

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Psyridou, Maria; Prezja, Fabi; Torppa, Minna; Lerkkanen, Marja-Kristiina; Poikkeus, Anna-Maija; Vasalampi, Kati

Title: Machine learning predicts upper secondary education dropout as early as the end of primary school

Year: 2024

Version: Published version

Copyright: © The Author(s) 2024

Rights: CC BY 4.0

Rights url: <https://creativecommons.org/licenses/by/4.0/>

Please cite the original version:

Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M.-K., Poikkeus, A.-M., & Vasalampi, K. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports*, 14, Article 12956. <https://doi.org/10.1038/s41598-024-63629-0>



OPEN

Machine learning predicts upper secondary education dropout as early as the end of primary school

Maria Psyridou^{1✉}, Fabi Prezja², Minna Torppa³, Marja-Kristiina Lerkkanen³, Anna-Maija Poikkeus³ & Kati Vasalampi⁴

Education plays a pivotal role in alleviating poverty, driving economic growth, and empowering individuals, thereby significantly influencing societal and personal development. However, the persistent issue of school dropout poses a significant challenge, with its effects extending beyond the individual. While previous research has employed machine learning for dropout classification, these studies often suffer from a short-term focus, relying on data collected only a few years into the study period. This study expanded the modeling horizon by utilizing a 13-year longitudinal dataset, encompassing data from kindergarten to Grade 9. Our methodology incorporated a comprehensive range of parameters, including students' academic and cognitive skills, motivation, behavior, well-being, and officially recorded dropout data. The machine learning models developed in this study demonstrated notable classification ability, achieving a mean area under the curve (AUC) of 0.61 with data up to Grade 6 and an improved AUC of 0.65 with data up to Grade 9. Further data collection and independent correlational and causal analyses are crucial. In future iterations, such models may have the potential to proactively support educators' processes and existing protocols for identifying at-risk students, thereby potentially aiding in the reinvention of student retention and success strategies and ultimately contributing to improved educational outcomes.

Keywords Machine learning, Education dropout, Longitudinal data, Upper secondary education, Comprehensive education, Kindergarten, Academic outcomes

Education is often heralded as the key to poverty reduction, economic prosperity, and individual empowerment, and it plays a pivotal role in shaping societies and fostering individual growth^{1–3}. However, the specter of school dropout casts a long shadow, with repercussions extending far beyond the individual. Dropping out of school is not only a personal tragedy but also a societal concern; it leads to a lifetime of missed opportunities and reduced potential alongside broader social consequences, including increased poverty rates and reliance on public assistance. Existing literature has underscored the link between school dropout and diminished wages, unskilled labor market entry, criminal convictions, and early adulthood challenges, such as substance use and mental health problems^{4–7}. The socioeconomic impacts, which range from reduced tax collections and heightened welfare costs to elevated healthcare and crime expenditures, signal the urgency of addressing this critical issue⁸. Therefore, understanding and preventing school dropout is crucial for both individual and societal advancement.

Beyond its economic impact, education differentiates individuals within the labor market and serves as a vehicle for social inclusion. Students' abandonment of the pursuit of knowledge translates into social costs for society and profound personal losses. Dropping out during upper secondary education disrupts the transition to adulthood, impedes career integration, and compromises societal well-being⁹. The strong link between educational attainment and adult social status observed in Finland and globally¹⁰ underscores the importance of upper secondary education as a gateway to higher education and the labor market.

An increase in school drop-out rates in many European countries¹¹ is leading to growing pockets of marginalized young people. In the European Union (EU), 9.6% of individuals between 18 and 24 years of age did not

¹Department of Psychology, University of Jyväskylä, 40014 Jyväskylä, Finland. ²Faculty of Information Technology, University of Jyväskylä, 40014 Jyväskylä, Finland. ³Department of Teacher Education, University of Jyväskylä, 40014 Jyväskylä, Finland. ⁴Department of Education, University of Jyväskylä, 40014 Jyväskylä, Finland. ✉email: maria.m.psyridou@ju.fi

engage in education or training beyond the completion of lower secondary education¹². This disconcerting statistic raises alarms about the challenge of preventing early exits from the educational journey. Finnish statistics¹³ highlight that 0.5% of Finnish students drop out during lower secondary school, but this figure is considerably higher at the upper secondary level, with dropout rates of 13.3% in vocational school and 3.6% in general upper secondary school. Amid this landscape, there is a clear and pressing need to not only support out-of-school youths and dropouts but also identify potential dropouts early on and prevent their potential disengagement. In view of the far-reaching consequences of school dropout for individuals and societies, social policy initiatives have rightly prioritized preventive interventions.

Machine learning has emerged as a transformative technology across numerous domains, particularly promising for its capabilities in utilizing large datasets and leveraging non-linear relationships. Within machine learning, deep learning¹⁴ has gained significant traction due to its ability to outperform traditional methods given larger data samples. Deep learning has played a significant role in advancements in fields such as medical computer vision^{15–20} and, more recently, in large foundation models^{21–24}. Although machine learning methods have significantly transformed various disciplines, their application in education remains relatively unexplored^{25,26}.

In education, only a handful of studies have harnessed machine learning to automatically classify between cases of students dropping out from upper secondary education or continuing in education. Previous research in this field has been constrained by short-term approaches. For instance, some studies have focused on collecting and analyzing data within the same academic year^{27,28}. Others have restricted their data collection exclusively to the upper secondary education phase^{29–31}, while one study has expanded its dataset to include data collection of student traits across both lower and upper secondary school years³². Only one previous study has focused on predicting dropout within the next three years following the collection of trait data³³, and another study aimed at predictions within the next 5 years³⁴. However, the process of dropping out of school often begins in early school years and is marked by a gradual disengagement and disassociation from education^{35,36}. These findings suggest that current machine learning models might need to incorporate data that spans further back into the past. In this study we extended this time horizon by leveraging a 13-year longitudinal dataset, utilizing features from kindergarten up to Grade 9 (age 15–16). In this study, we provide the first results for the automatic classification of upper secondary school dropout and non-dropout, using data available as early as the end of primary school.

Given that the process of dropping out of school often begins in early school years and may be influenced by a multitude of factors, our study utilized data from a comprehensive longitudinal study. We aimed to include a broad spectrum of traits that existing literature has shown to have a direct or indirect association with school dropout^{37–39}. From the available variables in the dataset, we incorporated features covering family background (e.g. parental education, socio-economic status), individual factors (e.g. gender, school absences, burn-out), behavioral patterns (e.g. prosocial behaviors, hyperactivity), motivation and engagement metrics (e.g. self-concept, task value, teacher-student relationships), experiences of bullying, health behaviors (e.g. smoking, alcohol use), media usage, and academic and cognitive performance (e.g. reading fluency, arithmetic skills). By incorporating this diverse set of features, we aimed to capture a holistic view of the students' educational journey from kindergarten through the end of lower secondary school.

This study is guided by two main research questions:

1. Can predictive models, developed using a comprehensive longitudinal dataset from kindergarten through Grade 9, accurately classify students' upper secondary dropout and non-dropout status at age 19?
2. How does the performance of machine learning classifiers in predicting school dropout compare when utilizing data up to the end of primary school (Grade 6; age 12–13) versus data up to the end of lower secondary school (Grade 9)? Can model predictions be made as early as Grade 6 without significantly compromising accuracy?

In response to these questions, we hypothesized that a comprehensive longitudinal dataset would facilitate the development of predictive models that could accurately classify dropout and non-dropout status as early as Grade 6. However, we acknowledge that the inherent variability in individual dropout factors may constrain the overall performance of these models. Additionally, we posit that while models trained with data up to Grade 9 are likely to demonstrate higher predictive accuracy than those trained with data only up to Grade 6, accurate model predictions could still be achieved with data up to Grade 6.

Methods

We trained and validated machine learning models, with a 13-year longitudinal dataset, to create classification models for upper secondary school dropout. Four supervised classification algorithms were utilized: Balanced Random Forest (B-RandomForest), Easy Ensemble (Adaboost Ensemble), RSBoost (Adaboost), and the Bagging Decision Tree. Six-fold cross-validation was used for the evaluation of performance. Confusion matrices were calculated for each classifier to evaluate performance. The methodological research workflow is presented in Fig. 1.

Sampling

This study used existing longitudinal data from the “First Steps” follow-up study⁴⁰ and its extension, the “School Path: From First Steps to Secondary and Higher Education” study⁴¹. The entire follow-up spanned a 13-year period, from kindergarten to the third (final) year of upper secondary education. In the “First Steps” study, approximately 2,000 children born in 2000 were followed 10 times from kindergarten (age 6–7) to the end of lower secondary school (Grade 9; age 15–16) in four municipalities around Finland (two medium-sized, one large, and one rural). The goal was to examine students' learning, motivation, and problem behavior, including their academic performance, motivation and engagement, social skills, peer relations, and well-being, in different

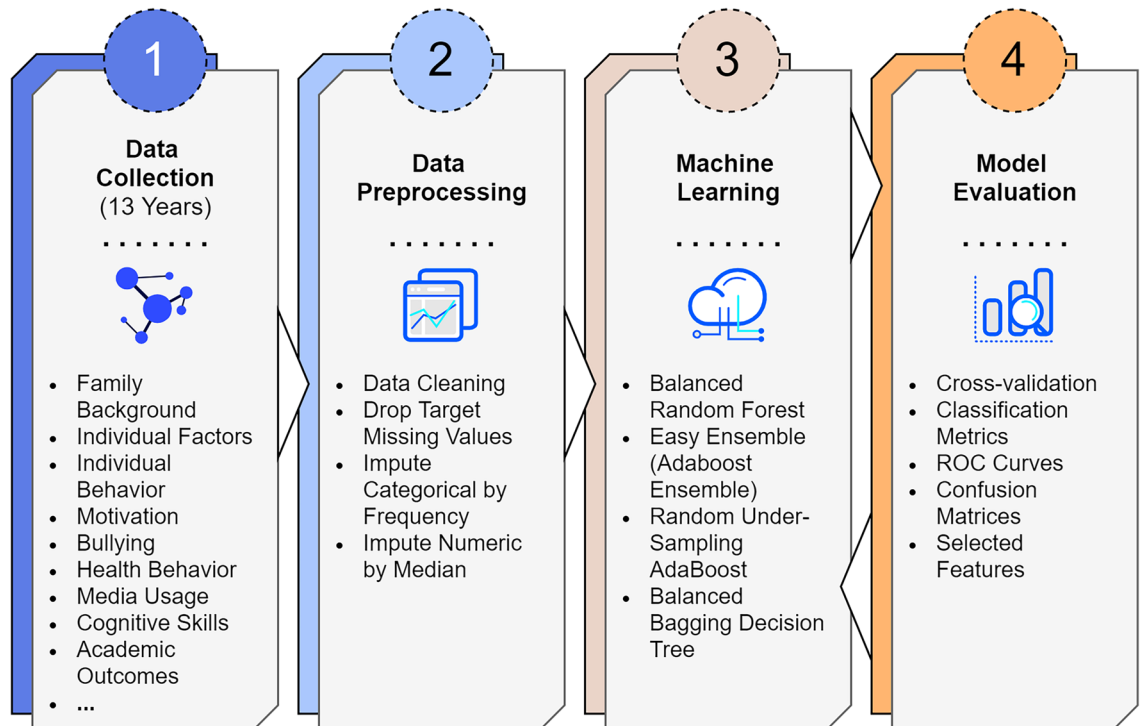


Figure 1. Proposed research workflow. Our process begins with data collection over 13 years, from kindergarten to the end of upper secondary education (Step 1), followed by data processing which includes cleaning and imputing missing feature values (Step 2). We then apply four machine learning models for dropout and non-dropout classification (Step 3), and evaluate these models using 6-fold cross-validation, focusing on performance metrics and ROC curves (Step 4).

interpersonal contexts. The rate at which the contacted parents agreed to participate in the study ranged from 78% to 89% in the towns and municipalities - depending on the town or municipality. Ethnically and culturally, the sample was very homogeneous and representative of the Finnish population, and parental education levels were very close to the national distribution in Finland⁴². In the “School Path” study, the participants of the “First Steps” follow-up study and their new classmates ($N = 4160$) were followed twice after the transition to upper secondary education: in the first year (Grade 10; age 16-17) and in the third year (Grade 12; age 18-19).

The present study focused on those participants who took part in both the “First Steps” study and the “School Path” study. Data from three time points across three phases of the follow-up were used. Data collection for Time 1 (T1) took place in Fall 2006 and Spring 2007, when the participants entered kindergarten (age 6-7). Data collection for Time 2 (T2) took place during comprehensive school (ages 7-16), which extended from the beginning of primary school (Grade 1; age 7-8) in Fall 2007 to the end of the final year of the lower secondary school (Grade 9; age 15-16) in Spring 2016. For Time 3 (T3), data were collected at the end of 2019, 3.5 years after the start of upper secondary education. We focused on students who enrolled in either general upper secondary school (the academic track) or vocational school (the vocational track) following comprehensive school, as these tracks represent the most typical choices available for young individuals in Finland. Common reasons for not completing school within 3.5 years included students deciding to discontinue their education or not fulfilling specific requirements (e.g. failing mandatory courses) during their schooling.

At T1 and T2, questionnaires were administered to the participants in their classrooms during normal school days, and their academic skills were assessed through group-administered tasks. Questionnaires were administered to parents as well. At T3, register information on the completion of upper secondary education was collected from school registers. In Finland, the typical duration of upper secondary education is three years. For the data collection in comprehensive school (T1 and T2), written informed consent was obtained from the participants’ guardians. In the secondary phase (T3), the participants themselves provided written informed consent to confirm their voluntary participation. The ethical statements for the follow-up study were obtained in 2006 and 2018 from the Ethical Committee of the University of Jyväskylä.

Measures

The target variable in the 13-year follow-up was the participant’s status 3.5 years after starting upper secondary education, as determined from the school registers. Participants who had not completed upper secondary education by this time were coded as having dropped out. Initially, we considered the assessment of 586 features. However, as is common in longitudinal studies, missing values were identified in all of them. Features with more than 30% missing data were excluded from the analysis, and a total of 311 features were used (with one-hot encoding) (see Supplementary Table S3). These features covered family background (e.g. parental education,

socio-economic status), individual factors (e.g. gender, absences from school, school burn-out), the individual's behavior (e.g. prosocial behavior, hyperactivity), motivation (e.g. self-concept, task value), engagement (e.g. teacher-student relationships, class engagement), bullying (e.g. bullied, bullying), health behavior (e.g. smoking, alcohol use), media usage (e.g. use of media, phone, internet), cognitive skills (e.g. rapid naming, raven), and academic outcomes (i.e. reading fluency, reading comprehension, PISA scores, arithmetic, and multiplication). Figure 2 presents an overview of the features used while Fig. 3 summarizes the features used in the models, the grades and the corresponding ages for each grade, and the time points (T1, T2, T3) at which different assessments were conducted. The Supplementary Table S3 provides details about the features included.

Data processing

In our study, we employed a systematic approach to address missing values in the dataset. Initially, the percentage of missing data was calculated for each feature, and features exhibiting more than 30% missing values were excluded. For categorical features, imputation was performed using the most frequent value within each feature, while a median-based strategy was applied to numeric features. To ensure unbiased imputation, imputation values were derived from a temporary dataset where the majority class (i.e. non-dropout cases) was randomly sampled to match the size of the positive class (i.e. dropout cases).

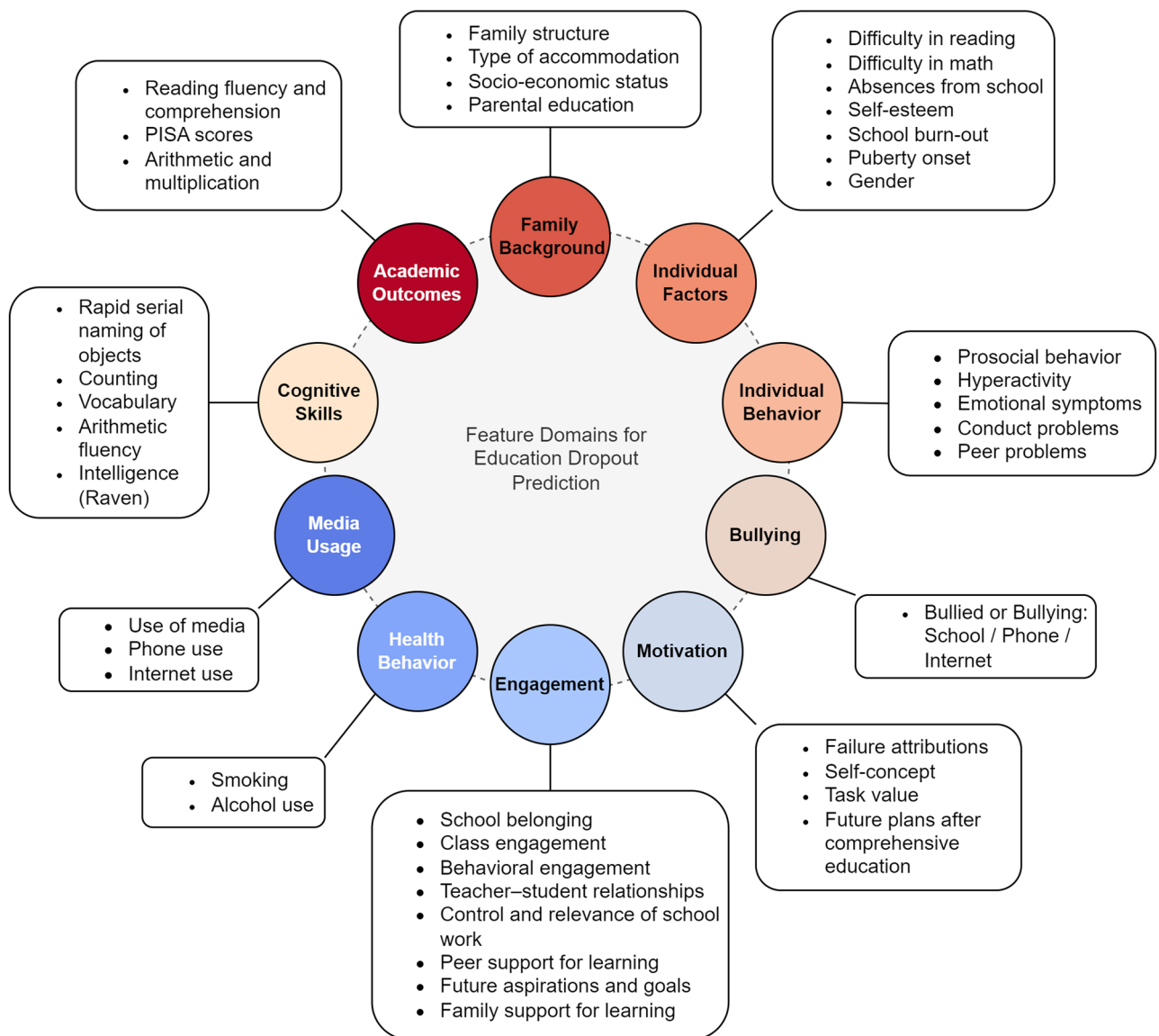


Figure 2. Features domains used for the classification of education dropout and non-dropout. The model incorporated a set of 311 features, categorized into 10 domains: family background, individual factors, behavior, motivation, engagement, bullying experiences, health behavior, media usage, cognitive skills, and academic outcomes. Each domain encompassed a variety of measures.

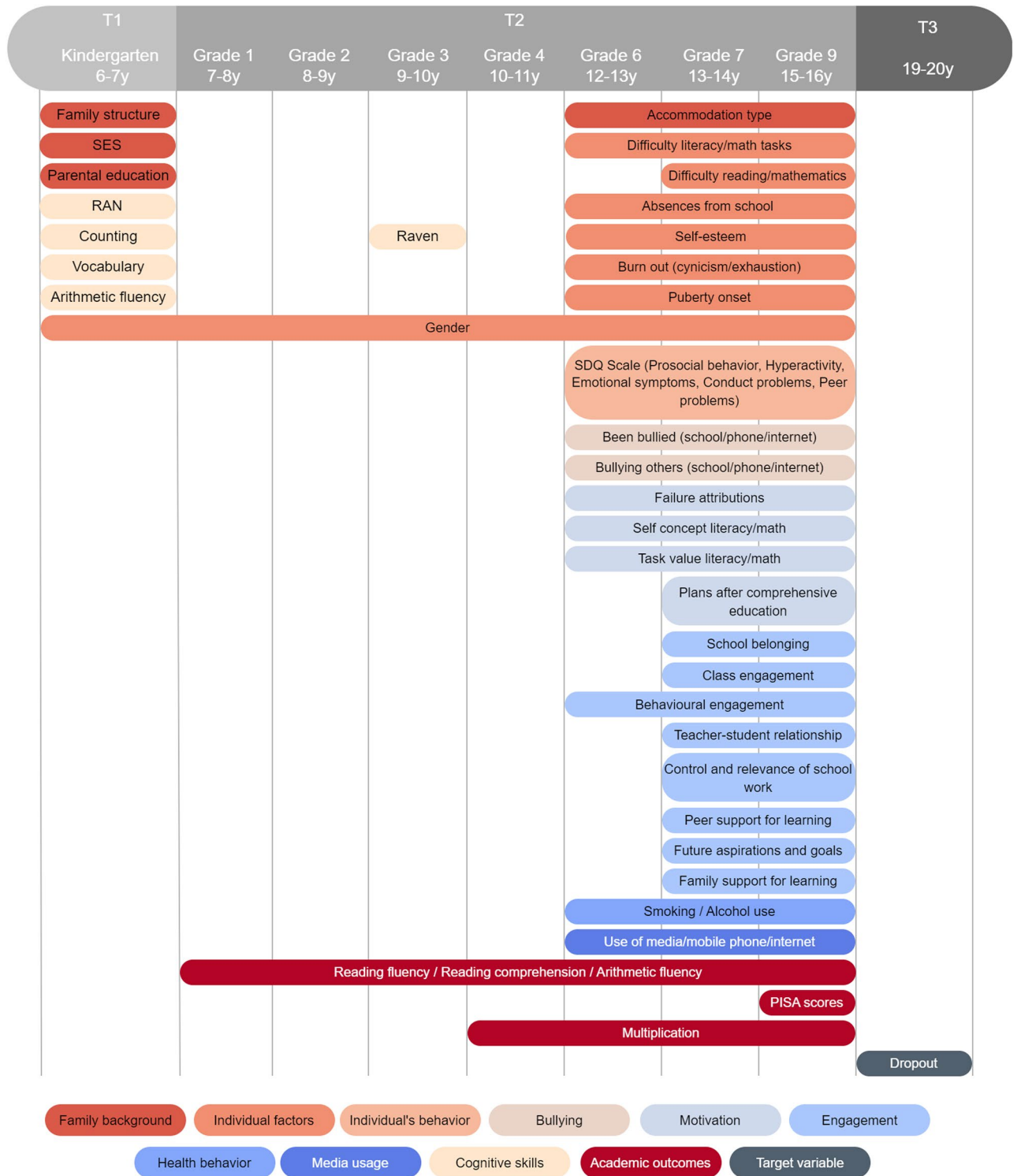


Figure 3. Gantt chart summarizing the features used in the models, the grades and the corresponding ages for each grade, and the time points (T1, T2, T3) at which different assessments were conducted. Assessments from Grades 7 and 9 were not included in the models predicting dropout with data up to Grade 6.

Machine learning

In our study, we utilized a range of balanced classifiers from the Imbalanced Learning Python package⁴³ for benchmarking. These classifiers were employed with their default hyperparameter settings. Our selection included Balanced Random Forest, Easy Ensemble (Adaboost Ensemble), RSBoost (Adaboost), and Bagging Decision Tree. Notably, the Balanced Random Forest classifier played a pivotal role in our study. We delve into

its performance, specific configuration, and effectiveness in the following section. Below are descriptions of each classifier:

1. *Balanced random forest*: This classifier modifies the traditional random forest⁴⁴ approach by randomly under-sampling each bootstrap sample to achieve balance. In our study, we refer to the classifier as “B-Random-Forest”.
2. *Easy ensemble (Adaboost ensemble)*: This classifier, known as EasyEnsemble⁴⁵, is a collection of AdaBoost⁴⁶ learners that are trained on differently balanced bootstrap samples. The balancing is realized through random under-sampling. In our study, we refer to the classifier as “E-Ensemble”.
3. *RSBoost (Adaboost)*: This classifier integrates random under-sampling into the learning process of AdaBoost. It under-samples the sample at each iteration of the boosting algorithm. In our study, we refer to the classifier as “B-Boosting”.
4. *Bagging decision tree*: This classifier operates similarly to the standard Bagging⁴⁷ classifier in the scikit-learn library⁴⁸ using decision trees⁴⁹, but it incorporates an additional step to balance the training set by using a sampler. In our study, we refer to the classifier as “B-Bagging”.

Each of these classifiers was selected for their specific strengths in handling class imbalances, a critical consideration of our study’s methodology. The next section elaborates on the performance and configurations of these classifiers, particularly B-RandomForest.

Random forest

The Random Forest (RF) method, introduced by Breiman in 2001⁴⁴, is a machine learning approach that employs a collection of decision trees for prediction tasks. This method’s strength lies in its ensemble nature, where multiple “weak learners” (individual decision trees) combine to form a “strong learner” (the RF). Typically, decision trees in an RF make binary predictions based on various feature thresholds. The mathematical representation of a single decision tree’s prediction, (T_d) for an input vector \mathbf{I} is given by the following formula:

$$T_d(\mathbf{I}) = \sum_{i=1}^n v_i \delta(f_i(\mathbf{I}) < t_i) \quad (1)$$

Here, n signifies the total nodes in the tree, v_i is the value predicted at the i -th node, $f_i(\mathbf{I})$ is the i -th feature of the input vector \mathbf{I} , t_i stands for the threshold at the i -th node, and δ represents the indicator function.

In an RF, the collective predictions from D individual decision trees are aggregated to form the final output. For regression problems, these outputs are typically averaged, whereas a majority vote (mode) approach is used for classification tasks. The prediction formula for an RF (F_D) on an input vector \mathbf{I} , is as follows:

$$F_D(\mathbf{I}) = \frac{1}{D} \sum_{d=1}^D T_d(\mathbf{I}) \quad (2)$$

In this equation, $T_d(\mathbf{I})$ is the result from the d -th tree for input vector \mathbf{I} , and D is the count of decision trees within the forest. Random Forests are particularly effective for reducing overfitting compared to individual decision trees because they average results across a plethora of trees. In our study, we utilized 100 estimators with default settings from the scikit-learn library⁴⁸.

Figures of merit

To evaluate the efficacy of our classification models, we employed a set of essential evaluative metrics, known as figures of merit.

The accuracy metric reflects the fraction of correct predictions (encompassing both true positive and true negative outcomes) in comparison to the overall number of predictions. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Notably, given the balanced nature of our target data, the accuracy rate in our analysis equated to the definition of balanced accuracy.

Precision, or the positive predictive value, represents the proportion of true positive predictions out of all positive predictions made. The equation to determine precision is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

Recall, which is alternatively called sensitivity, quantifies the percentage of actual positives that were correctly identified. The formula for calculating recall is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

Specificity, also known as the true negative rate, measures the proportion of actual negatives that were correctly identified. The formula for specificity is as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (6)$$

The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is imbalanced. The formula for the F1 Score is as follows:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

In these formulas, TP represents true positives, TN stands for true negatives, FP refers to false positives, and FN denotes false negatives.

The balanced accuracy metric, as referenced by Brodersen et al. in 2010⁵⁰, is a crucial measure in the context of classification tasks, particularly when dealing with imbalanced datasets. This metric is calculated as follows:

$$\text{BalAcc} = \frac{1}{2} \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) \quad (8)$$

Essentially, this equation is an average of the recall computed for each class. The balanced accuracy metric is particularly effective since it accounts for class imbalance by applying balanced sample weights. In situations where the class weights are equal, this metric is directly analogous to the conventional accuracy metric. However, when class weights differ, the metric adjusts accordingly and weights each sample based on the true class prevalence ratio. This adjustment makes the balanced accuracy metric a more robust and reliable measure in scenarios where the class distribution is uneven. In line with this approach, we also employed the macro average of F1 and Precision in our computations.

A confusion matrix is a vital tool for understanding the performance of a classification model. In the context of our study, the performance of each classification model was encapsulated by binary confusion matrices. One matrix was a 2×2 table categorizing the predictions into four distinct outcomes. In the columns of the matrix, the classifications predicted by the model are represented and categorized as Predicted Positive and Predicted Negative. The rows signify the actual classifications, which are labeled as Actual Positive and Actual Negative.

- The upper-left cell is the True Negatives (TN), which are instances where the model correctly predicted the negative class.
- The upper-right cell is the False Positives (FP), which are cases where the model incorrectly predicted the positive class for actual negatives.
- The lower-left cell is the False Negatives (FN), where the model incorrectly predicted the negative class for actual positives.
- Finally, the lower-right cell shows 'True Positives (TP)', where the model correctly predicted the positive class.

In our study, we aggregated the results from all iterations of the cross-validation process to generate normalized average binary confusion matrices. Normalization of the confusion matrix involves converting the raw counts of true positives, false positives, true negatives, and false negatives into proportions, which account for the varying class distributions. This approach allows for a more comparable and intuitive understanding of the model's performance, especially when dealing with imbalanced datasets. By analyzing the normalized matrices, we obtain a comprehensive view of the model's predictive performance across the entire cross-validation run, instead of relying on a single instance.

AUC score

The AUC score is a widely used metric in machine learning for evaluating the performance of binary classification models. Derived from the receiver operating characteristic (ROC) curve, the AUC score quantifies a model's ability to distinguish between two classes. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. By varying the threshold that determines the classification decision, the ROC curve illustrates the trade-off between sensitivity (TPR) and specificity ($1 - \text{FPR}$). The TPR and FPR are defined as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (10)$$

The AUC score represents the area under the ROC curve and ranges from 0 to 1. An AUC score of 0.50 is equivalent to random guessing and indicates that the model has no discriminative ability. On the other hand, a model with an AUC score of 1.0 demonstrates perfect classification. A higher AUC score suggests a better model performance in terms of distinguishing between the positive and negative classes.

Cross-validation

In this study, we employed the stratified K-fold cross-validation method with $K = 6$ to ascertain the robustness and generalizability of our approach⁵¹. This method partitions the dataset into k distinct subsets, or folds with an

even distribution of class labels in each fold to reflect the overall dataset composition. For each iteration of the process, one of these folds is designated as the test set, while the remaining folds collectively form the training set. This cycle is iterated k times, with a different fold used as the test set each time. This technique was crucial in our study to ensure that the model's performance would be consistently evaluated against varied data samples. One formal representation of this process with $K = 6$, is as follows:

$$CV(\mathcal{M}, \mathcal{D}) = \frac{1}{K} \sum_{k=1}^K \text{Eval}(\mathcal{M}, \mathcal{D}_k^{\text{train}}, \mathcal{D}_k^{\text{test}}) \quad (11)$$

Here, \mathcal{M} represents the machine learning model, \mathcal{D} is the dataset, $\mathcal{D}_k^{\text{train}}$ and $\mathcal{D}_k^{\text{test}}$ respectively denote the training and test datasets for the k -th fold, and Eval is the evaluation function (e.g. accuracy, precision, recall). Our AUC plots have been generated using the forthcoming version of utility functions from the Deep Fast Vision Python Library⁵².

Ethics declarations

Ethical approval for the original data collection was obtained from the Ethical Committee of the University of Jyväskylä in 2006 and 2018, ensuring that all experiments were performed in accordance with relevant guidelines and regulations.

Results

This study utilized a comprehensive 13-year longitudinal dataset from kindergarten through upper secondary education. We applied machine learning techniques with data up to Grade 9 (age 15–16), and subsequently with data up to Grade 6 (age 12–13), to classify registered upper secondary education dropout and non-dropout status. The dataset included a broad range of educational data on students' academic and cognitive skills, motivation, behavior, and well-being. Given the imbalance observed in the target, we trained four classifiers: Balanced Random Forest, or B-RandomForest; Easy Ensemble (AdaBoost Ensemble), or E-Ensemble; RSBoost (Adaboost), or B-Boosting; and Bagging Decision Tree, or B-Bagging. The performance of each classifier was evaluated using six-fold cross-validation, as shown in Fig. 4 and Table 1.

Our analysis using data up to Grade 9 (Fig. 4, Table 1), revealed that the B-RandomForest classifier was the most effective, as it achieved the highest balanced mean accuracy (0.61). It also showed a recall rate of 0.60 (i.e. dropout class) and a specificity of 0.62 (i.e. non-dropout class). While the other classifiers matched or exceeded the specificity (B-Bagging: 0.78, E-Ensemble: 0.64, B-Boosting: 0.62), they underperformed in classifying true positives (B-Bagging: 0.32, B-Boosting: 0.50, E-Ensemble: 0.56) and had higher false negative rates (B-Bagging: 0.68, B-Boosting: 0.48, E-Ensemble: 0.45). The B-RandomForest classifier demonstrated a mean area under the curve (AUC) of 0.65, which indicated good discriminative ability (Fig. 5).

We further obtained the feature scores for the B-RandomForest models across the six-fold cross-validation (Fig. 6; for the full list, refer to Supplementary Table S1). The top 20 rankings of the features (averaged across

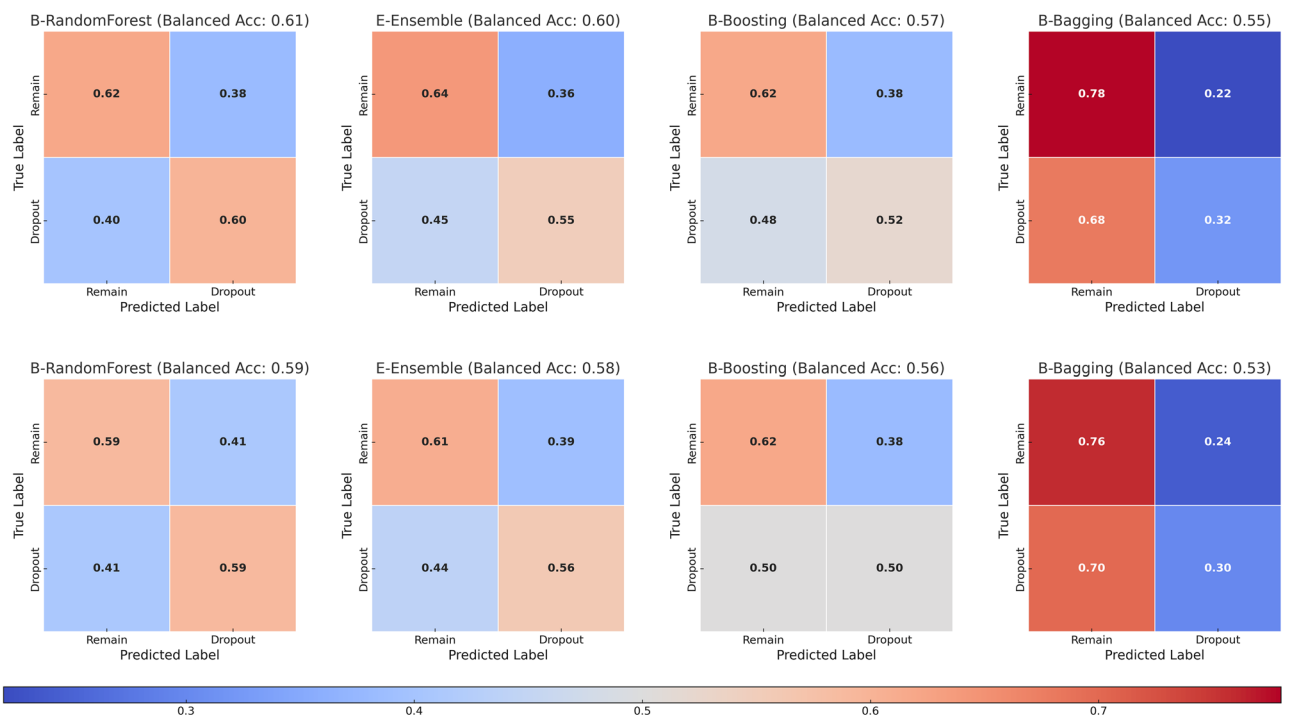


Figure 4. Confusion matrices for classifiers using data up to Grade 9 (first row) and up to Grade 6 (second row) averaged across all folds in six-fold cross-validation.

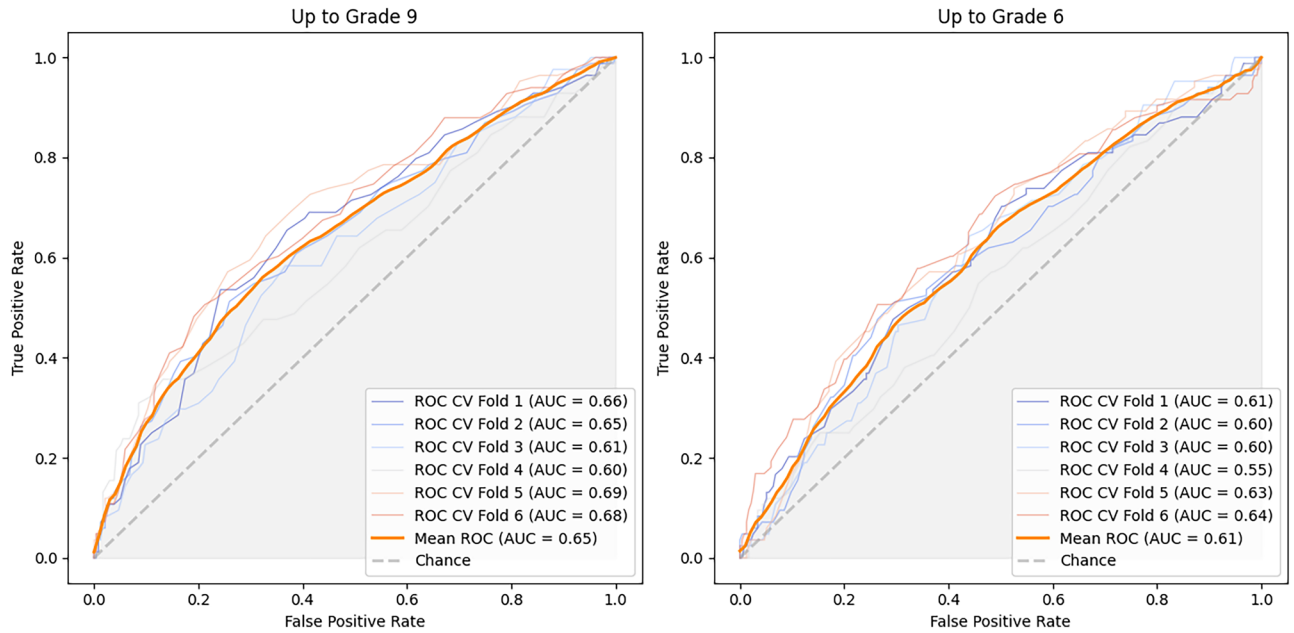


Figure 5. The ROC Curves for the B-RandomForest classifiers from cross-validation. Left: Curve for the B-RandomForest classifier trained using data up to Grade 9. Right: Curve for another classifier instance trained using data up to Grade 6.

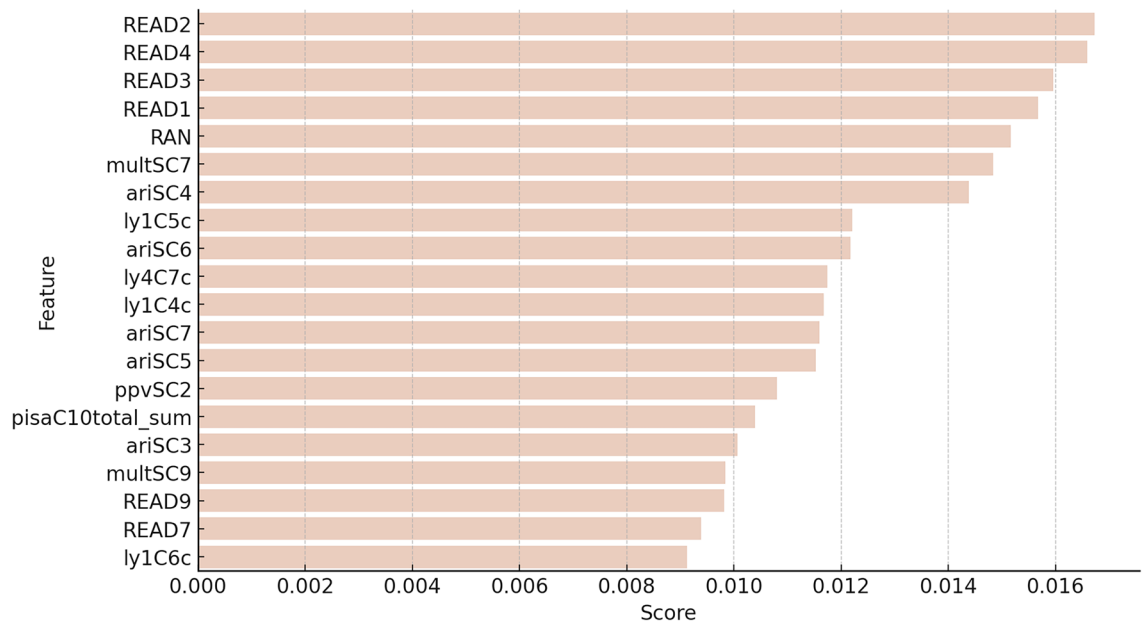


Figure 6. The top ranked 20 features for the B-RandomForest using data up to Grade 9. Features are listed in order of average score from top to bottom. The scores are averages from across all folds of the six-fold cross-validation. The features listed pertain to: READ2=Reading fluency, Grade 2; READ4=Reading fluency, Grade 4; READ3=Reading fluency, Grade 3; READ1=Reading fluency, Grade 1; RAN=Rapid Automatized Naming, Kindergarten; multSC7=Multiplication, Grade 4; ariSC4=Arithmetic, Grade 1 spring; ly1C5C=Reading comprehension, Grade 2; ariSC6=Arithmetic, Grade 3; ly4C7C=Reading comprehension, Grade 4; ly1C4C=Reading comprehension, Grade 1; ariSC7=Arithmetic, Grade 4; ariSC5=Arithmetic, Grade 2; ppvSC2=Vocabulary, Kindergarten; pisaC10total_sum=PISA, Grade 9; ariSC3=Arithmetic, Grade 1 fall; multSC9=Multiplication, Grade 7; READ9=Reading fluency, Grade 9; READ7=Reading fluency, Grade 7; ly1C6C=Reading comprehension, Grade 3.

Classifier	Accuracy	Balanced accuracy	Recall	Precision	F1-score
E-Ensemble	0.616	0.596	0.555	0.580	0.572
B-Boosting	0.596	0.571	0.517	0.550	0.547
B-Bagging	0.660	0.550	0.318	0.541	0.539
B-RandomForest	0.611	0.607	0.599	0.582	0.573

Table 1. Average performance metrics across six-fold cross-validation (data up to Grade 9).

folds) fell into two domains: cognitive skills and academic outcomes. The Supplementary Table S3 provides a detailed description of all features. Academic outcomes appeared as the dominant domain and included reading fluency skills in Grades 1, 2, 3, 4, 7, and 9, reading comprehension in Grade 1, 2, 3, and 4, PISA reading comprehension outcomes, arithmetic skills in Grades 1, 2, 3, and 4, and multiplication skills in Grades 4 and 7. Among the top ranked features were two cognitive skills assessed in kindergarten: rapid automatized naming (RAN) which involved naming a series of visual stimuli consisting of pictures of objects (e.g. a ball, a house) as quickly as possible and vocabulary.

Classifying school dropout using data up to grade 6

Using data from kindergarten up to Grade 6, we retrained the same four classifiers on this condensed dataset and evaluated their performance using a six-fold cross-validation method (Fig. 4, Table 2). The B-RandomForest classifier performed the highest, with a balanced mean accuracy of 0.59. It showed a recall rate of 0.59 (dropout class) and a specificity of 0.59 (non-dropout class). In comparison, the other classifiers had higher specificities (B-Bagging: 0.76, B-Boosting: 0.62, E-Ensemble: 0.61) but lower true positives (recall rates: B-Bagging: 0.30, B-Boosting: 0.50, E-Ensemble: 0.56) and exhibited higher false negative rates (B-Bagging: 0.70, B-Boosting: 0.50, E-Ensemble: 0.44). The B-RandomForest classifier demonstrated an AUC of 0.61 (Fig. 5). The performance of this classifier was slightly lower but comparable to that of the classifier that used the more extensive dataset up to Grade 9.

We obtained the feature scores for the B-RandomForest models across the six-fold cross-validation with data up to Grade 6 (Fig. 7; for the full list, refer to Supplementary S2). The top 20 feature ranks included four domains: cognitive skills, academic outcomes, motivation, and family background. The Supplementary Information contains a detailed description of all features (Table S3). Similarly to the previous models academic outcomes ranked highest, consisting of reading fluency skills in Grades 1, 2, 3, 4, and 6, reading comprehension in Grades 1, 2, 4, and 6, arithmetic skills in Grades 1, 2, 3, and 4, and multiplication skills in Grades 4 and 6. Motivational factors, parental education level and two cognitive skills assessed in kindergarten - RAN and vocabulary - were also included in the ranking.

Discussion

This study signifies a major advancement in educational research, as it provides the first predictive models leveraging data from as early as kindergarten to forecast upper secondary school dropout. By utilizing a comprehensive 13-year longitudinal dataset from kindergarten through upper secondary education, we developed predictive models using the Balanced Random Forest (B-RandomForest) classifier, which effectively predicted both dropout and non-dropout cases from as early as Grade 6.

The classifier's consistency was evident from its performance, which showed only a slight decrease in the AUC from 0.65 with data up to Grade 9 to 0.61 with data limited up to Grade 6. These results are particularly significant since they demonstrate predictive ability. Upon further validation and investigation, and by collecting more data, this approach may assist in the prediction of dropout and non-dropout as early as the end of primary school. However, it is important to note that the deployment and practical application of these findings must be preceded by further data collection, study, and validation. The developed predictive models offered some substantial indicators for future proactive approaches to help educators in their established protocols for identifying and supporting at-risk students. Such an approach could set a new precedent for enhancing student retention and success, potentially leading to transformative changes in educational systems and policies. While our predictive models marked a significant advancement in early automatic identification, it is important to recognize that this study is just the first step in a broader process.

The use of register data was a strength of this study because it allowed us to conceptualize dropout not merely as a singular event but as a comprehensive measure of on-time upper secondary education graduation. This

Classifier	Accuracy	Balanced accuracy	Recall	Precision	F1-score
E-Ensemble	0.593	0.581	0.557	0.564	0.552
B-Boosting	0.587	0.561	0.505	0.549	0.538
B-Bagging	0.639	0.531	0.302	0.532	0.531
B-RandomForest	0.589	0.588	0.587	0.569	0.554

Table 2. Average performance metrics across six-fold cross-validation (data up to Grade 6).

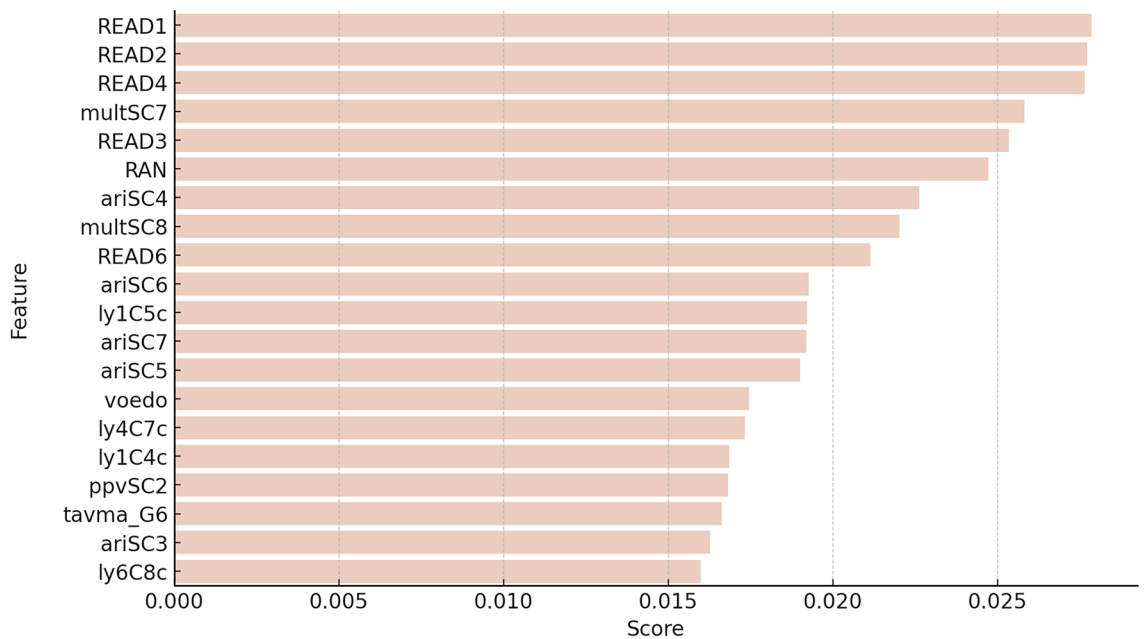


Figure 7. The top ranked 20 features for the B-RandomForest using data up to Grade 6. Features are listed in order of average score from top to bottom. The scores are averages from across all folds of the six-fold cross-validation. READ1=Reading fluency, Grade 1; READ2=Reading fluency, Grade 2; READ4=Reading fluency, Grade 4; multSC7=Multiplication, Grade 4; READ3=Reading fluency, Grade 3; RAN=Rapid Automated Naming, Kindergarten; ariSC4=Arithmetic, Grade 1 spring; multSC8=Multiplication, Grade 6; READ6=Reading fluency, Grade 6; ariSC6=Arithmetic, Grade 3; ly1C5c=Reading comprehension, Grade 2; ariSC7=Arithmetic, Grade 4; ariSC5=Arithmetic, Grade 2; voedo=Parental education; ly4C7c=Reading comprehension, Grade 4; ly1C4c=Reading comprehension, Grade 1; ppvSC2=Vocabulary, Kindergarten; tavma_g6=Task value for math, Grade 6; ariSC3=Arithmetic, Grade 1 fall; ly6C8c=Reading comprehension, Grade 6.

approach is particularly relevant for students who do not graduate by the expected time, as it highlights their high risk of encountering problems in later education and the job market and underscores the need for targeted supplementary support^{37,53}. This conceptualization of dropout offers several advantages⁵³ as it aligns with the nuanced nature of dropout and late graduation dynamics in educational practice. Additionally, it avoids mistakenly applying the dropout category to students who switch between secondary school tracks yet still graduate within the expected timeframe or drop out multiple times but ultimately graduate on time. From the perspective of the school system, delays in graduation incur substantial costs and necessitate intensive educational strategies. This nuanced understanding of dropout and non-dropout underpins the primary objective of our approach: to help empower educators with tools that can assist them in their evaluation of intervention needs.

In our study, we adopted a comprehensive approach to feature collection, acknowledging that the process of dropping out begins in early school years³⁵ and evolves through protracted disengagement and disassociation from education³⁶. With over 300 features covering a wide array of domains - such as family background, individual factors, behavior, motivation, engagement, bullying, health behavior, media usage, cognitive skills, and academic outcomes - our dataset presents a challenge typical of high-dimensional data: the curse of dimensionality. This phenomenon, where the volume of the feature space grows exponentially with the number of features, can lead to sparsity of data and make pattern recognition more complex.

To address these challenges, we employed machine learning classifiers like Random Forest, which are particularly adept at managing high-dimensional data. Random Forest inherently performs a form of feature selection, which is crucial in high-dimensional spaces, by building each tree from a random subset of features. This approach not only helps in addressing the risk of overfitting but also enhances the model's ability to identify intricate patterns in the data. This comprehensive analysis, with the use of machine learning, not only advances the methodology in automatic dropout and non-dropout prediction but also provides educators and policymakers with valuable tools and insights into the multifaceted nature of dropout and non-drop out identification from the perspective of machine learning classifiers.

In our study, the limited size of the positive class, namely the dropout cases, posed a significant challenge due to its impact on classification data balance. This imbalance steered our methodological decisions, leading us to forego both neural network synthesis and conventional oversampling techniques. Instead, we focused on using classification methods designed to handle highly imbalanced datasets. Our strategy was geared towards effectively addressing the issues inherent in working with severely imbalanced classification data.

Another important limitation to acknowledge pertains to the initial dataset and the subsequent handling of missing data. The study initially recruited around 2,000 kindergarten-age children and then invited their classmates to join the study at each subsequent educational stage. While this approach expanded the participant pool,

it also resulted in a significant amount of missing data in many features. To maintain reliability, we excluded features with more than 30% missing values. This aspect of our methodological approach highlights the challenges of managing large-scale longitudinal data. Future studies might explore alternative strategies for handling missing data or investigate ways to include a broader range of features for feature selection, while mitigating the impact of incomplete data and the curse of dimensionality.

Despite these limitations, this study confronts the shortcomings of current research, particularly the focus on short-term horizons. Previous studies that have used machine learning to predict upper secondary education dropout have operated within limited timeframes - by collecting data on student traits and dropout cases within the same academic year^{27,28}, limiting the collection of data on student traits to upper secondary education^{29–31}, and by collecting data on student traits during both lower and upper secondary school years³². Two previous studies have focused on predicting dropout within three years³³ and five years³⁴, respectively, of collecting the data. The present study has extended this horizon by leveraging a 13-year longitudinal dataset, utilizing features from kindergarten, and predicting upper secondary school dropout and non-dropout as early as the end of primary school.

Our study identified a set of top features from Grades 1 to 4 that were highlighted by the Random Forest classifier as influential in predicting school dropout or non-dropout status. These features included aspects like reading fluency, reading comprehension, and arithmetic skills. These top feature rankings did not significantly change with data up to Grades 9 and 6. It is important to note that these features were identified based on their utility in improving the model's predictions within the dataset and cross-validation and should not be interpreted as causal or correlational factors for dropout and non-dropout rates. Given these limitations, and considering known across-time feature correlations^{54–59}, we find it pertinent to suggest further speculative discussions of this ranking consistency between early and later academic grades. If, upon further data collection, validation, and correlational and causal analysis this kind of ranking profile is re-established and validated, it could indicate that early proficiency in these key academic areas could potentially be an important factor influencing students' educational trajectory and dropout risk.

In conclusion, this study represented a significant leap forward in educational research by developing predictive models that automatically distinguished between dropouts and non-dropouts as early as Grade 6. Utilizing a comprehensive 13-year longitudinal dataset, our research enriches existing knowledge of automatic school dropout and non-dropout detection and surpasses the time-frame confines of prior studies. While incorporating data up to Grade 9 enhanced predictive accuracy, the primary aim of our study was to predict potential school dropout status at an early stage. The Balanced Random Forest classifier demonstrated proficiency across educational stages. Although confronted with challenges such as handling missing data and dealing with small positive class sizes, our methodological approach was meticulously designed to address such issues.

The developed predictive models demonstrate potential for further investigation. Given that our study predominantly utilized data from the Finnish educational system, it is not clear how the classifiers would perform with different populations. Additional data, including data from populations from different demographic and educational contexts, and further validation using independent test sets are essential. Further independent correlational and causal analyses are also crucial. In future iterations, such models may have the potential to proactively support educators' processes and existing protocols for identifying at-risk students, thereby potentially aiding in the reinvention of student retention and success strategies, and ultimately contributing to improved educational outcomes.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 27 February 2024; Accepted: 30 May 2024

Published online: 05 June 2024

References

- Huisman, J. & Smits, J. Keeping children in school: Effects of household and context characteristics on school dropout in 363 districts of 30 developing countries. *SAGE Open* 5, 2158244015609666. <https://doi.org/10.1177/2158244015609666> (2015).
- Breton, T. R. Can institutions or education explain world poverty? An augmented Solow model provides some insights. *J. Socio-Econ.* 33, 45–69. <https://doi.org/10.1016/j.socsec.2003.12.004> (2004).
- The World Bank. *The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19* (The World Bank, 2021).
- Bäckman, O. High school dropout, resource attainment, and criminal convictions. *J. Res. Crime Delinq.* 54, 715–749. <https://doi.org/10.1177/0022427817697441> (2017).
- Bjerk, D. Re-examining the impact of dropping out on criminal and labor outcomes in early adulthood. *Econ. Educ. Rev.* 31, 110–122. <https://doi.org/10.1016/j.econedurev.2011.09.003> (2012).
- Campolieti, M., Fang, T. & Gunderson, M. Labour market outcomes and skill acquisition of high-school dropouts. *J. Labor Res.* 31, 39–52. <https://doi.org/10.1007/s12122-009-9074-5> (2010).
- Dragone, D., Migali, G. & Zucchelli, E. High school dropout and the intergenerational transmission of crime. *IZA Discuss. Paper* <https://doi.org/10.2139/ssrn.3794075> (2021).
- Catterall, J. S. The societal benefits and costs of school dropout recovery. *Educ. Res. Int.* 2011, 957303. <https://doi.org/10.1155/2011/957303> (2011).
- Freudentberg, N. & Ruglis, J. Reframing school dropout as a public health issue. *Prev. Chronic Dis.* 4, A107 (2007).
- Kallio, J. M., Kauppinen, T. M. & Erola, J. Cumulative socio-economic disadvantage and secondary education in Finland. *Eur. Sociol. Rev.* 32, 649–661. <https://doi.org/10.1093/esr/jcw021> (2016).
- Gubbels, J., van der Put, C. E. & Assink, M. Risk factors for school absenteeism and dropout: A meta-analytic review. *J. Youth Adolesc.* 48, 1637–1667. <https://doi.org/10.1007/s10964-019-01072-5> (2019).
- EUROSTAT. Early leavers from education and training (2021).
- Official Statistics of Finland (OSF). Discontinuation of education (2022).

14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. <https://doi.org/10.1038/nature21056> (2017).
16. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2) (2019).
17. Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S. & Ojala, T. Synthesizing bidirectional temporal states of knee osteoarthritis radiographs with cycle-consistent generative adversarial neural networks. Preprint at <http://arxiv.org/abs/2311.05798> (2023).
18. Prezja, F., Paloneva, J., Pölonen, I., Niinimäki, E. & Äyrämö, S. DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Sci. Rep.* **12**, 18573. <https://doi.org/10.1038/s41598-022-23081-4> (2022).
19. Prezja, F. *et al.* Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning. Preprint at <http://arxiv.org/abs/2310.16954> (2023).
20. Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56. <https://doi.org/10.1038/s41591-018-0300-7> (2019).
21. Wornow, M. *et al.* The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs. Preprint at <http://arxiv.org/abs/2303.12961> (2023).
22. Peng, Z. *et al.* Kosmos-2: Grounding multimodal large language models to the world. Preprint at <http://arxiv.org/abs/2306.14824> (2023).
23. Livne, M. *et al.* nach0: Multimodal natural and chemical languages foundation model. Preprint at <http://arxiv.org/abs/2311.12410> (2023).
24. Luo, Y. *et al.* Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. Preprint at <http://arxiv.org/abs/2308.09442> (2023).
25. Bernardo, A. B. I. *et al.* Profiling low-proficiency science students in the Philippines using machine learning. *Humanit. Soc. Sci. Commun.* **10**, 192. <https://doi.org/10.1057/s41599-023-01705-y> (2023).
26. Bilal, M., Omar, M., Anwar, W., Bokhari, R. H. & Choi, G. S. The role of demographic and academic features in a student performance prediction. *Sci. Rep.* **12**, 12508. <https://doi.org/10.1038/s41598-022-15880-6> (2022).
27. Krüger, J. G. C., Alceu de Souza, B. J. & Barddal, J. P. An explainable machine learning approach for student dropout prediction. *Expert Syst. Appl.* **233**, 120933. <https://doi.org/10.1016/j.eswa.2023.120933> (2023).
28. Sara, N.-B., Halland, R., Igel, C. & Alstrup, S. High-school dropout prediction using machine learning: A danish large-scale study. In *ESANN*, vol. 2015, 23rd (2015).
29. Chung, J. Y. & Lee, S. Dropout early warning systems for high school students using machine learning. *Child. Youth Serv. Rev.* **96**, 346–353. <https://doi.org/10.1016/j.childyouth.2018.11.030> (2019).
30. Lee, S. & Chung, J. Y. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Appl. Sci.* <https://doi.org/10.3390/app9153093> (2019).
31. Sansone, D. Beyond early warning indicators: High school dropout and machine learning. *Oxf. Bull. Econ. Stat.* **81**, 456–485. <https://doi.org/10.1111/obes.12277> (2019).
32. Aguiar, E. *et al.* Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proc. of the Fifth International Conference on Learning Analytics and Knowledge*, LAK '15, 93–102, <https://doi.org/10.1145/2723576.2723619> (Association for Computing Machinery, New York, NY, USA, 2015).
33. Colak, O. Z. *et al.* School dropout prediction and feature importance exploration in Malawi using household panel data: Machine learning approach. *J. Comput. Soc. Sci.* **6**, 245–287. <https://doi.org/10.1007/s42001-022-00195-3> (2023).
34. Sorensen, L. C. “Big Data” in educational administration: An application for predicting school dropout risk. *Educ. Adm. Q.* **55**, 404–446. <https://doi.org/10.1177/0013161X18799439> (2019).
35. Schoeneberger, J. A. Longitudinal attendance patterns: Developing high school dropouts. *Clear. House J. Educ. Strat. Issues Ideas* **85**, 7–14. <https://doi.org/10.1080/00098655.2011.603766> (2012).
36. Balfanz, R., Herzog, L., Douglas, I. & Mac, J. Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educ. Psychol.* **42**, 223–235. <https://doi.org/10.1080/00461520701621079> (2007).
37. Rumberger, R. W. *Why Students Drop Out of High School and What Can Be Done About It* (Harvard University Press, 2012).
38. De Witte, K., Cabus, S., Thyssen, G., Groot, W. & van Den Brink, H. M. A critical review of the literature on school dropout. *Educ. Res. Rev.* **10**, 13–28 (2013).
39. Esch, P. *et al.* The downward spiral of mental disorders and educational attainment: A systematic review on early school leaving. *BMC Psychiatry* **14**, 1–13 (2014).
40. Lerkkanen, M.-K. *et al.* The first steps study [alkuportaati] (2006–2016).
41. Vasalampi, K. & Aunola, K. The school path: From first steps to secondary and higher education study [koulupolku: Alkuportaita jatko-opintoihin] (2016).
42. Official Statistics of Finland (OSF). Statistical databases (2007).
43. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
44. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
45. Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. B (Cybernetics)* **39**, 539–550 (2008).
46. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
47. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
48. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
50. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, 3121–3124 (IEEE, 2010).
51. Kohavi, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* **14**, 1137–1145 (1995).
52. Prezja, F. Deep fast vision: A python library for accelerated deep transfer learning vision prototyping. Preprint at <http://arxiv.org/abs/2311.06169> (2023).
53. Knowles, J. E. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *J. Educ. Data Min.* **7**, 18–67. <https://doi.org/10.5281/zenodo.3554725> (2015).
54. Aunola, K., Leskinen, E., Lerkkanen, M.-K. & Nurmi, J.-E. Developmental dynamics of math performance from preschool to Grade 2. *J. Educ. Psychol.* **96**, 699–713. <https://doi.org/10.1037/0022-0663.96.4.699> (2004).
55. Ricketts, J., Lervåg, A., Dawson, N., Taylor, L. A. & Hulme, C. Reading and oral vocabulary development in early adolescence. *Sci. Stud. Read.* **24**, 380–396. <https://doi.org/10.1080/10888438.2019.1689244> (2020).
56. Verhoeven, L. & van Leeuwe, J. Prediction of the development of reading comprehension: A longitudinal study. *Appl. Cogn. Psychol.* **22**, 407–423. <https://doi.org/10.1002/acp.1414> (2008).

57. Khanolainen, D. *et al.* Longitudinal effects of the home learning environment and parental difficulties on reading and math development across Grades 1–9. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2020.577981> (2020).
58. Psyridou, M. *et al.* Developmental profiles of arithmetic fluency skills from grades 1 to 9 and their early identification. *Dev. Psychol.* **59**, 2379–2396. <https://doi.org/10.1037/dev0001622> (2023).
59. Psyridou, M. *et al.* Developmental profiles of reading fluency and reading comprehension from grades 1 to 9 and their early identification. *Dev. Psychol.* **57**, 1840–1854. <https://doi.org/10.1037/dev0000976> (2021).

Acknowledgements

The First Steps Study was funded by grants from the Academy of Finland (Grant numbers: 213486, 263891, 268586, 292466, 276239, 284439, and 313768). The School Path study was funded by grants from Academy of Finland (Grant numbers: 299506 and 323773). This research was also partly funded by the Strategic Research Council (SRC) established within the Academy of Finland (Grant numbers: 335625, 335727, 345196, 358490, and 358250 for the project CRITICAL and Grant numbers: 352648, 353392 for the project Right to Belong). In addition, Maria Psyridou was supported by the Academy of Finland (Grant number: 339418).

Author contributions

M.P. conceived the experiment, was involved in data curation, and analysed the results. F.P. was involved in data curation and analysed the results. M.K.L. was involved in data collection. A.M.P. was involved in data collection. M.T. conceived the experiment, was involved in data curation and data collection. K.V. conceived the experiment and was involved in data curation and data collection. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-63629-0>.

Correspondence and requests for materials should be addressed to M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024