Educational Data Mining Student Performance Research

Paper 1

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper1.pdf

Key Takeaways (For our Research)

- Used 450 students' grade data (65 courses, 9 semesters).
- Applied 5 ML algorithms: kNN, Random Forest, Decision Tree, Logistic Regression, Neural Network.
- Binary classification (Pass/Fail) performed much better than multi-class prediction.
- Neural Networks, kNN, and Random Forest gave the highest F1 scores (best ≈ 0.93).
- Used correlation-based feature selection — courses with correlation ≥ 0.3 improved prediction accuracy.
- Adding too many weakly correlated courses reduced performance.
- Used F1-score because dataset was imbalanced.
- Only academic grades were used (no LMS activity or behavioral data).

What's Useful for us

- Use binary + multi-class comparison in your paper.
- Apply correlation-based feature selection.
- Use F1-score instead of only accuracy.
- Improve by adding behavioral/LMS data (gap in this paper).

Paper 2

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper2.pdf

Key Takeaways (For our Research)

- Introduces **Federated Learning (FL)** for privacy-preserving Educational Data Mining.
- Instead of sharing student data, institutions share **model updates only**.
- Solves privacy issues related to sensitive data (grades, behavior, demographics).
- Handles multi-institution collaboration without centralized data storage.
- Highlights **non-IID data problem** (different institutions have different data distributions).
- Discusses privacy techniques:
  - Differential Privacy
  - Secure aggregation
  - Encryption
- Mentions trade-off between **privacy and model accuracy**.
- Suggests FL for:
  - Student performance prediction
  - Dropout detection
  - Personalized learning systems

What's Useful for us

- You can mention FL in **literature review or future work**.
- Add discussion on **privacy issues in EDM**.
- Highlight **data heterogeneity challenge** in student prediction.

Paper 3

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper3.pdf

Key Takeaways

- Used **UCI Student Performance dataset** (1044 students, 43 features).
- Performed both:
    - **Regression** (predict final grade)
    - **Classification** (pass/fail)
- **Ensemble methods performed best**:
    - Gradient Boosting → Best for regression (RMSE ≈ 3.38)
    - Random Forest → Best for classification (F1 ≈ 0.886)
- Excluded previous grades (G1, G2) to make prediction more realistic.
- Used **hyperparameter tuning (RandomizedSearchCV + 5-fold CV)**.
- Applied **SHAP for model explainability**.

Most Important Predictors

- Past failures (strongest factor)
- Study time
- School absences
- Alcohol consumption
- Parent education level

What's Useful for our Research

- Use **ensemble models** **(RF + Gradient Boosting)**.
- Compare **regression vs classification**.
- Add **feature engineering**.
- Include **explainable AI (SHAP)**.
- Mention dataset limitation and generalizability gap.

Paper 4

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper4.pdf

Key Takeaways

- Educational Data Mining mainly focuses on student performance and dropout prediction.
- Random Forest, SVM, Logistic Regression, and Ensemble models perform consistently well.
- Deep Learning (especially LSTM) is useful for temporal or sequential data.
- Handling imbalanced datasets (e.g., using SMOTE) significantly improves results.
- Explainable AI methods like SHAP and LIME are important for model transparency.
- Combining multiple data sources (grades, LMS logs, behavior) improves prediction accuracy.
- Data privacy and ethical concerns remain major challenges.

What Is Important for our Paper

- Clearly define the prediction problem (performance or dropout).
- Use proper preprocessing and imbalance handling techniques.
- Compare multiple models rather than using only one.
- Evaluate using metrics like Accuracy, F1-score, and AUC.
- Include explainability (e.g., SHAP) to strengthen the research contribution.
- Highlight the practical impact of your model for early intervention.

Paper 5

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper5.pdf

Key Takeaways

- Self-perception (especially math ability and academic expectations) is one of the strongest predictors of cognitive ability.
- Parental expectations and family support show strong and robust causal effects.
- Random Forest performed best among tested ML models.
- Using multiple explainability methods (SHAP, LIME, Morris, Feature Importance) reveals different feature priorities.
- Causal testing (PSM + robustness analysis) strengthens the reliability of findings.
- Relying on a single explainability method may give incomplete or biased interpretations.

What Is Important for our Paper

- Clearly justify why you are using explainable AI, not just prediction.
- Compare at least two explainability methods to strengthen validity.
- If possible, add causal analysis (e.g., PSM) to move beyond correlation.
- Highlight psychological and family-related variables, not just demographic factors.
- Emphasize transparency and interpretability as a research contribution.
- Discuss why model interpretability matters in educational decision-making.

Paper 6

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper6.pdf

Key Takeaways

- Model performance depends on matching feature selection with dataset characteristics.
- Information Gain + Decision Tree performed best (up to 96% accuracy).
- Chi-Square + Random Forest also showed strong results.
- Information Gain generally outperformed other feature selection methods.
- Feature selection improved accuracy and reduced redundancy.
- DE-FS achieved high accuracy (95.8%) with only 12 features.
- There is a trade-off between accuracy (RF, NN) and computation time (DT, NB faster).

Important for our EDM Paper

- Use and justify feature selection clearly.
- Compare multiple models and report accuracy, precision, recall, F1-score.
- Consider class imbalance handling.
- Highlight efficiency and interpretability along with prediction performance.

Paper 7

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper7.pdf

Main Useful Points

- The study used Educational Data Mining (clustering + regression analysis) to analyze elective course selection and student satisfaction.
- Four distinct student segments were identified: career-focused pragmatists, content enthusiasts, process-sensitive selectors, and balanced optimizers.
- Students differ significantly in how they select courses and what drives their satisfaction.
- Regression model explained 63% of satisfaction variance, showing strong predictive power.
- Satisfaction predictors vary across segments (e.g., career-focused students value career relevance most, process-sensitive students value administrative clarity).
- A five-layer framework was proposed: data collection, analytics, personalization, interface design, and feedback system.
- Emphasis on segment-aware personalization instead of one-size-fits-all systems.
- Ethical concerns highlighted: privacy, algorithmic fairness, transparency, and maintaining student autonomy.
- Digital divide and usability issues impact first- and second-year students more.

Most Important for our EDM Research Paper

- Use clustering to identify student segments before applying prediction models.
- Combine clustering + regression for stronger analytical contribution.
- Report variance explained (R^2) to show model effectiveness.
- Focus on explainable and personalized decision-support systems.
- Highlight information quality and expectation alignment as key educational variables.
- Link analytical findings to a practical framework or implementation model.
- Include ethical and fairness considerations in educational AI systems.

Paper 8

https://d.docs.live.net/ec55ae27f4c6f109/Desktop/EDM%20PROJECT/Paper8.pdf

Key Takeaways

- Educational Data Mining (EDM) extracts meaningful patterns from large student-related datasets.
- Uses techniques like classification, regression, clustering, association rules, and Bayesian models.
- Applications include student performance prediction, student modeling, behavior detection, and personalized learning.
- Follows a structured pipeline: data collection → preprocessing → modeling → validation.
- Major challenges: handling repeated student data (non-independence) and ensuring model generalization.

What Is Important for Your Paper

- Clearly define your research objective (e.g., student performance prediction).
- Explain your dataset and preprocessing steps properly.
- Justify the choice of machine learning technique.
- Use proper validation methods (e.g., cross-validation).
- Address limitations like data dependency and overfitting.
- Show practical impact (early intervention, decision support, or system improvement).