# Leveraging AI for Effective Teaching: A Machine Learning Approach to Tandem Learning in Mathematics

## Abstract

Educational institutions aim to offer quality education, employing diverse teaching methods like tandem learning. Recognizing the need for personalized approaches, institutions should use data mining techniques to extract insights from educational datasets for optimal predictive model selection for individual students or classrooms. The aim of this study was to evaluate the performance of machine learning (ML) algorithms for predicting student response to tandem learning.

A dataset comprising 89 high school students and 13 predictor variables was utilized. The focus was on a three-state variable that determined whether the student positively responded to the integration of tandem learning into the educational environment. The predictor variables included gender, class, teacher, previous mathematics grade, MBTI variables (extroversion-introversion, sensing-intuition, thinking-feeling, and judging-perceiving), mathematical anxiety, mathematical motivation, qualitative interaction in tandem learning, quantitative interaction in tandem learning, and whether the student outperformed their partner in tandem learning. Nine classification ML algorithms were implemented and the 5 by 2-fold cross-validation with stratified folds was utilized.

Using all predictor variables, Random Forest and K-Nearest Neighbors performed the best, having accuracies of 0.55, and 0.53, and macro F1 scores 0.37, and 0.36 respectively, which is fair considering data balance. Balancing the dataset and using only 2 outcome classes, the performance improved, with the best algorithm being Gradient boosting, performing moderately well (accuracy = 0.59; F1-score = 0.59).

The results imply that machine learning algorithms struggle to accurately predict students' responses to group learning in mathematics using the variables and sample size employed. As a result, they may not be appropriate for aiding teachers in making decisions about selecting teaching methods. Therefore, a simplified approach can yield more effective results, in our case from the transformation of the machine learning classification problem from three states to two.

### Keywords

Assessment, mathematics education, machine learning, tandem learning, teaching methods

### Math subject classification, MSC2020

97D40, 97D60, 62P99

## Introduction

Deep comprehension and knowledge retention are crucial in teaching mathematics (Adler et al., 2014). Recent educational models advocate for indirect learning processes, emphasizing the drawbacks of large classrooms (Mbofana & Banda, 2022; Olasen & Lawal, 2020) and traditional teacher-centered methods (Dervić et al., 2018; Lasry et al., 2014). Small-group learning, including tandem learning where two students collaborate on various activities (Stickler & Emke, 2011; G. Wilson & Blednick, 2011), has been shown to enhance academic achievement (Kalaian & Kasim, 2014), attitudes towards learning (Gaudet et al., 2010), and participation in STEM courses (Wieselmann et al., 2020). It is important to mention that this learning method is not only theoretical and a debate of research since it is used at some level by many teachers (Slavin et al., 2003).

Understanding and effectively mapping students' cognitive and non-cognitive traits, demographic factors, and teacher-related aspects is essential for implementing any teaching method (Gnesdilow et al., 2013; Kurniawati et al., 2023; Li et al., 2021; Puklek, 2001; Van Der Laan Smith & Spindle, 2007; Van Diggele et al., 2020). It is crucial to recognize the appropriate conditions for the application of such learning methods to maximize their positive impact on student outcomes. The recent advancements in artificial intelligence (AI), particularly machine learning (ML) based algorithms, offer unique opportunities for researchers to predict personalized education outcomes through the analysis of heterogeneous data (Sekeroglu et al., 2019; Yağcı, 2022). The success of these models relies on their interpretability, replicability, and generalizability, and there is a growing trend to enhance these aspects (Rastrollo-Guerrero et al., 2020).

While there is an increasing use of ML/AI models to predict overall student success in education (Ho et al., 2021; Ibarra-Vazquez et al., 2023; Luan & Tsai, 2021; Musso et al., 2020; Yağcı, 2022), their applicability, usability, and generalizability in the context of teaching method data remain significant research gaps. Therefore, this study aims to address this gap by developing machine-learning models using a substantial Slovene dataset (*N =* 89). These models will predict the personalized success of tandem learning based on various demographic factors, teacher-related factors, as well as cognitive and non-cognitive factors. The aim of the research is, therefore, to shed light on which ML algorithms perform the best and are the most accurate in predicting the effectiveness of tandem learning considering 13 features, including gender, class, teacher, previous mathematics grade, MBTI variables (extroversion-introversion, sensing-intuition, thinking-feeling, and judging-perceiving), mathematical anxiety, mathematical motivation, qualitative interaction in tandem learning, quantitative interaction in tandem learning, and whether the student outperformed their partner in tandem learning. Having a deeper understanding of which algorithm is the most accurate in predicting the effectiveness of applying the tandem learning format might help educators and researchers to adopt new methods, based on ML, to analyze data and implement tailored learning and teaching methods.

## Theoretical framework

### 2.1 Tandem learning

Recent trends to develop and use alternative, indirect forms of educational processes (Arias & Peralta, 2011; Blažič et al., 2003). This is driven by the belief that such approaches are more effective in fostering improved academic achievement (S. Wang et al., 2023). This is driven by the belief that such approaches are more effective in fostering improved academic achievement (Kalaian & Kasim, 2014), developing more positive attitudes toward learning (Gaudet et al., 2010; Hillyard et al., 2010), and enhancing persistence in STEM courses and programs (Kalaian et al., 2018; Micari et al., 2010; Wieselmann et al., 2020; S. Wilson & Varma-Nelson, 2016).

One notable small-group learning practice among these alternatives is tandem learning. This unique approach involves pairs of students conducting experiments together, formulating reports, solving problems, and so forth (Stickler & Emke, 2011; Tomić, 2002; G. Wilson & Blednick, 2011). Tandem learning stands out for its simplicity from an organizational standpoint, offering increased opportunities for active engagement compared to frontal and group teaching methods. Importantly, it provides a middle ground between the collaborative nature of group learning and the individualized approach of solo learning (Blažič et al., 2003).

The implementation of group learning serves to accomplish five significant objectives, as articulated by Peklaj (2001): (1) facilitating mutual understanding among students, (2) fostering a sense of group identity, (3) promoting peer support, (4) cultivating an appreciation for differences among group members, and (5) instilling characteristics of teamwork. This approach closely aligns with the five fundamental components of cooperative learning as outlined by Johnson et al. (1991), namely (1) positive interdependence, where students rely on each other for success; (2) face-to-face promotive interaction, encouraging constructive communication; (3) individual accountability and personal responsibility, ensuring active participation of each student; (4) regular utilization of interpersonal and small group social skills; and (5) consistent, periodic evaluation of group dynamics and performance. By embracing these principles, educators can better equip their students with the social and interpersonal competencies essential for success in the contemporary world. Slavin et al. (2003) identified four significant theoretical perspectives on the achievement effects of cooperative learning: (1) motivationalist, (2) social cohesion, (3) cognitive-developmental, and (4) cognitive-elaboration. The last two focus on the interaction among groups of students, and all four perspectives can be viewed as complementary.

### 2.2 Machine learning and classification

Data mining is the process of uncovering hidden patterns, relationships, or insights within vast datasets through techniques from statistics and database management (Baradwaj & Pal, 2012). It involves data preprocessing to prepare information for analysis and utilizes methods such as clustering and association rule mining (Singhal & Jena, 2013). In contrast, ML, a subset of AI, focuses on building predictive models by allowing computers to learn from data and make decisions or predictions (Candanedo et al., 2018). The sequences of steps identified in extracting knowledge from data are shown in Figure 1 below.

A black and white image of a square with a arrow pointing to the left

Description automatically generated with medium confidence

Figure 1: Knowledge discovery process.

ML encompasses various learning paradigms, including the classification process, and finds applications in areas like recommendation systems. Classification, a fundamental task in both data mining and ML, involves categorizing data into predefined classes, such as binary or three-state classification, based on patterns learned from labeled data. This classification process is employed in various domains, such as healthcare and sentiment analysis, which includes educational sciences (Shaik et al., 2023), to make data-driven decisions (Baradwaj & Pal, 2012).

Classification in ML is a pivotal process where the primary goal is to train a model to assign predefined labels or categories to new, unseen instances based on patterns identified during the learning phase (Charbuty & Abdulazeez, 2021). It relies on the extraction of features from the input data, which are then used to make predictions about the class to which a particular instance belongs. This process involves both supervised learning, where the model is trained on labeled data, and unsupervised learning, where the model must identify patterns in unlabeled data (Berry et al., 2020).

Sentiment analysis, as mentioned in the context of educational sciences, is a specific application of classification in natural language processing (Mite-Baidal et al., 2018). This involves determining the sentiment expressed in textual data, such as reviews, social media posts, or survey responses. By classifying sentiments into categories like positive, negative, or neutral, organizations can gain valuable insights into public opinion and tailor their strategies accordingly.

### 2.3 Connecting tandem learning and machine learning

Technologies can be considered in terms of whether they are mainly student teaching (with a primarily instructionist approach), student supporting (primarily constructivist approach), or teacher supporting (which primarily helps teachers do tasks they already do but faster or with less effort) (Holmes et al., 2019).

Beyond its broader applications, ML has been harnessed to precisely forecast student success (Abana, 2019; Kotsiantis et al., 2004; Ofori et al., 2020; Qazdar et al., 2019; Rastrollo-Guerrero et al., 2020; Yakubu & Abubakar, 2022), identify at-risk learners (Adnan et al., 2021; Chui et al., 2020), and tailor educational interventions (Luan & Tsai, 2021; Stimpson & Cummings, 2014; Tsai et al., 2020; Yang, 2021). The concept of “learning analytics” was introduced and demonstrated by other studies (Bhusal, 2021; Cortez & Silva, 2008; Siemens & Gasevic, 2012), before the widespread use of ML, which used weaker statistical methods. Aside from forecasting success, ML can help us identify the most important variables that affect said forecast (Lu et al., 2020; Luan & Tsai, 2021). Multiple studies have delved into the analysis of crucial features in the learning environment (Hodges, 2018; Humphrey et al., 2009; Moradi et al., 2018; Scribner & Donaldson, 2001), however only a few have harnessed the power of modern algorithms, such as ML, which hold the potential for significantly enhanced sight. Therefore, it is of paramount importance to explore in great detail the feature selection problem.

## Empirical work

The present research was conducted as the effectiveness of tandem learning in high school mathematics remains unclear due to the complexity of numerous variables influencing its success.

This study aims to leverage three-state classification machine learning algorithms to analyze multifaceted variables to determine the successfulness of a tandem learning environment.

Given the scarcity of comprehensive studies employing ML algorithms to assess teaching method selections, this research acknowledges the potential limitations in drawing definitive conclusions. While aiming to discern patterns of predicting the success of tandem learning environment, this study will consider the need for a cautious interpretation of results. Even in the absence of conclusive findings, insights gathered from the analysis will contribute to the ongoing discourse on the efficacy of tandem learning in mathematics education. Therefore we hypothesize that even when feeding ML models with diverse data sets encompassing student profiles and collaborative dynamics, some algorithms will perform sub par, while others may perform fair.

### 3.1 Methodology

In the present research, the causal non-experimental method of pedagogical research is applied.

### 3.2 Sample

In this retrospective study, we used a dataset gathered at a high school in Slovenia, which is publicly available at (Author, 2023). The sample comprised 89 11th- and 12th-grade students of a Slovenian Gymnasium (i.e., high school); 28 males and 61 females.

Participants in the study engaged in tandem learning for about a week during class time, with no randomization in pairings. The structure involved a mix of regular classroom activities and tandem learning sessions, with students seated in pairs at two-seat desks. The efficacy of tandem learning was measured across various factors, categorized into demographic, psychological, and tandem-learning-related variables (see section 3.5.1).

### 3.3 Data preprocessing and data analysis

The gathered data was analysed using Python programming language, primarily using *pandas* (version 2.1.3) and *scikit-learn* (version 1.3.2) libraries. ML code, as well as variable importance and other data analysis notebooks, are openly accessible at (Author, 2023).

Dataset was already in the form of tidy data (Wickham, 2014). Categorical variables, which were a priori encoded using label encoding were transformed into dummy variables using one-hot as most ML algorithms require numerical input and do not handle well numerical data, that is not continuous (Tan & Pu, 2023; Wu et al., 2020).

We implemented nine ML algorithms, chosen for their effectiveness in classification tasks and their ability to perform three-state classification (see references at each algorithm). Additionally, the analyses obtained by the usage of nine different ML algorithms could provide us with additional information about their comparability, i.e. their capability of predict correctly the effectiveness of tandem learning.

Naive Bayes (NB; Frank et al., 2000): it is a probabilitstic classifier based on Bayes' theorem and it assumes independence among features;

k-Nearest Neighbors (k-NN; Zhang, 2016): it is a non-parametric and instance-based learning algorithm which classifies data points on the majority class of their k-nearest neighbors. It is effective in classification and it is sensitive to the choice of the distance metric;

Decision Tree (Charbuty & Abdulazeez, 2021): it is a hierarchical tree-like structure that represents decisions based on features; it is easily interpretable, however it is prone to overfitting;

Logistic Regression Models with OvA strategy (LR; Sun et al., 2019): It is a binary classification algorithm used to predict the probability of an instance belonging to a particular class. However, when dealing with multi-class classification problems (more than two classes), the one-vs-all (OvA) strategy, also known as one-vs-rest (OvR), is often employed. The OvA strategy involves training three separate binary logistic regression models. Each model is designed to distinguish one class from the combination of the other two;

Linear Discriminant Analysis (LDA; Xanthopoulos et al., 2013): it is a classification algorithm seeking to find linear combinations of features that best separates classes; it is useful in dimensionality reduction and feature extraction, but it assumes normally distributed classes and equal covarances;

AdaBoost (with 1000 estimators; Schapire, 2003): it is an ensemble learning method that combines weak classifiers to create a strong classifier; it gives more weight to misclassified instances which improves the overall performance of the algorithm; it is also effective in boosting the performance of decision trees;

Gradient Boosting (GB with 1000 estimators; Natekin & Knoll, 2013): it is an ensemble technique that builds trees sequentially, with each tree correcting the errors of the previous one; it often emplyes decision trees as weak learners; it is a robust and widely used classification technique;

Support Vector Machine (SVM with RBF kernel; Suthaharan, 2016): it is a powerful algorithm for classification which constructs a hyperplane that maximally separates the classes; it is effective in high-dimensional spaces and can handle non-linear relationships through kernel functions;

Random Forest (RF with 1000 estimators; Rodriguez-Galiano et al., 2015): it is an ensemble learning method that builds multiple decision trees and combines their predictions while reducing overfitting and improving accuracy compared to individual trees; it is widely used in classification tasks.

Said algorithms are capable of performing 3-state classification. Models were evaluated using 5x2 cross-validation (CV) with stratified folds (Cieslak & Chawla, 2008) with the main metric of performance evaluation being macro F1-score through average fold repetition (Chicco & Jurman, 2020). The 5x2-CV is a technique that splits datasets into two subsets (folds), and the model is trained and tested on these subsets. The process is repeated five times, resulting in 10 overall evaluations. The final performance metric is the average of the performance metrics obtained from the ten evaluations. This method helps to reduce the variability that might arise from a single random split and provides more reliable estimates of the model’s performance. This method assesses how well a ML model generalizes unseen data and helps prevent overfitting. It is useful when the dataset is limited (Raschka, 2018). F1-score is a metric commonly used to evaluate the performance of a classification model, specifically when dealing with unbalanced datasets (Chicco & Jurman, 2020). It is the harmonic mean of precision () and recall (), ranging from 0 to 1, with higher values indicating better model performance (L. Wang et al., 2021). In the multi-class case, we must consider all the classes i.e. all the entries of the confusion matrix. To do so, we require a multi-class measure of precision and recall to be inserted into the harmonic mean. Such metrics may have two different specifications, giving rise to two different metrics: Micro F1-Score and Macro F1-Score (Opitz & Burst, 2019). The F1 macro score in a multi-class scenario involves computing Macro-Precision and Macro-Recall by averaging precision for each predicted class and recall for each actual class. In Macro F1-Score, all classes carry equal weight in the calculation, eliminating distinctions between highly and poorly populated classes (Grandini et al., 2020).

Additionally, the t-distributed stochastic neighbor embedding (t-SNE; Van der Maaten & Hinton, 2008) was employed in this research. The t-SNE is an ML algorithm commonly used for dimensionality reduction and visualization of high-dimensional data in a lower-dimensional space. It is particularly effective in revealing the underlying structure and patterns within the data, making it valuable for exploratory data analysis and visualization tasks. t-SNE works by modeling the similarity between data points in the high-dimensional space and the low-dimensional space. The algorithm focuses on preserving the pairwise similarities between data points, emphasizing the preservation of local relationships rather than global ones. t-SNE is a non-linear technique, meaning it can capture complex relationships between data points that linear methods might struggle to represent accurately. The t-SNE technique aims to retain the local structures and similarities between nearby data points in the original high-dimensional space. This is especially useful for visualizing clusters and patterns in the data.

### 3.4 Previous findings on the used dataset

In our research, we utilized a pre-analyzed dataset, as explored by (Author, 2024). The internal consistency of variables, assessed through instruments comprising multiple items, was previously examined with Cronbach alpha being used for variables (1) Anxiety, (2) Motivation, (3) Introversion, (4) Sensing, (5) Feeling, and (6) Judging all of them having fair results.

Moreover, feature importance was performed using Mutual information and recursive feature elimination in regards to logistic regression, highlighting that variables related to tandem learning held the utmost significance, followed by variables in a general context, and lastly, variables associated with students' psychological profiles. Key predictors for student responses were identified, with outperforming partner, class, and qualitative interaction within groups being the most influential, as indicated by mutual information scores. Recursive feature analysis underscored the importance of qualitative interaction, outperforming partner, and gender as primary predictors.

In addition, correlation analysis using Spearman ρ (all variables were analyzed with the Shapiro-Wilk test as normally distributed) unveiled meaningful relationships within the dataset. The three most positively correlated pairs were gender-anxiety, interaction qualitative-interaction quantitative, and grade-motivation, while the most negatively correlated pairs were motivation-anxiety, gender-feeling, and gender-motivation. All correlation coefficients exhibited a moderate strength of association.

### 3.5 Results

#### 3.5.1 Student sample and variables used

The target variable was a three-state measure of successfulness on a 3-levels Likert scale (“successful”, “neutral”, “not successful”), both regarding academic performance as well as general wellbeing during the implementation of tandem learning. The predictors were variables in (1) general sense (gender, class, teacher, previous mathematics grade), (2) psychological sense (MBTI variables: extroversion-introversion, sensing-intuition, thinking-feeling and judging-perceiving, and other variables: mathematical anxiety and motivation), and (3) tandem-learning-related variables (qualitative interaction, quantitative interaction, and whether student outperformed their partner). All variables provided were used for ML purposes. Their descriptive statistics can be found in Table 1, Table 2, and Figure 2.

Table 1: Dataset description of tandem-learning-related, and general variables used. The categorical variables, initially label encoded, were transformed using one-hot encoding, generating dummy variables, as that is needed for ML algorithms used (see references in section 3.3.). Ordinal variables assessed using Likert scales are also described utilizing measures of central tendency due to their inherent numerical order and range.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Successfulness | Grade | Interaction  quantitative | Interaction  qualitative | Outperforming  partner | Class | Teacher | Gender | |
| *M* | 2.4 | 3.4 | 2.2 | 2.1 | 2.1 | Categorical  (7 options) | Categorical  (4 options) | Categorical  (2 options) |
| *SD* | 0.6 | 1.0 | 0.7 | 0.7 | 0.6 |
| *min* | 1.0 | 2.0 | 1.0 | 1.0 | 1.0 |
| 25% | 2.0 | 3.0 | 2.0 | 2.0 | 2.0 |
| 50% | 2.0 | 3.0 | 2.0 | 2.0 | 2.0 |
| 75% | 3.0 | 4.0 | 3.0 | 3.0 | 2.0 |
| *max* | 3.0 | 5.0 | 3.0 | 3.0 | 3.0 |

Note. M = mean, SD = standard deviation.

Table 2: Dataset description of psychological variables. All variables belonging to this type were treated as continuous.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Anxiety | Motivation | Introversion | Sensing | Feeling | Judging |
| *M* | 25.8 | 20.4 | 20.6 | 22.7 | 23.3 | 22.8 |
| *SD* | 6.8 | 6.3 | 5.6 | 4.5 | 4.7 | 5.7 |
| *min* | 10.0 | 7.0 | 8.0 | 12.0 | 9.0 | 9.0 |
| 25% | 21.0 | 16.0 | 16.0 | 20.0 | 20.0 | 20.0 |
| 50% | 26.0 | 20.0 | 21.0 | 23.0 | 23.0 | 23.0 |
| 75% | 31.0 | 24.0 | 24.0 | 25.0 | 26.0 | 26.0 |
| *max* | 40.0 | 34.0 | 37.0 | 35.0 | 35.0 | 37.0 |

A group of graphs with different colored bars

Description automatically generated with medium confidence

Figure 2: Variables used.

For ML purposes that follow it is important to know that the target variable was not balanced with represented classes being 6, 39, and 44, indicating the population distribution across the ordinal classes 0, 1, and 2, respectively, on a Likert scale. Therefore, the usage of the F1-score for the assessment of the model’s performance was deemed suitable (Chicco & Jurman, 2020; Wang et al., 2021). For baseline reference, we consider a simple model that predicts the majority class for all instances, resulting in an accuracy of 44 (representing the largest class) out of 89 (total samples), equivalent to 49%. However, it is crucial to note that the evaluation metric after surpassing this baseline would be the F1-score, providing a more nuanced assessment of the model's performance in handling imbalanced class distributions. Subsequently, we transitioned to a two-state classification by merging classes 0 (“not successful”) and 1 (“neutral”). Finally, we reintroduced the three-state classification, but this time with a reduced set of variables, based on their predicting strength as seen in literature, to enhance the model performance. Notably, throughout this process, we refrained from dropping variables and merging classes simultaneously.

#### 3.5.2 Model performance

According to the F1-metric, the two models with generally better performances in our experiments were RF and k-NN. Their performances are fair, as their accuracy was still better than a number of samples in the majority class (49.4%). Therefore the models outperform the baseline. Other models performed worse, their poor performance indicating they might not be learning the underlying patterns effectively. The classification of individual ML techniques is presented in Table 3.

Table 3: Classification result.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision\_macro | Recall\_macro | F1\_macro |
| NB | 0.34 | 0.41 | 0.39 | 0.30 |
| k-NN | 0.53 | 0.36 | 0.38 | 0.36 |
| Decision Tree | 0.44 | 0.36 | 0.47 | 0.33 |
| LR | 0.48 | 0.33 | 0.34 | 0.33 |
| LDA | 0.44 | 0.31 | 0.31 | 0.31 |
| AdaBoost | 0.39 | 0.27 | 0.28 | 0.27 |
| GB | 0.46 | 0.32 | 0.33 | 0.32 |
| SVM | 0.45 | 0.16 | 0.30 | 0.21 |
| RF | 0.55 | 0.38 | 0.39 | 0.37 |

In analyzing the model's performance, the plotted confusion matrices, as shown in Figure 3 provide a visual representation of its classification accuracy, revealing the interplay between true positives, true negatives, false positives, and false negatives (Chicco et al., 2021).

A grid of blue squares

Description automatically generated

Figure 3: Confusion matrices of classification.

To counter the imbalance in our dataset, we opted for a binary approach, merging classes 0 and 1 to create a more balanced representation (Harangi et al., 2020). Therefore we got class 0 (“not successful”) with 45 (50.6%) and class 1 (“neutral”) with 44 (49.4%) representations. With this approach all algorithms, except RF and k-NN performed fair. GB performed the best, with moderate accuracy (0.59 averaged F1-score), as shown in Table 4 and Figure 4.

Table 4: Classification result of binary case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision\_macro | Recall\_macro | F1\_macro |
| NB | 0.54 | 0.54 | 0.54 | 0.51 |
| k-NN | 0.5 | 0.49 | 0.5 | 0.49 |
| Decision Tree | 0.56 | 0.58 | 0.56 | 0.55 |
| LR | 0.54 | 0.54 | 0.54 | 0.53 |
| LDA | 0.55 | 0.55 | 0.55 | 0.54 |
| AdaBoost | 0.55 | 0.55 | 0.55 | 0.54 |
| GB | 0.59 | 0.59 | 0.59 | 0.59 |
| SVM | 0.58 | 0.59 | 0.58 | 0.56 |
| RF | 0.49 | 0.48 | 0.49 | 0.48 |

A screenshot of a diagram

Description automatically generated

Figure 4: Confusion matrices for binary case.

In addition, we conducted ML algorithms with a reduced set of predictor variables, as previously discussed. Specifically, we excluded all six variables related to the psychological profile of students within our sample. It is crucial to note that this exclusion did not substantially alter the performance across all the algorithms, as shown in Table 5 and Figure 5. The omission of psychological variables did not lead to a significant impact on the models, and they exhibited similar predictive accuracy. This underscores the robustness of the models even in the absence of these psychological variables in our predictive modeling framework.

Table 5: Classification results with selected features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision\_macro | Recall\_macro | F1\_macro |
| NB | 0.34 | 0.4 | 0.38 | 0.28 |
| k-NN | 0.54 | 0.37 | 0.39 | 0.37 |
| Decision Tree | 0.54 | 0.42 | 0.41 | 0.41 |
| LR | 0.49 | 0.32 | 0.35 | 0.33 |
| LDA | 0.46 | 0.31 | 0.33 | 0.31 |
| AdaBoost | 0.37 | 0.28 | 0.39 | 0.31 |
| GB | 0.52 | 0.4 | 0.42 | 0.39 |
| SVM | 0.52 | 0.35 | 0.37 | 0.35 |
| RF | 0.52 | 0.36 | 0.36 | 0.34 |

A screenshot of a graph

Description automatically generated

Figure 5: Confusion matrices with dropped variables regarding student psychological profile.

Given the suboptimal performance of our model, we employed t-SNE analysis to visually explore and comprehend the underlying patterns and relationships within the data, aiming to uncover potential complexities or overlaps that might have impacted the model's performance (Bibal et al., 2023). These analyses in two and three dimensions can be found in Figure 6 and Figure 7, respectively, and they show that the patterns in data may be hard to distinguish between, as was speculated. On the three- dimensional plot, it is however seen, that the separation between class 0 and class 1 can be found.

A diagram of a graph

Description automatically generated with medium confidence

Figure 6: 2D t-SNE plot.

A diagram of a graph

Description automatically generated

Figure 7: 3D t-SNE plot.

## Discussion

Research shows various small-group learning methods should be considered in teaching (S. Wang et al., 2023), as they promote academic achievement (Kalaian et al., 2018) as well as a positive attitude towards learning (Gaudet et al., 2010; Hillyard et al., 2010). In this regard, tandem learning (Stickler & Emke, 2011; G. Wilson & Blednick, 2011) is an easy-to-implement method. Not everyone responds well to a one-size-fits-all method, and therefore, uncovering insights for predictive model selection tailored to individual students or classrooms becomes imperative for teaching institutions. Customizing the approach to accommodate individual students or classrooms is crucial for educational institutions, as a uniform method may not resonate with everyone (Ahmad et al., 2021; Pratt, 2002). Extracting valuable insights from the educational dataset requires the application of data mining techniques. Given the potential complexity of the relationships between various factors, machine learning methods are considered the most appropriate (Hilbert et al., 2021).

In this study, we used nine ML algorithms, specifically (1) NB, (2) k-NN, (3) Decision Tree, (4) LR, (5) LDA, (6) AdaBoost, (7) GB, (8) SVM, and (9) RF to predict the success of tandem in three-states of tandem learning dataset comprising of 89 students with 14 variables in various (continuous, categorical and ordinal) forms. Models performed poorly to fairly. The best accuracy and F1 score were achieved when the target variable was transformed into two states, however, performance was still not the best. When taking variable importance into account, the model did also not notably improve. Best models of classification were not consistent, which emphasizes certain algorithms are more suited to a particular problem (Alzubi et al., 2018), as well as metric chosen for model evaluation (Erickson & Kitamura, 2021).

The insights could contribute significantly to the development of tailored individual instructional strategies aimed at optimizing results and overall wellbeing during different teaching methods implementation. However, for the implementation of results, we would like for the models to be more accurate and consistent, therefore more insight regarding variables used is needed for model improvement. Based on the results of the present research, we suggest researchers to investigate the utilization of ML algorithms to measure the effectiveness of specific methodologies by applying different algorithms, which might perform very differently in different situations. Therefore, from the findings of the present research, we cannot state that there is a single ML method that should be preferred, nor that the description of the effectiveness of tandem learning might be described using the sole 13 predictors we employed in the present study. Nevertheless, the study highlights that some ML algorithms had fair accuracies, therefore their utilization for the assessment of pedagogical practices is encouraged.”

The application of supervised machine learning in predicting student response to tandem learning poses significant ethical challenges. While leveraging various general, psychological, and tandem learning-specific variables enhances predictive accuracy, it raises concerns regarding privacy, bias, and informed consent (Akgun & Greenhow, 2022; Starke et al., 2021). Collecting and analyzing personal data to predict individual behavior may compromise students' privacy, requiring strict adherence to data anonymization and consent protocols. Moreover, the inclusion of variables like cultural background or learning styles may inadvertently perpetuate biases or stereotypes (Marinucci et al., 2023). Transparency in model development, interpretability, and ongoing evaluation to mitigate biases remain pivotal.

## Conclusions and limitations

This study demonstrates that machine learning algorithms can be used for personalized predictive estimation of whether students respond well to new teaching methods. As using AI models is a fairly new construct, especially in education research, there is still a lot of room for improvement. Key factors that influence results must be investigated further to achieve better model outcomes. The potential incorporation of the developed model in regards to teacher assistance needs to be investigated further.

The present study does not include a prediction of whether tandem learning is overall effective or not. It simply includes information about predicting certain student performance, which is a limitation. Some of the variables, likely relevant were not used in our models because they were not obtained, which might have led to a slight bias in our study. The dataset exhibited a slight imbalance, with merely 6.7% of students reporting that the approach was unsuccessful, which could pose a challenge to the accuracy of the model. Therefore, we suggest researchers to explore the usage of ML algorithms to predict the efficacy of a teaching method by including additional predictor variabbles that might have a bigger weight on the effectiveness.

## References

Abana, E. (2019). A Decision Tree Approach for Predicting Student Grades in Research Project using Weka. *International Journal of Advanced Computer Science and Applications*, *10*(7). https://doi.org/10.14569/IJACSA.2019.0100739

Adler, J., Hossain, S., Stevenson, M., Clarke, J., Archer, R., & Grantham, B. (2014). Mathematics for teaching and deep subject knowledge: Voices of Mathematics Enhancement Course students in England. *Journal of Mathematics Teacher Education*, *17*(2), 129–148. https://doi.org/10.1007/s10857-013-9259-y

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

Ahmad, A., Zeeshan, F., Marriam, R., Samreen, A., & Ahmed, S. (2021). Does one size fit all? Investigating the effect of group size and gamification on learners’ behaviors in higher education. *Journal of Computing in Higher Education*, *33*(2), 296–327. https://doi.org/10.1007/s12528-020-09266-8

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Arias, R., & Peralta, H. (2011). La enseñanza, una puerta para la complejidad y la crítica. *Estudios Pedagógicos (Valdivia)*, *37*(1), 293–302. https://doi.org/10.4067/S0718-07052011000100017

Author. (2023). *Online repository* (1.0) [dataset].

Author. (2024). *In press*.

Baradwaj, B. K., & Pal, S. (2012). *Mining Educational Data to Analyze Students’ Performance* (arXiv:1201.3417). arXiv. https://doi.org/10.48550/arXiv.1201.3417

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2020). *Supervised and Unsupervised Learning for Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2

Bhusal, A. (2021). *Predicting Student’s Performance Through Data Mining*. https://doi.org/10.48550/ARXIV.2112.01247

Bibal, A., Delchevalerie, V., & Frénay, B. (2023). DT-SNE: T-SNE discrete visualizations as decision tree structures. *Neurocomputing*, *529*, 101–112. https://doi.org/10.1016/j.neucom.2023.01.073

Blažič, M., Ivanuš-Grmek, M., Kramar, M., & Strmčnik, F. (2003). *Didaktika: Visokošolski učbenik*. Visokošolsko središče, Inštitut za raziskovalno in razvojno delo.

Candanedo, I. S., Nieves, E. H., González, S. R., Martín, M. T. S., & Briones, A. G. (2018). Machine Learning Predictive Model for Industry 4.0. In L. Uden, B. Hadzima, & I.-H. Ting (Eds.), *Knowledge Management in Organizations* (Vol. 877, pp. 501–510). Springer International Publishing. https://doi.org/10.1007/978-3-319-95204-8\_42

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(1), 13. https://doi.org/10.1186/s13040-021-00244-z

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*, *107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032

Cieslak, D. A., & Chawla, N. V. (2008). Learning Decision Trees for Unbalanced Data. In W. Daelemans, B. Goethals, & K. Morik (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 5211, pp. 241–256). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-87479-9\_34

Cortez, P., & Silva, A. (2008). *Using data mining to predict secondary school student performance*.

Dervić, D., Salibašić Glamočić, D., Gazibegović-Busuladžić, A., & Mešić, V. (2018). Teaching physics with simulations: Teacher-centered versus student-centered approaches. *Journal of Baltic Science Education*, *17*(2), 288–299. https://doi.org/10.33225/jbse/18.17.288

Erickson, B. J., & Kitamura, F. (2021). Magician’s Corner: 9. Performance Metrics for Machine Learning Models. *Radiology: Artificial Intelligence*, *3*(3), e200126. https://doi.org/10.1148/ryai.2021200126

Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Naive Bayes for regression. *Machine Learning*, *41*(1), 5–25. https://doi.org/10.1023/A:1007670802811

Gaudet, A. D., Ramer, L. M., Nakonechny, J., Cragg, J. J., & Ramer, M. S. (2010). Small-Group Learning in an Upper-Level University Biology Class Enhances Academic Performance and Student Attitudes Toward Group Work. *PLoS ONE*, *5*(12), e15821. https://doi.org/10.1371/journal.pone.0015821

Gnesdilow, D., Evenstone, A. L., Rutledge, J., Sullivan, S., & Puntambekar, S. (2013). *Group Work in the Science Classroom: How Gender Composition May Affect Individual Performance*. 34–37. https://doi.org/10.13140/2.1.1718.5285

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview*. https://doi.org/10.48550/ARXIV.2008.05756

Harangi, B., Baran, A., & Hajdu, A. (2020). Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support. *Biomedical Signal Processing and Control*, *62*, 102041. https://doi.org/10.1016/j.bspc.2020.102041

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), e3310. https://doi.org/10.1002/rev3.3310

Hillyard, C., Gillespie, D., & Littig, P. (2010). University students’ attitudes about learning in small groups after frequent participation. *Active Learning in Higher Education*, *11*(1), 9–20. https://doi.org/10.1177/1469787409355867

Ho, I. M. K., Cheong, K. Y., & Weldon, A. (2021). Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques. *PLOS ONE*, *16*(4), e0249423. https://doi.org/10.1371/journal.pone.0249423

Hodges, L. C. (2018). Contemporary Issues in Group Learning in Undergraduate Science Classrooms: A Perspective from Student Engagement. *CBE—Life Sciences Education*, *17*(2), es3. https://doi.org/10.1187/cbe.17-11-0239

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education. Promise and Implications for Teaching and Learning.*

Humphrey, N., Lendrum, A., Wigelsworth, M., & Kalambouka, A. (2009). Implementation of primary Social and Emotional Aspects of Learning small group work: A qualitative study. *Pastoral Care in Education*, *27*(3), 219–239. https://doi.org/10.1080/02643940903136808

Ibarra-Vazquez, G., Ramírez-Montoya, M. S., Buenestado-Fernández, M., & Olague, G. (2023). Predicting open education competency level: A machine learning approach. *Heliyon*, *9*(11), e20597. https://doi.org/10.1016/j.heliyon.2023.e20597

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Cooperative learning: Increasing college faculty instructional productivity*. School of Education and Human Development, George Washington University.

Kalaian, S., & Kasim, R. (2014). A Meta-Analytic Review of Studies of the Effectiveness of Small-Group Learning Methods on Statistics Achievement. *Journal of Statistics Education*, *22*(1), 2. https://doi.org/10.1080/10691898.2014.11889691

Kalaian, S., Kasim, R., & Nims, J. (2018). Effectiveness of Small-Group Learning Pedagogies in Engineering and Technology Education: A Meta-Analysis. *Journal of Technology Education*, *29*(2), 20–35. https://doi.org/10.21061/jte.v29i2.a.2

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students’ performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411–426. https://doi.org/10.1080/08839510490442058

Kurniawati, A. D., Genarsih, T., & Nurhidayati, M. (2023). Motivation to Learn Mathematics on Different Personality Types. *Sainstek : Jurnal Sains Dan Teknologi*, *15*(1), 36. https://doi.org/10.31958/js.v15i1.8622

Lasry, N., Charles, E., & Whittaker, C. (2014). When teacher-centered instructors are assigned to student-centered classrooms. *Physical Review Special Topics - Physics Education Research*, *10*(1), 010116. https://doi.org/10.1103/PhysRevSTPER.10.010116

Li, Q., Cho, H., Cosso, J., & Maeda, Y. (2021). Relations Between Students’ Mathematics Anxiety and Motivation to Learn Mathematics: A Meta-Analysis. *Educational Psychology Review*, *33*(3), 1017–1049. https://doi.org/10.1007/s10648-020-09589-z

Lu, D.-N., Le, H.-Q., & Vu, T.-H. (2020). The Factors Affecting Acceptance of E-Learning: A Machine Learning Algorithm Approach. *Education Sciences*, *10*(10), 270. https://doi.org/10.3390/educsci10100270

Luan, H., & Tsai, C.-C. (2021). A Review of Using Machine Learning Approaches for Precision Education. *Educational Technology & Society*, *24*(1), 250–266.

Marinucci, L., Mazzuca, C., & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: State of the art and challenges with a focus on gender. *AI & SOCIETY*, *38*(2), 747–761. https://doi.org/10.1007/s00146-022-01474-3

Mbofana, A., & Banda, S. (2022). The effects of class size on the delivery of quality mathematics learning in secondary schools. *Humanities Southern Africa*, *2*(1). http://ir.gzu.ac.zw:8080/xmlui/handle/123456789/556

Micari, M., Pazos, P., Streitwieser, B., & Light, G. (2010). Small-group learning in undergraduate STEM disciplines: Effect of group type on student achievement. *Educational Research and Evaluation*, *16*(3), 269–286. https://doi.org/10.1080/13803611.2010.520860

Mite-Baidal, K., Delgado-Vera, C., Solís-Avilés, E., Espinoza, A. H., Ortiz-Zambrano, J., & Varela-Tapia, E. (2018). Sentiment Analysis in Education Domain: A Systematic Literature Review. In R. Valencia-García, G. Alcaraz-Mármol, J. Del Cioppo-Morstadt, N. Vera-Lucio, & M. Bucaram-Leverone (Eds.), *Technologies and Innovation* (Vol. 883, pp. 285–297). Springer International Publishing. https://doi.org/10.1007/978-3-030-00940-3\_21

Moradi, S., Faghiharam, B., & Ghasempour, K. (2018). Relationship Between Group Learning and Interpersonal Skills With Emphasis on the Role of Mediating Emotional Intelligence Among High School Students. *SAGE Open*, *8*(2), 215824401878273. https://doi.org/10.1177/2158244018782734

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*, *80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Ofori, F., Maina, E., & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Students’ Performance and Improve Learning Outcome: A Literature Based Review. *Journal of Information and Technology*, *4*(1), Article 1.

Olasen, V. M., & Lawal, D. D. (2020). Experimenting the Effect of Class Size on Mathematics Based Performance: A Case Study of Selected Public Secondary School in Akure, Nigeria. *Higher Education of Social Science*, *18*(2), Article 2. https://doi.org/10.3968/11691

Opitz, J., & Burst, S. (2019). *Macro F1 and Macro F1*. https://doi.org/10.48550/ARXIV.1911.03347

Peklaj, C. (2001). *Sodelovalno učenje ali Kdaj več glav več ve* (1. izd., 1. natis). DZS.

Pratt, D. D. (2002). Good Teaching: One Size Fits All? *New Directions for Adult and Continuing Education*, *2002*(93), 5–16. https://doi.org/10.1002/ace.45

Puklek, M. (2001). Skupinsko delo: Kako ga oceniti? *Didakta*, *11*(60/61), 47–51.

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24*(6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. https://doi.org/10.48550/ARXIV.1811.12808

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students’ Performance by Means of Machine Learning: A Review. *Applied Sciences*, *10*(3), 1042. https://doi.org/10.3390/app10031042

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

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear Estimation and Classification* (Vol. 171, pp. 149–171). Springer New York. https://doi.org/10.1007/978-0-387-21579-2\_9

Scribner, J. P., & Donaldson, J. F. (2001). The Dynamics of Group Learning in a Cohort: From Nonlearning to Transformative Learning. *Educational Administration Quarterly*, *37*(5), 605–636. https://doi.org/10.1177/00131610121969442

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 7–11. https://doi.org/10.1145/3318396.3318419

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. https://doi.org/10.1016/j.nlp.2022.100003

Siemens, G., & Gasevic, D. (2012). Guest Editorial—Learning and Knowledge Analytics. *Educational Technology and Society*, *15*(1–2).

Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative Technology and Exploring Engineering*, *2*(6), 250–253.

Slavin, R. E., Hurley, E. A., & Chamberlain, A. (2003). Cooperative Learning and Achievement: Theory and Research. In I. B. Weiner (Ed.), *Handbook of Psychology* (1st ed., pp. 177–198). Wiley. https://doi.org/10.1002/0471264385.wei0709

Starke, G., De Clercq, E., Borgwardt, S., & Elger, B. S. (2021). Computing schizophrenia: Ethical challenges for machine learning in psychiatry. *Psychological Medicine*, *51*(15), 2515–2521. https://doi.org/10.1017/S0033291720001683

Stickler, U., & Emke, M. (2011). Tandem Learning in Virtual Spaces: Supporting Non-formal and Informal Learning in Adults. In P. Benson & H. Reinders (Eds.), *Beyond the Language Classroom* (pp. 146–160). Palgrave Macmillan UK. https://doi.org/10.1057/9780230306790\_12

Stimpson, A. J., & Cummings, M. L. (2014). Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. *IEEE Access*, *2*, 78–87. https://doi.org/10.1109/ACCESS.2014.2303071

Sun, Y., Zhang, Z., Yang, Z., & Li, D. (2019). Application of Logistic Regression with Fixed Memory Step Gradient Descent Method in Multi-Class Classification Problem. *2019 6th International Conference on Systems and Informatics (ICSAI)*, 516–521. https://doi.org/10.1109/ICSAI48974.2019.9010220

Suthaharan, S. (2016). Support Vector Machine. In S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification* (Vol. 36, pp. 207–235). Springer US. https://doi.org/10.1007/978-1-4899-7641-3\_9

Tan, S., & Pu, Y. (2023). Frac-Vector: Better Category Representation. *Fractal and Fractional*, *7*(2), 132. https://doi.org/10.3390/fractalfract7020132

Tomić, A. (2002). *Spremljanje pouka* (1. natis). Zavod Republike Slovenije za šolstvo.

Tsai, S.-C., Chen, C.-H., Shiao, Y.-T., Ciou, J.-S., & Wu, T.-N. (2020). Precision education with statistical learning and deep learning: A case study in Taiwan. *International Journal of Educational Technology in Higher Education*, *17*(1), 12. https://doi.org/10.1186/s41239-020-00186-2

Van Der Laan Smith, J., & Spindle, R. M. (2007). The impact of group formation in a cooperative learning environment. *Journal of Accounting Education*, *25*(4), 153–167. https://doi.org/10.1016/j.jaccedu.2007.09.002

van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, *9*(nov), 2579–2605.

Van Diggele, C., Burgess, A., & Mellis, C. (2020). Planning, preparing and structuring a small group teaching session. *BMC Medical Education*, *20*(S2), 462. https://doi.org/10.1186/s12909-020-02281-4

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023). When adaptive learning is effective learning: Comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, *31*(2), 793–803. https://doi.org/10.1080/10494820.2020.1808794

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Wieselmann, J., Dare, E., Ring‐Whalen, E., & Roehrig, G. (2020). “I just do what the boys tell me”: Exploring small group student interactions in an integrated STEM unit. *Journal of Research in Science Teaching*, *57*(1), 112–144. https://doi.org/10.1002/tea.21587

Wilson, G., & Blednick, J. (2011). *Teaching in tandem: Effective co-teaching in the inclusive classroom*. ASCD.

Wilson, S., & Varma-Nelson, P. (2016). Small Groups, Significant Impact: A Review of Peer-Led Team Learning Research with Implications for STEM Education Researchers and Faculty. *Journal of Chemical Education*, *93*(10), 1686–1702. https://doi.org/10.1021/acs.jchemed.5b00862

Wu, H., Liu, Y., & Wang, J. (2020). Review of Text Classification Methods on Deep Learning. *Computers, Materials & Continua*, *63*(3), 1309–1321. https://doi.org/10.32604/cmc.2020.010172

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear Discriminant Analysis. In P. Xanthopoulos, P. M. Pardalos, & T. B. Trafalis, *Robust Data Mining* (pp. 27–33). Springer New York. https://doi.org/10.1007/978-1-4419-9878-1\_4

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

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

Yang, S. (2021). Guest Editorial: Precision Education - A New Challenge for AI in Education. *Educational Technology and Society*, *24*(1), 105–108.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218–218. https://doi.org/10.21037/atm.2016.03.37