Computers and Education: Artificial Intelligence 7 (2024) 100301

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models

Abderrafik Laakel Hemdanou *, Mohammed Lamarti Sefian, Youssef Achtoun, Ismail Tahiri

Applied Mathematics and Computer Science Team, Higher Normal School, Avenue Mohamed V, BP 209, Martil, 93150, Tetouan, Morocco

ARTICLE INFO

Keywords:
Selection and extraction features
Machine learning
Educational data mining
Student performance

ABSTRACT

Education is at the core of developmental progress, necessitating the exploration and implementation of diverse contemporary methods to ensure the success of students across multiple levels. However, impediments to this success exist, categorized into three primary groups which are, individual factors, family factors and social factors. These factors can manifest as absenteeism and boredom, posing a threat to the future of both students and society at large. Addressing these challenges proves to be a complex task for educators and pedagogues, given the unique problems each student faces. This paper aims to employ a broad spectrum of feature selection and extraction methods, some unconventional in the education sector but proven reliable in other domains. By integrating machine learning (ML) and deep learning (DL) models, we seek to predict student performance based on these identified factors. Subsequently, a comparative analysis will be conducted to determine the most effective model, considering the relevance of various factors.

1. Introduction

Academic success stands as the paramount objective of all educational systems, with student's performance serving as the crucial determinant of this achievement. Nevertheless, certain students encounter academic challenges, prompting teachers, parents, and counselors to collaborate in earnest efforts to address and overcome these issues. Student academic performance serves as a metric through which individuals can track their academic progress, leading to certification. Various types of evaluations, including oral presentations and practical training, contribute to assessing a student's level of academic competence. In a broader sense, the measurement of student performance relies on chosen objectives and criteria, utilizing diverse standards such as GPA (Grade Point Average). The GPA is the basis for studies on academic performance, which used as a tool to simplify a set of chores, solve a collection of problems or as a tool to help with decision-making.

Shahiri et al. (2015) has proved that the decision trees are less efficient than neural networks at predicting new data sets, although they are still simpler than many other approaches. Because neural networks can identify every potential combination of features, they are very useful. Nevertheless, to mitigate biased data, caution is essential, as highlighted by Rashid and Aziz (2016). Their study suggests that a neural network comprising four neurons may suffice for predicting student performance, incorporating various factors such as the student's department and tutor. Conversely, certain factors such as GPA, appear to carry less significance in this predictive model. However, since it is based on a certain student group, it can be prejudiced. Artificial neural networks (ANNs), as in the study of Isong et al. (2018), provide a reliable technique for modeling, forecasting and prediction with an accuracy of over 99 on samples of computer science students. According to a study by Trakunphutthirak et al. (2019), Random Forest demonstrates the highest accuracy rate among various machine learning techniques. However, the comparison with neural networks remains inconclusive due to a lack of information regarding the number of nodes in the neural network model, making it challenging to determine their relative performance. A variety of classification algorithms, including Naïve Bayes and the Decision tree, are used in other research to predict student performance based on different criteria, while Naïve Bayes seems to produce superior results, concerns arise regarding the small size of the dataset. More dynamic data, such homework progress and classroom conduct, are crucial for forecasting student achievement, according to Alireza Ahadi et al. (Ihantola et al., 2015), they come to the conclusion that Random Forest provides the best accuracy. Data mining is emphasized in the paper by Mohammed Afzal Ahamed et al. (2019) as an essential first step before utilizing machine

* Corresponding author.
E-mail addresses: abderrafik.laakelhemdanou@etu.uae.ac.ma (A. Laakel Hemdanou), lamarti.mohammed.sefian@uae.ac.ma (M. Lamarti Sefian), achtoun44@outlook.fr (Y. Achtoun), istahiri@uae.ac.ma (I. Tahiri).

https://doi.org/10.1016/j.caeai.2024.100301
Received 15 March 2024; Received in revised form 10 September 2024; Accepted 15 September 2024
Available online 1 October 2024

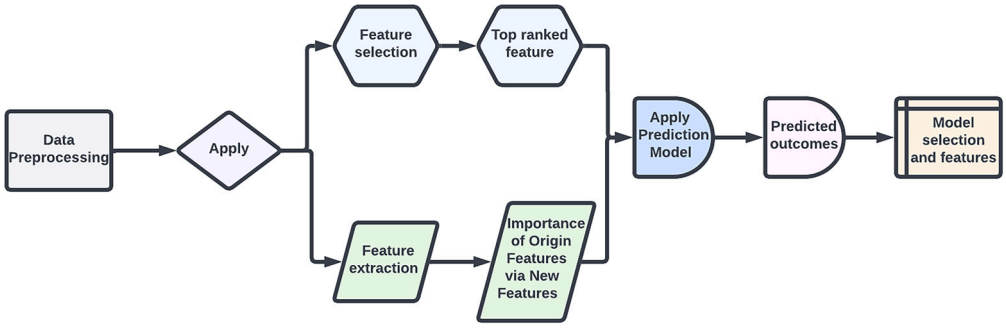

Fig. 1. The process to be used.

learning techniques. The article by Arto Hellas et al. (2018) provides a summary of a large body of research on the prediction of academic performance, it emphasizes the significance of metrics like retention, GPA and individual course grades, but it also makes the argument that machine learning techniques are frequently not appreciably superior to other approaches in this field.

Our research involves leveraging established feature selection and extraction methods from diverse real-world datasets and adapting them for application within the educational context. Specifically, we've tailored FSMRMR, Autoencoder, and GAN models to identify crucial factors influencing student performance, comparing them against commonly used models. Our objective is to interpret the significance of each factor and validate their impact using various ML and DL methods. Visually depicted within our research strategy, we aim to effectively predict student grades, providing concrete evidence of the effectiveness of our selected elements and models. Broadly, our research is grounded in a diverse range of methods often overlooked in educational studies, with a focus on applying ML and DL algorithms known for their predictive capabilities. The strategy that will be used in this research is shown in Fig. 1.

From our research objectives, several fundamental questions emerge:

- What strategic design decisions can be proposed to facilitate the adoption and widespread use of various ML and DL models in modeling high school students academic performance
- How do decision trees, random forests, K-nearest neighbors (KNN), Support Vector Regression (SVR), Deep Neural Networks (DNN), Gated Recurrent Units (GRU), and Transformers perform in predicting student performance when applied to features selected and extracted through specific methods?
- What are the comparative abilities of these models in processing various feature types for predicting student performance?
- How can different statistical methods and machine learning algorithms be utilized to identify and analyze key factors influencing high school students academic achievement, and what is the impact of various feature selection and extraction techniques on the accuracy of predictive models for designing personalized educational strategies and informing educational policy?

2. Methodology

2.1. Selection of features

In the areas of machine learning and data analysis, using various dimension reduction techniques has several benefits, these consist of reduced complexity and increased generalizability of machine learning models can be achieved by lowering the dimension of the data, which also helps to minimize noise and redundancy. Dimension reduction minimizes model complexity by removing features that are not included in the model, which can assist decrease overfitting and assure a faster learning by lowering the number of dimensions, we can process less data, which can hasten learning and model training. This is particularly significant for big datasets. Saving IT resources such memory and processing time by minimizing the quantity of data. By converting data into a more understandable set of features, dimension reduction can help analysts better grasp the relationships between variables. Some dimension reduction techniques allow for the selection of the most crucial features, which can be helpful for removing noise or pinpointing the elements that are most pertinent to a given task. Some dimension reduction techniques can be used with particular machine learning techniques, such as logistic regression and regression analysis. To ensure the suitability of dimension reduction methods, several conditions need to be taken into account, the most important of which is the distribution of the data, dimension reduction methods often assume that the data are distributed linearly or non-linearly, and to determine the linearity of the relationships, a statistical analysis and study was carried out, finding that the negative skewness coefficient, such as the value is -1.993, suggests that the distribution has a longer tail on the left side compared to the tail on the right. In other words, there is a greater concentration of high values on the right-hand side of the distribution, pulling the mean to the left. A skewness coefficient of -1.993 indicates a relatively high degree of skewness. And for the kurtosis coefficient is 8.820 indicates positive kurtosis, meaning that the distribution has thicker tails and higher peaks than the normal distribution. A positive kurtosis generally indicates a leptokurtic distribution, meaning that values are more concentrated around the mean than in a normal distribution, and distribution tails are thicker, which may result in more frequent extreme values than expected by a normal distribution. The Durbin-Watson test is 1.918, we can say that the statistic approaches 2, indicating low auto correlation in the residuals. The Jarque-Bera (JB) test is 819.029, this value is relatively high. In general, a high JB value suggests that the data do not follow a normal distribution. The "Prob(JB)" value associated with the Jarque-Bera (JB) test: 1.41e-178 is an extremely small p-value, close to zero. A very small p-value suggests strong evidence against the null hypothesis of normality. In other words, the probability of observing such an extreme JB statistic in a normal distribution is practically zero.

In order to reduce the dimensions and identify the most relevant features by eliminating noise (Al-Zawqari et al., 2022), two approaches can be used, either selection or extraction of features (Htun et al., 2023), both approaches have the same objectives but differ in their methods.

Feature selection involves identifying and choosing a subset of pertinent features (variables, predictors) for constructing a model, aims to retain useful features by eliminating those that are not useful, thereby enhancing the performance of machine learning models by reducing overfitting and improving accuracy. On the other hand, feature extraction creates new features from the original data by transforming or combining the original set of features to produce more meaningful ones, selection methods are usually classified (Gong et al., 2022) into four categories: filtering, wrapper methods, embedded methods and information-theoretic methods. The filtering methods (Beckham et al., 2023) rank the variables according to their relevance using statistics such as correlation or distance between the factors and the output variables. For this purpose, the PCC (Pearson correlation coefficient) method was selected, with a threshold of 0.5, as it has proven to be effective.

There are also other correlation coefficients such as the Spearman coefficient. The general PCC formula is given as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where:

- cov is covariance,
- σ_X is standard deviation of X,
- σ_Y is standard deviation of Y.

Wrapping methods integrate feature selection into the learning process of an ML algorithm, they use the performance of the predictor to obtain an optimal subset of features and thus improve the accuracy of the predictions (Htun et al., 2023).

However, these methods are known to be computationally expensive as they require several iterations, to avoid over fitting, a wrapper-like feature selection technique called RFE (Recursive feature elimination) was chosen, this technique relies on an iterative procedure to train an ML model at each iteration, RFE computes the ranking criterion for all features, eliminates those with the lowest importance score and then trains the model again based on the newly selected feature set, the algorithm which will be used is Algorithm 1, such that X is the feature matrix $(n \times p)$, y is the target vector $(n \times 1)$ and β is the vector of regression $(p \times 1)$.

Algorithm 1: Recursive feature elimination with elastic net regression.

Require: X , y , $p\_features\_to\_select$ (number of features to select)
Ensure: Selected subset of features
  Normalize X
  while number of features in $X > p\_features\_to\_select$ **do**
    Fit Elastic Net to X, y to obtain vector β
    Calculate feature importance using the magnitude of β
    Identify the feature i with the smallest $|\beta_i|$
    Remove feature i from X
  end while
  return Remaining features in X

Embedded methods combine the advantages of filtering and wrapping methods by integrating feature selection into the learning process, they use an algorithm for analysis and feature selection simultaneously as part of the learning process (Htun et al., 2023), integrating both algorithmic modeling and feature selection (Villa-Blanco et al., 2023). For this category, we chose to use the deep adversarial generative model due to its success in feature selection in recent years. GAN (Liu et al., 2021) (Generative Adversarial Network) is composed of two deep neural networks in competition the generator and the discriminator, the generator $G(z)$ takes a noise z as input and generates an output similar to the real data while the discriminator $D(x)$ takes as input the data x either real or generated and the loss function implies a competition between generator and discriminator such the generator seeks to minimize the fact that the discriminator can distinguish between real and generated data, while $D(x)$ seeks to maximize its ability to distinguish, we use cross-entropy, so we define the discriminator and generator as follows:

$$\max_D J(D) = E_{x\sim p\ \mathrm{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z\ (z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right], \tag{2}$$

$$\min_G J(G) = E_{z\sim p_z\ (z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right], \tag{3}$$

the GAN can be defined as:

$$\min_G \max_D J(D,G) = E_{x\sim p\ \mathrm{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z\ (z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]. \tag{4}$$

Information-theoretic methods use mutual information (MI) to evaluate the importance of each feature. In this category, we applied the FSMRMR (Minimum redundancy maximum relevance) method (Gong et al., 2022), which demonstrated superior performance to the CMIM (Conditional Mutual Information Maximization) method on several data sets.

Algorithm 2: Algorithmic description of FSMRMR.

Require: Define series X and Y; feature set S; class C
Ensure: Selected feature subset S_{best}
1: Mutual Information:
2: $\mathrm{MI}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$
3: Correlation $D(S)$:
4: $D(S) = \frac{1}{|S|}\sum_{f_m\in S}\mathrm{MI}(f_m;C)$
5: Redundancy $R(S)$:
6: $R(S) = \frac{1}{|S|^2}\sum_{f_m,f_n\in S}\mathrm{MI}(f_m;f_n)$
7: mRMR Criterion:
8: $\phi(S) = D(S) - R(S)$
9: Feature Selection:
10: $S_{\mathrm{best}} = \arg\max_S \phi(S)$

2.2. Feature extraction

The task of feature extraction methods is to reduce the number of features in a dataset by creating new features that synthesize most of the information contained in the original feature set (Htun et al., 2023), two types of feature extraction techniques were identified in the reviewed studies which are statistical techniques and optimization-based techniques. A common statistical method for dimensionality reduction is principal component analysis (PCA) (Jolliffe, 2002), it transforms a high-dimensional feature vector into a low-dimensional feature vector with uncorrelated components (Htun et al., 2023), this technique calculates the eigenvectors of the covariance matrix of the original features to create new components that summarize most of the information contained in the original feature set, in our study, we examined the eigenvectors or coefficients associated with each principal component to understand how the original features contribute to each component. The features with the highest absolute coefficients in the first principal components are generally the most important. And for optimization-based methods we have chosen the autoencoder (AE), it is an unsupervised learning model based on a neural network developed by Kramer in 1991 (Nguyen & Quanz, 2021), reconstructs the inputs of the neural network in its output layer, composed of an encoder and a decoder, the AE reduces the dimension of the input to a codeword and then uses this dimension to reconstruct the input to its original size using the decoder. For our work, we have chosen a variational autoencoder (VA), in general, its architecture combines autoencoder and probabilistic generative models, in our study, when the VA model is trained, latent variable analysis is used to understand which features of the original data are captured in the latent space. The most important factors were determined by analyzing correlations between latent dimensions and original features to identify important relationships. VA is defined by encoder $q_{\varphi(z|x)}$ which is the product of a probabilistic distribution q on the latent space z as a function of the input x, and generator $p_{\theta(x|z)}$ which is the product of probabilistic distribution p on the data space x as a function of the latent variable z, the loss function which comprises two components, the Kullback-Leibler divergence between the distributions $q_{\varphi(z|x)}$ and $p(z)$ a priori on z and a measure of the reconstruction quality of $p_{\theta(x|z)}$ with respect to the input x,

$$L = E_{q_{\varphi(z|x)}}\left[\log p_{\theta(x|z)}\right] - KL\left[q_{\varphi(z|x)}||p(z)\right], \tag{5}$$

$$\mathrm{KL}\left(q(x)\parallel p(x)\right) = \int \log\frac{q(x)}{p(x)}\, q(x)\,\mathrm{d}x = E_{q(x)}\left[\log(q(x)) - \log(p(x))\right]. \tag{6}$$

Let f be the function representing the decoder. In our application, we wish to minimize the reconstruction error between the decoder output and the encoder input, the global cost function of the variational auto-encoder is the solution to the following optimization problem:

$$f^* = \arg\max_f E\left[\varphi(z \mid x)\right]\left[\log\left(p\left(x \mid z\right)\right)\right] + \mathrm{KL}\left(q_\varphi\left(z \mid x\right) \parallel p(z)\right) \tag{7}$$

$$f^* = \arg\min_f E\left[\varphi(z \mid x)\right]\left[\|x - f(z)\|\right] + \mathrm{KL}\left(q_\varphi\left(z \mid x\right) \parallel p(z)\right) \tag{8}$$

2.3. Predictive model

2.3.1. ML models

The volume of data is adequate to train a predictive model (Priyambada et al., 2023; Qiu et al., 2022), the data are clean and without missing values, the models chosen are appropriate for regression (Okoye et al., 2024), we considered that the data are sequential such that the use of models must be respected according to the specific order of the data in the original data base, in order to adapt certain sequential models and use them in prediction, data normalization and scaling is a very important step before the essential prediction step, to avoid over-fitting we used cross-validation to evaluate the performance of our models and divided our data into training, validation and test sets to evaluate model generalization. Using various feature extraction and feature selection techniques, we will attempt to identify the main variables that affect students performance, we will then use each set of variables to predict the grades of students using various machine learning and deep learning models to demonstrate the true impact of these characteristics on students performance, it should encourage everyone involved in education to focus on these factors in order to improve educational outcomes.

Decision tree: The decision tree is an inductive algorithm that is rarely used, which is divided into two categories, classification trees and regression trees according to the explanation of the value based on continuous or discrete variables (Beckham et al., 2023), it is non-linear and non-parametric, the objective of this algorithm is to partition according to the prediction of variable explains.

Random forest: Random forest is one of the ensemblistic methods, the principle of Random forest is to assemble a set of estimators into a very complex estimator, such that each estimator has a fragmented view of the problem to solve it. Let be a data (X, Y), the construction of Radom Forest is realized by drawing randomly a sample B noted (X_B, Y_B), for each couple of data (X_B, Y_B), a decision tree is trained, When we apply these B trees on new data, we obtain B answers. The final answer is determined by taking the majority of the B answers obtained. We introduce a random draw on the variables to be used, by default $\sqrt{n}$ for an n variable problem, feature sampling is a major idea that contributes very strongly to reducing the variance of the set created. Indeed, an average of B independent and identically distributed variables, each with variance σ^2 has a variance of $\frac{\sigma^2}{B}$ if we exclude the assumption of independence of the variables which is often the case in reality, the variance of the set is (noting the correlation coefficient ρ of the pairs of variables), we assume the random forest regression at new point x is defined by,

$$\hat{f}_{\mathrm{rf}}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{9}$$

where $\{T_b\}_1^B$ is the ensemble of trees chosen by the algorithm.

KNN: The principle of the KNN (K-nearest neighbors) algorithm is to classify data according to their distance from a number k of training samples. This type of algorithm is commonly called “base instance” because it uses no parameters for training, the model assumes that the distance is sufficient to make an inference, without making any assumptions about the underlying data or its distribution. The optimal choice of the value of k is crucial for the KNN algorithm. However, the use of this method can be hampered by the “curse of dimensions”, where the distance measurement can become inaccurate in high dimensions. This may be due to the fact that the differences between the nearest and farthest neighbors are reduced, which can make classification more difficult.

Let be m classes C_j such $j \in \{1, \ldots m\}$ and let be n learning points named P_i each belonging to a class C_i for $i \in \{1, \ldots n\}\}$ and let a target point note T of unknown class, so the KNN algorithm consists in calculating the distance between each point P_i and T, such that for the k points closest to T will count the number of occurrences of each class and assign to T the most frequent class.

To measure the distance between two points $\mathrm{X}\left(X_1, Y_1\right)$ and $\mathrm{Y}\left(X_2, Y_2\right)$, we can use the Euclidean distance

$$d_{X,Y} = \sqrt{\left(X_2 - X_1\right)^2 + \left(Y_2 - Y_1\right)^2} \text{ in } \mathbb{R}^2, \tag{10}$$

we can use other distances such as manhatan, chebychev, mahalanobis, haversine, hamming, canberra, braycurtis, jaccard, matching, dice, rogerstanimoto, russellrao, sokalmichener and sokalsneath.

SVM: The Support Vector Machine (SVM) algorithm is very efficient for detecting complex and non-linear patterns, it uses two fundamental concepts which are maximizing the distance between the decision boundary and choosing a separating hyperplane in a space of non-linear combinations between variables. This space allows a linear separation of the data, making the classification easier and more accurate, let the vector θ and α be the parameters of our model, as well as the inputs x_i and the labels y_i, in the case of linear inseparability, SVM is defined in a formulation where the constraint is integrated into the objective function, this formulation is called dual or Lagrangian, where the objective of SVM is to solve the following optimization problem:

$$\min\ \pounds(\theta, \alpha) = \frac{1}{2}\theta^T\theta + \sum_{i=1}^{m}\alpha_i y_i\ \left(\theta^T x_i - 1\right) \tag{11}$$

constrained $\alpha_i \geq 0\ \ \forall\, i \in \{1, \ldots, m\}$

After optimizing the hyper parameters of SVR, the kernel used is the RBF (Radial Basis Function) kernel, also known as the Gaussian or radial kernel.

2.3.2. DL models

DNN: A deep neural network (DNN) (Nabil et al., 2021; El Fouki et al., 2019) is a type of artificial neural network (Rodríguez-Hernández et al., 2021) made up of several layers of neurons. The more layers a network contains, the deeper it is. These networks are used in the field of deep learning to solve complex tasks such as pattern recognition, classification and natural language processing. Learning in DNNs is generally done by backpropagation and optimization, where the network adjusts its weights and biases to minimize the error between predictions and true target values. DNNs have shown excellent performance in many applications, the loss function is chosen for regression, for parameter updating the stochastic gradient descent algorithm is used, weights and biases are updated using the SGD optimization (Mehmood et al., 2023).

RNN-GRU (Recurrent Neural Networks-Gated Recurrent Units): Recurrent neural networks (RNN) are a class of neural network architectures that are designed to process sequential or temporal data, unlike classical neural networks, RNNs are able to take into account the context and history of data by using recurrent connections, but there are problems encountered such as problem of the exploding or disappearing gradient, difficulty in capturing long-term dependencies, Sensitivity to the order of the sequence, LSTM (Long Short-Term Memory) architecture has solved these problems, it based on the three gates of the LSTM (Forget gate, Input gate, Output gate) are replaced by the reset gate and the update gate, as for the LSTM, these gates use sigmoid activations which constrain their values to be between 0 and 1. The reset gate determines how much of the previous state we want to keep in memory, while the update gate allows us to control how much the new state is a replica of the old state, in a GRU (Zhang et al., 2021), these gates take into account both the current time step input and the hidden state of

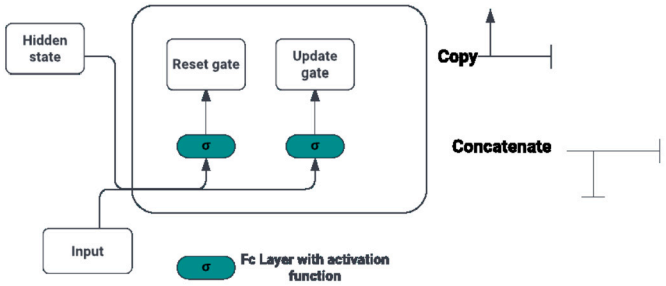

Fig. 2. The GRU architecture.

Table 1
Distribution of characteristics for methods.

PCC	RFE	GAN	FSMRMR	PCA	VA
Medu	age	failures	age	age	sex
higher	failures	schoolsup	Medu	Fedu	age
Fedu	freetime	paid	Fedu	reason	studytime
Failures	goout	internet	Mjob	studytime	failures
Age	Dalc	romantic	failures	activities	activities
goout	Walc	health	higher	famrel	nursery
reason	health	absences	romantic	freetime	higher
romantic	absences	G1	goout	goout	Dalc
G1	G1	G2	G1	health	G1
G2	G2	freetime	G2	absences	G2

the previous time step, as shown in Fig. 2, the outputs of both gates are computed using fully connected layers with a sigmoid activation function.

In the subject of mathematics, let's say that at a given time step t, the input is a minibatch with the following properties, minibatch $X_t \in \mathbb{R}^{n\times d}$ (number of examples n, number of inputs d) and the hidden state of the previous time step is $H_{t-1} \in \mathbb{R}^{n\times h}$ (number of hidden units h). Then, the reset gate $R_t \in \mathrm{IR}^{n\times h}$ and update gate $Z_t \in \mathbb{R}^{n\times h}$ are computed as follows

$$R_t = \sigma(X_t W_{\mathrm{xr}} + H_{t-1} W_{\mathrm{hr}} + b_r), Z_t = \sigma(X_t W_{\mathrm{xz}} + H_{t-1} W_{\mathrm{hz}} + b_z), \quad (12)$$

where

$$\sigma(X) = \frac{1}{1+e^{-X}}, \quad W_{\mathrm{xr}}, W_{\mathrm{xz}} \in \mathbb{R}^{d\times h},$$

$W_{\mathrm{hr}}, W_{\mathrm{hz}} \in \mathbb{R}^{h\times h}$ are weight parameters,

$\boldsymbol{b}_r, \boldsymbol{b}_z \in \mathbb{R}^{1\times h}$ are bias parameters.

Tansformer Attention: The transformer is a very efficient algorithm (Satake et al., 2021), used in most cases for natural language processing, this model is a network fully connected with the attention mechanism The attention mechanism can look backward and forward so that the model knows the context of the current input features the transformer (Kusumawardani & Alfarozi, 2023) consists of two main elements the encoder and the decoder, for our model, an input layer is used for each feature, then these layers are concatenated to form the model's global input, from which multi-headed attention layers are added to capture the interactions between the different features, feedforward layers to transform the information. Finally, a dense output layer to predict the student's performance (Liu et al., 2023).

The general architecture of the transformer (Vaswani et al., 2017) is shown in the Fig. 3. In the Transformer model, we distill the core functionality into two primary equations: one for the encoder and another for the decoder. Let's start with the encoder's formula:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V\ . \quad (13)$$

In this equation Q, K, and V stand for the query, key, and value matrices, correspondingly. The term d_k represents the dimensionality of each key vector.

Moving on to the decoder's critical equation, it leverages the encoder's output represented as $H = \{h_1, h_2, \ldots . h_n\}$ and the sequence of target outputs up to the current step, denoted as $Y_{<t} = \{y_1, y_2, \ldots . y_{t-1}\}$. The attention mechanism in the decoder mirrors the encoder's with an adjustment to ensure it does not access future information:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V\ . \quad (14)$$

Where $Q{=}Y_{<t}W^Q$, $K{=}HW^K$ and $V{=}HW^v$, such that W^Q, W^K and W^v are learned weight matrices. This setup in the decoder facilitates focusing on relevant parts of the input sequence without peeking ahead into the future tokens, ensuring each output is generated based on the preceding context.

The methodology of this study is explained and visualized in the two Figs. 4 and 5, which represent the research objectives.

3. Result

The dataset used is a dataset of students from two schools in Portugal, which is downloaded from https://archive.ics.uci.edu/ml/datasets/~student 2Bperformance, the data consists of 395 students (Beckham et al., 2023) and 33 features are represented in the table. To evaluate these various models (Beckham et al., 2023), we used RMSE and MAE which are defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{y}_i - \hat{\mathbf{y}}_i\right|, \quad (15)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}, \quad (16)$$

where $\hat{y}$ are the predicted values, y values are the observed values and n is the number of observations. After pre-processing, preparation and cleaning of the data, the factors with a strong correlation with the target variable $G3$ were determined using different methods of feature selection and extraction.

The Table 1, in the form of a distribution, summarizes the different results obtained by using selection and extraction of features. This distribution, which can allow us to identify the most significant qualities that greatly influence the student's performance, was derived from the results of the methods used for dimension reduction and the calculation of the importance of each variable for the target variable (Bilal et al., 2022). This distribution can be seen in the Table 1.

In the results section of our study, we present a comprehensive analysis of the feature selection and extraction techniques applied to the dataset concerning high school students' academic performance. The techniques include Pearson Correlation Coefficient (PCC), Recursive Feature Elimination (RFE), Generative Adversarial Networks (GAN), Feature Selection using Mutual Information Maximization (FSMRMR), Principal Component Analysis (PCA) and Variational Autoencoder (VA). Our findings indicate a diverse range of features identified as significant by the different methods. For instance, PCC highlighted 'Medu' (mother's education level) 'G1' (first period grade) and 'G2' (second period grade), and 'failures' (number of past class failures), with 'failures' being consistently recognized across multiple methods, denoting its strong influence on student performance. The color in the matrix represents the frequency and significance of features selected by each method, with 'age' and 'failures' frequently appearing in blue and green, respectively, suggesting their prominence in the predictive models.

RFE and GAN identified 'age' and 'failures' as well, but also brought attention to 'goout' (frequency of going out with friends), which is marked in red, indicating a potential correlation with academic success. Interestingly, RFE, PCA, and VA all identified 'age' as a key feature, which may point to the importance of developmental stages in academic

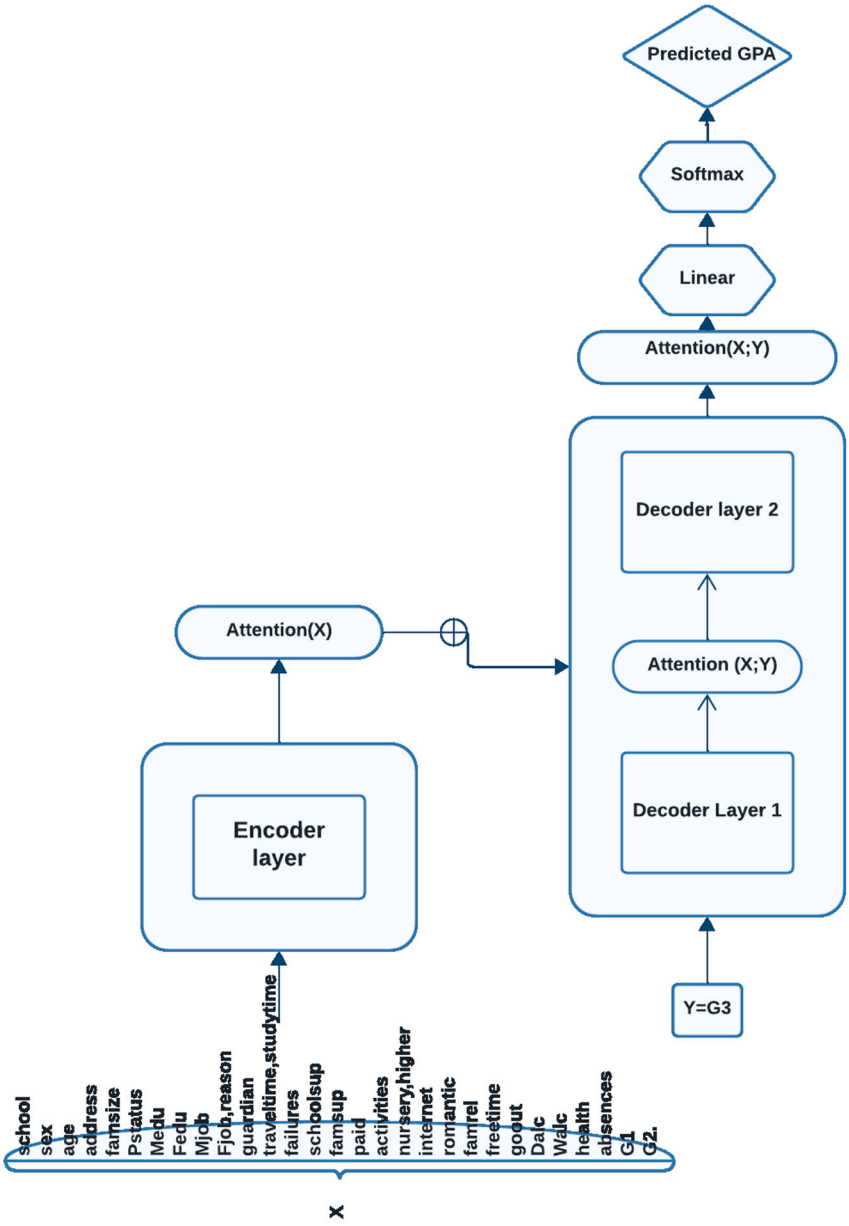

Fig. 3. The transformer architecture model.

achievement. The advanced algorithms like GAN and FSMRMR have also identified 'G1' (first period grade) and 'G2' (second period grade), highlighted in cyan and yellow, as critical predictors of student performance. This underscores the importance of continuous assessment and the cumulative nature of academic achievement.

In summary, our analysis suggests that while there are commonalities in feature selection across methods, each technique has its unique contributions, revealing a multifaceted set of factors that influence academic performance. These results can guide the development of targeted interventions and personalized pedagogical approaches to support student achievement. The integration of these findings into educational policy can help address the underlying causes of school failure and promote student engagement and success.

From Table 2, we delve into the significance of various characteristics that influence student performance, as determined by our feature selection and extraction methods. The table provided summarizes the importance of each characteristic, expressed as a percentage, and the cumulative sum of these percentages. The analysis reveals that the combination of 'Age', 'G1' (first period grade), and 'G2' (second period grade), and 'Failures' is the most significant, accounting for 8.33% of importance individually and collectively contributing to 33.32% when considering their combined impact, which represents the most important percentage in the table. This suggests that both the maturity level of the students, their previous academic track record and continuous academic performanceare critical factors in predicting their academic performance.

Table 2
Distribution of characteristics for methods.

Characteristics	Importance in %	Sum
Age-Failures-G1-G2	**8.33%**	**33.32%**
Go out	6.66%	6.66%
Absences-Health-Romantic-Fedu-Freetime-Higher	5%	30%
Dalc-Medu-Studytime-Activities-Reason	3.33%	16.66%
Walc-Schoolsup-Paid internet-Mjob-Famrel-Nursery	1.66%	10%

The next set of characteristics, which includes 'Go out', holds a combined importance of 6.66%, leading to a total of 6.66%. This indicates that social habits are also strong indicators of overall student success. A

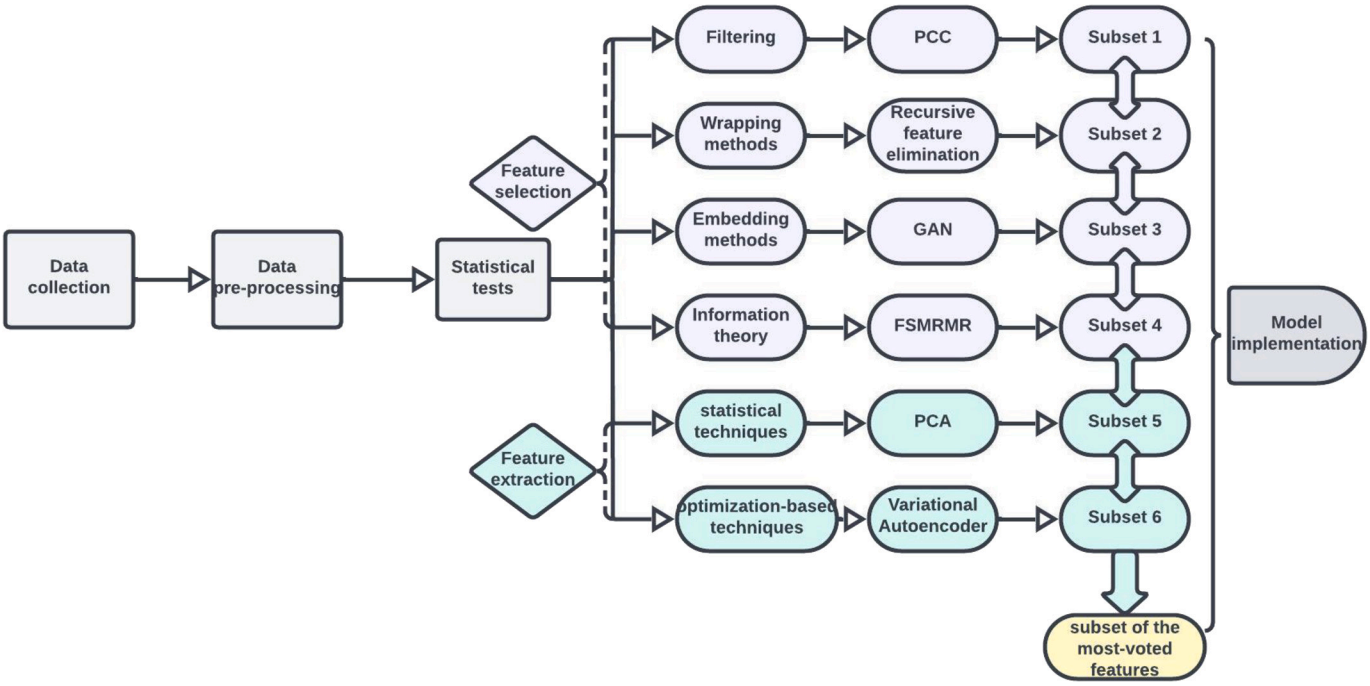

Fig. 4. Feature selection and extraction.

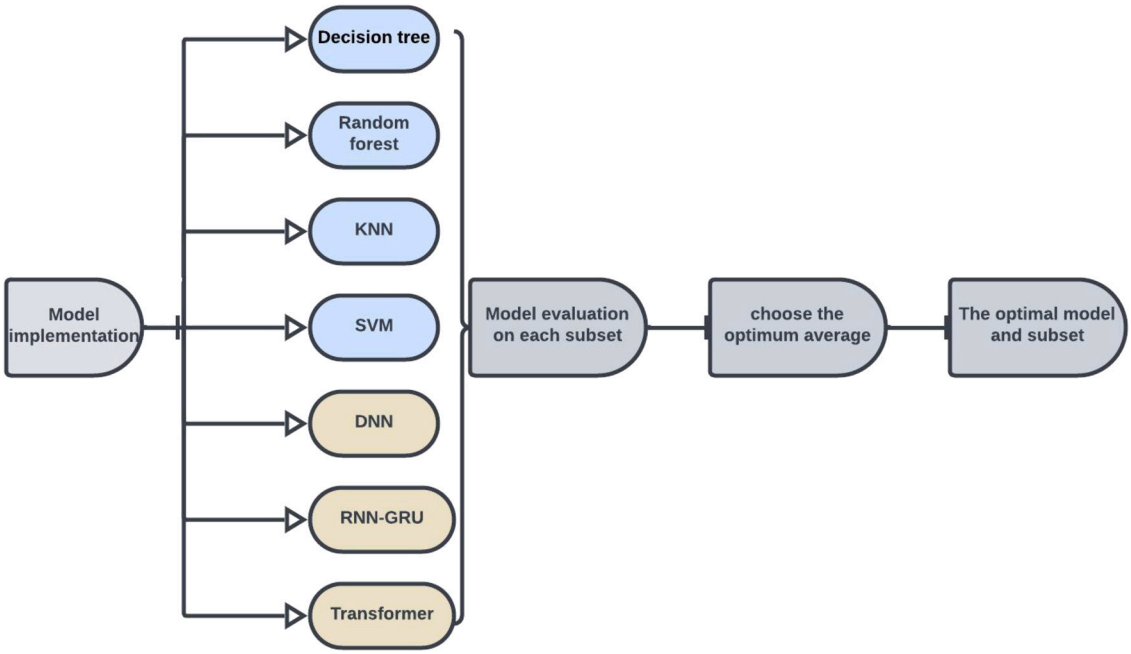

Fig. 5. ML and DL implementation and evaluation.

diverse group of characteristics, namely 'Absences', 'Health', 'Romantic', 'Fedu' (father's education level), 'Freetime', and 'Higher' (desire for higher education), each accounts for 5% of importance. Together, they contribute a significant 30%, underscoring the multifaceted nature of factors that can affect a student's academic journey. Characteristics such as 'Dalc' (weekday alcohol consumption), 'Medu' (mother's education level), 'Studytime', 'Activities', and 'Reason' for choosing the school, each have an importance of 3.33%, summing up to 16.66%. These factors suggest that both parental influence and personal habits play a role in educational outcomes.

Finally, the characteristics with the least individual importance at 1.66% include 'Walc' (weekend alcohol consumption), 'Schoolsup' (extra educational support), 'Paid' (extra paid classes), 'Internet' access, 'Mjob' (mother's job), 'Famrel' (quality of family relationships), and 'Nursery' (attended nursery school). Collectively, they account for 10%, indicating that while they may have a smaller individual impact, they still contribute to the overall picture of student performance.

Therefore, our results section highlights the varying degrees of importance of different student characteristics in predicting academic performance. The findings emphasize the need for a holistic approach in educational strategies, considering a wide array of factors that can influence a student's ability to succeed academically. From this distribution, which represents the results of all the methods used for the selection and extraction of features, the characteristics most selected by these methods and with a high correlation with student performance which are student's age, the number of past class failures, going out with friends,

Table 3
Metric Evaluation of all models on Datasets of each method.

	PCC	RFE	GAN	FSMRMR	PCA	VA	SF
Decision tree	5.78-4.42	2.48-1.30	2.65-1.54	2.70-1.43	2.25-1.20	2.79-1.64	2.52-1.31
Random forest	4.57-3.59	1.83-1.04	2.14-1.25	2.41-1.53	1.97-1.17	2.06-1.32	1.95-1.15
KNN	4.67-3.71	2.14-1.39	2.05-1.30	2.33-1.44	2.32-1.59	2.15-1.35	1.88-1.24
SVM	4.31-3.41	2.02-1.20	2.01-1.21	2.09-1.13	2.03-1.20	2.08-1.13	2.09-1.14
DNN	4.23-3.43	2.25-1.58	2.25-1.53	2.31-1.64	2.20-1.47	2.28-1.51	2.20-1.40
GRU	4.41-3.60	2.57-1.37	2.54-1.35	2.64-1.39	2.57-1.42	2.59-1.43	2.67-1.38
Transformer	0.23-0.08	0.161-0.04	0.168-0.04	**0.09-0.01**	0.11-0.02	0.10-0.02	0.12-0.09

Table 4
Average Metric Evaluation of each model on all subsets and each method subset on all models.

Methods	PCC	RFE	GAN	FSMRMR	PCA	VA	SF
$A-mt$	4.02-3.17	1.92-1.13	1.97-1.17	2.08-1.22	1.92-1.15	2.0-1.20	**1.91-1.10**
Models	Decision tree	Random forest	KNN	SVM	DNN	GRU	Transformer
$A-mo$	3.03-1.83	2.41-1.57	2.50-1.71	2.37-1.48	2.53-1.79	2.85-1.70	**0.13-1.70**

first period grade, second period grade, the number of school absences, the current health status, with a romantic relationship, father's education, free time after school, wants to take higher education, this subset is noted by SF. The Table 3 summarizes the results of the various models on the different sets of features obtained from selection or extraction and on SF with the first value representing the RMSE and the second the MAE, we denote A_{mt} by the average obtained by each selection and extraction method on the sets of models and A_{mo} by the average obtained by a model on a subset of the methods.

Our analysis reveals that the Transformer model significantly outperforms all other models on the datasets, with the lowest RMSE and MAE values across the board. Notably, on the FSMRMR method's dataset, the Transformer model achieves an exceptionally low RMSE of 0.09 and MAE of 0.01, indicating its superior predictive capability. The Random Forest and Decision Tree models also show strong performance on the SF subset, with RMSE values of 1.95 and 2.52, respectively, and MAE values of 1.15 and 1.31, respectively. This suggests that these models are particularly adept at handling the SF feature set, which contains the most critical characteristics for predicting student performance. The K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Deep Neural Network (DNN), and Gated Recurrent Unit (GRU) models exhibit moderate performance, with their respective RMSE and MAE values indicating reasonable predictive accuracy. However, they are outclassed by the Transformer model's exceptional results. In conclusion, our results highlight the effectiveness of the Transformer model in handling complex feature sets for the prediction of student performance. The SF subset, which encapsulates the most influential characteristics, allows for improved model accuracy across various machine learning algorithms. These insights can be instrumental in guiding the development of predictive models in educational settings, enabling educators and policymakers to better understand and enhance student achievement.

From Table 4 and based on $A-mt$, we note that the subset SF represents the most optimal set among the other subsets, which justifies the validity of this set and this approach, and based on $A-mo$, Transformer model significantly outperforms other models, standing out for its ability to capture complex patterns and incorporate contextual information relevant to school performance prediction. This superiority is particularly noticeable in situations where feature selection and extraction techniques are crucial for refining models, underlining the applicability and effectiveness of transform in the specific context of predicting student performance. In general, and based on the results in Table 2, we can conclude that the Transformer model remains the best-performing of the models tested, we instead of what's interesting is that the best result obtained is with the FSMRMR method, which shows the usefulness of this method in uncovering the most important factors in predicting student performance, so we can deduce that the student's age and mother's education, have a significant impact on student performance, father's training mother's occupation, the quantity of prior academic failures, the desire to pursue higher education while engaging in romantic relationships hanging out with pals they've highest correlation with G3 is as follows. Figs. 6–54 show the learning curves of different models on different subsets of selection and extraction features.

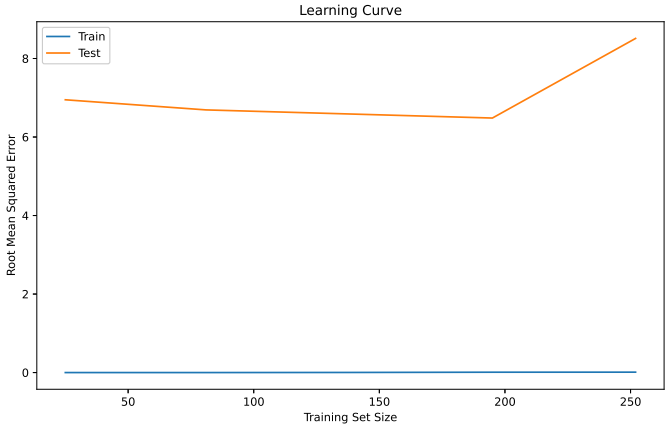

Fig. 6. Decision tree on Autoencoder set.

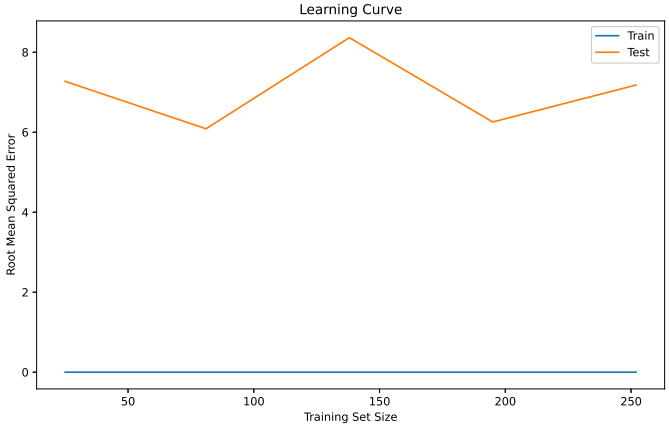

Fig. 7. Decision tree on FSMRMR set.

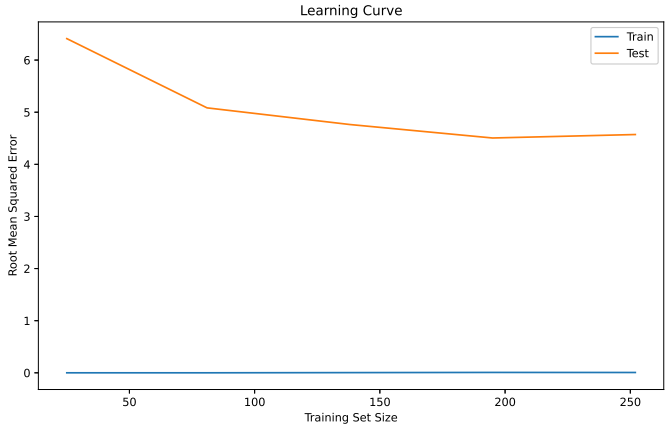

Fig. 8. Decision tree on GAN set.

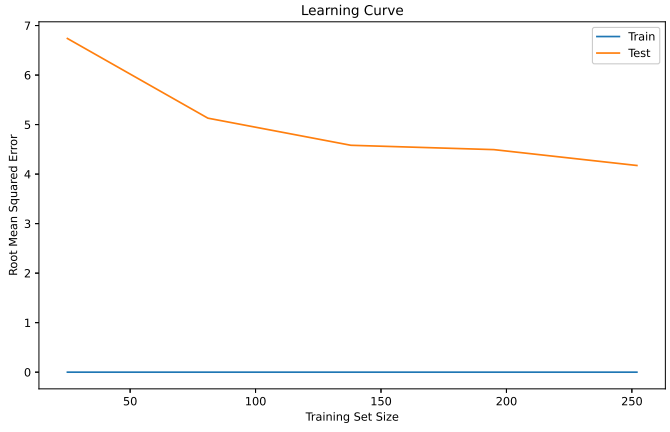

Fig. 9. Decision tree on PCA set.

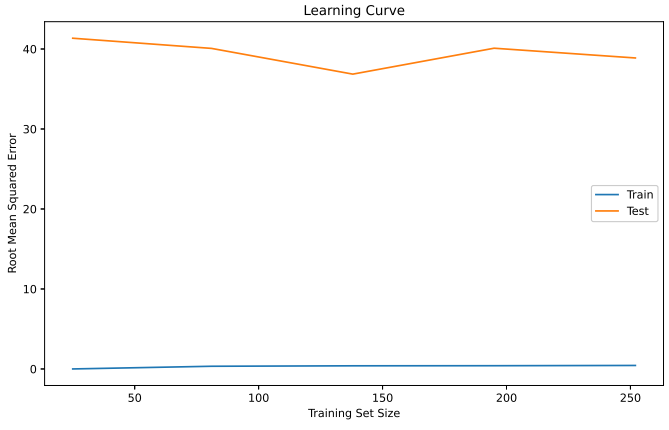

Fig. 10. Decision tree on PCC set.

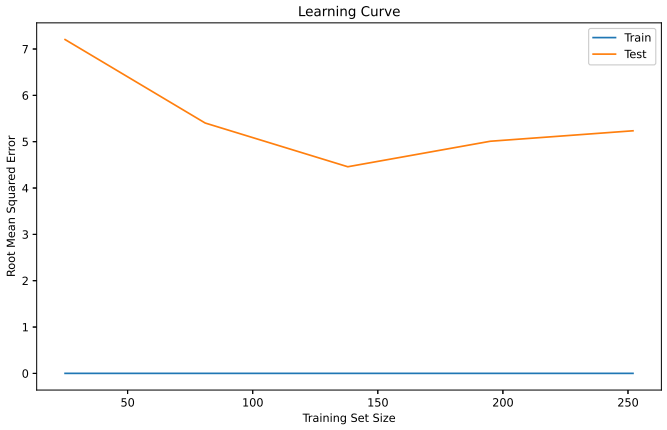

Fig. 11. Decision tree on RFE set.

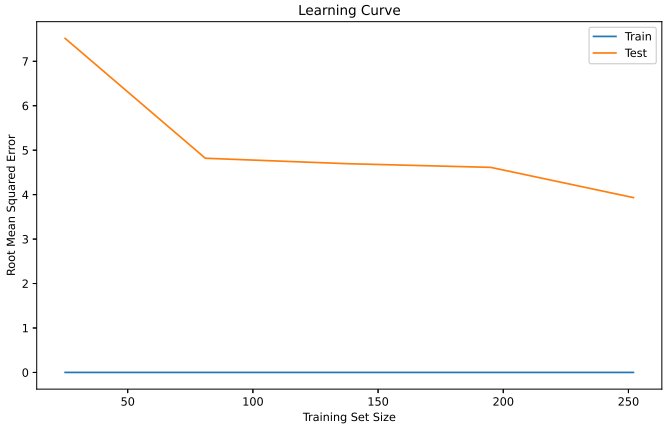

Fig. 12. Decision tree on SF set.

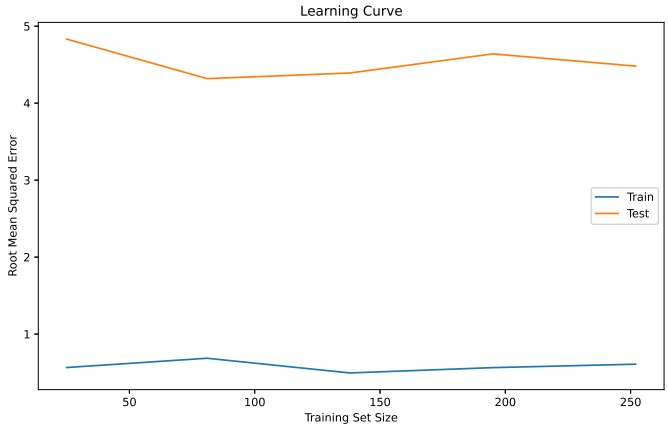

Fig. 13. Random forest on Autoencoder set.

4. Discussion

4.1. The importance of unconventional ML/DL techniques in educational data analysis

In the realm of machine learning (ML) and deep learning (DL), analyzing educational data necessitates the application of diverse techniques to fully understand predictive performance. Accurately forecasting student outcomes and identifying the key influencing factors are crucial. Traditional ML and DL techniques, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision trees and Deep Neural Networks (DNN) are widely adopted due to their proven effectiveness. However, the complexity and uniqueness of educational data warrants the exploration of less common ML/DL techniques. This section examines the reasons for incorporating unconventional approaches alongside popular techniques for prediction performance, feature selection and extraction in the analysis of educational data. Lesser-known techniques in machine learning (ML) and deep learning (DL) often introduce innovative approaches that can offer surprising insights. For example, Generative Adversarial Networks (GAN) and Variational Autoencoder (VA) improve representation learning, crucial for feature extraction and selection, challenging traditional paradigms.

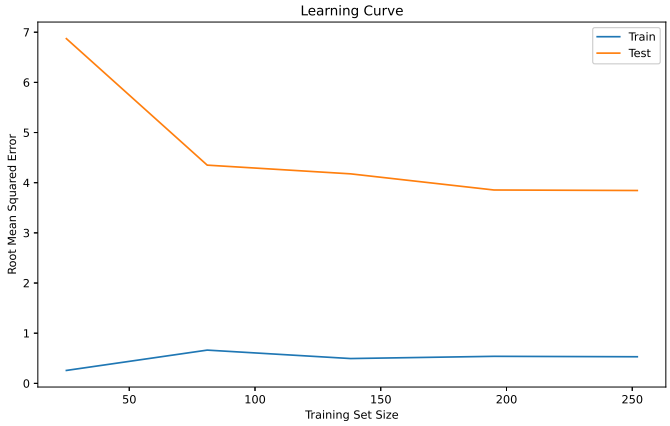

Fig. 14. Random forest on FSMRMR set.

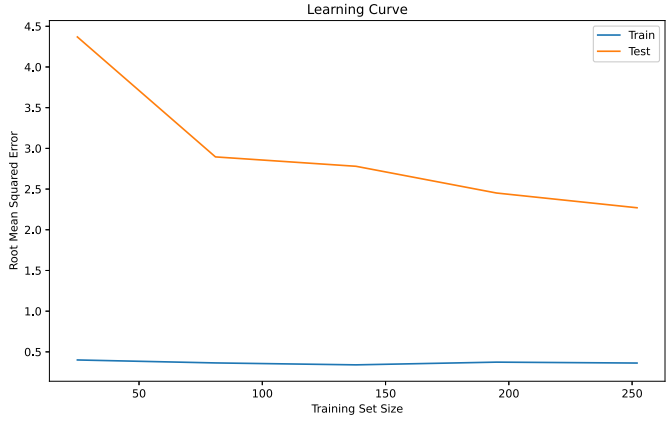

Fig. 15. Random forest on GAN set.

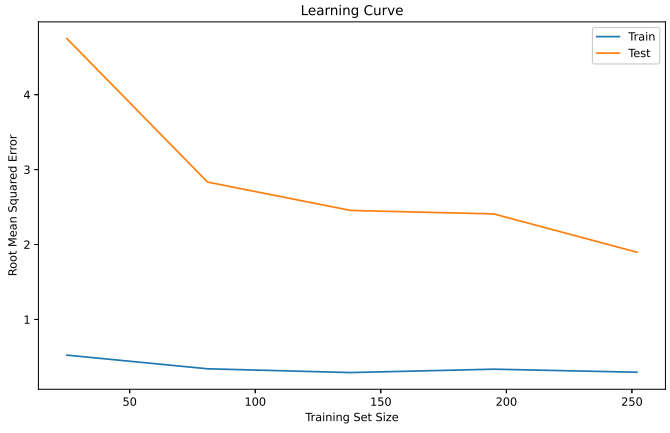

Fig. 16. Random forest on PCA set.

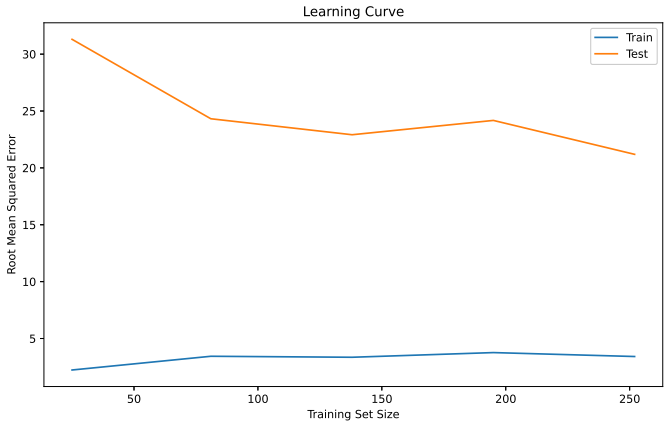

Fig. 17. Random forest on PCC set.

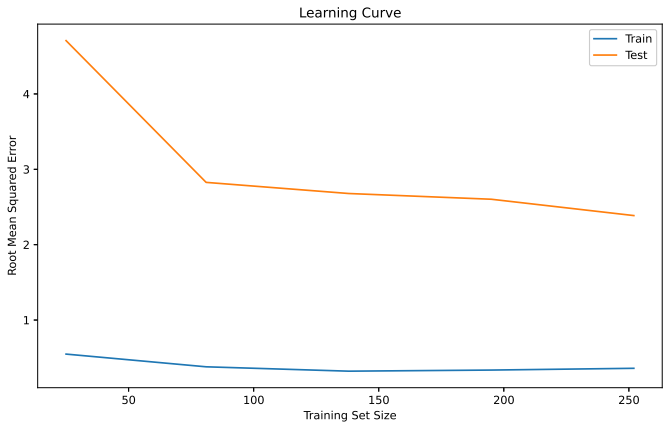

Fig. 18. Random forest on RFE set.

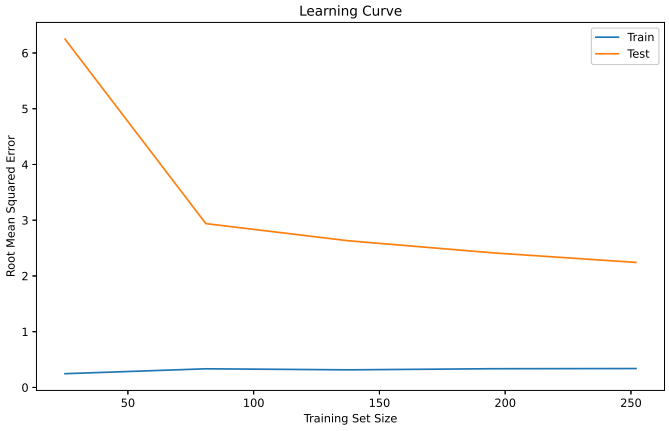

Fig. 19. Random forest on SF set.

These techniques often reveal patterns or relationships in the data that standard methods might ignore, leading to greater understanding and new discoveries. By isolating important factors, they improve the clarity of model results, making it easier for researchers to interpret results and understand model behavior. Integrating a mix of mainstream and less mainstream ML/DL techniques allows for a more comprehensive analysis of educational data. Techniques such as GAN, Feature Selection using Minimum Redundancy Maximum Relevance (FSMRMR), Principal Component Analysis (PCA) and VA offer unique insights and analytical capabilities that complement traditional methods. The complexity of educational data often lies in its high dimensionality and the complex relationships between features. Techniques such as PCA, FSMRMR and VA excel at identifying the most relevant features and reducing dimensionality, improving model performance and interpretability. Recursive feature elimination (RFE) further refines feature selection, ensuring that models focus on the most impactful variables.

Educational data can have non-linear relationships and complex data distributions. Techniques such as GAN and VA are able to capture these complexities, enabling more accurate predictions and insights. Their ability to generate synthetic data can also prove invaluable in solving problems of data sparsity and imbalance.

Going beyond traditional predictive models, RNN-GRU and Transformers offer advanced capabilities for managing sequential and time-

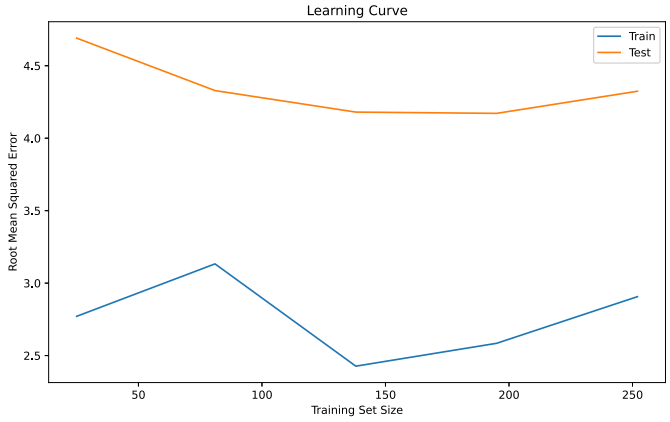

Fig. 20. KNN on Autoencoder set.

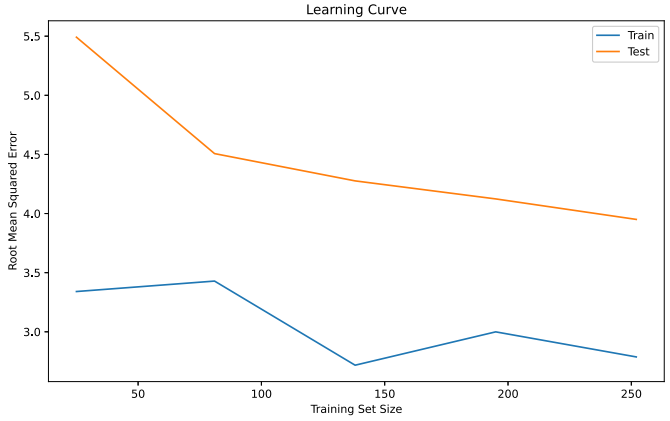

Fig. 21. KNN on FSMRMR set.

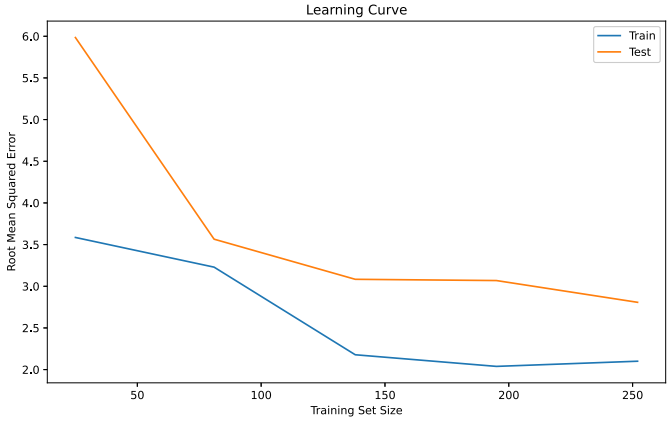

Fig. 22. KNN on GAN set.

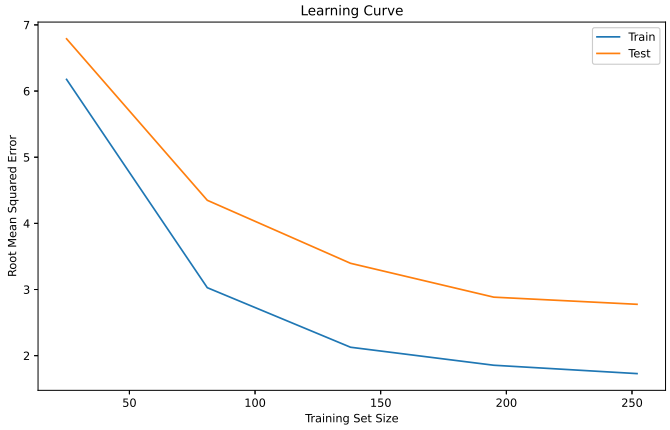

Fig. 23. KNN on PCA set.

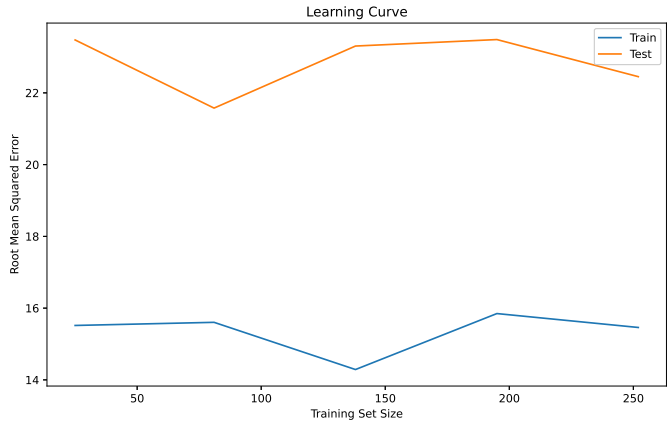

Fig. 24. KNN on PCC set.

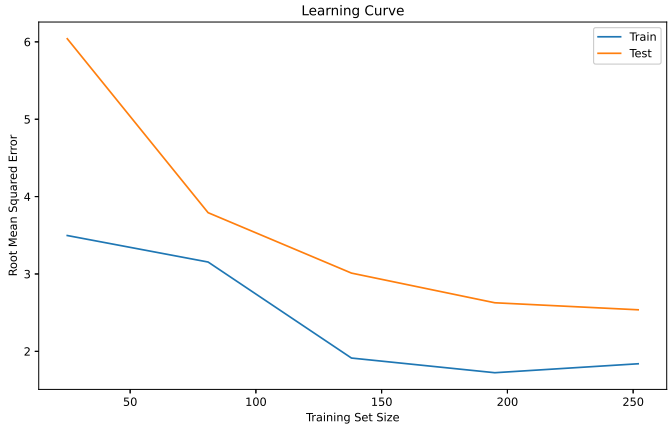

Fig. 25. KNN on RFE set.

series data, common in educational environments. These models can capture dependencies and temporal patterns over time, providing a deeper understanding of student performance trajectories. The use of unconventional techniques encourages exploratory analysis and can lead to new insights into the factors affecting student performance. By taking advantage of the generative capabilities of GAN or the deep learning of VA representations, researchers can uncover hidden patterns and relationships that are not readily apparent using conventional methods. Less common techniques offer complementary perspectives that can enrich the analysis. For example, GAN can be used to generate synthetic data to improve the robustness of models, while VA allow for more flexible and non-linear dimensional reduction. Lesser-known techniques may offer alternative perspectives that could overcome the limitations inherent in the more popular methods. For example, while traditional models can have difficulties with high-dimensional or unbalanced datasets.

In general, the inclusion of less common ML/DL techniques alongside popular ones in the analysis of educational data is not simply a search for novelty but a strategic approach to harnessing the full analytical potential of ML/DL. By adopting a diverse toolkit, researchers can more effectively navigate the complexities of educational data, uncover deeper insights and stimulate innovation in educational research and practice. This holistic approach ensures that analysis remains accurate, versatile and capable of addressing the multifaceted challenges inher-

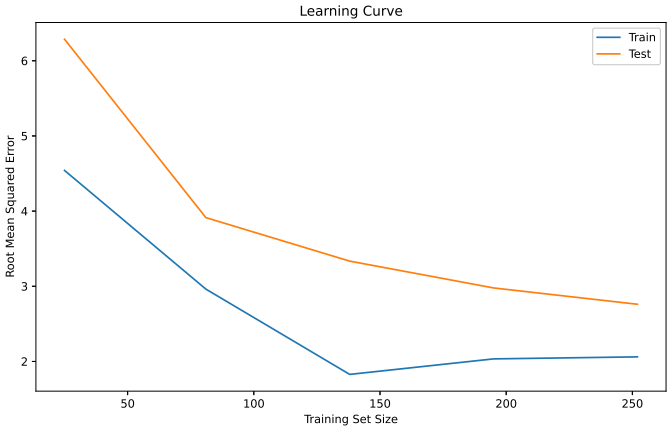

Fig. 26. KNN on SF set.

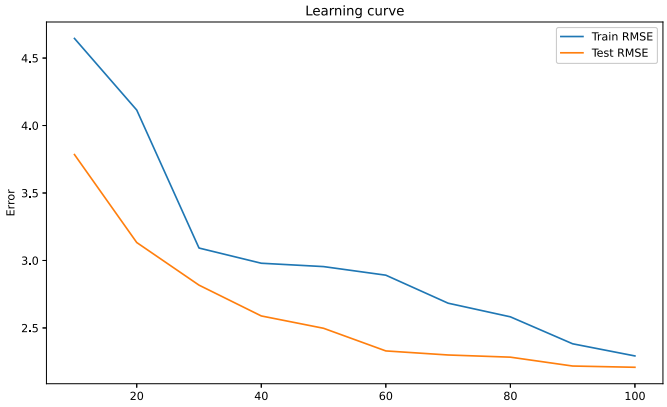

Fig. 27. SVM on Autoencoder set.

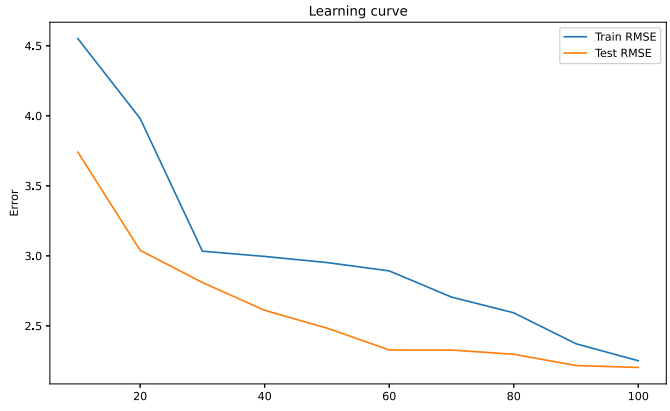

Fig. 28. SVM on FSMRMR set.

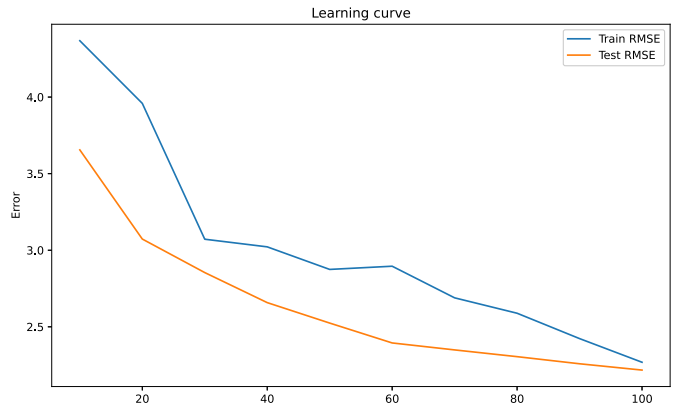

Fig. 29. SVM on GAN set.

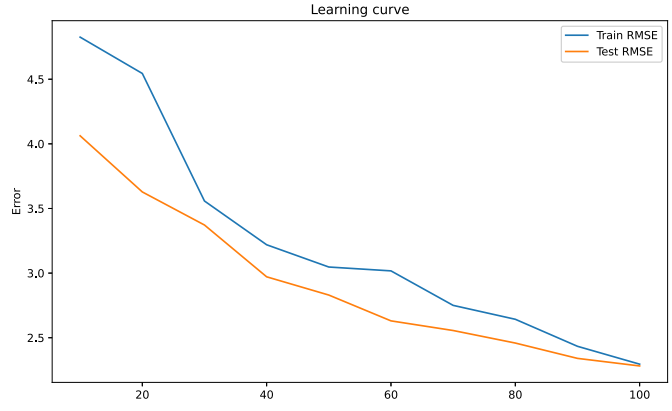

Fig. 30. SVM on PCA set.

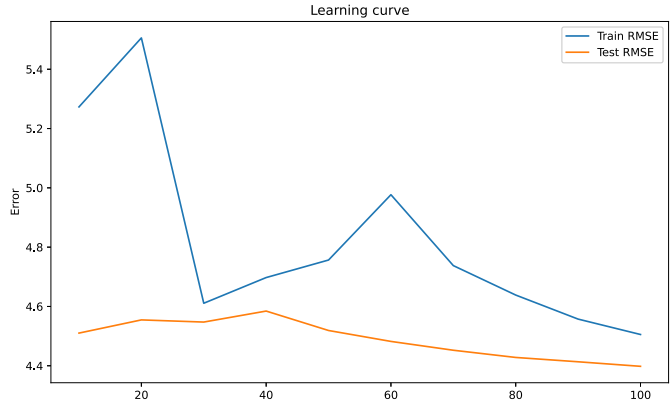

Fig. 31. SVM on PCC set.

ent in predicting student performance and determining the importance of various factors.

4.2. Analysis of correlations between student data features and performance outcomes

In this section we will discuss a study to determine the 10 factors most correlated with the final score using the Pearson, Spearman and Kendall correlation methods (2022), we applied these three statistical techniques exhaustively to the available data. Each method offers a distinct perspective on the relationships between the features, allowing for a comprehensive and accurate assessment of correlations. Our method is based on a combination of these three methods to determine these factors, Figs. 55–57 shows the three heatmaps obtained after applying these three methods to the student data set.

The first correlation used is the Pearson correlation, which is already defined in the Feature selection section, factors such as 'G1', 'G2', 'Studytime', Failures', and 'Medu' showed high correlation coefficients with the final score 'G3'. For example, 'G1' and 'G2' show very strong correlations of 0.79 and 0.86 respectively, indicating that past performance is a significant important factors in predicting the final score.

The second correlation used is Spearman's rank correlation coefficient measures the strength and direction of the monotonic relationship between two ranked variables. It can be used for both continuous and ordinal data. Like Pearson, it ranges from -1 to 1. Its formula is:

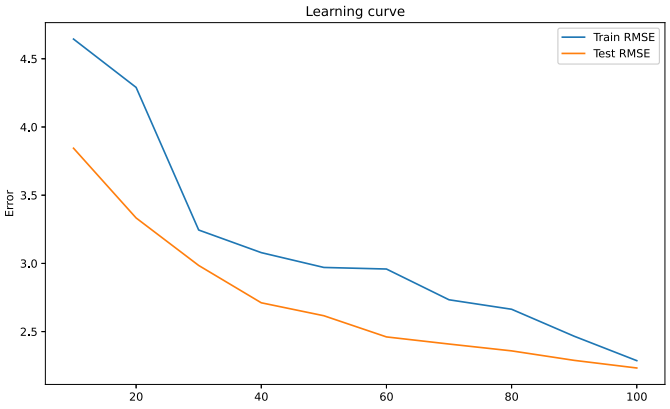

Fig. 32. SVM on RFE set.

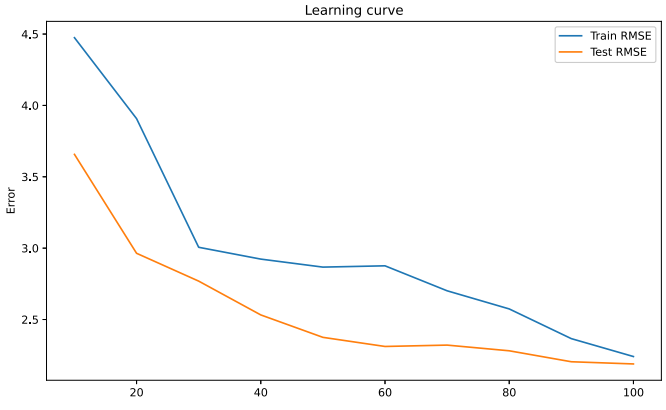

Fig. 33. SVM on SF set.

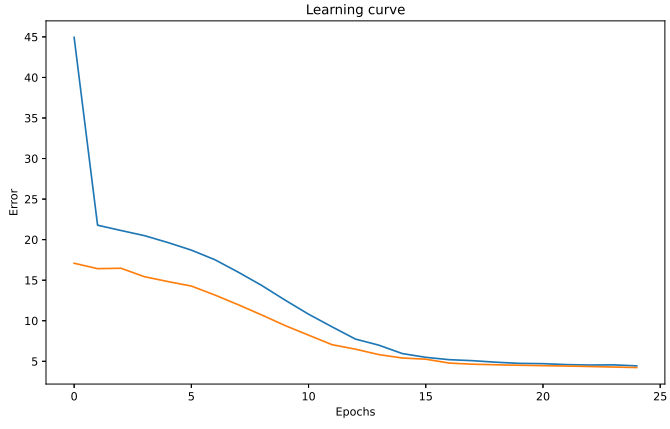

Fig. 34. DNN on Autoencoder set.

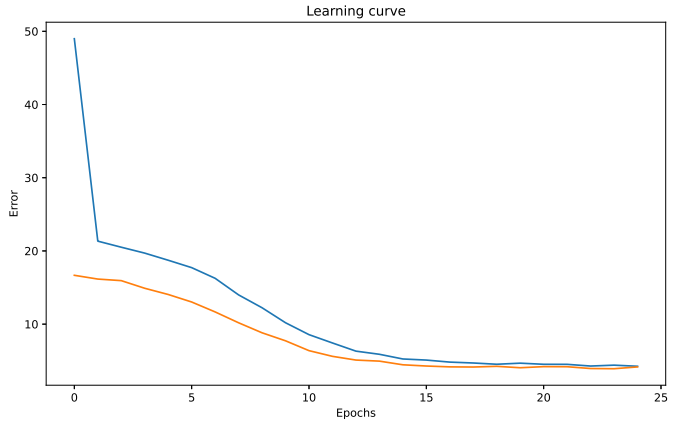

Fig. 35. DNN on FSMRMR set.

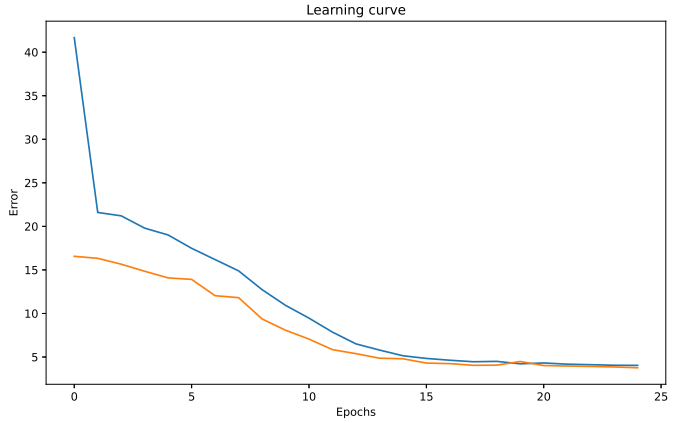

Fig. 36. DNN on GAN set.

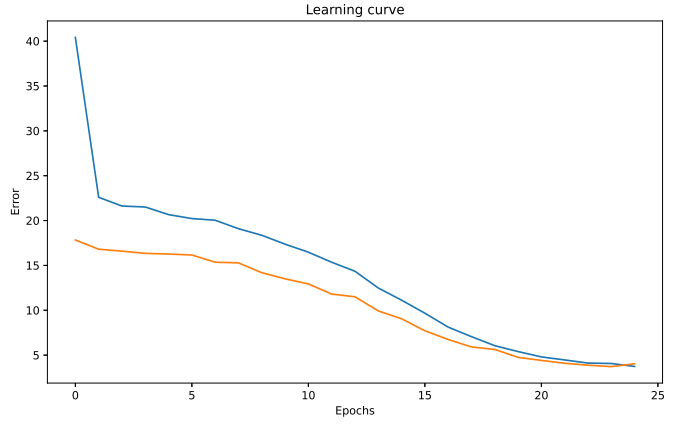

Fig. 37. DNN on PCA set.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{17}$$

Where:

- d_i is the difference between the ranks of corresponding variables.
- n is the number of observations.

Spearman's analyses revealed similar correlations to Pearson's for the features 'G1' and 'G2', with coefficients of 0.78 and 0.85 respectively. However, Spearman also highlighted significant monotonic relationships with features such as 'Higher' (0.33) and 'Freetime' (-0.16), which may not be captured by Pearson.

The third correlation used is Kendall's Tau measures the association between two variables based on the ranks of the data. It is based on the number of concordant and discordant pairs. This technique is particularly accurate for small sample sizes and linked data. Its formula is:

$$\tau = \frac{(C-D)}{\sqrt{(C+D+T_x)(C+D+T_y)}} \tag{18}$$

Where: - C is the number of concordant pairs. - D is the number of discordant pairs. - T_x is the number of ties only in the x variable. - T_y is the number of ties only in the y variable. The Kendall analyses confirmed the results obtained with Pearson and Spearman for the 'G1' and 'G2' variables with coefficients of 0.70 and 0.80 respectively, while

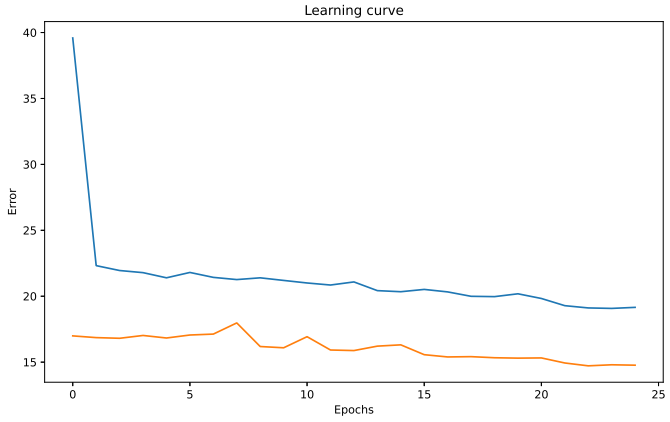

Fig. 38. DNN on PCC set.

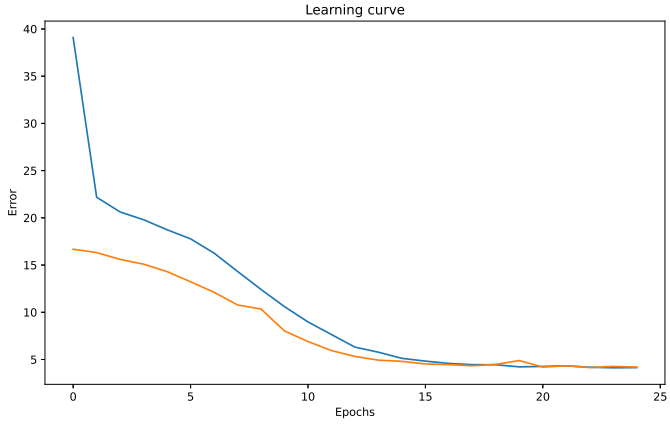

Fig. 39. DNN on RFE set.

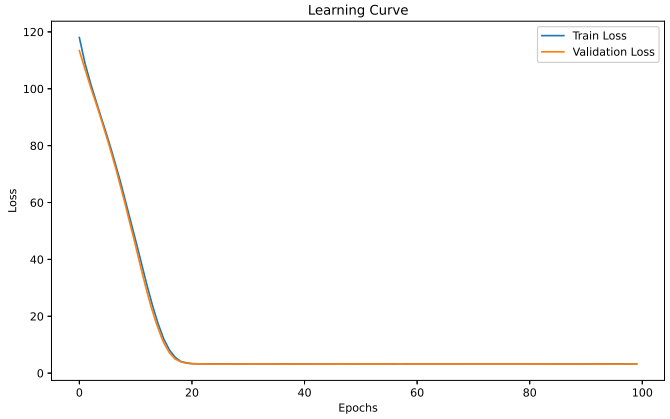

Fig. 40. DNN on SF set.

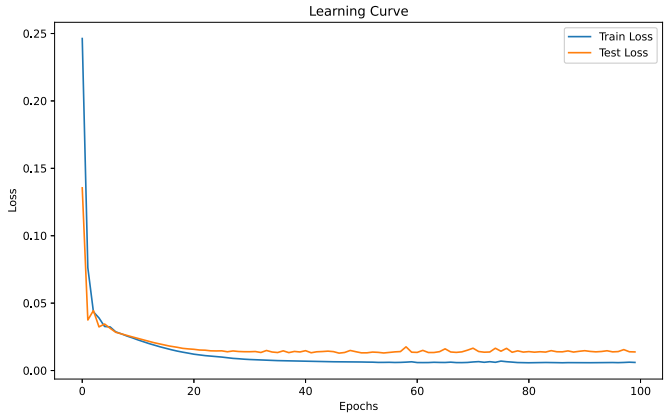

Fig. 41. GRU on Autoencoder set.

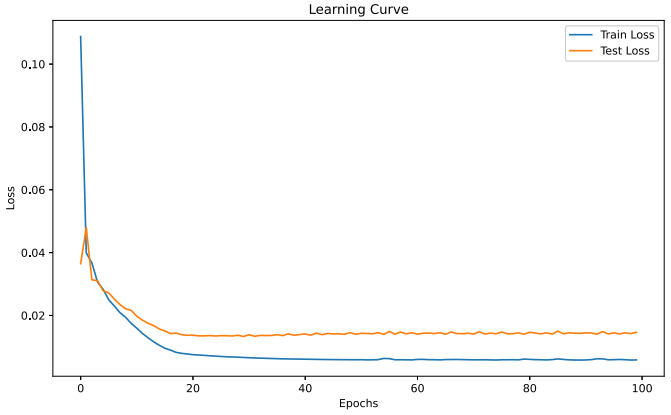

Fig. 42. GRU on FSMRMR set.

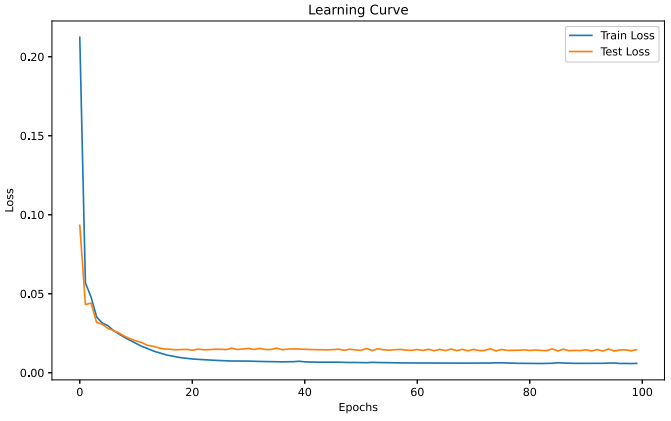

Fig. 43. GRU on GAN set.

identifying significant relationships with variables such as 'studytime' and 'failures', thus providing accurate cross-validation. By combining the results of the three methods, we have identified the 10 factors most correlated with the final 'G3' score:

- 'G2' (Pearson: 0.86, Spearman: 0.85, Kendall: 0.80)
- 'G1' (Pearson: 0.79, Spearman: 0.78, Kendall: 0.70)
- Studytime (Pearson: 0.15, Spearman: 0.16, Kendall: 0.14)
- Failures (Pearson: -0.36, Spearman: -0.36, Kendall: -0.30)
- Higher (Pearson: 0.33, Spearman: 0.33, Kendall: 0.24)
- Medu (Pearson: 0.22, Spearman: 0.20, Kendall: 0.16)
- Fedu (Pearson: 0.18, Spearman: 0.17, Kendall: 0.14)
- Goout (Pearson: -0.18, Spearman: -0.19, Kendall: -0.15)
- 'Schoolsup' (Pearson: -0.081, Spearman: -0.082, Kendall: -0.078)
- 'Freetime' (Pearson: -0.16, Spearman: -0.16, Kendall: -0.13)

These results show strong consistency between the methods for the main correlated features, while revealing important nuances in the relationships captured by each approach. The combined use of Pearson, Spearman and Kendall thus provides an in-depth understanding of the factors influencing academic performance, offering valuable insights for targeted educational interventions.

More in-depth feature selection, extraction and correlation methods offer a more nuanced and detailed view of the factors influenc-

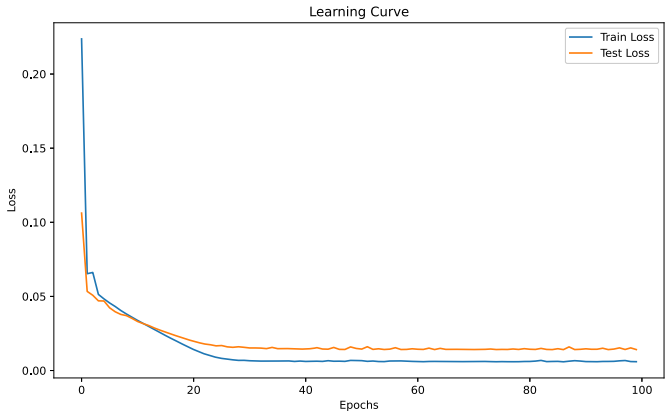

Fig. 44. GRU on PCA set.

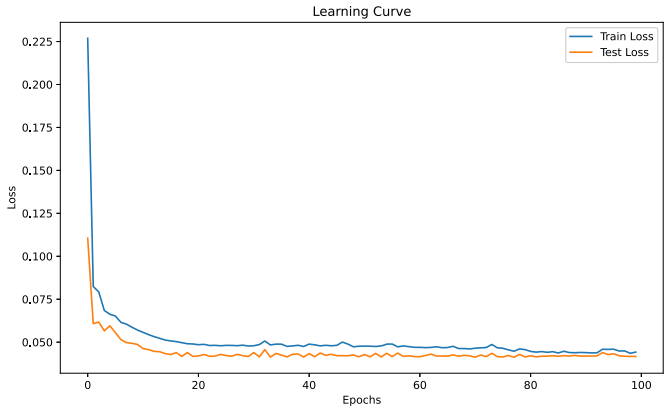

Fig. 45. GRU on PCC set.

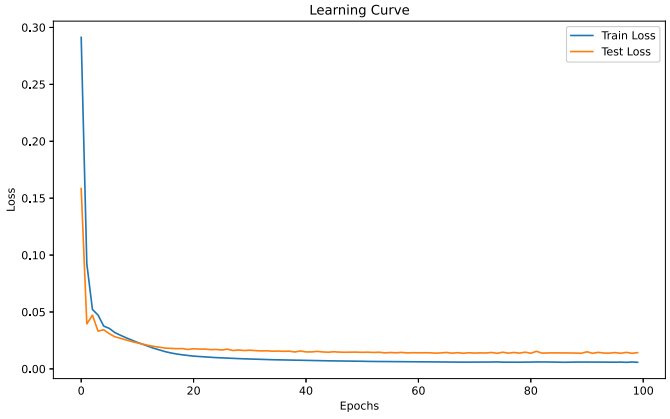

Fig. 46. GRU on RFE set.

ing students' academic performance (Begum & Padmannavar, 2022). By combining traditional approaches such as PCC with advanced techniques such as GANs and PCA (Kumar & Ahuja, 2022), educators and researchers can gain valuable insights for developing targeted and effective interventions. Case studies demonstrate how these methods can be applied in practical ways to improve educational outcomes and support at-risk students.

4.3. Comparison of feature importance across different studies

To analyze and discuss the first objective of our research, we have chosen two studies on the same dataset. The first study is by Sean M. Shiverick (2019). We need to examine the features deemed important

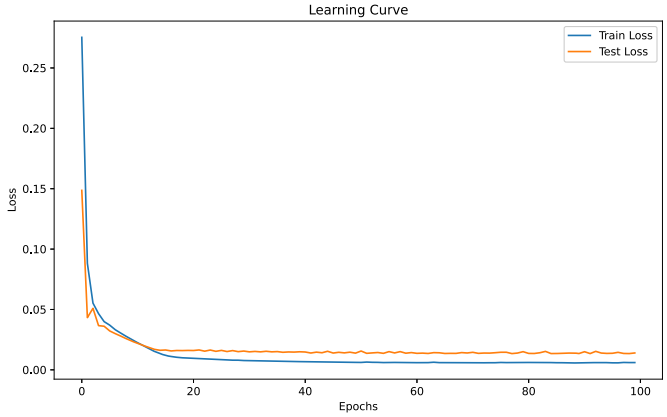

Fig. 47. GRU on SF set.

Fig. 48. Transformer on Autoencoder set.

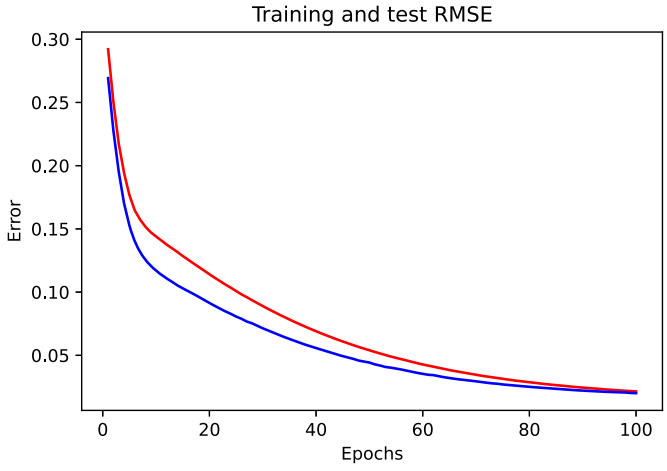

Fig. 49. Transformer on FSMRMR set.

by the random forest regression and gradient boosting models used in this study, then compare them to the features selected by our methods (PCC, RFE, GAN, FSMRMR, PCA, VA).

According to Table 6 from Sean M. Shiverick's article (Shiverick, 2019), which represents the importance of variables, there are two methods. The first method is Random Forests, where features are evaluated according to the increase in MSE (Mean Squared Error) when they are omitted. The most important characteristics include mother's education, absences, environment (urban/rural), and higher education. The

Table 5
Importance of various characteristics.

The study was conducted by Sean M. Shiverick			
Random Forests		Gradient Boosting	
Characteristics	Increase MSE	Characteristics	Relative Importance
Mother's Education	1.368	Absences	16.754
Absences	0.884	Course	11.151
Area (Urban/Rural)	0.592	Mother's Education	7.926
Higher Education	0.575	Age	7.460
Course	0.481	Go out w/ Friends	5.566
Weekly Alcohol	0.473	Study Time	4.618
Mother's Job	0.463	Health	3.942
Father's Education	0.399	Free Time	3.284
Go Out w/ Friends	0.389	Family Relations	3.213
Daily Alcohol	0.373	Weekly Alcohol	3.150
Our Study		Study conducted by Chuang Liu et al.	
Characteristics	Importance in %	Characteristics	Importance
Age-Failures-G1-G2	**8.33%**	G2	403.0
Go out	6.66%	G1	252.0
Absences-Health-Romantic-Fedu-Freetime-Higher	5%	Health	141.0
Dalc-Medu-Studytime-Activities-Reason	3.33%	Absences	133.0
Walc-Schoolsup-Paid internet-Mjob-Famrel-Nursery	1.66%	Famrel	103.0
–	–	Mjob	103.0
–	–	Walc	102.0
–	–	Medu	95.0
–	–	Age	80.0
–	–	Freetime	78.0

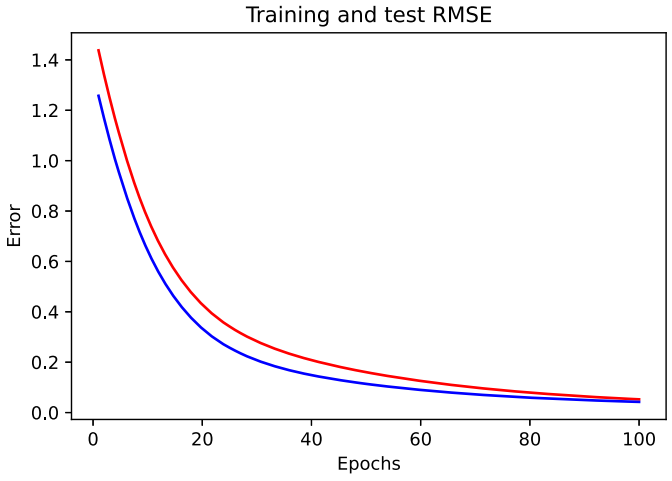

Fig. 50. Transformer on GAN set.

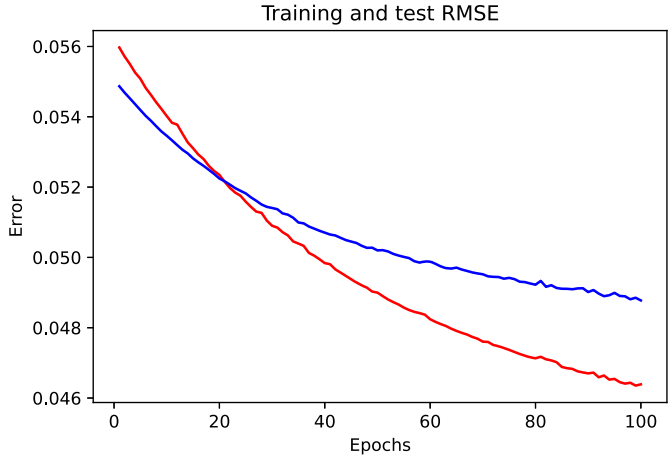

Fig. 52. Transformer on PCC set.

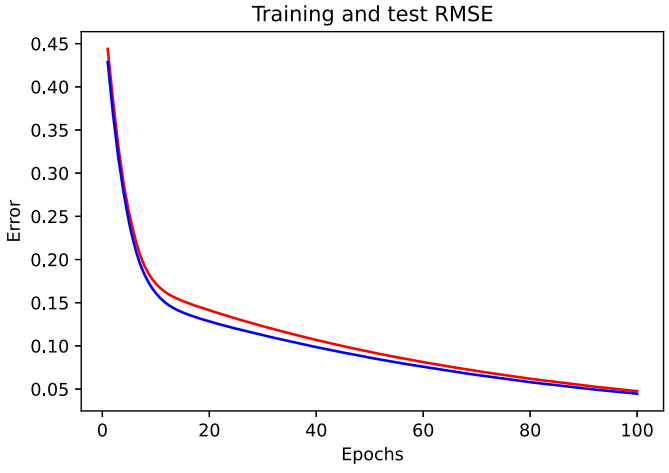

Fig. 51. Transformer on PCA set.

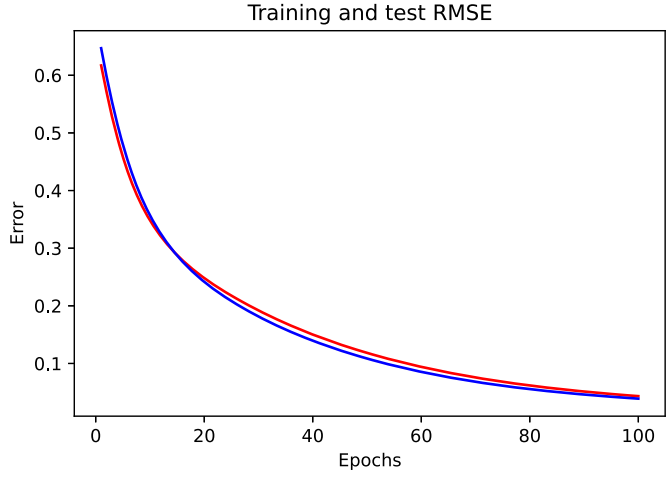

Fig. 53. Transformer on RFE set.

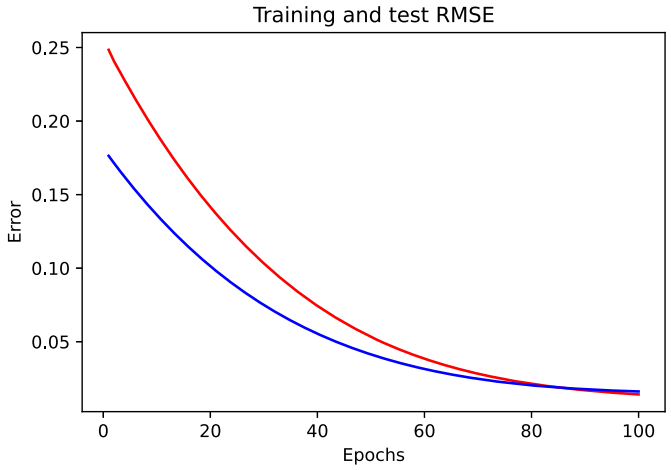

Fig. 54. Transformer on SF set.

second method is Gradient Boosting, where characteristics are evaluated according to their relative importance. The most important characteristics are absences, course taken, mother's education and age.

Comparing these two methods with those used in our research, we can deduce the following:

- **Mother's Education (Medu)**: This feature is considered important in all models and methods, indicating a significant influence on student performance.
- **Absences**: This is very important for the gradient boosting model and is present in our work via RFE and VA, suggesting a notable impact on performance prediction.
- **Age**: This feature appears in several of our selection methods (PCC, RFE, FSMRMR, VA) and is also considered important by the gradient boosting model.
- **Higher Education**: This feature is present in our feature selection via RFE and FSMRMR, and is considered quite important by both random forests and gradient boosting.
- **Weekly/Daily Alcohol Consumption**: This is mentioned in Table 6 for both models but is less present in our research, which may indicate differences in how the impact of alcohol consumption is assessed.
- **Going Out with Friends**: This is important for the gradient boosting model and is present in our feature selection via RFE and VA.

In general, there are similarities between the characteristics selected by the models in Table 5 and our methods, such as mother's education and absences, which are consistent across the different approaches. This reinforces the idea that these factors are important predictors of student performance. However, there are also differences, such as the importance attached to alcohol consumption and extracurricular activities, which may vary depending on the feature selection method and model used. These differences may be due to the nature of the data, the assumptions of the models, or the way in which each method assesses the importance of characteristics.

For the second work we compared, which is by Chuang Liu et al. (2022), they used a heatmap to determine the importance of each feature. Comparing the results of this research with the results of our research, we can deduce the following:

- **G2 and G1**: These two features have the highest importance scores and also appear in our methods. This confirms their importance in predicting student performance.
- **Failures**: This feature is frequently recognized in our research and also has a high importance score in the heatmap method, which underlines its significance.
- **Mother's Education (Medu)**: This feature is common to both selections, with high importance in the heatmap selection and a recurring presence in our work.
- **Age**: This feature is valued in both selections, suggesting that it has an important impact on student performance.
- **Go Out with Friends and Free Time**: These characteristics are present in both feature selections but appear to have a more moderate importance in the heatmap compared to their frequency of appearance in our results.
- **Health and Absences**: These features have very high importance scores in the heatmap, indicating their potential impact on student performance. They are also present in our work.

This table represents the most important results of our work and other research work.

In summary, there is a correlation between the characteristics deemed important in our research and those with high importance scores in the heatmap. This validates the significance of these characteristics in predicting student performance. However, some characteristics may be valued differently depending on the feature selection method or model used, indicating that the context and specific criteria of each method may influence the perceived importance of different characteristics (Begum & Padmannavar, 2022). The characteristics most frequently selected and deemed important across different methods and models are: Mother's Education (Medu), Absences, Age, Higher Education, G2 and G1, Failures, and Health. These characteristics show a strong correlation across different feature selection approaches and predictive models, confirming their significant influence on student performance.

4.4. Comparison of prediction models for student performance

For the second research objective, which consists in predicting the performance of students with the most important predictors, we compared our research with two studies. The first study is by Mr. Riki Apriyadi et al. (2023), and the second study is by Herliyani Hasanah et al. (2022).

Concerning the comparison with the research of Mr. Riki Apriyadi et al. (2023), the initial results show the RMSEs for different machine learning algorithms without specifying the feature selection or extraction techniques used:

- SVR (Support Vector Regression): 2.09
- Naïve Bayes: 2.01
- Neural Networks: 2.05
- Decision Tree: 1.94
- Random Forest: 1.75

Comparing the two sets of results, we can observe that Random Forest seems to be the better performing algorithm in both cases, with an RMSE of 1.75 in the initial results and a range of 1.95-1.15 in our work, indicating that feature selection and extraction can further reduce prediction error. Decision Tree also shows significant improvement with feature selection and extraction techniques, moving from an RMSE of 1.94 to a range of 1.20-2.70. Feature selection and extraction techniques have a variable impact on the different algorithms. The Transformer shows exceptional improvement with extremely low RMSE values, suggesting that this combination of algorithm and techniques is particularly effective for this dataset. It's important to note that the RMSE ranges in our work indicate the sensitivity of the algorithms to the different techniques used (Zhao, 2024). A narrower range suggests that the algorithm is less sensitive to the feature selection or extraction technique, while a wider range indicates greater sensitivity.

In general, feature selection and extraction play a crucial role in improving the accuracy of prediction models. The results show that the use of appropriate techniques can significantly reduce prediction error, and that the choice of prediction algorithm needs to be tailored

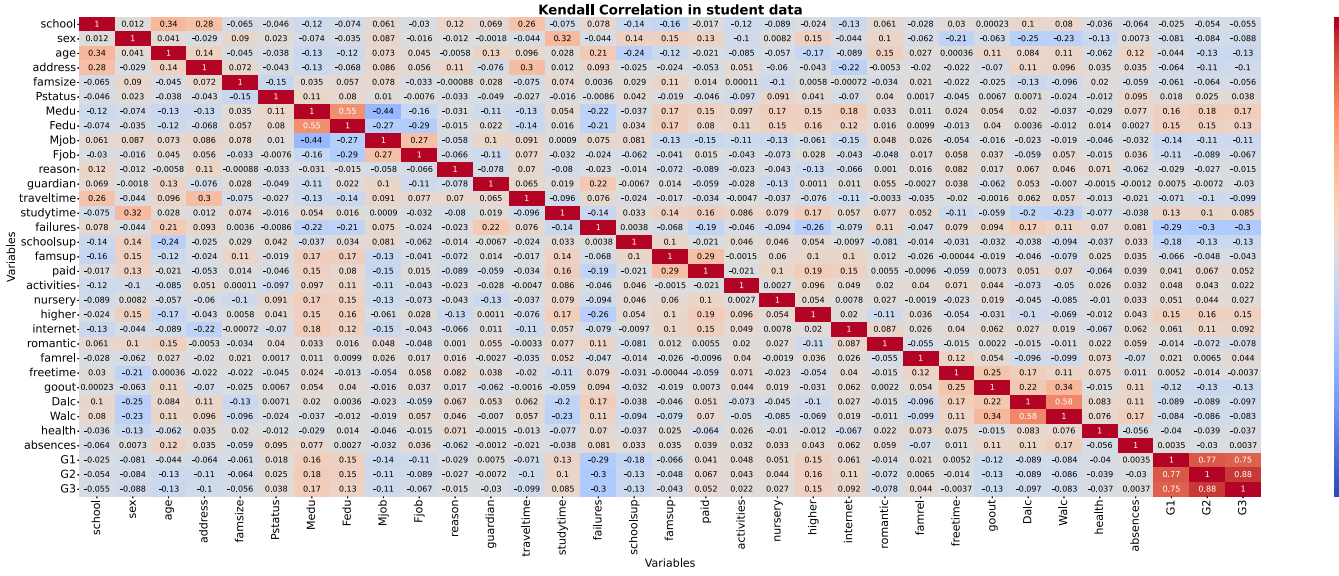

Fig. 55. Heatmap of the Kendall correlation matrix.

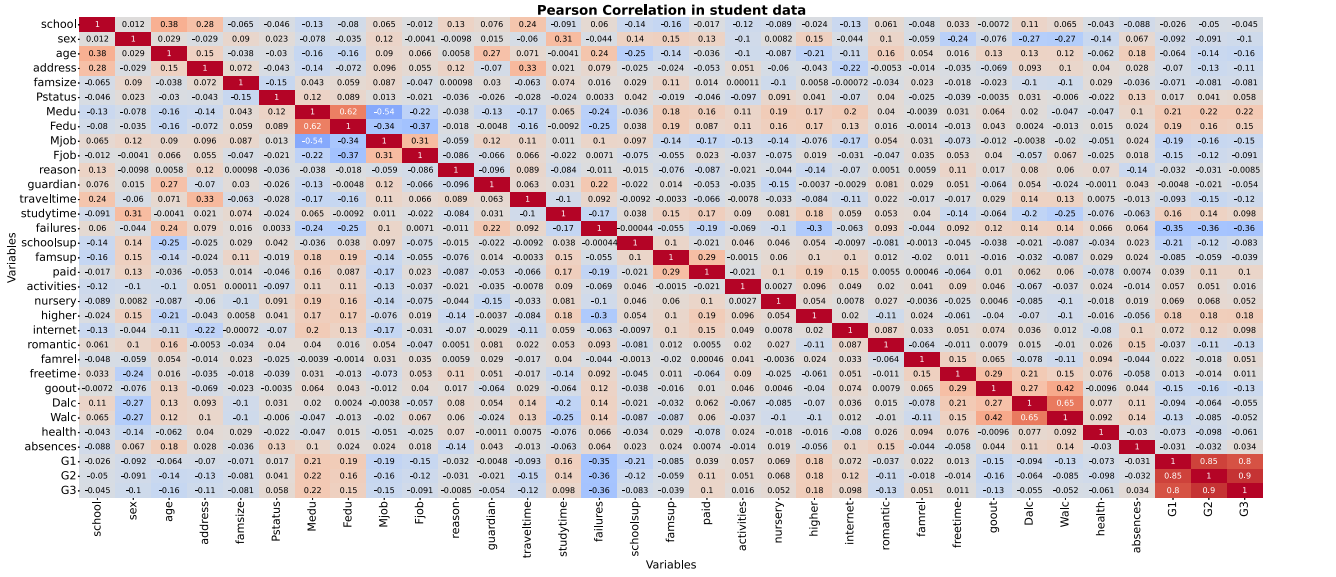

Fig. 56. Heatmap of the Pearson correlation matrix.

to the specific characteristics of the dataset to achieve the best results. The algorithm that shows the best performance improvement with feature selection and extraction techniques appears to be the Transformer. RMSE values for the Transformer are exceptionally low (ranging from 0.23 to 0.01), indicating a significant improvement over the other algorithms listed.

However, it should be noted that the extremely low RMSE values for the Transformer are validated through cross-validation. Apart from the Transformer, the Random Forest and Decision Tree algorithms also show notable improvement in terms of RMSE reduction when combined with feature selection and extraction techniques, with RMSE ranges reaching lower values than those obtained without these techniques. This indicates that these algorithms also benefit from the application of these methods to improve prediction performance.

The second study by Herliyani Hasanah et al. (2022) used four linear regression models following the correlation with G3. This study remains very limited because the sample used is only 10. The calculations determine the influence of each attribute on the final value of the period. The results obtained are as follows:

- Model 2 (failures): The magnitude of the influence of the independent variable (failures) on the dependent variable is 63.8
- Model 1 (study time): The influence is 37.1
- Model 3 (free time): The influence is 32.0
- Model 4 (absences): The influence is 0.01

Scientifically, the performance prediction results for the research students of Herliyani Hasanah et al. (2022) are as follows:

- Model 1: RMSE of 1.5133 and MAE of 0.11
- Model 2: RMSE of 1.148 and MAE of 0.98
- Model 3: RMSE of 1.573 and MAE of 0.11
- Model 4: RMSE of 1.908 and MAE of 0.147

Comparing the RMSE and MAE of these models with those of our work, we can see that the Transformer has extremely low RMSE and MAE values, suggesting superior performance compared with the other models Begum and Padmannavar (2022). In terms of consistency of results, the Transformer results are remarkably consistent and well below those

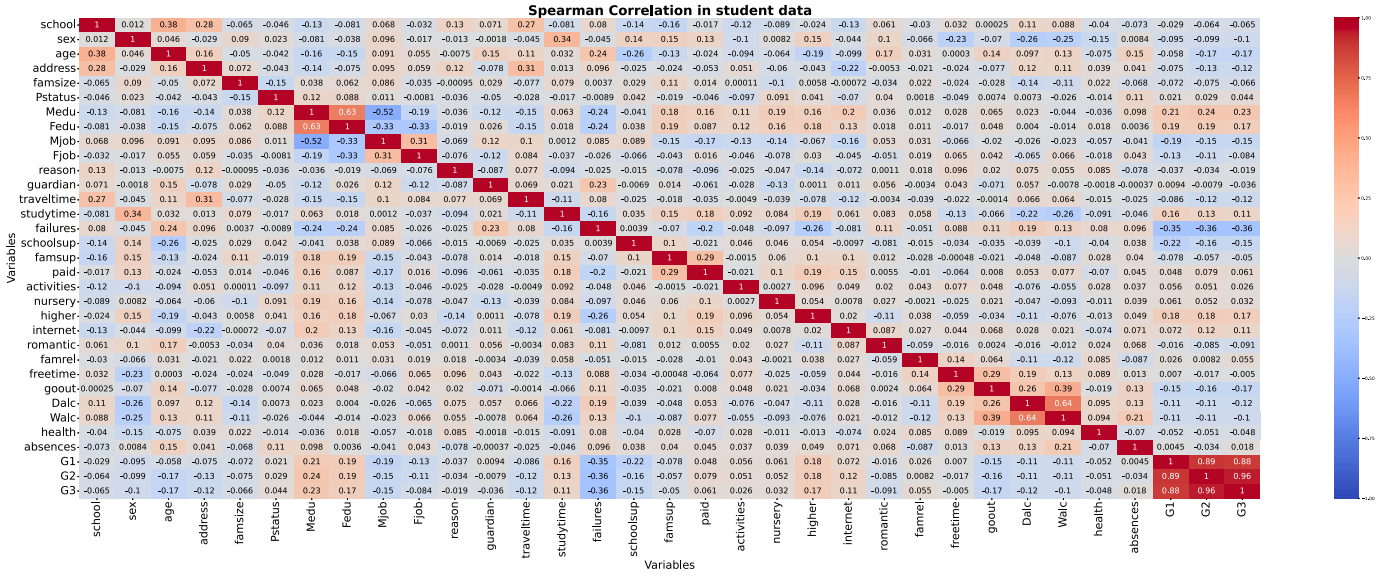

Fig. 57. Heatmap of the Spearman correlation matrix.

Table 6
Comparison of optimal values across different studies.

Our Study with Optimal Values		Study by Herliyani Hasanah et al.		Study by Mr. Riki Apriyadi et al.	
Decision Tree	2.25-1.20	Model 1	1.5133-11.1%	SVR	2.09
Random Forest	1.83-1.04	Model 2	1.148-9.82%	Naïve Bayes	2.01
KNN	1.88-1.24	Model 3	1.573-11.1%	Neural Networks	2.05
SVM	2.01-1.21	Model 4	1.908-14.7%	Decision Tree	1.94
DNN	2.20-1.40	-	-	Random Forest	1.75
GRU	2.54-1.35	-	-	-	-
Transformer	**0.09-0.01**	-	-	-	-

of the other models, which could indicate a better generalization capability or a better fit of the model to the data.

Regarding the variability between models, the models in Table 5 of the article by Herliyani Hasanah et al. show less variability in the RMSE and MAE results than those in our work, where a wider range of values is observed. This could be due to the use of different feature selection and extraction techniques that influence model performance. Following the impact of pre-processing techniques, the feature selection and extraction techniques used in our work, such as PCC, RFE, GAN, FSMRMR, PCA, VA, and SF, appear to have an important impact on model performance, particularly for the Transformer.

Table 6 represents the most important results of our work and other research on predictive models.

In conclusion, the results suggest that the Transformer, combined with the specified feature selection and extraction techniques, offers a significant performance improvement over other models. Since the Transformer is the best-performing model, and as part of the evaluation and validation of its performance, it is crucial to distinguish between random and systematic errors. The method used for this distinction consists of examining the distribution of residuals generated by the model. Analysis of the distribution of residuals can reveal whether the errors introduced during prediction are random in nature or follow a systematic pattern.

To assess the distribution of residuals, we use graphical representations like Q-Q (quantile-quantile) plots. A Q-Q graph compares the quantiles of the residuals with those of a theoretical normal distribution. We have noticed that the points on the Q-Q graph line up approximately on a straight line, indicating that the residuals follow a normal distribution. This normal distribution of residuals is often interpreted as an indication that the errors introduced by the model are mainly random, meaning that there is no significant systematic bias affecting predictions (Figs. 58–60).

To accomplish our research objectives and reach a general conclusion, we employed a variety of methods and techniques for feature selection and extraction. For each class of these methods, we selected the most effective ones from the literature and attempted to demonstrate that these factors actually have an impact on student performance by predicting student performance as a function of these features using various models, including Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Deep Neural Network (DNN), Gated Recurrent Unit (GRU), and Transformer models. We tested the most recent models for comparison and argumentation, which were well chosen for the task of regressing student scores because they are generally the most widely used in prior work.

For decision tree regression, random forest regression, KNeighbors regression, and support vector machine regression, we used standard implementations. For deep neural networks, we adapted the architecture for regression, creating a three-layer DNN. The first layer is a dense layer (fully connected) with 128 units and 4224 trained parameters, using the “relu” (Rectified Linear Unit) activation function defined as $ReLU(x) = \max(0, x)$. The second layer is another dense layer with 64 units and trained parameters, also using the “relu” activation function. The third layer is a dense layer with a single unit, serving as the output layer for regression. We used the Adam optimizer for training. We also modified the RNN-GRU architecture for this task, creating a two-layer RNN-GRU. The first layer is a GRU layer with 64 units and 12864 trained parameters. The second layer is a dense layer with only one unit, serving as the output layer for regression. We compiled the model using the Adam optimizer.

Finally, we adapted the Transformer architecture with an encoder-decoder mechanism. The model starts with an encoder, which is a dense layer with 64 units and the “relu” activation function. Dropout layers are used for regularization to prevent overfitting. The decoder consists of a dense layer with 64 units and the “relu” activation function. The

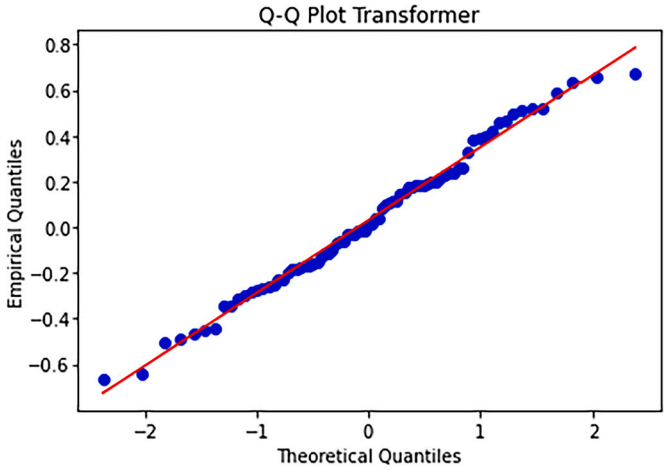

Fig. 58. Q-Q plot Transformer on PCA set.

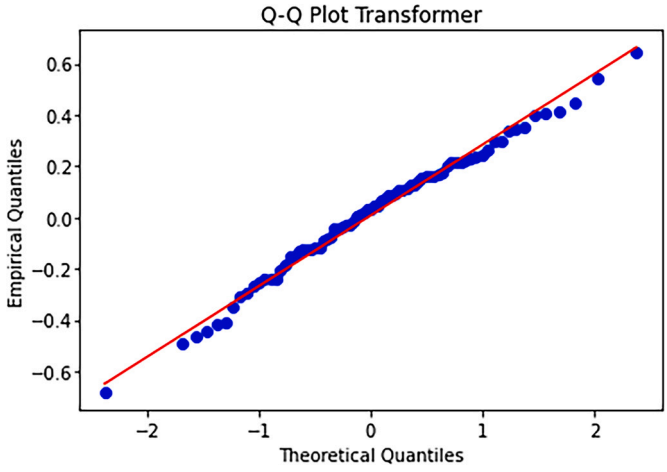

Fig. 59. Q-Q plot Transformer on FSMRMR set.

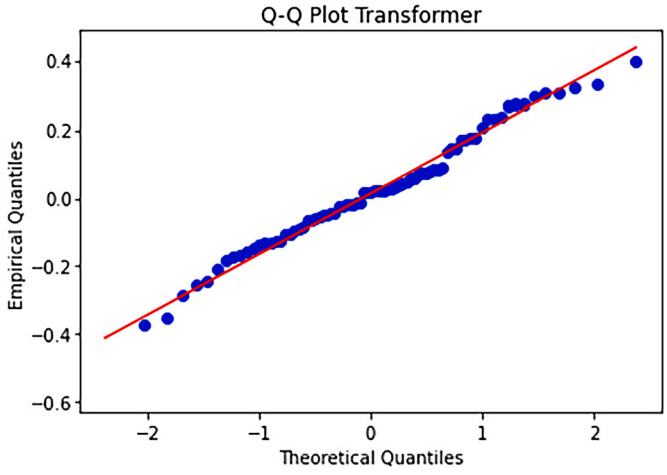

Fig. 60. Q-Q plot Transformer on SF set.

output layer is a dense layer with a single unit and a "linear" activation function, generating the final prediction.

4.5. Research findings

Our results showed that the Random Forest model exhibited consistently strong performance, likely due to its resilience to overfitting and its adeptness at managing numerous input features. In contrast, the Transformer model's performance, particularly when paired with the FSMRMR feature selection method, was distinct from the rest, suggesting a possible incongruity between the model's capacity and the chosen features, or an opportunity for further model refinement. The study's limitations, such as dataset size and diversity, could influence the generalizability of our findings. Future research directions include testing additional models, incorporating a broader array of features, and applying the models to varied educational settings to corroborate and expand upon our results. Moreover, a more in-depth exploration of the Transformer model's performance, with potential fine-tuning or architectural modifications, is warranted.

Through this comprehensive analysis, we deepen our understanding of the determinants of student performance and elucidate how various machine learning models can be effectively harnessed to forecast educational outcomes. The ML models used in this study, such as decision trees, random forests, KNNs, and SVMs, are simple and easy to interpret. They are crucial to process these data thanks to reduced dimensionality, characterized by ease of optimization. However, their simplicity can also be a drawback, as they may not capture complex, non-linear relationships without sophisticated educational data pre-processing or feature engineering, which could limit their application in scenarios where the data is inherently complex or highly dimensional.

In contrast, DL models, including DNNs, GRUs, and Transformers, are capable of handling large datasets and extracting complex data representations autonomously, thanks to their deep layered structures. Their strength lies in their ability to model non-linearity and high-level interactions between features without the need for extensive manual feature engineering. Nevertheless, the deployment of these models is not without its problems. The substantial computational resources required for training, such as GPUs or TPUs, may not be readily available to all researchers, especially those with limited access to state-of-the-art computing infrastructure. In addition, the propensity of DL models to overfit, especially when trained on small datasets, has forced us to apply regularization, dropout, and data augmentation techniques to ensure generalizability.

Interpretability remains a major concern, as ML models are more transparent and easier to explain. DL models, due to their "black box" nature, pose considerable problems in terms of explicability, which may hinder their acceptance in the education field where understanding the model's reasoning is as important as its predictive performance. Training time is another critical factor, as DL models require much longer training periods, which can hamper rapid prototyping and iterative development. Furthermore, the technical expertise required to implement DL models is not trivial, as it demands a thorough understanding of neural network architectures, activation functions, optimization techniques, and regularization strategies.

Consequently, although DL models offer advanced capabilities for modeling complex data, their implementation is associated with significant challenges in terms of computing resources, the risk of overfitting, interpretability, and technical skills required. Conversely, traditional ML models, despite their potential limitations in handling complex data, offer advantages in terms of simplicity, speed of learning, and explainability. This provides another avenue for research to compare ML and DL models according to data complexity and size, and choose the most optimal.

4.6. The significance of our research to the field of education

The significance of this research in education is multifaceted. It not only contributes to the theoretical understanding of machine learning (ML) and deep learning (DL) applications in educational contexts (2022; Rong et al., 2023), but also has practical implications for educators, policymakers, and educational institutions. The research demonstrates how various ML and DL models can enhance predictive analytics to forecast educational outcomes, which is valuable for identifying at-risk students, personalizing learning experiences, and improving educational plan-

ning and efficient resource allocation. In practical terms, the results of this research can be applied to identify at-risk students and personalize learning experiences. Early identification of at-risk students using ML and DL models enables rapid interventions, such as additional tutoring, counseling, or personalized support programs. These models can help create personalized learning pathways, adaptive learning environments, and optimize resource allocation by identifying effective interventions. Information from ML and DL models can inform program development, provide detailed feedback to students and parents, and guide the professional development of educators. Ongoing data analysis enables schools to monitor and refine interventions, promoting inclusion and equity in educational opportunities.

To apply these findings effectively, educational institutions must have access to quality data and the infrastructure needed to implement and maintain these predictive systems. Collaboration between data scientists, educators, and policymakers is crucial to ensure the ethical and constructive use of predictive tools in the educational environment. Predictive models can help educators and administrators provide targeted support to at-risk students through intervention programs, personalized education plans, parental involvement, professional development, and ongoing monitoring of interventions. Schools can also implement early warning systems, social and emotional support, career and college readiness programs, and foster a supportive school culture to improve student success. By comparing the performance of different models (Assegie et al., 2024; Zoralioğlu et al., 2023), the research provides information on which algorithms might be best suited to specific educational datasets, helping researchers and practitioners select the most appropriate tools. The study highlights the importance of feature selection methods and model refinement in achieving optimal performance, guiding future efforts to develop more accurate and efficient predictive systems. It also recognizes the computational demands of DL models and the potential obstacles for institutions with limited resources, arguing for more equitable access through resource-efficient models or shared computing infrastructure. The research emphasizes the need for transparent, comprehensible models to build trust and facilitate their integration into educational practice. Furthermore, the research highlights the technical expertise required to implement DL models, underlining the need for professional development and training for educators and researchers. It also underscores the necessity for policies and ethical guidelines to ensure the responsible use of predictive analytics in education, avoiding the reinforcement of biases or inequalities. By outlining limitations and potential directions for future research, the study encourages continued research and innovation in adapting ML and DL to educational contexts.

ML and DL models can identify various indicators of risk in student data by analyzing patterns and correlations. While these indicators are useful, they must be part of a holistic approach that takes into account the full context of each student. Ethical use of student data, avoidance of bias and respect for privacy are essential to the implementation of these predictive analyses. The research also has important implications for underserved communities. Simplified ML models that require less computing power can be particularly useful in these communities, ensuring that the benefits of predictive analytics are available to all schools, closing equity gaps.

Ultimately, this research advances understanding of how data-driven models can improve educational outcomes, providing a framework for selecting and optimizing these models and highlighting challenges and considerations for their effective and equitable use. ML and DL model predictions must be integrated into a holistic approach to student support, involving the essential human element of educators, counselors and administrators to interpret data and make informed decisions. Ensuring the ethical and respectful use of predictive analytics is paramount to supporting at-risk students and improving educational practices.

5. Future research

To extend the work presented in this study to better understand the factors that influence student performance and improve the predictive models used in educational institutions, future work will aim to integrate more diverse datasets, including longitudinal data to track changes over time, socio-economic factors, psychological assessments, and data from educational technologies used by students. This integration would provide a more comprehensive view of the factors influencing educational outcomes. Additionally, we will explore more sophisticated feature engineering techniques to extract new insights. Specifically, we will employ text analysis algorithms on student essays and sentiment analysis on teacher and peer feedback to uncover qualitative factors that influence performance. We will also closely integrate educational theories to interpret the meaning of features identified by machine learning models.

The factors currently used to predict students' future success remain insufficient. With the availability of new data sources, such as students' interactions with online learning platforms, video annotations of their classroom activities, and recorded facial expressions and emotions, we can enhance our understanding. Multimodal machine learning, which combines different types of data such as text, images, and video, could enable a better understanding of each student's background and development. We will explore this promising avenue of research into multimodal machine learning models to more accurately predict students' future success, considering both the cognitive and socio-emotional aspects of learning. This approach could help personalize student support and improve educational outcomes overall.

6. Conclusion

In conclusion, this study conducted an in-depth investigation into predictive modeling of student performance using various machine learning algorithms and feature selection methods. A set of these methods are never explored for this set of data and epically in education, as the FSMRMR, the results have shown the importance of this approach. The main objective was to identify the most influential features and the most effective models for predicting academic performance. The most critical finding of this research is the identification of a subset of features, denoted as SF, which includes student's age, number of past class failures, frequency of going out with friends, grades from previous periods, school absences, current health status, romantic relationship status, father's education, free time after school, and the aspiration for higher education. This subset has been found to have a high correlation with student performance and has consistently shown to enhance the predictive accuracy of the models tested. Among the various machine learning models evaluated, the Transformer model demonstrated superior performance, achieving the lowest RMSE and MAE values across the different feature sets. This indicates that the Transformer model is particularly adept at capturing the complexities and nuances of the data, making it a valuable tool for educational data analysis.

The study also compared the average performance of each feature selection and extraction method across all models and the average performance of each model across all feature subsets. The SF subset emerged as the most effective feature set, while the Transformer model stood out as the most accurate predictor. The implications of these findings are significant for the field of educational data mining. By pinpointing the most relevant features and the most effective predictive models, educators and policymakers can better understand the factors that contribute to student success and can develop targeted interventions to support students at risk of under performing. The contribution of this study lies in its systematic approach to feature selection and model evaluation, providing a clear benchmark for future research in the field. Additionally, the study offers practical insights that can be applied to enhance the design of educational programs and policies, ultimately aiming to improve

student outcomes and the quality of education. Overall, this research advances our knowledge of educational predictive analytics and sets the stage for more personalized and effective educational strategies that can adapt to the needs of individual students.

Statement on open data and ethics

The dataset used in this study is publicly available and can be downloaded from https://archive.ics.uci.edu/dataset/320/student+performance. All materials (dataset and associated codes and scripts) are available from the corresponding author on request.

CRediT authorship contribution statement

Abderrafik Laakel Hemdanou: Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mohammed Lamarti Sefian:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Conceptualization. **Youssef Achtoun:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition. **Ismail Tahiri:** Writing – review & editing, Writing – original draft, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table 7
Features table and description of each feature.

	Features	Description
1	School	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	Sex	Student's sex (binary: 'F' - female or 'M' - male)
3	Age	Student's age (numeric: from 15 to 22)
4	Address	Student's home address type (binary: 'U' - urban or 'R' - rural)
5	Famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8	Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9	Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	Reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	Guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
13	Traveltime	Home to school travel time (numeric: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour)
14	Studytime	Weekly study time (numeric: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - > 10 hours)
15	Failures	Number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
16	Schoolsup	Extra educational support (binary: yes or no)
17	Famsup	Family educational support (binary: yes or no)
18	Paid	Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	Activities	Extra-curricular activities (binary: yes or no)
20	Nursery	Attended nursery school (binary: yes or no)
21	Higher	Wants to take higher education (binary: yes or no)
22	Internet	Internet access at home (binary: yes or no)
23	Romantic	With a romantic relationship (binary: yes or no)
24	Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	Freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
26	Goout	Going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	Health	Current health status (numeric: from 1 - very bad to 5 - very good)
30	Absences	Number of school absences (numeric: from 0 to 93)
31	G1	First period grade (numeric: from 0 to 20)
32	G2	Second period grade (numeric: from 0 to 20)
33	G3	Final grade (numeric: from 0 to 20, output target)

References

Ahamed, M. A., Chaisanit, P., & Mahesh, T. (2019). Performance of student prediction. *International Journal of Computer Science and Information Technology*, *8*(6).

Al-Zawqari, A., Peumans, D., & Vandersteen, G. (2022). A flexible feature selection approach for predicting students' academic performance in online courses. *Computers and Education: Artificial Intelligence*, *3*, Article 100103.

Apriyadi, M. R., Ermatita, & Rini, D. P. (2023). Hyperparameter optimization of support vector regression algorithm using metaheuristic algorithm for student performance prediction. *International Journal of Advanced Computer Science and Applications*. https://api.semanticscholar.org/CorpusID:257396760.

Assegie, T. A., Salau, A. O., Chhabra, G., Kaushik, K., & Braide, S. L. (2024). Evaluation of random forest and support vector machine models in educational data mining. In *2024 2nd international conference on advancement in computation & computer technologies (InCACCT)* (pp. 131–135). https://api.semanticscholar.org/CorpusID:270396009.

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2023). Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science*, *216*, 597–603.

Begum, S., & Padmannavar, S. S. (2022). Genetically optimized ensemble classifiers for multiclass student performance prediction. *International Journal of Intelligent Engineering and Systems*. https://api.semanticscholar.org/CorpusID:247163927.

Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports*, *12*(1), Article 12508.

El Fouki, M., Aknin, N., & El Kadiri, K. (2019). Multidimensional approach based on deep learning to improve the prediction performance of dnn models. *International Journal of Emerging Technologies in Learning*, *14*(2).

Gong, L., Xie, S., Zhang, Y., Wang, M., & Wang, X. (2022). Hybrid feature selection method based on feature subset and factor analysis. *IEEE Access*, *10*, 120792–120803.

Hasanah, H., Farida, A., & Yoga, P. P. (2022). Implementation of simple linear regression for predicting of students' academic performance in mathematics. *Jurnal Pendidikan Matematika (Kudus)*. https://api.semanticscholar.org/CorpusID:250193787.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175–199).

Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, *9*(1), 26.

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., Isohanni, E., Korhonen, A., Petersen, A., Rivers, K., et al. (2015). Educational data mining and learning analytics in programming: literature review and case studies. In *Proceedings of the 2015 ITiCSE on working group reports* (pp. 41–63).

Isong, E., Kingsley, U., & Ansa, G. (2018). *Cognitive factors in students' academic performance evaluation using artificial neural networks*. *Information and knowledge management: Vol. 8* (pp. 57–71). Number 5.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Kumar, S., & Ahuja, B. (2022). Analysis and prediction of student performance by using a hybrid optimized bfo-alo based approach. *International Journal on Recent and Innovation Trends in Computing and Communication*. https://api.semanticscholar.org/CorpusID:255665919.

Kusumawardani, S. S., & Alfarozi, S. A. I. (2023). Transformer encoder model for sequential prediction of student performance based on their log activities. *IEEE Access*, *11*, 18960–18971.

Liu, C., Wang, H., Du, Y., & Yuan, Z. (2022). A predictive model for student achievement using spiking neural networks based on educational data. *Applied Sciences*. https://api.semanticscholar.org/CorpusID:248139774.

Liu, T., Zhang, M., Zhu, C., & Chang, L. (2023). Transformer-based convolutional forgetting knowledge tracking. *Scientific Reports*, *13*(1), Article 19112.

Liu, X., Li, T., Zhang, R., Wu, D., Liu, Y., & Yang, Z. (2021). A gan and feature selection-based oversampling technique for intrusion detection. *Security and Communication Networks*, *2021*, 1–15.

Mehmood, F., Ahmad, S., & Whangbo, T. K. (2023). An efficient optimization technique for training deep neural networks. *Mathematics*, *11*(6), 1360.

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

Nguyen, N., & Quanz, B. (2021). Temporal latent auto-encoder: A method for probabilistic multivariate time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence: Vol. 35* (pp. 9117–9125). Number 10.

Okoye, K., Nganji, J. T., Escamilla, J., & Hosseini, S. (2024). Machine learning model (rg-dmml) and ensemble algorithm for prediction of students' retention and graduation in education. *Computers and Education: Artificial Intelligence*, *6*, Article 100205.

Priyambada, S. A., Usagawa, T., & Mahendrawathi, E. (2023). Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge. *Computers and Education: Artificial Intelligence*, *5*, Article 100149.

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P-k. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1), 453.

Rashid, T. A., & Aziz, N. K. (2016). Student academic performance using artificial intelligence. *ZANCO Journal of Pure and Applied Sciences*, *28*(2).

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, Article 100018.

Rong, Y., He, T., Li, Z., & Liu, G. (2023). Achievement performance prediction model based on deep neural network of student similarity. In *Proceedings of the 4th international conference on modern education and information management, ICMEIM*. https://api.semanticscholar.org/CorpusID:265645206.

Satake, A., Fujiyoshi, H., Yamashita, T., Hirakawa, T., & Shimada, A. (2021). Performance prediction and importance analysis using transformer. In *29th international conference on computers in education conference, ICCE 2021–proceedings* (pp. 538–543).

Shahiri, A. M., Husain, W., et al. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422.

Shiverick, S. M. (2019). Predictive models of student performance for data-driven learning analytics. In *Proceedings of the international conference on learning analytics and knowledge*. https://api.semanticscholar.org/CorpusID:197524537.

Sorkar, N. U., & Sorkar, R. U. (2022). Determinants and prediction of secondary students performance in mathematics in Portugal using machine learning. *Computer Engineering and Intelligent Systems*. https://api.semanticscholar.org/CorpusID:253320443.

Trakunphutthirak, R., Cheung, Y., & Lee, V. C. (2019). A study of educational data mining: evidence from a Thai university. In *Proceedings of the AAAI conference on artificial intelligence: Vol. 33* (pp. 734–741). Number 01.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Villa-Blanco, C., Bielza, C., & Larrañaga, P. (2023). Feature subset selection for data and feature streams: a review. *Artificial Intelligence Review*, 1–52.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. arXiv preprint. arXiv:2106.11342.

Zhao, Y. (2024). Improvement of applicability in student performance prediction based on transfer learning. https://api.semanticscholar.org/CorpusID:271270133.

Zoralioğlu, Y., Gül, M. F., Azizoğlu, F., Azizoglu, G., & Toprak, A. N. (2023). Predicting academic performance of students using machine learning techniques. In *2023 innovations in intelligent systems and applications conference (ASYU)* (pp. 1–6). https://api.semanticscholar.org/CorpusID:264881420.