Navigating Educational Paths: Optimizing Student Prediction Models via Feature Selection Analysis

Mousumi Banerjee[a], Priyanka Dhar[a], Piyali Naha[a], Supriyo Saha[a]

[a]Department of Computer Science, Asutosh College, Kolkata, 700026, India

{banerjeemousumi538, priyankadhar812, piyalinaha9, supriyosaha108}@gmail.com

1 Introduction and Motivation

The Cortez and Silva [1] use of data mining techniques in the education sector has gained significant interest in recent years. Data mining involves extracting valuable insights and discovering new, potentially useful information from large datasets, aiming to identify new trends and patterns through the application of various classification algorithms [2]. In the contemporary educational landscape, accurately predicting and understanding student performance is of great importance. This provides educators with insights into learning patterns and empowers students to identify areas for improvement, fostering an effective learning environment. This study explores the application of machine learning techniques for predicting academic performance, addressing a crucial gap in current educational research.

Technology-enhanced learning (TEL) platforms are among the primary sources of data within the academic environment. For instance, virtual campus management platforms, which are now commonplace in education, generate data from a variety of educational management tasks such as the virtualization of teaching, monitoring of student academic progress, storing teaching materials, and tracking student interactions with them. These data and the tools that produce them present significant opportunities for innovative research that can benefit not only the educational community but also contribute to the advancement of knowledge. Research areas such as predicting students' behaviour, developing new learning support tools, recommending resources, preventing dropouts, and enhancing activities can all be pursued using these data sources. To achieve these goals, computer and data sciences offer advanced methods for processing and analyzing data to extract knowledge. Techniques from fields such as data mining, big data, machine learning, deep learning, collaborative filtering, and recommender systems enable the development of advanced methods that provide substantial potential for these research purposes, resulting in new applications and more effective approaches to academic analysis and prediction.

Educational data mining (EDM) is an emerging field that utilizes data mining techniques to analyze educational data. Data is collected from learning platforms and examined to obtain insights into current educational practices with the goal of improving existing capabilities. These insights can involve tracking individual student progress, setting up alerts for planning actions that boost student retention, or monitoring course performance [3]. By leveraging user-generated data, institutions can better plan and implement targeted intervention strategies.

Implementing a student prediction system offers multifaceted benefits to educational institutions.

Firstly, it empowers educators with data-driven insights to personalize learning experiences, enabling them to identify and address students' individual needs effectively. Secondly, it fosters a proactive approach to student support, allowing institutions to intervene early and prevent potential academic challenges before they escalate. Moreover, by leveraging predictive analytics, institutions can optimize resource allocation, ensuring that students receive the right level of support at the right time. Ultimately, investing in a student prediction system reflects a commitment to student success and continuous improvement, positioning educational institutions at the forefront of educational innovation.

In this study, we examined recent real-world data from two Portuguese secondary schools. The data came from two sources: mark reports and questionnaires. Given the limited information in the mark reports, which included only grades and number of absences, the data was supplemented with information from questionnaires. This provided various demographic, social, and school-related attributes (e.g. student's age, alcohol consumption, and mother's education). The primary goal is to predict student achievement and, if possible, identify key variables that influence educational success or failure. We modelled the two core subjects (Mathematics and Portuguese) under three data mining objectives:

I. Binary classification (pass or fail); and

II. Classification into five levels (ranging from I, denoting very good or excellent, to V, indicating insufficient); and

III. Regression, producing a numeric output ranging from zero (0%) to hundred (100%).

For each of these approaches, the data is pre-processed and a percentage of it is selected to minimize computation and remove unnecessary information. This reduced data is then used for testing different machine-learning algorithms. Additionally, an explanatory analysis is conducted on the best-performing models to identify the most relevant features, providing insights into which factors contribute most significantly to the model's success.

2 Literature Review

Recent research on student prediction systems focuses on harnessing data analytics and machine learning techniques to forecast student performance and identify potential areas of concern. These systems aim to predict various outcomes, such as academic success, dropout rates, and the need for academic interventions. Cortez and Silva [1] utilized predictive analytics in Portuguese secondary schools to forecast student success and identify key influencing factors, employing classification and regression supervised learning methods. Techniques such as decision trees, random forests, neural networks, and support vector machines were leveraged, focusing on factors like demographic attributes, prior assessments, attendance, and socio-economic status. However, they used all features without justifying the importance of different features. On the other hand, Osmanbegovi and Suljic [4] explored the influence of socio-demographic factors, high school performance, and attitudes towards studying on the success of first-year students at the University of Tuzla, utilizing Naive Bayes, Multilayer Perceptron, and C4.5 decision tree algorithms. The study highlighted the superior performance of the Naive Bayes classifier in a traditional classroom setting, post-data collection. Similarly, Ahmed and Elaraby [5] aimed to enhance student performance predictions using data mining techniques such as classification and decision trees on data from 2005 to

2010, employing the ID3 algorithm for constructing decision trees and using entropy and information gain for optimal classification. Another study by Lakkaraju et al. [6] in U.S. district schools compared machine learning models like SVM, RF, LR, Adaboost, and DT, finding Random Forest to outperform others in accuracy and recall. The research followed rigorous procedures including the FP-Growth algorithm for pattern identification and risk score-based student ranking. In addition, Al-Shehri et al. [7] analyzed two Portuguese secondary schools' data to compare the effectiveness of SVM and KNN in predicting math performance, highlighting SVM's superior performance with a correlation coefficient of 0.96. Another study by Mansur and Yusof [8] applied k-means clustering to enhance student performance and behaviour in e-learning systems, focusing on teacher-student interactions and identifying key attributes affecting student outcomes. A comparative study and research designed by Ramaswam et al. [9] assessed the effectiveness of EDM using data from Xorro-Q, employing Naïve Bayes, LR, and kNN classifiers, showing a remarkable 88% accuracy with RF outperforming other models. In another study by Pratama and Husnayaini [10], k-means clustering was used to analyze PISA results across 17 Asian countries, aiming to discern patterns that could guide improvements in educational strategies within the region. The following year, a study by Mustapha [11] focused on the Open University Learning Analytics dataset, employing advanced regression, classification, and feature engineering techniques. The use of models like XGBoost and Neural Networks underscored the growing integration of deep learning methods to predict educational outcomes with high accuracy and precision. Looking ahead to 2024, research by Gupta et al. [12] was set to perform a comprehensive comparative analysis of diverse classification, regression, and clustering algorithms on varied datasets. This work further refined the effectiveness of predictive models in education.

3 Methodology

In this section, we describe the proposed method for predicting student achievement and, when feasible, identifying key factors that influence educational success or failure. The proposed model selects important features from the actual dataset and disregards less relevant ones. By applying various machine learning algorithms to the chosen features, we aim to emphasize the most impactful aspects. The following subsections outline each module of the proposed system.

3.1 Preprocessing

Before proceeding with further processing, several crucial preprocessing steps have been completed to ensure data quality and suitability for analysis. The dataset has been checked for duplicates and cleaned to maintain integrity. Data types of each column have been verified and adjusted as necessary to align with the expected formats. The number of unique values in each column has been assessed to understand the distribution and diversity of data. This step helps identify areas for potential grouping or encoding. Descriptive statistics like mean, median, standard deviation, and percentiles have been examined to understand data tendencies and variability. This analysis helps identify outliers and areas needing normalization or scaling.

Finally, categorical columns have been reviewed to understand category range and frequency. This insight assists in choosing suitable encoding techniques and deciding whether any categories need merging or redefining. These preprocessing steps have prepared the dataset for advanced analysis or modelling, ensuring it is clean, structured, and ready for use.

3.2 Random Forest Regression

Random Forest Regression is an ensemble method used in machine learning for both regression and classification tasks. It builds multiple decision trees and combines their outputs through a method known as Bootstrap Aggregation, or "bagging". This approach enhances accuracy by reducing reliance on individual decision trees and mitigating overfitting. In this technique, each tree is constructed using a randomly selected subset of the data, known as bootstrap sampling, encompassing both row and feature selection. This diversity in training samples boosts the model's robustness and generalization ability.
Random Forest aggregates the predictions from individual trees by averaging them in regression tasks or using majority voting in classification, leading to more accurate and stable results. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables.

The main steps involved in the Random Forest algorithm :

1. In the Random Forest algorithm, each decision tree is built from a randomly selected subset of data. Specifically, from a dataset containing k records, n random instances and m features are chosen for each tree. It ensures for no two alike trees to enhance model's diversity.

2. Separate decision trees are constructed using each of the samples derived in the first step. It uses only a fraction of available features for each tree thus preventing overfitting and improving model performance.

3. Every individual tree in the ensemble is independent allowing for the parallel processing of trees and predicts an outcome based on the input data.

4. Random Forest inherently partitions the data, keeping approximately 30% as an unseen dataset that functions similarly to a test set, which helps in validating the model without a separate train-test split.

5. The final output of the Random Forest is determined by using Majority Voting for classification tasks or Averaging for regression tasks. This method aggregates the predictions of all trees to produce a single, more accurate prediction.

Feature importance : In Random Forest Regressor, we have used this to give feature importance and quantifies the relative contribution of each input variable to the predictive power of the model. This metric identifies which features most strongly influence the target variables to focus on the most impactful data and reduce the model's complexity by eliminating less important features.

The final feature importance, at the Random Forest level, is it's average over all the trees :

$$RFfi_i = \frac{\sum_{j \in \text{ all trees}} \text{ norm } fi_{ij}}{T} \tag{1}$$

$RFfi_i$ = the i calculated from all trees in the Random Forest model,
$normfi_{ij}$ = the normalized feature importance for i in tree j,
T = total number of trees

3.3 Classification

In our classification section, we have employed a rich array of machine learning algorithms tailored to address diverse challenges with finesse and precision. Among these algorithms, AdaBoost Regressor (ABR) stands out for its sequential training approach, wherein it iteratively improves predictive performance by focusing on instances that were misclassified in previous iterations, thus enhancing overall accuracy. CatBoosting Regressor (CBR), distinguished by its robust handling of categorical features and ability to automatically handle missing data, offers a powerful solution for tackling complex datasets with high-dimensional categorical variables. Random Forest Regressor (RFR), an ensemble method renowned for its versatility and capability to handle non-linear relationships and interactions among features, has been instrumental in delivering reliable predictions across various scenarios. Additionally, K-Neighbours Regressor (KNR) has been leveraged for its simplicity and effectiveness in capturing local patterns in the data, making it particularly suited for situations where the underlying data distribution is not easily characterized by parametric models. By harnessing the complementary strengths of these diverse algorithms, our classification framework is poised to excel in a wide range of applications, ensuring robust performance and adaptability to evolving data landscapes.

4 Results and Analysis

In this section, we examine the performance of the proposed detection method and conduct an in-depth analysis of the results. The upcoming subsections offer detailed insights into our experiment and its findings.

4.1 Dataset description

To assess the effectiveness of our proposed detection approach, we carried out experiments on Student Performance [1] which is publicly accessible. The data focuses on student achievement in secondary education across two Portuguese schools. It includes attributes such as student grades, demographic details, social factors, and school-related information, which were gathered from school reports and questionnaires. There are two datasets available, each covering a different subject: Mathematics (mat) and Portuguese language (por). The datasets comprises multivariate data with 30 features and 649 instances in Portuguese and 395 instances in Mathematics. We divided the dataset using 80% for training and 20% for testing.

4.2 Comparison of results with and without feature selection

In the 5-level classification conducted utilizing Table 1, each grade corresponds to a distinct level of performance, ranging from excellent/very good to fail in the subject. Grades A through V represent a spectrum of achievement, with grade A indicative of outstanding performance and grade V signifying a failure. The scores associated with each grade provide a quantitative measure of performance, facilitating the classification process and enabling a nuanced assessment of student achievement across different proficiency levels. This classification scheme offers a comprehensive framework for evaluating performance in the subject, accommodating a wide range of abilities and outcomes.

Table 1: The system with five distinct levels of classification.

	I	II	III	IV	V
Country	(excellent/very good)	(good)	(satisfactory)	(sufficient)	(fail)
Portugal	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

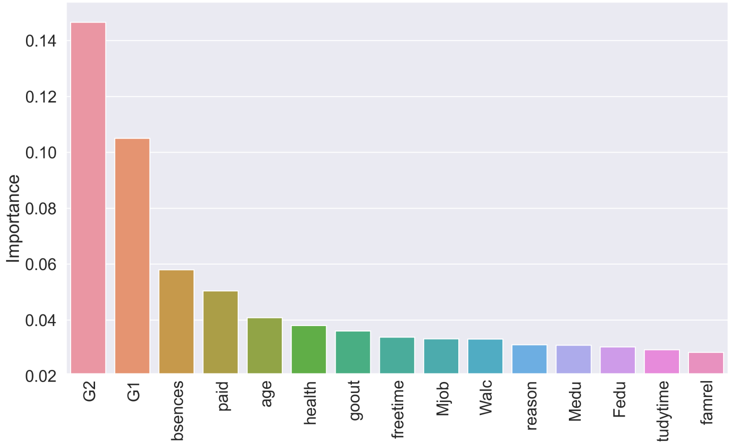


Figure 1: Top 15 features and their importance scores.

The findings of our study underscore the significance of feature selection in enhancing the predictive performance of classification models applied to both the Mathematics and Portuguese datasets. The comparison between Table 2, which represents the performances without feature selection, and Table 3, which showcases the results after employing the top 15 selected features (Fig. 1 shows the top 15 features with their score), highlights a notable improvement ranging from 2-3% in accuracy. This enhancement not only validates the effectiveness of our feature selection approach but also demonstrates its utility in optimizing model generalization and robustness. The observed increase in accuracy post-feature selection reaffirms the notion that prioritizing informative features can mitigate the curse of dimensionality and improve the models' ability to discern meaningful patterns within the data. Overall, these results underscore the pivotal role of feature selection in refining classification models and achieving higher predictive accuracies in both binary and multi-level classification tasks across diverse datasets. The graph depicted in Fig. 2 illustrates the impact of feature selection on the accuracy of both Mathematics and Portuguese datasets. Across most instances, feature selection has notably improved accuracy values. However, it's

Table 2: Performance comparison of different setups without utilizing features selection, with accuracy values expressed in percent (%).

	Mathematics				Portuguese			
Input Setup	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Binary Classification	67.2	65.1	60.92	60.01	41.68	44.02	45.59	50.27
5-class Classification	62.14	64.7	62.85	53.21	38.00	45.34	48.62	43.31
Performance Prediction	80.44	81.06	80.92	77.86	71.73	80.11	83.62	77.33

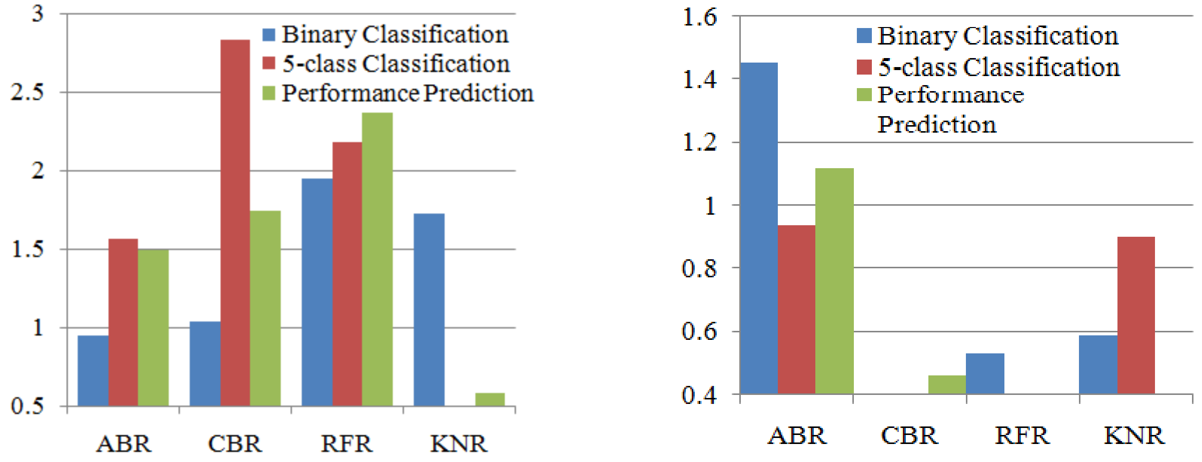
Table 3: Performance comparison of different setups utilizing top 15% features, with accuracy values expressed in percent (%).

	Mathematics				Portuguese			
Input Setup	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Binary Classification	68.15	66.14	62.87	61.74	43.13	43.86	46.12	50.86
5-class Classification	63.70	67.54	65.04	53.22	38.94	45.25	44.67	44.21
Performance Prediction	81.93	82.81	83.29	78.45	72.85	80.57	79.91	75.35

worth noting that for certain scenarios within the Portuguese dataset, the accuracy did not witness an increase post-feature selection. Fig. 3 depicts the confusion matrices for binary classification and 5-level prediction in both the Mathematics and Portuguese datasets. Upon examination of these figures, it becomes evident that the model demonstrates satisfactory performance across both categories. The visualization provides insight into the model's ability to accurately classify instances within the datasets. Overall, the results suggest that the model exhibits competence in handling both binary and multi-class predictions for Mathematics and Portuguese datasets.

4.3 Ablation study

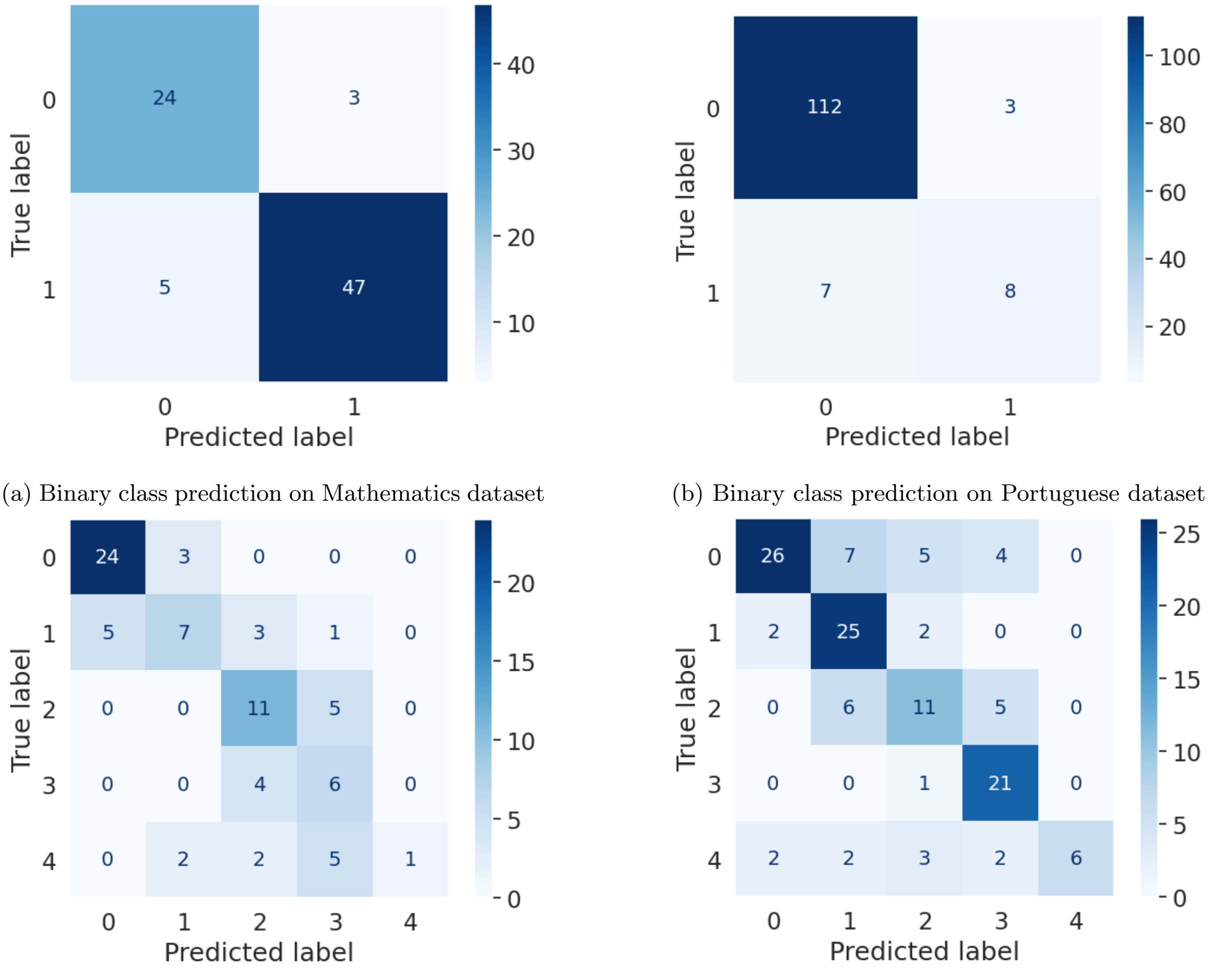
We conducted an ablation study on Performance Prediction to assess the impact of removing key features, such as G1 and G2, on our results. Initially, we removed G1 and predicted G3 using the top 14 features, then repeated the process by removing G2 and predicting G3. Through meticulous analysis, we systematically eliminated the G2 and G3 columns, revealing the varying influence of these variables on our analysis. Surprisingly, our findings highlighted the crucial role of the second grading period (G2) over the first (G1). Removing G2 resulted in a notably greater decline in results compared to removing G1, emphasizing its significance in outcome prediction. This observation not only underscores the importance of temporal progression in academic performance but also emphasizes the need for comprehensive assessment frameworks. The disparate impacts of G2 and G1 shed light on the intricate dynamics within our



(a)

(b)

Figure 2: Accuracy increased after feature selection in case of (a) Mathematics and (b) Portuguese datasets.



(a) Binary class prediction on Mathematics dataset

(b) Binary class prediction on Portuguese dataset

(c) 5-class prediction on Mathematics dataset

(d) 5-class prediction on Portuguese dataset

Figure 3: Confusion matrices under different setup.

dataset, providing valuable insights for enhancing predictive models and educational strategies. Table 4 exclusively presents the study results for Performance Prediction.

Table 4: Performance comparison of different setups with and without some features. Accuracy values expressed in percent (%).

	Mathematics				Portuguese			
Input Setup	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Without G1 only	78.52	79.26	80.98	75.42	70.07	77.93	77.68	74.44
Without G2 only	69.15	74.25	74.47	67.48	56.40	68.20	65.64	62.70

5 Future Implementation Plan

Education plays a vital role in our society. Business Intelligence (BI) and Data Mining (DM) techniques offer intriguing opportunities in the education sector by extracting high-level insights from raw data. Many studies have applied these methods to improve educational quality and optimize resource management. This paper focuses on predicting secondary student grades in two key subjects, Mathematics and Portuguese, using past academic performance, demographic, social, and other school-related data. In our future work, we aim to enhance the student prediction system by merging various feature engineering processes for more comprehensive data insights. Additionally, we will test different modern machine learning algorithms to identify those with the best performance and generalization capabilities. This approach seeks to develop a more effective model for guiding targeted interventions and supporting student success.

References

[1] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[2] R. Shaun, J. De Baker, and P. Inventado, *Chapter 4: Educational data mining and learning analytics*, 2014.

[3] D. J. A. Lewis, "The smart university: The transformational role of learning analytics," *Information and Learning Science*, vol. 119, no. 12, pp. 758–760, 2018.

[4] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," 2012.

[5] A. B. E. D. Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," 2014.

[6] H. Lakkaraju, C. S. Everaldo Aguiar, N. B. David Miller, R. Ghani, and K. L. Addison, "A machine learning framework to identify students at risk of adverse academic outcomes," 2015.

[7] H. Al-Shehri, L. A.-S. Amani Al-Qarni, H. B. Arwa Batoaq, J. A. Saleh Alrashed, and S. O. Olatunji, "Student performance prediction using support vector machine and k-nearest neighbor," 2017.

[8] A. B. F. Mansur and N. Yusof, "The latent of student learning analytic with k-mean clustering for student behaviour classification," 2018.

[9] G. Ramaswami, T. Susnjak, and P. G. Anuradha Mathrani James Lim, “Using educational data mining techniques to increase the prediction accuracy of student academic performance,” 2019.

[10] D. Pratama and I. Husnayaini, “Program for international student assessment (pisa) analysis of asian countries using k-mean clustering algorithms,” 2022.

[11] S. M. F. D. S. Mustapha, “Predictive analysis of students’ learning performance using data mining techniques: A comparative study of feature selection methods,” 2023.

[12] S. Gupta, B. Kishan, and P. Gulia, “Comparative analysis of predictive algorithms for performance measurement,” 2024.