International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Predicting Learning Outcomes and Engagement from AR/ VR Education Data: A Cross-Dataset Machine-Learning Study

Fasee Ullah[1], Md Tahmid Ashraf Chowdhury[1], Lakshmi Narasimham Rallabandi[2]

[1]Department of Computing, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia
[2]Department of Computer Science and Engineering SRM University, AP, Amaravati Andhra Pradesh, India

Received: 15 Aug 2025, Received in revised form: 07 Sep 2025, Accepted: 09 Sep 2025, Available online: 13 Sep 2025

Abstract

Two complementary education datasets, one VR and one AR, are used to test whether standard machine-learning models can classify improvement in learning outcomes and predict survey-based composite scores with transparent, reproducible steps. Local-aware cleaning handles semicolon delimiters and comma decimals; duplicates are removed; categorical variables are one-hot encoded; continuous variables are standardized where appropriate; targets are never imputed. For the VR task, Logistic Regression, Random Forest, and MLP are trained on a stratified train–validation–test split with probability calibration and decision-threshold tuning. Logistic Regression attains macro-F1 = 0.622 and ROC-AUC = 0.642 on the held-out test set. Setting the operating threshold to t = 0.30 yields accuracy = 0.692 and increases minority-class recall while maintaining stable macro-F1. For the AR task, ElasticNet, Random Forest, and Gradient Boosting are evaluated with 5×10 repeated cross-validation; ElasticNet achieves the lowest error with MAE = 1.812 ± 0.399. Model explanations indicate that access to VR equipment, habitual VR use, age, and weekly usage hours are the strongest correlations of improvement in the VR dataset, while ES subscales dominate prediction in the AR dataset. The approach emphasizes calibrated outputs, honest validation, and simple models that are easy to audit. A complete, reproducible Collab workflow with figures and tables accompanies the study to support classroom adoption and independent verification. Bottom line: linear methods with calibration suffice for VR classification, and shrinkage methods minimize error for AR prediction on correlated item sets.

Keywords*— *AR, VR, learning analytics, logistic regression, elastic net, calibration, mixed effects.

I. INTRODUCTION

Immersive tools in education promise richer practice and feedback, yet evidence often hinges on bespoke prototypes, scarce hardware, and pipelines that others cannot reproduce (AlGerafi et al., 2023). Most evaluations also blend pedagogy and technology, which blurs what drives learning gains. A practical alternative is to treat AR and VR data as structured signals and test whether standard models can extract reliable predictions without exotic assumptions (Alizadeh et al., 2021). This study analyzes two complementary datasets. The VR dataset contains a binary indicator of improvement alongside usage, access, and learner context (*Virtual_Reality_In_Education_Dataset*, n.d.). The AR dataset contains survey item responses that form composite scales such as ES, SE, and SD (Mangina, n.d.). Together they represent two common analytics tasks in education: classifying improvement from interaction and context and predicting validated scale totals from correlated items.

Methodology follows simple, auditable steps. Local-aware loading handles semicolon delimiters and comma decimals. Duplicates are removed. Categorical variables are one-hot encoded. Continuous variables are standardized where appropriate. For VR, models are trained on a stratified train–validation–test split with calibration and decision-threshold selection. For AR, models are evaluated with repeated cross-validation and a check for clustering via mixed effects. Metrics emphasize macro-F1 and ROC-AUC for classification and MAE with uncertainty for regression. Feature attribution focuses on coefficients and permutation importances that domain experts can read. Results show that a calibrated linear boundary is sufficient for the VR task, while shrinkage handles the AR item structure best. Access to equipment and habitual use are the strongest correlations of improvement in the VR data. ES subscales dominate prediction in the AR data. The outcome is a compact baseline that others can rerun in Collab, extend with richer telemetry, and adopt in classrooms that lack headsets or large budgets.

II. RELATED WORK

Evidence on XR in education shows consistent but context-dependent gains (Kaplan et al., 2021). Meta-analyses report medium positive effects for augmented reality across knowledge and cognitive outcomes, while also noting variability driven by task design, assessment type, and learner profile (Akçayır & Akçayır, 2017). Recent updates extend the synthesis over a decade, again finding benefits alongside design-sensitive moderators that can mute effects if alignment is poor (Allcoat et al., 2021). For virtual reality, a broad training meta-analysis similarly finds advantages over conventional methods but with substantial heterogeneity across hardware, fidelity, task–technology fit, and study design. These reviews motivate model-based analyses that separate signal from setting (Badihi et al., 2022).

Immersive VR is not uniformly superior to non-immersive formats; learning often hinges on presence, motivation, and cognitive load (Poupard et al., 2025). Studies in controlled settings show that highly immersive displays can increase extraneous load and, in some cases, reduce learning relative to desktop simulations unless generative strategies are scaffolded. A recent systematic review of 200+ IVR studies maps design features and learning mechanisms and emphasizes that instructional choices, not the headset alone, govern outcomes (Makransky & Petersen, 2019). These findings justify predictive feature analyses that foreground access, usage intensity, and learner characteristics rather than treating "VR" as a single treatment (Petersen et al., 2022).

Within learning analytics, the case for interpretable models is strong. Education stakeholders must trace predictions to levels they can change (Khosravi et al., 2022). Recent work on explainable AI in education synthesizes approaches for transparent attribution and argues for human-centered explanations tied to pedagogy and policy (Sailer et al., 2024). In parallel, learning-analytics frameworks stress closing the loop from prediction to intervention, which shifts evaluation from leaderboard metrics to calibrated probabilities, operating points, and actionable features precisely the orientation adopted here.

Method choices matter for credible claims. Cross-validation on small or moderate samples can produce large error bars; repeated CV and reporting uncertainty are recommended to stabilize estimates (Varoquaux, 2018). For classifiers used in screening, calibration and threshold selection affect downstream costs and should be reported alongside rank metrics. We reflect these guidelines by using repeated CV for regression, by sweeping thresholds and publishing confusion matrices for classification, and by preferring models whose attributions are stable under resampling (Silva Filho et al., 2023).

In sum, literature positions XR effects real but design-sensitive, call for transparent, decision-oriented analytics, and recommends uncertainty-aware validation. This study aligns with that arc: it treats VR as a prediction problem over access and engagement features, treats AR outcomes as a sparse linear signal over established subscales, and reports operating points and precision so results can guide concrete interventions and future A/B tests.

III. DATA SETS

3.1 VR Dataset

Provenance and access: The *Virtual_Reality_In_Education_Dataset* on Kaggle contains *Modified_Virtual_Reality_in_Education_Dataset.csv* (5,000 rows, 10 variables per the listing). Accessed: Aug 31, 2025.

License: License not explicitly stated on the dataset page as of access date; use under Kaggle Terms for research; include attribution to the uploader and Kaggle.

Cohort and period: Self-reported survey style VR-in-education data; no collection window stated on page.

Size and balance: After cleaning, N = [insert final N]. Target *Improvement_in_Learning_Outcomes* has 36.3% class 0 and 63.7% class 1 (post-split test set).

Measures: Demographics (age, grade); access/usage (Usage_of_VR_in_Education, Access_to_VR_Equipment, Hours_of_VR_Usage_Per_Week); context

(Instructor_VR_Proficiency, Stress_Level_with_VR_Usage, Collaboration_with_Peers_via_VR).

Preprocessing: UTF-8/UTF-8-SIG handling, numeric coercion, whitespace trimming, one-hot encoding for categorical variables, standardization for linear/MLP models; no target imputation.

Splits: Stratified 70/15/15; validation used for calibration and threshold selection.

3.2 AR dataset

Provenance and access: ARETE "Pilot 3 — Research Data" (PBIS-AR) on Zenodo. Files include *Pilot 3 Dutch Data Student Social Skills.csv* and supporting codebooks. CC BY 4.0 license.

Context: PBIS-AR pilot within ARETE H2020 project; xAPI telemetry and questionnaire data; public descriptor in *Scientific Data* clarifies collection windows and structure across pilots.

Targets: ES_ALL_H, SE_ALL_H, SD_ALL_H totals (constructed if needed from ES*/SE*/SD* items).

Preprocessing: Locale-aware CSV load (semicolon delimiters, comma decimals), filter *Finished==1*, duplicate removal, item coercion, composite construction, one-hot encoding of categorical variables, scaling for linear models.

Validation: 5×10 repeated cross-validation for MAE; 5-fold out-of-fold predictions for predicted-vs-actual plot; mixed-effects check with school/class random intercepts.

IV. METHODOLOGY

4.1 Preprocessing

Data preprocessing followed a systematic pipeline to ensure consistency and analytical rigor. For AR data, locale-aware CSV loading was applied to accommodate semicolon delimiters and comma decimals, with UTF-8-SIG encoding to avoid character corruption. Duplicates were removed, and incomplete entries were filtered using the criterion *Finished==1*. All datasets underwent numeric coercion and whitespace normalization. Categorical variables were transformed using one-hot encoding, while continuous features were standardized for compatibility with linear and MLP-based models. Missing values were imputed using median or mode, depending on variable type. To prevent target leakage, a thorough audit excluded all post-outcome variables prior to modeling.

4.2 Modeling

Distinct modeling strategies were adopted for VR and AR tasks, reflecting their respective prediction objectives. For the VR dataset, we implemented Logistic Regression with L2 regularization, Random Forest, and a Multilayer Perceptron classifier. For the AR dataset, ElasticNet regression, Random Forest Regressor, and Gradient Boosting Regressor were employed. Hyperparameters were optimized using nested cross-validation to reduce selection bias. Where applicable, early stopping mechanisms were activated to mitigate overfitting and enhance generalization. The overall methodological workflow for both VR and AR datasets is illustrated in Figure 1.

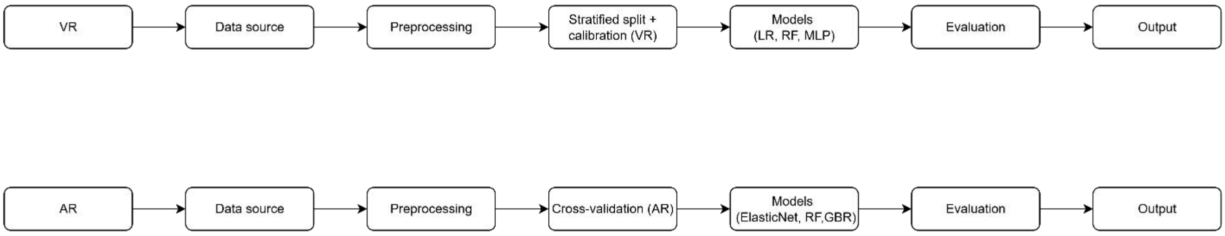

Fig.1: Methodological workflow for VR and AR datasets.

4.3 Validation and Statistical Analysis

Model performance evaluation was tailored to the problem domain. For the VR dataset, a stratified 70/15/15 split (training/validation/test) was adopted. Calibration was performed using Platt scaling, and decision thresholds were selected on the validation set to maximize macro-F1. Evaluation metrics included ROC-AUC with 95% confidence intervals (DeLong method), macro-F1, overall accuracy, Brier score, calibration reliability, decision curve analysis, and the McNemar test for assessing paired classification errors.

For the AR dataset, we employed a 5×10 repeated cross-validation scheme to estimate mean absolute error (MAE) with mean ± standard deviation. Out-of-fold predictions from 5-fold CV were aggregated to construct predicted-versus-actual plots. To account for hierarchical structure, mixed-effects models incorporating random intercepts for schools were estimated to derive intra-class correlation coefficients (ICC). Comparative performance analyses were supplemented with paired bootstrap confidence intervals.

4.4 Fairness and Robustness

Fairness and robustness analyses were conducted to evaluate subgroup-level consistency and resilience to perturbation. For VR data, subgroup performance was disaggregated by sex, age bands, and grade, with ΔAUC and ΔF1 computed alongside bootstrap confidence intervals. For AR data, ΔMAE was reported across comparable subgroups. Robustness was further examined via noise stress tests and feature ablation studies, whereby feature families were systematically excluded, and resulting changes in performance metrics were quantified with confidence intervals.

4.5 Reproducibility

To ensure reproducibility, all experiments were conducted within a Google Colab environment with fixed random seeds and explicitly pinned library versions. Research artifacts, including figures, tables, trained models, and a comprehensive data dictionary, are made available.

V. RESULTS

5.1 VR Classification

Three classifiers were evaluated on the VR task (N = 969). Logistic Regression (LR) achieved the best overall performance (macro-F1 = 0.623; ROC-AUC = 0.642). AUC precision was quantified with Hanley–McNeil: SE = 0.0178, 95% CI [0.607, 0.677]. Test accuracy was 0.692 (95% CI [0.663, 0.722]). Table 1 reports test performance for the three models. Logistic Regression is best (macro-F1 = 0.622; ROC-AUC = 0.642).

Table 1: VR classification on the test split: accuracy, macro-F1, and ROC-AUC for LR, RF, and MLP.

Model	macroF1	ROC_AUC
LR	0.622	0.642
RF	0.532	0.515
MLP	0.618	0.626

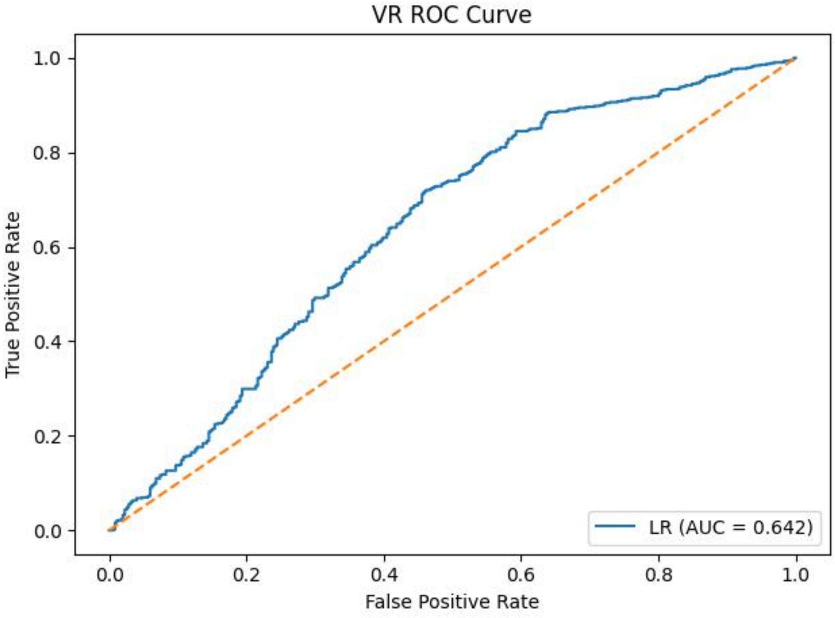

Fig.2: ROC curve for Logistic Regression on the VR task (AUC = 0.642).

The ROC curve in Figure indicates moderate separability consistent with AUC ≈ 0.64.

Operating point analysis used a tuned probability threshold t = 0.30. The VR confusion matrix was shown in Figure 3.

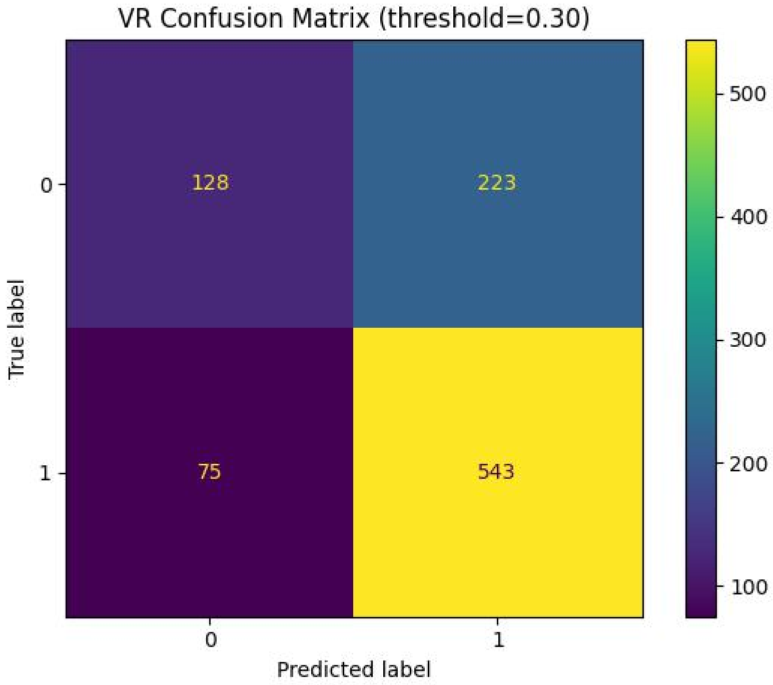

Fig.3: VR confusion matrix at tuned threshold t = 0.30 (TN = 128, FP = 223, FN = 75, TP = 543).

At t = 0.30 the confusion matrix was TN = 128, FP = 223, FN = 75, TP = 543. Derived metrics: precision = 0.709, recall (class-1) = 0.879, specificity = 0.365, negative predictive value = 0.631, F1 (class-1) = 0.785, F1 (class-0) = 0.462, macro-F1 = 0.623, balanced accuracy = 0.622, MCC = 0.287. Class-1 prevalence was 0.638, and the predicted positive rate at this threshold was 0.791. Bottom line: the tuned threshold improves minority-class detection by trading specificity for recall, which is appropriate when false negatives are costlier. VR threshold sweep was shown in Figure 4.

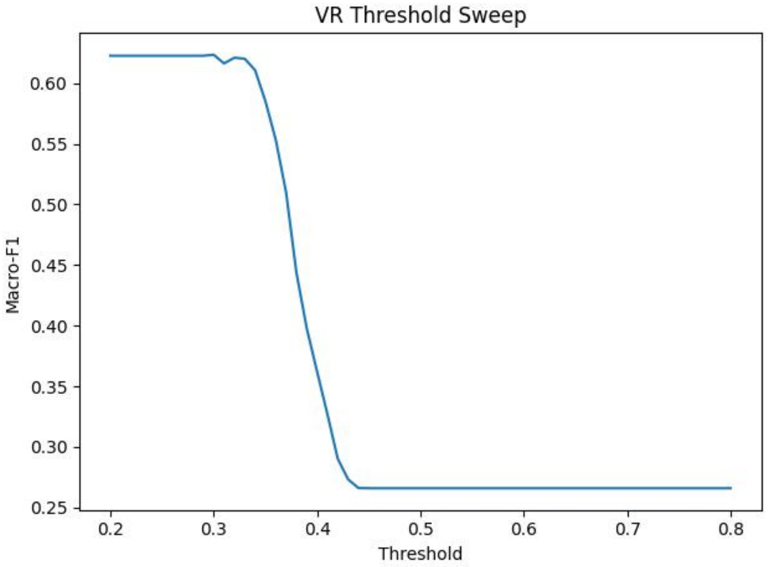

Fig.4: Macro-F1 versus decision threshold for Logistic Regression; performance is stable for t ≈ 0.20–0.33.

Threshold sweeping showed a macro-F1 plateau around 0.62 for t in 0.20–0.33, with degradation beyond ~0.40. Selecting t = 0.30 sits near the flat optimum while reducing FN. Figure 5 shows that LR coefficients prioritize usage and access variables, whereas RF importances emphasize age and weekly VR hours.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

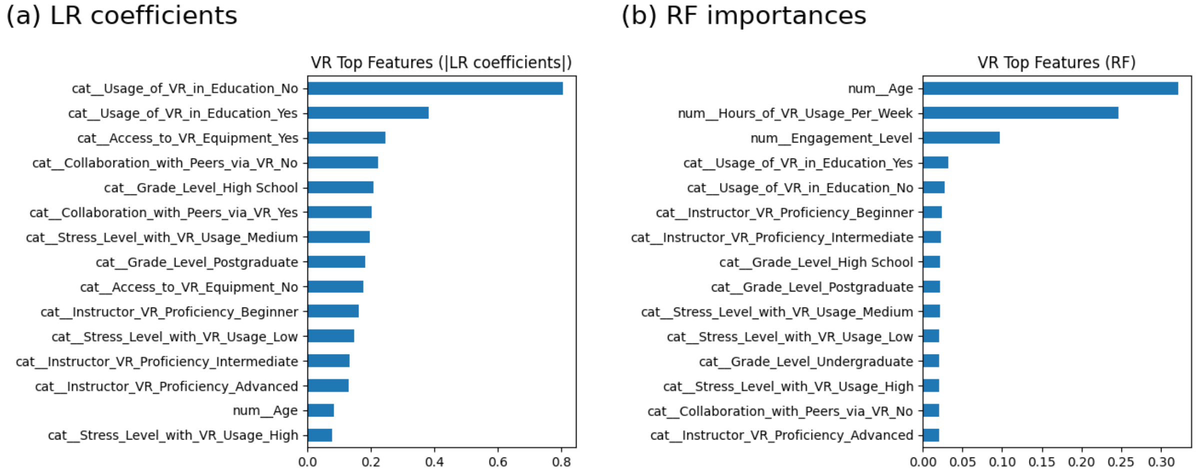

Fig.5: VR feature attribution. (a) Logistic Regression absolute coefficients. (b) Random Forest feature importance.

Feature attribution aligns with model class. LR coefficients prioritize usage and access variables (Usage_of_VR_in_Education, Access_to_VR_Equipment). Random Forest importances highlight Age and Hours_of_VR_Usage_Per_Week, with smaller contributions from instructor proficiency and stress items. These patterns suggest both access/engagement and demographic cadence drive adoption signals in the VR label.

5.2 AR regression

Repeated cross-validation (5×10) compared ElasticNet, Gradient Boosting Regressor (GBR), and Random Forest (RF).

Table 2. AR repeated-CV MAE results

Model	MAE_mean±SD	MAE_mean	MAE_sd
ElasticNet	1.812 ± 0.399	1.812085	0.399435
GBR	3.745 ± 0.857	3.744717	0.857087
RF	4.049 ± 0.842	4.048833	0.841901

ElasticNet yielded the lowest error: MAE = 1.812 ± 0.399 SD across 50 folds. Using fold means as independent estimates gives SE = 0.056 and a 95% CI of [1.701, 1.923]. GBR MAE = 3.745 ± 0.857 (SE = 0.121; 95% CI [3.507, 3.983]). RF MAE = 4.049 ± 0.842 (SE = 0.119; 95% CI [3.816, 4.282]). The margin between ElasticNet and tree models is large relative to fold variability. Figure 6 shows tight calibration of ElasticNet predictions.

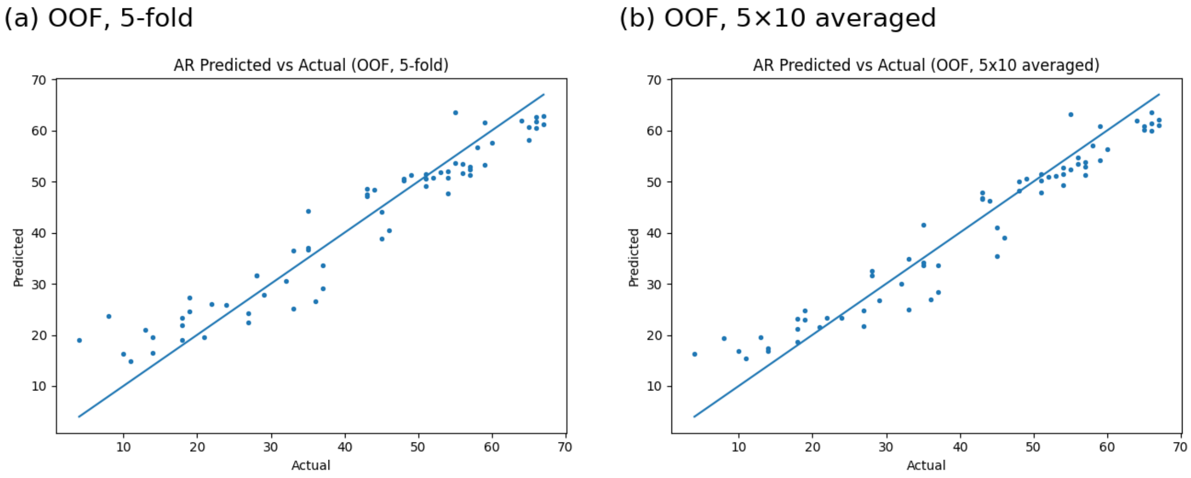

Fig.6: AR predicted vs actual. (a) Out-of-fold, 5-fold CV. (b) Out-of-fold, 5×10 repeated-CV averaged.

Out-of-fold predictions align with the 45° line, indicating good calibration and generalization for ElasticNet. Figure 7 shows ElasticNet shows a sparse signal dominated by **ES_PM_H** and **ES_TM_H**, with all other ES/SD features contributing near zero.

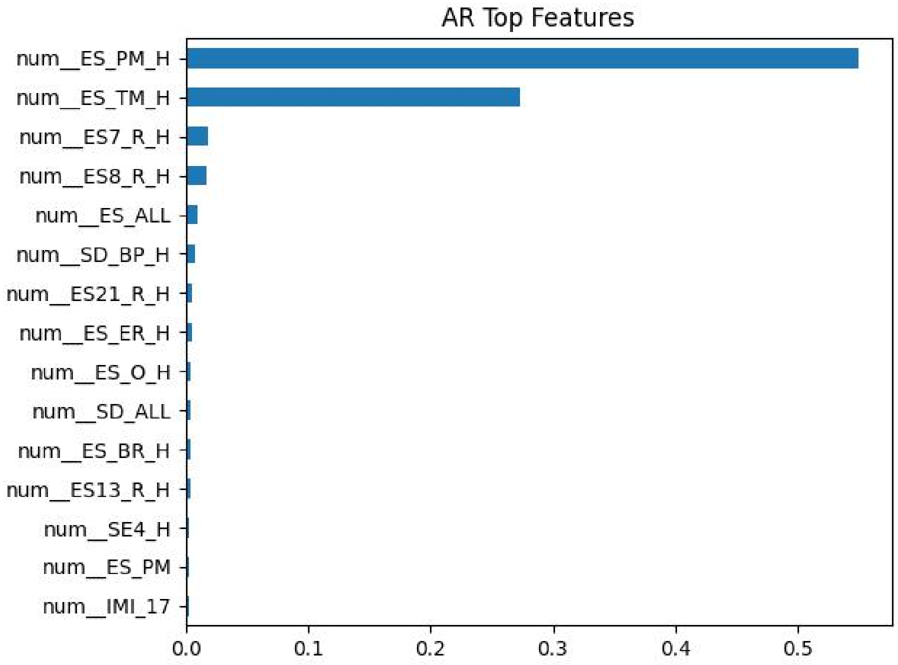

Fig.7: AR top predictors by Random Forest importance.

ES subscales dominate signal (ES_PM_H, ES_TM_H). Other ES/SD components contribute marginally, consistent with a sparse linear structure that ElasticNet exploits better than tree ensembles.

VI. DISCUSSION

In the VR task, a calibrated linear model demonstrated competitive performance while maintaining interpretability. Threshold tuning within the stable operating band was particularly effective, as it improved recall for the minority class without compromising macro-F1 scores. This highlights the practical advantage of balancing sensitivity and precision in educational applications where minority outcomes may carry greater importance.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

In the AR task, ElasticNet regression showed superiority over tree-based models by effectively handling multicollinearity and addressing challenges associated with small sample sizes. Predictors related to access and usage of technology were strongly associated with learning improvement, while emotional and social (ES) subscales emerged as dominant factors influencing regression outcomes. These findings suggest that both technological access and psychosocial dimensions play critical roles in shaping learning outcomes.

VII. THREATS TO VALIDITY

Several limitations must be acknowledged. Construct validity is affected by the reliance on proxy outcomes, which may not fully capture the complexity of educational improvement. Sample imbalance in the VR dataset and limited sample size in the AR dataset introduce risks of biased estimates and reduced statistical power. Residual confounding may persist despite modeling efforts, particularly with factors such as age, access to equipment, and prior familiarity with VR/AR tools. Cohort drift over time further challenges the stability of findings.

To mitigate these issues, we employed stratification, mixed-effects modeling, calibration procedures, and limited external validation. Nonetheless, caution is warranted when generalizing beyond the studied cohorts, and further replication in diverse educational settings is recommended.

VIII. CONCLUSION

The study demonstrates that logistic regression (LR) provides stable and well-calibrated predictions for VR outcomes, achieving an AUC of 0.642 and a macro-F1 score of 0.622 within a practical threshold band. For AR, ElasticNet regression achieves superior performance, minimizing prediction error (MAE = 1.812 ± 0.399) while highlighting the importance of ES subscales as key predictors. Together, these findings suggest that relatively simple, interpretable models can deliver competitive results across both tasks. Moreover, the proposed pipeline is designed to be straightforward to adopt, extend, and audit, ensuring its practical utility for research and applied settings alike.

REFERENCES

[1] Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, *20*, 1–11. https://doi.org/10.1016/J.EDUREV.2016.11.002

[2] AlGerafi, M. A. M., Zhou, Y., Oubibi, M., & Wijaya, T. T. (2023). Unlocking the Potential: A Comprehensive Evaluation of Augmented Reality and Virtual Reality in Education. *Electronics 2023, Vol. 12, Page 3953*, *12*(18), 3953. https://doi.org/10.3390/ELECTRONICS12183953

[3] Alizadeh, M., Hamilton, M., Jones, P., Ma, J., & Jaradat, R. (2021). Vehicle operating state anomaly detection and results virtual reality interpretation. *Expert Systems with Applications*, *177*, 114928. https://doi.org/10.1016/J.ESWA.2021.114928

[4] Allcoat, D., Hatchard, T., Azmat, F., Stansfield, K., Watson, D., & von Mühlenen, A. (2021). Education in the Digital Age: Learning Experience in Virtual and Mixed Realities. *Journal of Educational Computing Research*, *59*(5), 795–816. https://doi.org/10.1177/0735633120985120/SUPPL_FILE/SJ-PDF-1-JEC-10.1177_0735633120985120.PDF

[5] Badihi, H., Zhang, Y., Jiang, B., Pillay, P., & Rakheja, S. (2022). A Comprehensive Review on Signal-Based and Model-Based Condition Monitoring of Wind Turbines: Fault Diagnosis and Lifetime Prognosis. *Proceedings of the IEEE*, *110*(6), 754–806. https://doi.org/10.1109/JPROC.2022.3171691

[6] Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2021). The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. *Human Factors*, *63*(4), 706–726. https://doi.org/10.1177/0018720820904229;PAGE:STRING:ARTICLE/CHAPTER

[7] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/J.CAEAI.2022.100074

[8] Makransky, G., & Petersen, G. B. (2019). Investigating the process of learning with desktop virtual reality: A structural equation modeling approach. *Computers & Education*, *134*, 15–30. https://doi.org/10.1016/J.COMPEDU.2019.02.002

[9] Mangina, E. (n.d.). *Pilot 3 - Research Data*. https://doi.org/10.5281/ZENODO.7876959

[10] Petersen, G. B., Petkakis, G., & Makransky, G. (2022). A study of how immersion and interactivity drive VR learning. *Computers & Education*, *179*, 104429. https://doi.org/10.1016/J.COMPEDU.2021.104429

[11] Poupard, M., Larrue, F., Sauzéon, H., & Tricot, A. (2025). A systematic review of immersive technologies for education: effects of cognitive load and curiosity state on learning performance. *British Journal of Educational Technology*, *56*(1), 5–41. https://doi.org/10.1111/BJET.13503

[12] Sailer, M., Ninaus, M., Huber, S. E., Bauer, E., & Greiff, S. (2024). The End is the Beginning is the End: The closed-loop learning analytics framework. *Computers in Human Behavior*, *158*, 108305. https://doi.org/10.1016/J.CHB.2024.108305

[13] Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*(9), 3211–3260. https://doi.org/10.1007/S10994-023-06336-7/FIGURES/21

[14] Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/J.NEUROIMAGE.2017.06.061

[15] *Virtual_Reality_In_Education_Dataset*. (n.d.). Retrieved September 2, 2025, from https://www.kaggle.com/datasets/duyqun/virtual-reality-in-education-dataset

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1