q2 being classified as Q3 and 34 instances of Q3 being classified as Q4. Despite this, the GLMNet model shows strong predictive performance overall.

Fig. 4.5 Support Vector Machine (SVM)

The SVM confusion matrix illustrates a fair level of accuracy but also highlights several areas of confusion, especially between Q1 and Q2 and Q3 and Q4. For instance, Q2 is often misclassified as Q3 (107 instances), and Q3 is misclassified as Q4 (90 instances). This suggests that while SVM performs adequately, it has difficulty differentiating between certain categories.

Fig. 4.6 Decision Tree

The Decision Tree confusion matrix shows that the model is relatively accurate in classifying the different categories, with Q1 and Q4 having high correct classification rates. However, there is noticeable confusion between Q2 and Q3, with significant misclassifications observed. This indicates that while Decision Trees are generally effective, they might not be as robust as some other models for this task.

Fig. 4.7 Convolutional Neural Network (CNN)

The CNN confusion matrix indicates a strong performance, especially in predicting Q1 and Q3, with minimal misclassifications. The most notable confusion occurs between Q2 and Q3, where 80 instances of Q2 were classified as Q3 and 48 instances of Q4 were classified as Q3. Overall, the CNN model demonstrates high accuracy and robustness in its predictions.

4.1.3 ROC Curves and AUC

The area under the curve (AUC) of ROC curves, which show the trade-off between true positive rate and false positive rate for each model, summarizes the overall performance.

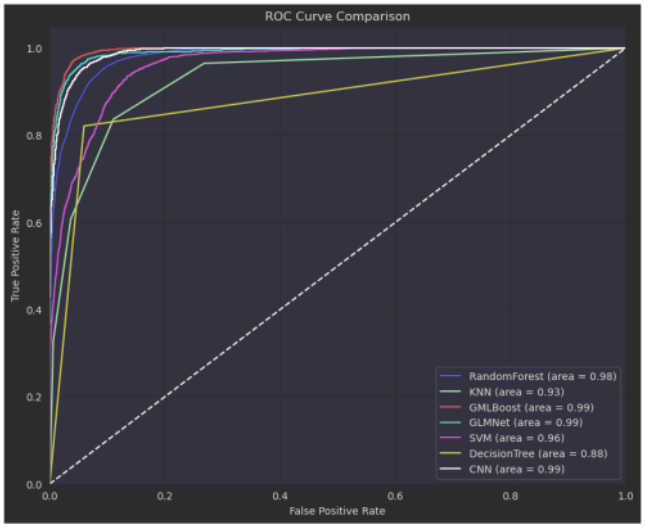

Fig 4.8 ROC of all models

Overall, the GLMBoost, GLMNet, and CNN models exhibit the highest AUC values (0.99), indicating exceptional performance in distinguishing between the classes. The RandomForest model also performs very well with an AUC of 0.98. The SVM model shows strong performance with an AUC of 0.96. The KNN model, with an AUC of 0.93, demonstrates good performance but is outperformed by the previously mentioned models. The DecisionTree model has the lowest AUC of 0.88, indicating comparatively lower performance in classification tasks.

This comparison highlights the superior predictive capabilities of the GLMBoost, GLMNet, and CNN models for this dataset, with RandomForest also being a reliable model.

4.2 Feature Importance

Evaluating feature importance helps identify the key variables that influence model predictions. Techniques such as permutation importance, SHAP values, and LIME were used to analyze feature importance.

4.2.1 Permutation Importance

Permutation importance quantifies the drop in model performance that occurs when the values of a feature are shuffled at random. Socioeconomic characteristics and university-level competencies were recognized as the top features.

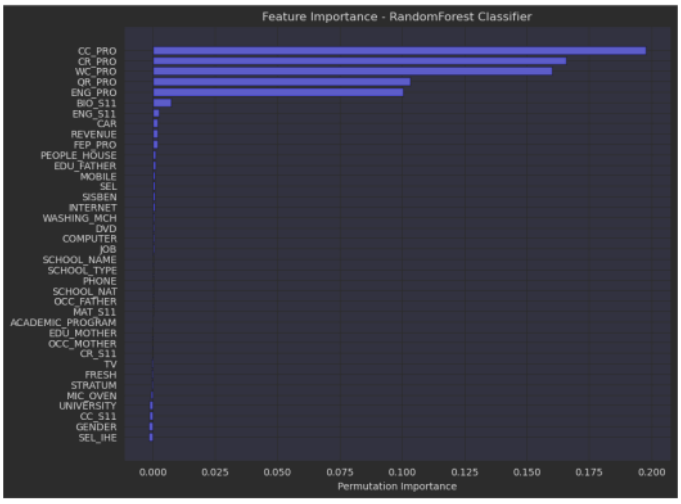

Fig 4.9 Permutation Importance for Random Forest

- **Top Features:** Critical Reading (CR_PRO), Citizen Competencies (CC_PRO), Quantitative Reasoning (QR_PRO), and socioeconomic status.

4.2.2 SHAP Values

SHAP values provide a detailed explanation of each feature's contribution to the model's predictions. The SHAP summary plot and dependence plots offer insights into the impact of various features.

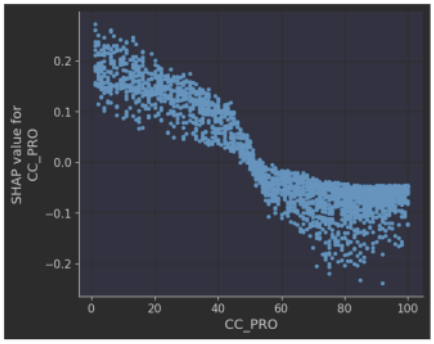

Fig 4.10 Dependence Plot for Q1

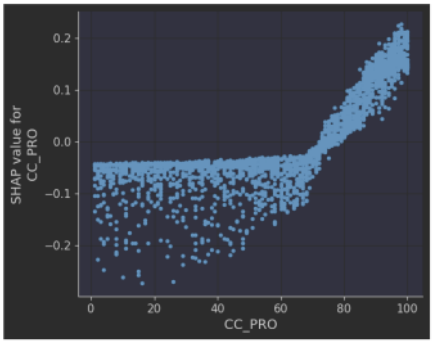

Fig 4.11 Dependence Plot for Q4

Key Insights: The SHAP dependence plots for Citizen Competencies (CC_PRO) in both Q1 and Q4 show a consistent transition from negative to positive SHAP values as CC_PRO scores increase. Lower scores in CC_PRO negatively impact predictions for both quartiles, while higher scores positively influence academic performance. This transition zone, around CC_PRO scores of 50-70, highlights the critical role of Citizen Competencies. Higher CC_PRO scores are linked to better academic outcomes, underscoring the importance of these competencies in predicting and enhancing student performance.

4.2.3 LIME Explanations

LIME focuses on explaining individual predictions, providing a locally faithful interpretable model around each prediction.

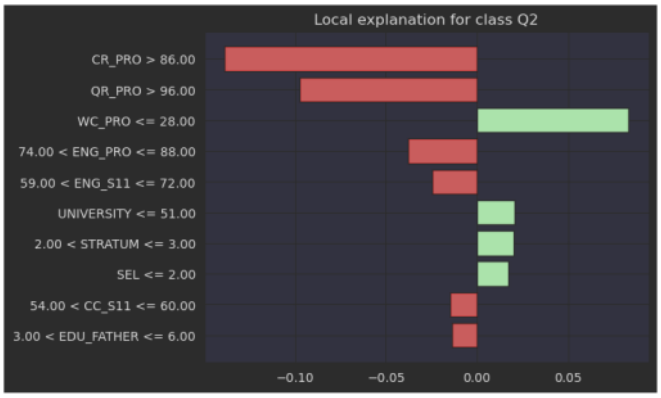

Fig 4.12 LIME Plot Explanation for a single instance

- **Specific Features:** Critical Reading (CR_PRO), Quantitative Reasoning (QR_PRO), and Written Communication (WC_PRO) significantly influence predictions for individual students.

4.3 Comparative Analysis

A comparative analysis of the results highlights the strengths and weaknesses of each model, providing a comprehensive understanding of their performance.

- **Gradient Boosting:** Consistently high accuracy and minimal misclassifications, making it the best overall performer.
- **CNN:** Strong predictive power, particularly excelling in Q1 and Q4.
- **Random Forest:** Robust performance with clear feature importance insights.
- **GLMNet:** High accuracy and interpretability, making it a balanced choice for predicting student performance.

Table 4.2 Comparative Analysis

Study	Dataset	Model	Accuracy
Current Research	ICFES, Colombia (Standardized Test Scores, Socioeconomic)	Gradient Boosting Machine (GBM)	92 %
Alhazmi and Shenea mer (2023) [8]	Jazan University, Saudi Arabia (Admission Scores, First-Year Grades)	Gaussian Naive Bayes (GNB)	74%
Soto-Acevedo et al. (2023) [2]	ICFES, Colombia (Standardized Test Scores, Socioeconomic)	Generalized Linear Network Model (GLMNet)	82%
Poudyal et al. (2022) [17]	OULAD (LMS Interaction, Demographics, Assessment)	Hybrid 2D CNN	88%
Aljaloud et al. (2022) [21]	University of Ha'il, Saudi Arabia (Blackboard Interaction)	CNN-LSTM	94%

4.4 Findings and Insights

The analysis reveals several key findings:

- **University-Level Competencies:** Critical Reading (CR_PRO) and Citizen Competencies (CC_PRO) are crucial for predicting academic success.

- **Socioeconomic Factors:** These play a significant role, with varying degrees of influence across different models.
- **Model Performance:** Ensemble methods like Gradient Boosting and Random Forest are highly effective for complex predictive tasks in educational data mining.

These findings underscore the importance of targeted educational interventions and policies to support students' academic journeys, particularly those from diverse socioeconomic backgrounds. The insights gained from this study can inform the development of data-driven strategies to promote equity and excellence in engineering education.

By leveraging advanced machine learning techniques and interpretability tools, this research provides a comprehensive understanding of the factors influencing academic performance in engineering education, offering practical implications for educators and policymakers.

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Conclusion

This study aimed to investigate the factors influencing academic performance in engineering students by leveraging a comprehensive dataset comprising academic, social, and economic information from 12,411 students in Colombia. Using various machine learning models, we predicted student performance and analyzed the importance of different features. Our findings provide valuable insights into the determinants of academic success and offer practical implications for enhancing educational outcomes in engineering programs.

5.1.1 Key Findings

The key findings of this research are summarized as follows:

- **University-Level Academic Competencies:** Critical Reading (CR_PRO) and Citizen Competencies (CC_PRO) emerged as the most significant predictors of student success, highlighting the importance of these skills in engineering education.
- **Secondary-Level Performance:** Strong performance in secondary-level subjects, particularly mathematics and science, also contributed to higher academic achievement at the university level.
- **Socioeconomic Factors:** Socioeconomic status, parental education, and access to resources like internet and household appliances played significant roles in influencing student performance. These factors demonstrated varying degrees of impact across different models.
- **Model Performance:** Among the machine learning models evaluated, Gradient Boosting and Convolutional Neural Networks (CNN) achieved the highest predictive accuracy, followed closely by Logistic Regression (GLMNet) and Random Forest. These models demonstrated robust performance in predicting student performance quartiles.
- **Interpretability:** Techniques such as SHAP values and LIME provided valuable insights into the contributions of different features, enhancing the interpretability of the models, and enabling a deeper understanding of the factors influencing academic performance.

5.1.2 Implications for Engineering Education

The findings of this study have several implications for educators and policymakers:

- **Curriculum Development:** Emphasizing the development of critical reading, problem-solving, and communication skills within engineering curricula can better prepare students for academic success and future careers.
- **Targeted Interventions:** Implementing early interventions and support systems for students struggling in key areas identified by the feature importance analysis, such as critical reading and citizen competencies, can help improve educational outcomes.
- **Equity in Education:** Addressing the potential impact of socioeconomic disparities on academic achievement through initiatives that provide equitable access to resources and support for students from disadvantaged backgrounds is crucial for promoting equity in engineering education.
- **Data-Driven Decision Making:** Utilizing data mining and machine learning techniques can provide educators and policymakers with a deeper understanding of student performance patterns, enabling the development of evidence-based strategies to enhance learning outcomes.

5.2 Future Work

While this study provides valuable insights into the factors influencing academic performance in engineering education, several areas warrant further investigation. Future research could build on the findings of this study by exploring the following directions:

5.2.1 Dataset Specificity and Generalizability

The findings of this study are based on data from Colombia and may not be directly generalizable to other contexts. Future research could explore similar analyses in different countries and educational systems to validate the results and identify context-specific factors influencing academic performance.

5.2.2 Regression Models for Continuous Performance Prediction

This study utilized classification models to predict student performance quartiles. Exploring regression models to predict continuous performance scores could provide more granular insights into the factors influencing academic achievement and help develop more precise interventions.

5.2.3 Longitudinal Studies

Tracking student performance over time through longitudinal studies can provide a deeper understanding of how different factors interact and evolve throughout the academic journey. This approach can help identify critical periods for intervention and support.

5.2.4 Incorporating Additional Factors

Exploring the influence of additional variables such as learning styles, motivation, engagement, and psychological factors on academic performance could provide a more comprehensive understanding of the determinants of student success.

5.2.5 Personalized Learning and Adaptive Systems

Developing adaptive learning systems and interventions tailored to individual student needs and learning styles based on predictive models and data analysis can enhance educational outcomes. Future research could explore the integration of personalized learning approaches within engineering education.

5.2.6 Causal Inference

Utilizing advanced statistical techniques to establish causal relationships between identified factors and academic performance can provide a deeper understanding of the underlying mechanisms influencing student outcomes. Moving beyond correlation-based analyses to causal inference can help develop more effective strategies for improving educational practices.

5.3 Closing Remarks

This research contributes to the growing field of educational data mining and offers practical implications for improving student support systems and educational outcomes in higher education. By leveraging advanced machine learning techniques and interpretability tools, this study provides a comprehensive understanding of the factors influencing academic performance in engineering education. The insights gained from this research can inform the development of data-driven strategies to promote equity and excellence in engineering education, ensuring that all students can succeed regardless of their socioeconomic background.

Through continued exploration and innovation in educational data mining, researchers and practitioners can develop more effective and targeted interventions, ultimately enhancing the quality and accessibility of engineering education worldwide.

References

[1] Delahoz-Domínguez, E., Zuluaga-Ortiz, R., Herrera, T.J.F.: Dataset of academic performance evolution for engineering students. Data in Brief 30, 105537 (2020).

[2] Soto-Acevedo, M., Abuchar-Curi, A.M., Zuluaga-Ortiz, R., Delahoz-Domínguez, E.: A machine learning model to predict standardized tests in engineering programs in Colombia. IEEE-RITA 18(3), 211–218 (2023).

[3] Khan, M., Naz, S., Khan, Y., Zafar, M., Khan, M., & Pau, G., "Utilizing machine learning models to predict student performance from LMS activity logs," IEEE Access, vol. 11, pp. 86953–86962, (2023)

[4] Rizvi, S., Rienties, B., & Khoja, S. A., "The role of demographics in online learning; A decision tree based approach," Computers and Education/Computers & Education, vol. 137, pp. 32–47, (2019)

[5] Rivas, A., Gonzalez-Briones, A., Hernandez, G., Prieto, J., Chamoso, P., "Artificial neural network analysis of the academic performance of students in virtual learning environments," Neurocomputing, vol. 423, pp. 713–720, (2021)

[6] Zohair, Abu, and Lubna Mahmoud. "Prediction of Student's performance by modelling small dataset size." International Journal of Educational Technology in Higher Education 16.1 (2019): 1-18.

[7] Azizah, E.N.; Pujianto, U.; Nugraha, E. Comparative performance between C4. 5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment. In Proceedings of the 4th International Conference on Education and Technology (ICET), Malang, Indonesia, pp. 18–22. (2018)

[8] Alhazmi, E., Sheneamer, A.: Early predicting of students performance in higher education. IEEE Access 11, 27579–27589 (2023).

[9] Sun, D., et al.: A university student performance prediction model and experiment based on Multi-Feature Fusion and Attention Mechanism. IEEE Access 11, 112307–112319 (2023).

[10] Delahoz-Domínguez, E., Zuluaga-Ortiz, R., Herrera, T.J.F.: Dataset of academic performance evolution for engineering students. Data in Brief 30, 105537 (2020).

[11] Ramesh, V., Parkavi, P., Ramar, K.: Predicting Student Performance: A statistical and data mining approach. International Journal of Computer Applications 63(8), 35–39 (2013).

[12] Bydžovská, H.: A Comparative Analysis of Techniques for Predicting Student Performance. International Educational Data Mining Society (2016).

[13] Kabakchieva, D.: Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies 13(1), 61–72 (2013).

[14] Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A., Durán-Domínguez, A.: Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences 10(3), 1042 (2020).

[15] Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F., Alowibdi, J.S.: Predicting Student Performance using Advanced Learning Analytics. In: 26th International Conference on World Wide Web Companion, pp. 1-2 (2017).

[16] Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., Schmidt-Thieme, L.: Recommender system for predicting student performance. Procedia Computer Science 1(2), 2811–2819 (2010).

[17] Poudyal, S., Mohammadi-Aragh, M.J., Ball, J.E., "Prediction of student academic performance using a hybrid 2D CNN model," Electronics, vol. 11, no. 7, p. 1005, (2022)

[18] Chau, V.T.N., Phung, N.H., "Enhanced CNN models for binary and multiclass student classification on temporal educational data at the program level," Vietnam Journal of Computer Science, vol. 08, no. 02, pp. 311–335, (2020)

[19] S. Kanani, S. Patel, R. K. Gupta, A. Jain, and J. C.-W. Lin, "An AI-Enabled ensemble method for rainfall forecasting using Long-Short term memory," Mathematical Biosciences and Engineering, vol. 20, no. 5, pp. 8975–9002, (2023)

[20] Nabil, A., Seyam, M., Abou-Elfetouh, A., "Prediction of students' academic performance based on courses' grades using deep neural networks," IEEE Access, vol. 9, pp. 140731–140746, (2021)

[21] Aljaloud, A.S., Uliyan, D.M., Alkhalil, A., Abd Elrhman, M., Alogali, A.F.M., Altameemi, Y.M., Altamimi, M., Kwan, P., "A deep learning model to predict student learning outcomes in LMS using CNN and LSTM," IEEE Access, vol. 10, pp. 85255–85265, (2022)

[22] Pallathadka, Harikumar, et al. "Classification and prediction of student performance data using various machine learning algorithms." *Materials today: proceedings* 80 (2023): 3782-3785.

[23] Sweeney, Mack, et al. "Next-term student performance prediction: A recommender systems approach." *arXiv preprint arXiv:1604.01840* (2016).

[24] Kim, Byung-Hak, Ethan Vizitei, and Varun Ganapathi. "GritNet: Student performance prediction with deep learning." *arXiv preprint arXiv:1804.07405* (2018).

[25] Delianidi, Marina, et al. "Student performance prediction using dynamic neural models." *arXiv preprint arXiv:2106.00524* (2021).

[26] Su, Yu, et al. "Exercise-enhanced sequential modeling for student performance prediction." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

[27] Pandey, Mrinal, and Vivek Kumar Sharma. "A decision tree algorithm pertaining to the student performance analysis and prediction." *International Journal of Computer Applications* 61.13 (2013): 1-5.

[28] Yang, Fan, and Frederick WB Li. "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining." *Computers & Education* 123 (2018): 97-108.

ytrts

ORIGINALITY REPORT

6%	4%	6%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Emerging Trends and Applications in Artificial Intelligence", Springer Science and Business Media LLC, 2024 Publication	1%
2	www.ncbi.nlm.nih.gov Internet Source	1%
3	fastercapital.com Internet Source	1%
4	www.oer.unn.edu.ng Internet Source	1%
5	"Computer Information Systems and Industrial Management", Springer Science and Business Media LLC, 2022 Publication	<1%
6	dokumen.pub Internet Source	<1%
7	www.mdpi.com Internet Source	<1%
8	deepnote.com Internet Source	<1%

9	link.springer.com Internet Source	<1 %
10	Bayan Alnasyan, Mohammed Basheri, Madini Alassafi. "The Power of Deep Learning Techniques for Predicting Student Performance in Virtual Learning Environments: A Systematic Literature Review", Computers and Education: Artificial Intelligence, 2024 Publication	<1 %
11	Submitted to Kaunas University of Technology Student Paper	<1 %
12	Liu Zhen, Alina Bărbulescu. "Comparative Analysis of Convolutional Neural Network-Long Short-Term Memory, Sparrow Search Algorithm-Backpropagation Neural Network, and Particle Swarm Optimization-Extreme Learning Machine Models for the Water Discharge of the Buzău River, Romania", Water, 2024 Publication	<1 %
13	Máté Baranyi, Marcell Nagy, Roland Molontay. "Interpretable Deep Learning for University Dropout Prediction", Proceedings of the 21st Annual Conference on Information Technology Education, 2020 Publication	<1 %

14	acikbilim.yok.gov.tr Internet Source	<1%
15	pdfs.semanticscholar.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 15 words
Exclude bibliography	On		