PREDICTING STUDENTS' ACHIEVEMENTS USING MACHINE LEARNING METHODS

VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

AUTHORS:

Hande Safiye Erdal
hande.safiye.erdal@vdu.lt

Monika Katilienė
monika.katiliene@card-ai.eu

Tomas Krilavičius
tomas.krilavicius@vdu.lt

Tomas Urbonas
tomas@sonaro.lt

1 GOAL

- The **aim** of this research is develop a predictive model for student achievement using machine learning methods.
- For this, multiple regression models and ensemble methods were evaluated at various time horizons.

2 DATASET

- 11 577 students
- 12 schools
- 6 781 795 records
- 2017–2023 years

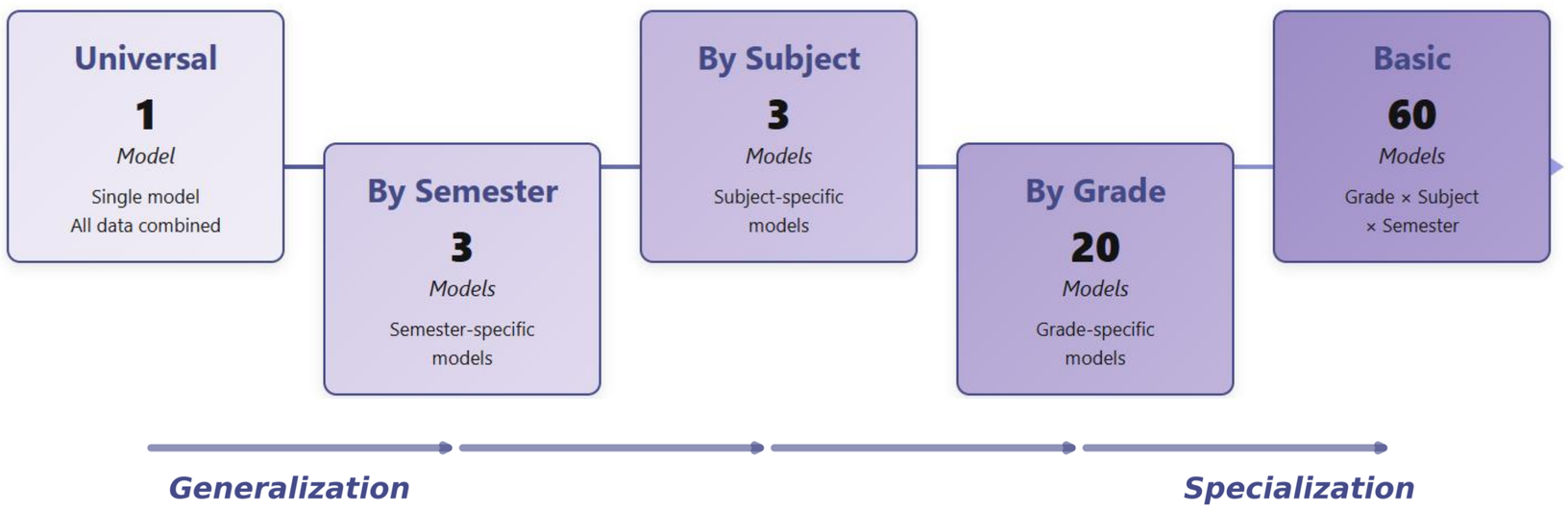
The dataset captures students' grades and attendance across different semesters, classes, and subjects. It also includes information on teachers' professional characteristics and qualifications, as well as school-level attributes.

3 METHODS

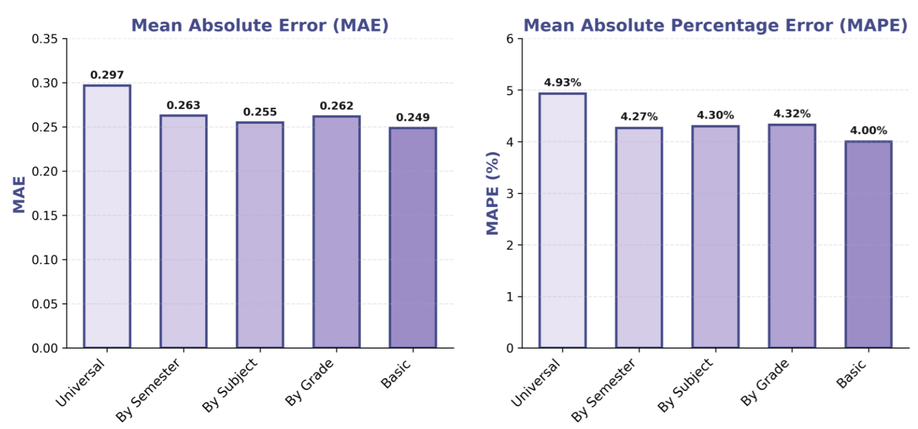
A variety of machine learning models were tested to predict student performance, including tree-based ensembles (Random Forest, Extra Trees, Gradient Boosting, XGBoost, CatBoost, LightGBM, AdaBoost, Bagging), individual Decision Trees, as well as regression and other classical methods (Linear Regression, Polynomial Linear Regression, SVR, KNN). Hyperparameter tuning was performed using Optuna, RandomSearch, and GridSearch. Feature importance was analyzed using PredictionValuesChange, LossFunctionChange, and SHAP values. Model performance was evaluated using RMSE, MAE, MAPE, MSE, and R^2.

4 MODEL STRATEGIES

Five approaches were evaluated: Universal uses a single model for all data. By Semester, By Subject, and By Grade create separate models for each semester, subject, or grade level, respectively. Basic creates individual models for each grade-subject-semester combination.



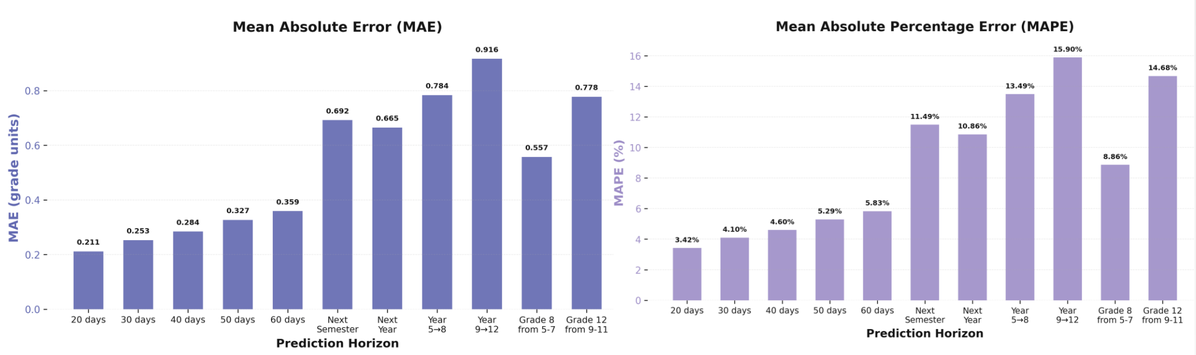
To determine the optimal approach, predictions were made 30 days before semester end. CatBoost consistently outperformed others across all approaches. For non-Universal strategies, results were averaged across individual models. Given minimal performance differences and superior generalization, Universal was selected for further development.



5 FORECASTING PERIOD SELECTION

We performed predictions for various scenarios of student performance:

- **Current semester final grades** (20, 30, 40, 50, 60 days before end of the semester)
- **Next semester final grades**
- **Next academic year final grades** based on previous years
- **Grade 8** from grades 5–7
- **Grade 12** from grades 9–11
- **Grade 8** from grade 5
- **Grade 12** from grade 9



6 CONCLUSIONS

We applied a universal modeling strategy to predict students' final grades across different semesters and academic levels, identifying CatBoost with Optuna hyperparameter tuning as the most effective method. Evaluation of all models demonstrated high accuracy for current semester final grade predictions (MAPE ranged from %3.42 to %5.83). The results show that shorter prediction horizons are associated with better performance. Moreover, model trained on longer observation period (grades 9,10,11 → 12) do not necessarily outperform model trained using only the previous year's data, likely due to the larger training dataset available for this model, which improve overall prediction accuracy.

Co-funded by the European Union

forest 4.0

sonaro

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT