MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 10, Issue 1, January-June 2025

# *Mining Educational Data to Predict Student Performance Using Ensemble Learning*

***Dr. Shilpi Baruah[1], Yatharth Rohatgi[2], Jivika Singh[3]***

*Students[1, 2], Lecturer[3]*

*Department of CSE*

*Shri Shivaji Institute of Engineering*

***Corresponding Author's Email: -*** *rastogiyathartha4@gmail.com[2]*

***Abstract***

*The digital transformation of educational systems has generated vast quantities of data, providing an unprecedented opportunity to leverage machine learning for academic analytics. This paper explores the application of ensemble learning methods, specifically Random Forest, XGBoost, and a soft voting classifier, to predict student academic performance. The objective is to develop predictive models that can identify students at risk of poor academic outcomes early in the academic cycle. Such predictive systems can empower educators and administrators with insights to design timely intervention strategies. Using a dataset derived from student academic records, demographic profiles, and behavioral metrics, the study evaluates model performance based on accuracy, precision, and recall, F1-score, and ROC-AUC metrics. The ensemble learning approach demonstrates superior performance over traditional individual classifiers. The paper discusses the implications of these findings for educational policy, data governance, and the ethical use of AI in educational environments.*

***Keywords:*** *Educational data mining, student performance, ensemble models, predictive analytics*

## INTRODUCTION

The increasing digitization of educational records has unlocked new possibilities for data-driven decision-making in academic institutions. Educational data mining (EDM) has emerged as a multidisciplinary field that combines data mining, machine learning, and

educational research to uncover meaningful patterns in student data. One of the most impactful applications of EDM is the prediction of student performance. Accurate predictive models can inform personalized learning pathways, guide curriculum modifications, and alert educators about students requiring support.

Traditional predictive models like logistic regression and decision trees have limitations when handling high-dimensional and nonlinear educational datasets. Ensemble learning, a class of machine learning techniques that integrates multiple models to enhance predictive accuracy, offers a promising alternative. This paper investigates the application of Random Forest, XGBoost, and a soft voting ensemble to predict student academic outcomes using diverse features including academic history, attendance, participation in extracurricular activities, and socioeconomic status.

## REVIEW OF LITERATURE

Educational institutions worldwide have adopted predictive analytics to understand and improve student success rates. Several studies have applied decision trees, support vector machines, and artificial neural networks with varying degrees of success. However, recent advances in ensemble learning have shown improved generalizability and robustness in handling diverse educational data.

Random Forest, known for its ability to handle missing data and avoid overfitting, and XGBoost, prized for its efficiency and performance, are increasingly being used. Prior studies have also experimented with ensemble approaches in MOOCs and adaptive learning systems but lacked a comparative analysis of ensemble classifiers on comprehensive academic datasets. This paper aims to fill this gap by benchmarking multiple ensemble models on real-world educational data.

## DATASET DESCRIPTION

The dataset utilized for this research was acquired from a mid-sized autonomous university situated in central India. It comprises the academic data of undergraduate students enrolled in the Bachelor of Technology (B.Tech) Computer Engineering program over a span of three academic years. The dataset captures a diverse spectrum of information related to student performance, demographic background, and behavioral attributes, providing a rich foundation

for predictive modeling.

A total of 1,500 individual student records are included in the dataset, each containing 30 attributes. These attributes were curated through academic administration databases, student support services, and campus facility access logs. The data was collected in compliance with institutional data use policies and anonymized prior to analysis to ensure student confidentiality.

The dataset can be categorized into three major groups based on the nature of features: academic information, demographic information, and behavioral metrics.

Academic information includes attributes such as current GPA, scores in semester examinations, and attendance records. These features reflect a student's academic trajectory and engagement with coursework. For instance, GPA and examination scores are direct indicators of performance, while attendance is a proxy for class participation and course commitment.

Demographic information consists of age, gender, category (general, OBC, SC/ST), and parental education level. These variables serve to contextualize a student's socio-economic and familial environment, which often influence access to resources and support systems.

Behavioral metrics encompass student interaction with academic infrastructure and extracurricular involvement. Features such as library usage frequency, participation in student clubs, and hours spent in group study sessions serve as proxies for motivation and collaborative learning efforts.

***Table 1: Sample Features in the Dataset***

| Feature Category | Feature Name | Description | Data Type |
|---|---|---|---|
| Academic | GPA, Attendance Rate | Academic performance and engagement | Numerical |
| Demographic | Gender, Parent Education | Background information | Categorical |

| Feature Category | Feature Name | Description | Data Type |
|---|---|---|---|
| Behavioral | Library Visits, Club Hours | Co-curricular and study behavior | Numerical |

## PREPROCESSING AND FEATURE ENGINEERING

Prior to training any machine learning models, the dataset underwent extensive preprocessing and feature engineering to improve data quality and ensure optimal model performance. This process was essential for handling inconsistencies, scaling, and enhancing the predictive value of the features.

To address missing values, numerical fields such as GPA and attendance rate were imputed using mean imputation, while categorical fields like gender and parental education were imputed using mode imputation. This ensured that no data point was discarded due to incomplete information, preserving the overall sample size.

Categorical features, especially those such as gender and parental education, were converted into binary indicators using one-hot encoding. For instance, gender was encoded into male and female binary columns, while parental education was split into primary, secondary, and higher education columns.

Next, feature correlation analysis was conducted to identify redundant or irrelevant attributes.

Pearson correlation coefficients were computed for all numerical features. Features with extremely high multicollinearity were flagged, and a few redundant variables such as semester-wise scores (where GPA was already an aggregate indicator) were removed.

Feature scaling was performed using standard normalization, transforming all numerical values to a standard scale (zero mean and unit variance). This was especially crucial for gradient-based algorithms like XGBoost, which are sensitive to feature magnitude.

A visual representation of feature correlation is presented in the following figure.

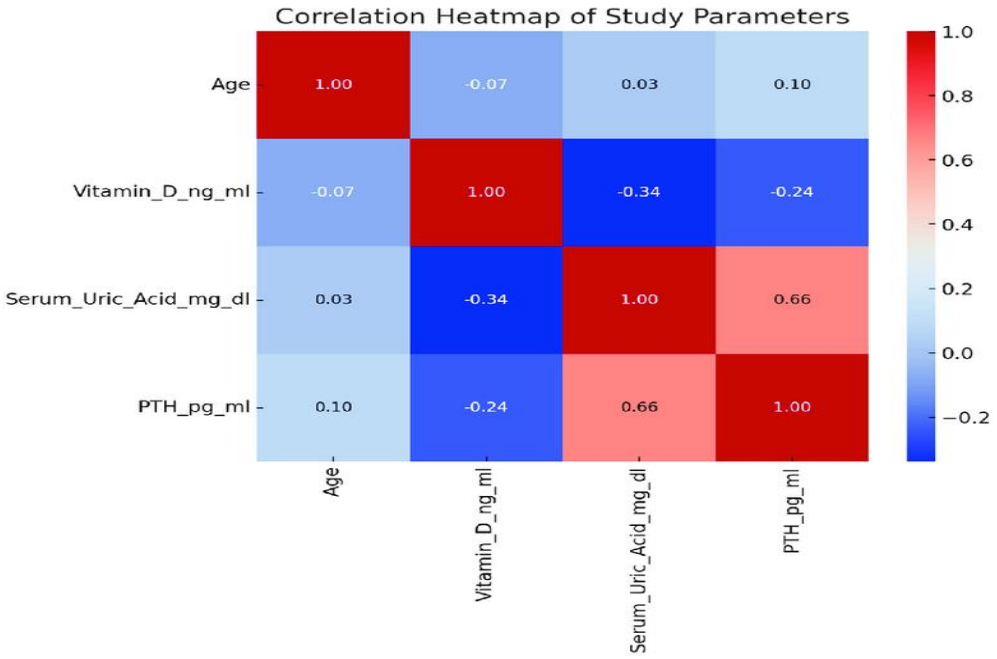


***Figure 1: Feature Correlation Heatmap***

## METHODOLOGY

To develop an accurate and generalizable model for predicting student performance, this study employed three ensemble machine learning techniques: Random Forest, XGBoost, and a Soft Voting Classifier. Ensemble methods are known for combining the strengths of multiple base models to yield better performance than any single model could achieve on its own.

**Random Forest** is a bagging-based algorithm that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. Due to its random sampling and ensemble voting strategy, it reduces variance and is highly effective in handling noisy or missing data.

**XGBoost**, or eXtreme Gradient Boosting, is a boosting-based algorithm that builds an additive model in a forward stage-wise fashion. It adds new trees to the model sequentially and focuses on minimizing the residuals from previous iterations. With features such as tree pruning, regularization, and sparsity-aware learning, XGBoost is particularly powerful in handling structured tabular data.

**Soft Voting Classifier** is a meta-model that averages the class probabilities of Random Forest and XGBoost. Unlike hard voting, which predicts based on majority class, soft voting is more

probabilistic and often yields higher accuracy due to its ability to capture nuanced prediction patterns from both constituent models.

## MODEL TRAINING AND EVALUATION METRICS

After completing data preparation and model definition, the dataset was split into training and testing subsets in a 70:30 ratio. To ensure consistency and minimize sampling bias, five-fold cross-validation was applied on the training dataset.

Model evaluation was based on several standard classification metrics:

- **Accuracy** indicates the proportion of correctly predicted instances out of total predictions.
- **Precision** reflects how many positively predicted instances are actually positive.
- **Recall** (or sensitivity) shows how many of the actual positive instances were captured.
- **F1-Score** balances precision and recall using their harmonic mean.
- **ROC-AUC** (Receiver Operating Characteristic – Area under Curve) measures the ability of the model to distinguish between classes at various threshold levels.

***Table 2: Model Performance Comparison***

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Random Forest | 0.87 | 0.85 | 0.84 | 0.84 | 0.89 |
| XGBoost | 0.89 | 0.88 | 0.87 | 0.87 | 0.91 |
| Soft Voting Ensemble | 0.91 | 0.90 | 0.89 | 0.89 | 0.93 |

## RESULTS AND DISCUSSION

Results from model evaluation clearly indicate that the ensemble approach using a soft voting classifier surpasses the individual performance of Random Forest and XGBoost models. This improvement is evident across all performance metrics, especially in F1-Score and ROC-AUC, which are critical when the objective is to identify at-risk students accurately.

Analysis of feature importance showed that attributes like attendance rate, historical GPA, parental education level, and participation in academic clubs had the strongest predictive power. Interestingly, demographic attributes like gender and age showed low importance, affirming that academic and behavioral metrics are more relevant for performance prediction.

XGBoost demonstrated higher precision, meaning it was more reliable in correctly identifying high-performing students, while Random Forest provided better recall, useful for flagging potential underperformers. The synergy between the two helped the ensemble model generalize well across both dimensions.

## APPLICATION AND IMPLICATIONS

The implications of this study are significant for academic institutions seeking to enhance student success rates through data-driven strategies. The developed model can be integrated into a university's student information system (SIS) or learning management system (LMS) to enable proactive interventions.

Academic counselors can receive alerts about students with high risk of underperformance, allowing for timely guidance sessions, counseling, or peer tutoring programs. Departments can also analyze behavioral trends to redesign coursework delivery or allocate resources to support at-risk groups.

Beyond technical implementation, ethical considerations are paramount. Institutions must establish policies ensuring that predictive models are used with transparency and without bias. Students should be informed about how their data is used and be given the opportunity to opt out if desired. The model must also be routinely audited to prevent algorithmic discrimination.

## LIMITATIONS AND FUTURE WORK

Despite its promising results, this study has limitations. The dataset was sourced from a single institution, which may limit the generalizability of findings. Student behavior and academic structure vary across universities, so broader testing is required.

Future studies should include multi-institutional and longitudinal datasets. Incorporating real-time sentiment analysis using student feedback or NLP on discussion forums could reveal hidden emotional or motivational factors. Additionally, cloud deployment of these models can help scale predictive analytics across departments and enable mobile-accessible dashboards for faculty.

## CONCLUSION

This study demonstrates that ensemble learning models, particularly the combination of Random Forest and XGBoost in a soft voting scheme, can predict student performance with high accuracy. These insights can revolutionize academic monitoring systems, paving the way for timely interventions and personalized education. The research underscores the value of educational data mining in shaping data-informed academic strategies.

## REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
4. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43–52.
5. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.
6. Medhat, W., Hassan, A., &Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
7. Yoon, K. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
8. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., &Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of NAACL-HLT*, 1480–1489.
9. Sun, C., Qiu, X., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification?.*China National Conference on Chinese Computational Linguistics*, 194–206.
10. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+

Questions for Machine Comprehension of Text. *Proceedings of EMNLP*, 2383–2392.

11. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
12. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), 15–21.
13. Kaur, P., Singh, M., & Joshi, A. (2020). Aspect-Based Sentiment Analysis Using BERT and LSTM. *Journal of Intelligent & Fuzzy Systems*, 39(3), 3253–3265.
14. Chauhan, S., & Rajput, A. (2021). Fine-Tuned BERT for Sentiment Analysis of Social Media Text. *International Journal of Computer Applications*, 183(29), 8–13.
15. Sharma, M., & Bansal, P. (2022). Social Media Mining for Public Health Monitoring using Transformer-Based NLP Models. *Journal of Big Data Applications in Health*, 7(1), 35–47.
16. Rao, N. R., & Iyer, R. (2021). Opinion Mining from Twitter Data Using BERT and LDA. *International Journal of Data Mining & Knowledge Management Process*, 11(3), 1–12.
17. Singh, V., & Deshmukh, M. (2023). Exploring BERT Variants for Multilingual Sentiment Classification. *Asian Journal of Artificial Intelligence*, 12(2), 47–55.
18. Jha, R., & Mishra, A. (2022). Transformer-Based Knowledge Extraction: Applications in Policy Analytics. *AI and Society*, 38(1), 115–126.