PREDICT STUDENT PERFORMANCE FROM GAME PLAY

ALEJANDRO PENAGOS - HENRY RICAURTE - GERMAN AYA

UNIVERSIDAD DISTRITAL

INTRODUCTION

Game-Based Learning (GBL) is an interactive educational method that uses games to enhance learning. This study seeks to evaluate student performance in GBL environments by analyzing data from a Kaggle competition, applying a systems thinking approach to design a robust data processing architecture.

METHODOLOGY

Systems Thinking Approach

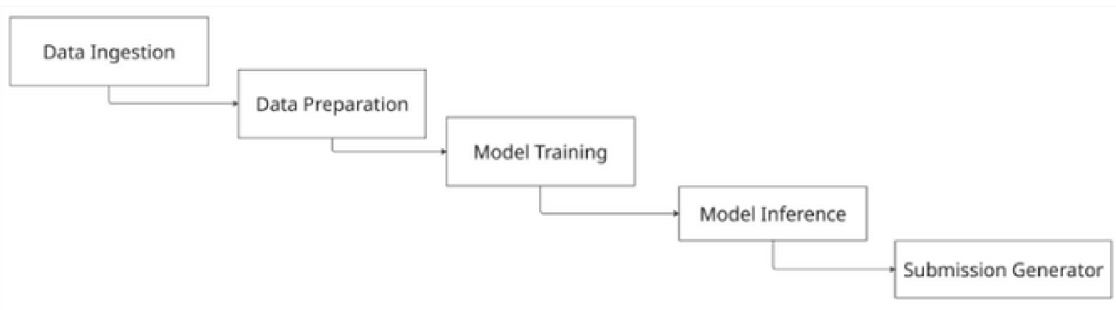
- Modeled the game as a complex system:
  - Components: user interactions (clicks/hovers), levels, settings
  - Relationships: player progression → difficulty scaling
  - Nonlinear dynamics: sensitivity to chaotic inputs (e.g., student focus)

Solution

- Analysis pipeline:
  - Capture interaction events (x,y coordinates + timestamps)
  - Predictive modeling (XGBoost/Random Forest)
  - Cross-validation

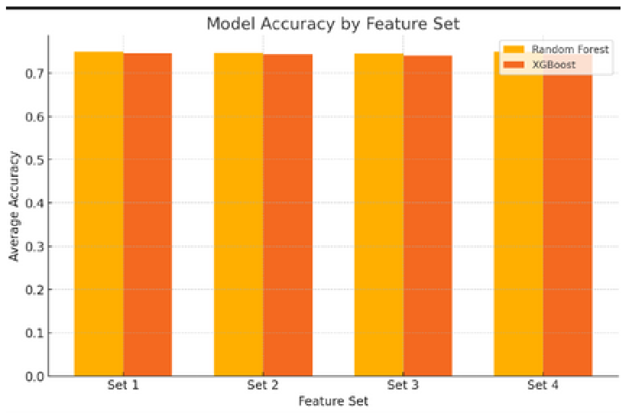
Technology Stack

- Python + TensorFlow Decision Forests



RESULTS

Both Random Forest (RF) and XGBoost achieved near-identical peak performance using Feature Set 1, with RF attaining a marginally higher accuracy of 0.750 (75.0%) compared to XGBoost's 0.747 (74.7%)—a minimal difference of 0.003. This parity confirms both models' suitability for predicting student performance, though RF holds a slight edge in optimal configuration.



While RF exhibited tighter consistency (accuracy range: 0.739–0.750, Δ=1.1%), XGBoost showed broader variability (0.731–0.747, Δ=1.6%). Crucially, Feature Set 1 outperformed other sets, underscoring that feature engineering drives reliability. For deployment, RF with Feature Set 1 is recommended as the most robust solution (0.750 accuracy), balancing performance and stability despite chaotic user inputs.

DISCUSSION

Both models (Random Forest and XGBoost) showed comparable performance in predicting student performance (~75% accuracy), with a slight advantage for Random Forest (0.750 vs. 0.747). This result, together with the lower variability of Random Forest (Δ1.1% vs. Δ1.6% for XGBoost), confirms its greater robustness to chaotic environments. It critically highlights the role of Feature Set 1—focused on spatio-temporal interactions—as a driver of peak performance, demonstrating that feature engineering outweighs the importance of algorithm choice. For practical implementations, Random Forest with dynamic features is recommended.

CONCLUSION

Adopting a systems perspective shifts focus beyond isolated data points to model the entire learning environment as interconnected components. This holistic approach accelerates problem decomposition, enabling scalable and precise predictive solutions.