International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, Page 29-35
Received Date: Jul 27, 2025
Accepted Date: Oct 15, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.86968.

Predicting Secondary Student Academic Performance Using Stacked Regression Ensembles on UCI Datasets

Kamal Shrestha[1], Bhagirath Padhya Aryal[2*]

[1]*Department of Electronics and Computer Engineering, Thapathali Engineering Campus, Thapathali, Nepal,*
kamalshrestha17104@gmail.com
[2]*Department of Electronics and Computer Engineering, Thapathali Engineering Campus, Thapathali, Nepal,*
bhagiratharyal2@gmail.com

Abstract

This study explores the use of stacked regression ensembles to predict secondary student academic performance using the UCI Portuguese student dataset. While previous works have focused on individual models such as Random Forest and XGBoost, this research investigates whether combining multiple regressors under a Theil-Sen meta-learner improves prediction accuracy. Among 15 models evaluated through cross-validation, XGBoost achieved the highest individual performance with R^2 of 0.8420 and RMSE of 1.2665. To further improve accuracy, stacking ensembles were created using 2 to 4 base regressors. The best-performing ensemble comprising XGBoost, ExtraTrees, LinearRegression, and Lasso achieved a cross-validated R^2 of 0.8553 and RMSE of 1.2082. These findings show that stacking diverse models offers enhanced predictive power and generalization, providing a robust solution for student performance prediction.

Keywords: Ensemble Learning, Gradient Boosting (XGBoost), Stacking Ensemble, Student Performance Prediction.

1. Introduction

Predicting student academic performance is an important goal in educational data mining. It helps teachers and administrators identify students who may need support and design early interventions. By using machine learning on datasets that include grades, demographic details, school factors, and behavioral records, institutions can make more informed, data-driven decisions to improve student outcomes.

The UCI Student Performance datasets (Cortez & Silva, 2008) contain information about Portuguese and mathematics students and are widely used for this type of research. Earlier studies used models such as Decision Trees, Random Forests, and Support Vector Regression, achieving RMSE values around 1.3 when predicting final grades (G3). These studies showed that combining early-term grades (G1, G2) with demographic and behavioral features can explain much of the variation in students' final performance.

Recent works have introduced more advanced models, including hybrid approaches that merge tree-based methods with attention mechanisms from neural networks. Some of these achieved RMSE values close to 0.83 on the Portuguese dataset (Yuan et al., 2025).

However, most of these studies evaluated each model separately, without exploring how combining multiple models might improve accuracy. Stacked ensemble learning, which combines the outputs of several diverse models through a meta-learner, provides a way to capture the strengths of different algorithms and reduce their individual weaknesses.

In this study, we evaluated fifteen regression algorithms, including tree-based, kernel-based, and linear models. We then developed stacking frameworks using the Theil-Sen Regression model as the meta-learner. Using repeated k-fold cross-validation and hyperparameter tuning, we found that combining top-performing models such as XGBoost, ExtraTrees, Linear Regression, and Lasso increased prediction accuracy. The best ensemble achieved an R^2 of 0.8553 and an RMSE of 1.2082, showing that stacked regression can provide a reliable and accurate approach for predicting student performance.

* *Corresponding author*

2. Literature Review

Cortez and Silva (2008) were the first to apply machine learning methods to the student-por.csv dataset, using Decision Trees, Random Forests, Neural Networks, and Support Vector Regression to predict final grades (G3). Random Forest achieved the lowest RMSE of approximately 1.32, establishing an early benchmark. Their study demonstrated that combining prior grades (G1, G2) with demographic, family, and behavioral factors could explain a substantial portion of variance in academic performance.

Alamri et al. (2020) extended this work by comparing Random Forest and SVM on both classification and regression tasks. In regression experiments, bagged-tree ensembles matched or slightly outperformed kernel-based methods, yielding RMSEs between 1.3 and 1.5, confirming the robustness of tree-based models for continuous-grade prediction.

Apriyadi, Ermatita & Rini (2023) investigated metaheuristic hyperparameter tuning of SVR using Particle Swarm Optimization and Genetic Algorithms. Their optimized SVR achieved an RMSE of 1.608, showing that parameter tuning improves predictive accuracy, though tree-based ensembles remain more effective.

Yu & Liu (2022) illustrated the effectiveness of stacking in classification by combining CART, Random Forest, XGBoost, and LightGBM under a LightGBM meta-learner, achieving 84% accuracy on an online learning dataset. While their work targeted categorical outcomes, it highlights the potential of heterogeneous ensembles, motivating similar approaches for continuous-grade prediction.

Airlangga et al. (2024) compared multiple neural architectures including MLP, CNN, and LSTM with Attention for predicting student test scores. CNNs outperformed other models, emphasizing the relevance of deep learning approaches alongside ensemble methods.

Yuan et al. (2025) introduced the TGEL-Transformer, a theory-guided transformer model incorporating behavioral, cognitive, and environmental factors. The model achieved RMSE = 1.87 and R^2 = 0.75, with attention weights aligned to Bronfenbrenner's ecological systems theory, demonstrating the growing trend of integrating theory and interpretability into predictive models.

Tong & Li (/2025) applied SHAP analysis to a stacking ensemble of six base learners: KNN, Naive Bayes, Random Forest, Gradient Boosting Decision Trees, XGBoost, and MLP, on the XuetangX dataset. Their ensemble achieved 98.53% accuracy, with SHAP revealing that behavioral engagement (video usage, quiz participation) was a stronger predictor than demographic factors, highlighting the value of explainable AI in student performance prediction.

3. Related Theory

This study is grounded in the principles of ensemble learning, particularly stacking, which seeks to improve prediction accuracy by combining multiple diverse regression models. Rather than relying on a single estimator, stacking builds a two-level learning structure. In the first level, multiple base regressors are trained independently, and in the second level, a meta-regressor is trained to learn from their predictions. This structure helps mitigate individual model limitations and captures complementary patterns in the data.

The key base models used in this study include XGBoost, ExtraTrees, Linear Regression, and Lasso. These were selected for their strong individual performance and diversity in learning mechanisms:

- XGBoost is a gradient-boosted decision tree algorithm that excels at capturing complex non-linear relationships through sequential model improvements and regularization.
- ExtraTrees is an ensemble of decision trees where both features and split thresholds are selected randomly, leading to increased variance reduction and speed.
- Linear Regression assumes a direct linear relationship between the input features and the target, offering simplicity and interpretability.
- Lasso Regression extends linear regression by applying L1 regularization, which helps prevent overfitting and automatically selects relevant features.

At the second level, the meta-regressor used was Theil–Sen Estimator, a robust linear model known for its resistance to outliers and reliable slope estimation based on the median of pairwise slopes.

This ensemble design benefits from the bias-variance tradeoff: while tree-based models tend to have low bias and high variance, linear models provide higher bias but lower variance. The stacking mechanism balances these traits, yielding better generalization. To assess model reliability, repeated k-fold cross-validation was employed, offering a stable estimation of generalization error using metrics such as R^2 and RMSE.

4, Methodology

4.1. Dataset Overview

We used the Portuguese subset of the UCI Student Performance dataset, which contains 649 records collected from two Portuguese secondary schools. Each record includes 33 attributes covering demographic information (such as age and sex), family background (e.g., parental education and family support), school-related factors (such as school and address), behavioral data (e.g., absences), and academic history (including prior grades G1 and G2). The target variable is the final grade (G3), which is treated as a continuous numeric value for regression tasks. Although the dataset also contains a mathematics subset with 395 records, we excluded it to maintain a consistent subject context throughout the analysis. The dataset is publicly available via the UCI Machine Learning Repository.

4.2. Dataset Preprocessing

All binary yes/no features (e.g., schoolsup, famsup) were encoded as 0 and 1. Other binary categorical variables such as sex and Pstatus were similarly mapped. Nominal categorical features including school, address, and guardian were transformed using one-hot encoding with the drop_first=True option to avoid multicollinearity. Numerical features such as G1, G2, age, absences, and parental education level were standardized using z-score normalization to ensure consistent scaling across all inputs. For the initial evaluation and hyperparameter tuning phases, the dataset was randomly split into 80% training and 20% test sets. In the final stacking experiments, we used the entire dataset under a repeated cross-validation setting to ensure robust performance estimation.

4.3 Model Training and Evaluation

To evaluate model performance and assess the benefit of stacking, we implemented a structured three-phase experimental pipeline: (1) baseline evaluation with default hyperparameters, (2) hyperparameter tuning using repeated cross-validation, and (3) construction and evaluation of stacking ensembles.

4.3.1. Baseline Evaluation Using Default Hyperparameters

Initially, we performed a simple 80/20 train-test split of the preprocessed Portuguese dataset. Each of the 15 regressors including tree-based, kernel/instance-based, and linear/robust models was trained using default hyperparameters. We then evaluated their performance on the held-out test set using two standard regression metrics: coefficient of determination (R^2) and root mean square error (RMSE). This step provided a reference baseline for how each model performs without optimization.

4.3.2. Hyperparameter Tuning with Repeated Cross-Validation

To improve upon the default performance, we tuned each model using a grid search over predefined hyperparameter spaces. The tuning process was applied only on the 80% training portion from the initial split. For each regressor, we used GridSearchCV with 5-fold cross-validation repeated 5 times (i.e., 25 total evaluations per model) and selected the configuration that yielded the highest average cross-validated R^2. The tuned models , each trained with their optimal parameters, were used for later ensemble construction.

4.3.3. Stacking Ensemble Construction and Evaluation

To leverage the complementary strengths of diverse regression models, we constructed stacked ensembles using Theil–Sen Regression as the meta-learner. This choice was based on empirical testing: we evaluated all 15 models individually as meta-learners in stacking configurations using a fixed set of strong base regressors. Among them, Theil–Sen Regression achieved the best performance in terms of RMSE and R^2, outperforming the others consistently. To preserve diversity and avoid redundancy, Theil–Sen was excluded from the base learner pool.

We systematically generated all possible combinations of 2, 3, and 4 base regressors from the remaining 14 tuned models. For each combination, we created a StackingRegressor in which the base learners were trained with their best hyperparameters, and Theil-Sen served as the final estimator. The meta-learner was trained on out-of-fold predictions produced internally by the stacking mechanism using cross-validation, ensuring unbiased inputs and avoiding data leakage.

Each stacking configuration was evaluated using repeated 5-fold cross-validation (with 5 repetitions) over the full dataset. Final results were computed as the average R² and RMSE across all folds. This exhaustive approach allowed us to identify ensemble combinations that achieved superior predictive accuracy compared to any individual regressor.

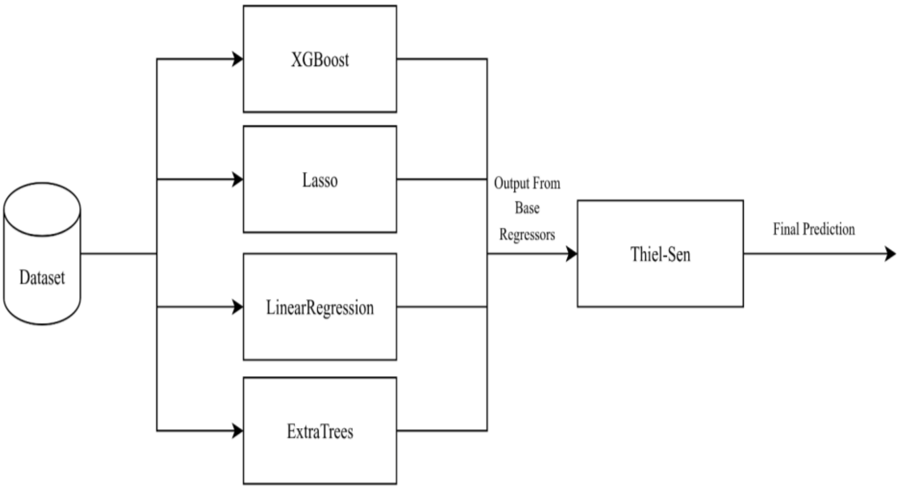

Fig 1. Stacking Ensemble Architecture (4 Base Regressors)

5. Results and Discussions

5.1 Baseline Model Performance

Table 1 presents the train and test-set R² and RMSE for fifteen regression algorithms under default hyperparameters. On the test set, Huber regression achieved the highest R² (0.8707) and lowest RMSE (1.1229), closely trailed by Lasso (R² = 0.8662, RMSE = 1.1423) and Theil–Sen (R² = 0.8638, RMSE = 1.1525). In contrast, tree ensembles (ExtraTrees, XGBoost, Random Forest) exhibited near-perfect training fits (train R² ≈ 1.00) but substantially lower test performance, indicating overfitting: for example, ExtraTrees dropped from train R² = 1.0000 to test R² = 0.8455 (RMSE from 0 to 1.2276). Mid-tier methods (Gradient Boosting, AdaBoost) balanced moderate fit with fair generalization, while K-Nearest Neighbors and SVR registered the weakest test scores (R² < 0.77). These default results establish a baseline for subsequent hyperparameter tuning and ensemble experiments

Table 1. Baseline performance of fifteen regression algorithms with default hyperparameters

Model	Train R²	Test R²	Train RMSE	Test RMSE
Random Forest	0.9782	0.8370	0.4789	1.2608
ExtraTrees	1.0000	0.8455	0.0000	1.2276
XGBoost	1.0000	0.8294	0.0056	1.2899
AdaBoost	0.9033	0.8227	1.0087	1.3148
GradientBoosting	0.9566	0.8152	0.6758	1.3426
SVR	0.7944	0.7657	1.4709	1.5115
KNN	0.7507	0.7378	1.6198	1.5990
LinearRegression	0.8583	0.8487	1.2211	1.2149
Ridge	0.8583	0.8488	1.2212	1.2143
Lasso	0.8428	0.8662	1.2864	1.1423

BayesianRidge	0.8575	0.8498	1.2249	1.2102
Huber	0.8480	0.8707	1.2649	1.1229
PassiveAggressive	0.7990	0.7869	1.4545	1.4416
RANSAC	0.8032	0.8382	1.4392	1.2560
Theil-Sen	0.8550	0.8638	1.2356	1.1525

5.2 Cross-Validated Hyperparameter Tuning

After hyperparameter tuning using repeated five-fold cross-validation, performance improved for most models. XGBoost achieved the highest mean R² (0.8420) and the lowest RMSE (1.2665), confirming its strong generalization ability. Lasso (R² = 0.8376, RMSE = 1.2848) and Gradient Boosting (R² = 0.8361, RMSE = 1.2873) followed closely.

Ensemble methods such as Random Forest (R² = 0.8301, RMSE = 1.3092) and ExtraTrees (R² = 0.8283, RMSE = 1.3190) also performed well. On the other hand, SVR and KNN remained the least effective models.

Table 2. Cross-validated R² and RMSE

Model	R²	RMSE
Random Forest	0.8301	1.3092
ExtraTrees	0.8283	1.3190
XGBoost	0.8420	1.2665
AdaBoost	0.7801	1.4920
GradientBoosting	0.8361	1.2873
SVR	0.7086	1.7365
KNN	0.6235	1.9701
LinearRegression	0.8217	1.3501
Ridge	0.8269	1.3303
Lasso	0.8376	1.2848
BayesianRidge	0.8266	1.3316
Huber	0.8301	1.3165
PassiveAggressive	0.8162	1.3706
RANSAC	0.7818	1.4867
Theil-Sen	0.8268	1.3333

5.3 Stacked Ensemble Performance

The stacking ensemble results demonstrate that combining models with different learning biases such as tree-based ensembles and robust linear regressors consistently improves predictive performance. As more complementary models are added, the ensemble becomes more effective at capturing complex relationships in the data. Among the stacking configurations evaluated, the four-model ensemble showed the highest accuracy, indicating its strong potential for student grade prediction within this study.

Table 3 shows the performance of two-model stacking ensembles. Among the tested combinations, the stack of XGBoost and Huber achieved the best results, with a CV R² of 0.8539 and RMSE of 1.2153. This demonstrates that even a simple ensemble of strong and diverse models can significantly enhance predictive accuracy.

Table 3. Cross-validated performance of stacking ensembles with 2 base regressors

Model	R^2	RMSE
XGBoost , Huber	0.853875	1.215296
XGBoost, Lasso	0.853868	1.214356
XGBoost, PassiveAggressive	0.851505	1.225564
GradientBoosting, Lasso	0.851332	1.224042
XGBoost, ExtraTrees	0.851289	1.225540

Table 4 presents the results of three-model ensembles. The best configuration XGBoost, Linear Regression, and Lasso achieved a CV R^2 of 0.8547 and RMSE of 1.2110. Adding a third complementary model improved the ensemble's ability to generalize by blending linear and nonlinear perspectives.

Table 4. Cross-validated performance of stacking ensembles with 3 base regressors

Model	R^2	RMSE
XGBoost, LinearRegression, Lasso	0.854703	1.210958
XGBoost, ExtraTrees, Huber	0.854699	1.210727
XGBoost, Lasso, Huber	0.854614	1.211493
XGBoost, Lasso, BayesianRidge	0.854576	1.211418
XGBoost, Ridge, Lasso	0.854572	1.211410

Table 5 reports the performance of four-model ensembles. The combination of XGBoost, ExtraTrees, Linear Regression, and Lasso yielded the highest accuracy overall, with a CV R^2 of 0.8553 and RMSE of 1.2082. This configuration demonstrates the benefit of incorporating both tree-based and linear regressors for optimal predictive performance.

Table 5. Cross-validated performance of stacking ensembles with 4 base regressors

Model	R^2	RMSE
XGBoost, ExtraTrees, LinearRegression, Lasso	0.855255	1.208234
XGBoost, ExtraTrees, BayesianRidge, Huber	0.855189	1.209271
XGBoost, ExtraTrees, LinearRegression, Huber	0.855188	1.209473
XGBoost, ExtraTrees, Ridge, Huber	0.855149	1.209401
XGBoost, ExtraTrees, Ridge, Lasso	0.855034	1.209020

5.4 Discussion on Recent Baselines

While this study primarily compared the proposed stacking ensemble methods against traditional regression models such as Support Vector Regression (SVR), Lasso, and Random Forest, recent advances in machine learning have introduced more powerful baselines that warrant consideration. Notably, gradient boosting frameworks such as LightGBM and CatBoost extend the XGBoost paradigm by improving computational efficiency and providing advanced handling of categorical features. These models have demonstrated competitive accuracy and significantly faster training on large-scale structured datasets.

Additionally, deep learning architectures, including multilayer perceptrons and transformer-based models like the TGEL-Transformer, have gained traction in educational data mining, offering strong predictive capabilities by incorporating attention mechanisms and educational theories. For example, TGEL-Transformer achieved RMSE of 1.87 and R^2 of 0.683 on a large-scale student dataset, highlighting its potential in educational contexts.

However, these advanced models generally require substantially larger datasets (on the order of thousands of samples) and involve trade-offs in interpretability and computational complexity. In contrast, the UCI Portuguese dataset used in this study comprises 649 samples, a size wherein traditional tree-based and ensemble methods remain highly effective. The stacking framework presented here balances strong predictive performance (achieving an R^2 of 0.8553 and RMSE of 1.2082) with interpretability essential for educational stakeholders. By combining heterogeneous base learners under a regression-based meta-learner, the model maintains transparency while capturing complex patterns within smaller tabular educational data. Thus, while more recent ensemble and deep learning methods offer promising directions, the proposed approach is better suited for medium-scale educational datasets where interpretability and computational feasibility are critical.

This positioning also aligns with recent empirical findings that tree-based gradient boosting models outperform deep learning alternatives on medium-sized tabular datasets, reinforcing the choice of stacking ensembles for practical student performance prediction tasks.

Conclusion

The proposed stacking framework, integrating XGBoost, ExtraTrees, Linear Regression, and Lasso, achieved strong predictive performance ($R^2 = 0.8553$, RMSE = 1.2082) on the UCI Portuguese student dataset. However, the generalizability of these results is constrained by the dataset's limited size (n = 649), narrow geographic scope restricted to two Portuguese schools, and its dated nature, as the data were collected in 2008. These factors may affect the transferability of findings to other subjects, educational systems, or demographic contexts. Nevertheless, the methodological pipeline comprising preprocessing, hyperparameter optimization, and stacking remains broadly applicable. Future research should validate this framework on larger, more diverse, and contemporary datasets, such as the UCI Mathematics subset or multi-institutional records, to assess robustness and ensure wider applicability.

Acknowledgements

We would like to express our gratitude to the UCI Machine Learning Repository for providing open access to the Student Performance dataset, and to Paulo Cortez and Alice Silva for their valuable contribution in creating the dataset.

References

Cortez, P. and Silva, A., 2008. Using data mining to predict secondary school student performance. *EUROSIS*.2

Alamri, L., Almuslim, R., Alotibi, S., Alkadi, D., Khan, I. and Aslam, N., 2020. Predicting Student Academic Performance using Support Vector Machine and Random Forest. pp.100–107. https://doi.org/10.1145/3446590.3446607.

Apriyadi, M.R., Ermatita & Rini, D.P., 2023. *Student performance prediction modeling with hybrid particle swarm optimization-support vector regression (PSO-SVR) and genetic algorithm-support vector regression (GA-SVR)*. Doctoral Dissertation, Faculty of Engineering, Sriwijaya University, Palembang, Indonesia.

Yu, F. and Liu, X., 2022. Research on Student Performance Prediction Based on Stacking Fusion Model. *Electronics*, 11, p.3166. https://doi.org/10.3390/electronics11193166.

Airlangga, G., 2024. Predicting Student Performance Using Deep Learning Models: A Comparative Study of MLP, CNN, BiLSTM, and LSTM with Attention. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4, pp.1561–1567. https://doi.org/10.57152/malcom.v4i4.1668.

Gong, Y., Wang, F., Zhang, Y. and Geng, J., 2025. TGEL-transformer: Fusing educational theories with deep learning for interpretable student performance prediction. *PLOS One*, 20. https://doi.org/10.1371/journal.pone.0327481.

Tong, T. and Li, Z., 2025. Predicting learning achievement using ensemble learning with result explanation. *PLOS ONE*, 20. https://doi.org/10.1371/journal.pone.0312124.