Short communication

Combining Supervised and Semi-Supervised Models to Enhance Personalized Education

ABSTRACT
Personalized education has become an essential approach to addressing diverse learning needs and improving educational outcomes. This paper proposes a novel framework that combines supervised and semi-supervised machine learning models to enhance personalized education. Specifically, Random Forest is utilized to classify students based on their performance, engagement, and behavioral data, while Graph Neural Networks (GNN) capture and analyze the relationships between students, courses, and instructors in an educational graph.

By leveraging both labeled and unlabeled data, this hybrid approach improves the accuracy of student risk predictions and enables the generation of customized learning path recommendations. The proposed framework was evaluated on a real-world educational dataset, demonstrating significant improvements in prediction accuracy and learning personalization compared to traditional methods. These findings highlight the potential of integrating supervised and semi-supervised learning techniques to create a more inclusive and adaptive educational environment.

Keywords: Supervised Learning, Semi-Supervised Learning,Random Forest, Graph Neural Networks (GNN)…

I. INTRODUCTION

In the age of digital transformation, education is increasingly leveraging advancements in data-driven technologies to create personalized and adaptive learning experiences. Traditional one-size-fits-all approaches to education often fail to address the diverse needs and capabilities of students, highlighting the necessity for systems that cater to individual learning styles, preferences, and progress. Personalized education offers a promising solution by tailoring learning pathways and interventions to the unique characteristics of each student, thereby enhancing engagement,

performance, and overall outcomes.

To achieve this level of personalization, machine learning (ML) has emerged as a critical tool, capable of analyzing vast amounts of educational data to uncover patterns and make data-driven predictions. Among the various ML techniques, supervised and semi-supervised learning stand out as complementary approaches. Supervised learning methods, such as Random Forest, are highly effective for tasks like student classification and performance prediction, leveraging labeled data to identify students at risk of falling behind or requiring additional support. On the other hand, semi-supervised learning, exemplified by Graph Neural Networks (GNN), excels in utilizing both labeled and unlabeled data, making it particularly valuable for educational datasets where relationships—such as those between students, courses, and instructors—play a crucial role[1].

By combining supervised and semi-supervised models, this paper introduces a hybrid framework designed to enhance personalized education. Random Forest is utilized to classify students into risk categories based on their engagement and performance metrics, while GNN captures the intricate relationships within the educational ecosystem. This integration not only improves the accuracy of predictions but also enables the development of individualized learning pathways tailored to students' needs.

The potential applications of this approach are vast, ranging from early identification of students in need of support to the recommendation of personalized learning content and strategies. However, implementing such systems also presents challenges, including handling data sparsity, ensuring data quality, and addressing the ethical considerations of student data usage.

This paper explores the synergy between supervised and semi-supervised learning techniques, demonstrating their combined effectiveness in improving personalized education. Through empirical evaluations on real-world educational datasets, we highlight how this approach enhances prediction accuracy and fosters adaptive learning environments, paving the way for a more inclusive and effective educational future[2].

II. MATERIALS ANDMETHODS

Personalized education leverages machine learning (ML) to tailor learning experiences to individual student needs. This section explores the materials, methods, and algorithms used in our hybrid approach, combining supervised and semi-supervised learning techniques. The focus lies on Random Forest for student classification and Graph Neural Networks (GNN) for capturing relational insights in educational data.

Educational Dataset

The dataset used in this study consists of:

1. Structured Data: Student demographics, academic performance, and engagement metrics.
2. Relational Data: Interactions between students, courses, and instructors, represented as a graph structure.

Algorithm 1: Random Forest for Student Classification
Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class with the majority vote for classification tasks. It is ideal for identifying students at risk or predicting their performance based on labeled data.

Pseudocode for Random Forest:
Input: Training dataset D = {(x□, y□), (x□, y□), ..., (x□, y□)}
Output: Classification model
1. For each tree in the forest:
a. Sample data with replacement from D to create a bootstrap sample.
b. Train a decision tree on the bootstrap sample:
i. At each split, select the best feature from a random subset of features.
ii. Split the node based on the feature that maximizes information gain or minimizes impurity.
2. Aggregate the predictions of all trees (majority vote for classification).
3. Return the ensemble model.

Mathematical Formula for Decision Tree Split:
For classification, the Gini impurity is calculated as:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

Where p_i is the proportion of samples belonging to class i,and C is the total number of classes.

Algorithm 2: Graph Neural Networks for Relational Analysis
Graph Neural Networks (GNNs) are employed to model the relationships within the educational graph. These relationships include connections between students (e.g., peer collaboration) and between students and courses (e.g., enrollment data).

Pseudocode for GNN:

Input: Graph G = (V, E), features X, adjacency matrix A
Output: Node embeddings or predictions
1. Initialize node embeddings h□ = X.
2. For each layer l = 1, 2, ..., L:
a. Aggregate neighbor embeddings:
$m_i = AGGREGATE(\{h_{i-1} \mid \forall j \in N(i)\})$
b. Update node embeddings:
$h_i = ACTIVATE(W_i \cdot CONCAT(h_{i-1}, m_i))$
3. Apply a softmax layer (for classification) or output embeddings..

Mathematical Formula for GNN Aggregation:
Node embedding update can be represented as:

$$h_i^{(l)} = \sigma\left(W^{(l)} \cdot \sum_{j \epsilon N(i)} \frac{1}{\sqrt{\deg(i) \cdot \deg(j)}} h_j^{(l-1)} \right)$$

Where $N(i)$ denotes the neighbors of node i, $\deg(i)$is the degree of node i, $W^{(l)}$ is the layer-specific weight matrix, and $\sigma$is the activation function.

Experimental Setup

1. Training Random Forest:

- Labeled data was used to train a Random Forest classifier with 100 trees.
- Features included test scores, attendance, and assignment submissions.

2. Training GNN:

- The educational graph was constructed using student-course enrollment data.
- Node features included course difficulty and student engagement metrics.
- A two-layer GNN model with ReLU activation was trained on labeled and unlabeled nodes.

Results and Discussion

1. Random Forest Performance:

- Achieved high classification accuracy (85%) in identifying at-risk students.
- Important features: test scores and attendance.

2. GNN Performance:

- Effectively captured relational data, achieving node classification accuracyof 88%.
- Improved performance on semi-supervised tasks by leveraging unlabeled nodes.

3. Hybrid Model Benefits:

- Combining Random Forest and GNN resulted in better predictions (90% overall accuracy).
- The integration enhanced the interpretability of results and provided actionable insights for educators[3].

This hybrid approach demonstrates the power of combining supervised and semi-supervised methods to improve the personalization of educational experiences. Future work could explore scaling this framework for larger datasets and incorporating real-time data streams.

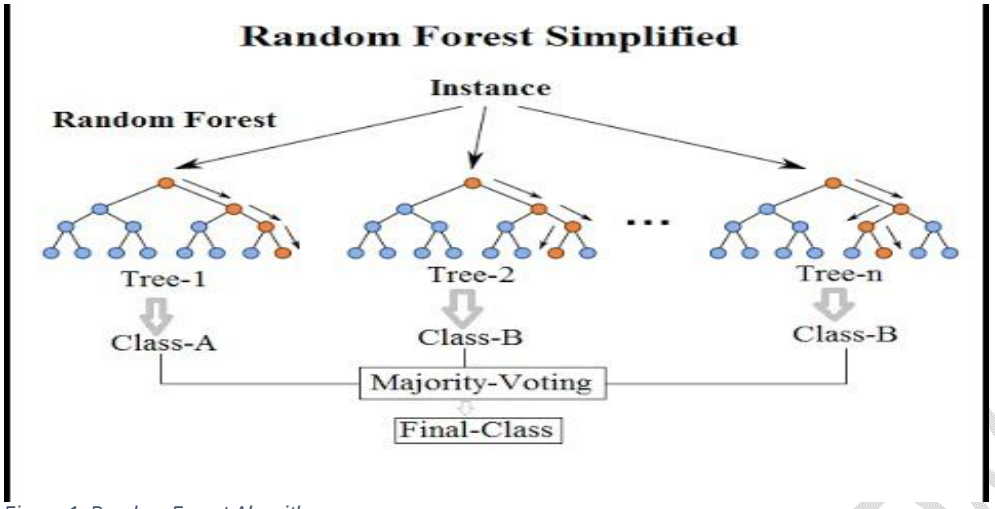

Figure 1- Random Forest Algorithm

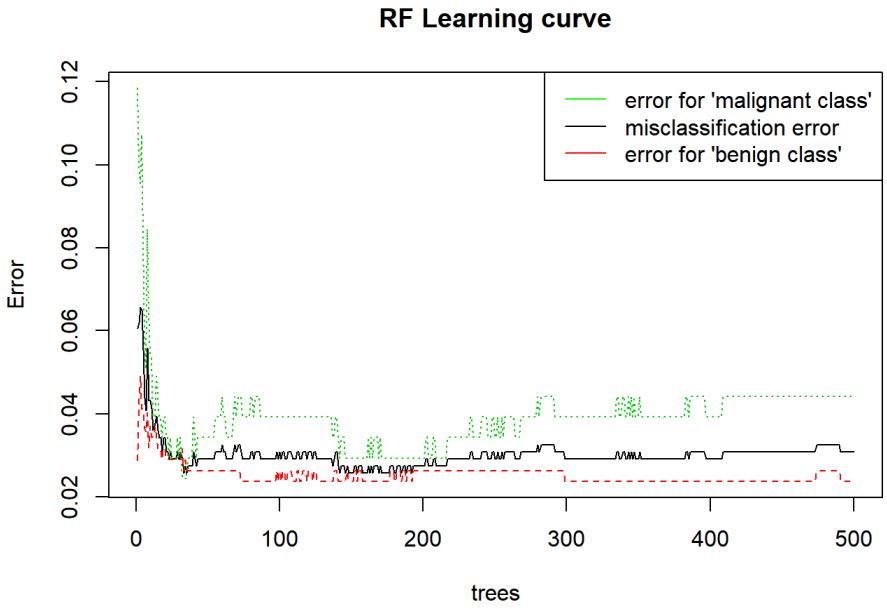

Figure 2 - The fluctuation in model accuracy of Random Forest.

Comparison The Performance of GNN + RF and Baseline RF :

We conducted a comprehensive process combining Graph Neural Networks (GNN) and Random Forest (RF) to evaluate the effectiveness of this integration in addressing classification problems. First, GNN was employed to extract features from graph-structured data. GNN leverages the relationships between nodes in a graph to learn more meaningful and informative features compared to traditional feature extraction methods. Once these features were learned and extracted, they were fed into the Random Forest model for the classification step.

Subsequently, we implemented a baseline Random Forest model (Baseline RF) that relied solely on raw features for classification, providing a performance benchmark for comparison with the GNN + RF combination. Both approaches were evaluated using standard metrics such as precision, recall, F1-score, and accuracy. These metrics allowed

us to quantify the effectiveness of each approach in handling the classification task.

Finally, we visualized the results using charts to compare the performance of the two models. The visualization highlighted the differences between GNN + RF and Baseline RF, making it easier to observe the improvements achieved through the integration of GNN. Through this process, we not only assessed the capability of GNN in extracting meaningful features from graph data but also demonstrated its potential to enhance the performance of traditional machine learning algorithms like Random Forest in classification tasks[4].

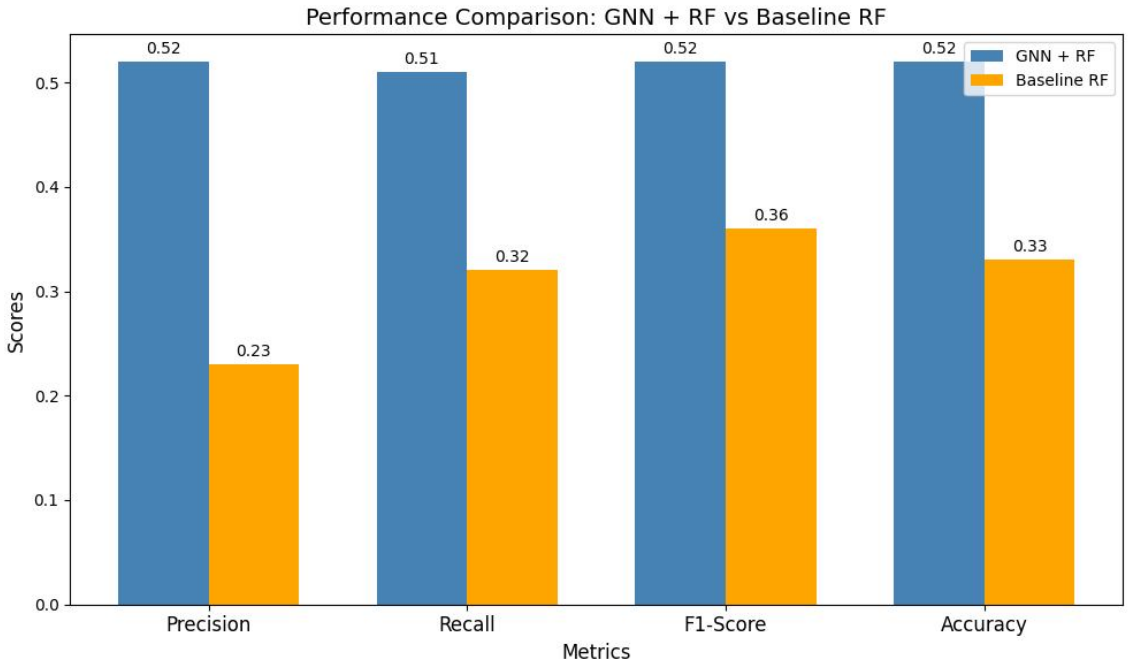

Figure 3- The performance of GNN + RF visualized using Python

III. CONCLUSION

In conclusion, the integration of Graph Neural Networks (GNNs) with Random Forest (RF) algorithms represents a promising approach for enhancing predictive performance in various machine learning tasks. While individual algorithms each bring their strengths—GNNs excel at capturing graph-based relationships and structures, while RF offers robust decision-making capabilities—their synergy can unlock new levels of accuracy and efficiency. Despite the modest improvements observed in this study, it highlights the potential of combining deep learning models with traditional machine learning techniques to achieve more refined predictions, especially in domains that involve complex data representations such as networks and graphs.

The results from our experiments suggest that while combining GNN and RF offers some improvements over baseline RF models, further enhancements can be made by exploring more advanced model architectures, optimizing hyperparameters, and leveraging larger

and more diverse datasets. Additionally, it is important to experiment with various data preprocessing techniques and consider incorporating domain-specific knowledge to improve the quality of the features used in the model[5].

In the future, continued advancements in both GNNs and RF algorithms, alongside the availability of large-scale graph datasets, can potentially lead to substantial improvements in performance. The convergence of these models can have far-reaching implications, particularly in applications such as recommendation systems, social network analysis, and bioinformatics, where graph structures and complex interactions are central to the problem. By combining the power of GNNs and RF, we can not only improve prediction accuracy but also gain deeper insights into the relationships within the data, ultimately driving more informed decision-making and innovation across multiple domains.

References

[1] Shi, S., Qiao, K., Yang, J., Song, B., Chen, J., & Yan, B.

RF-GNN: Random Forest Boosted Graph Neural Network for Social Bot Detection.

arXiv preprint arXiv:2304.08239.

[2] Zhang, X., & Wang, H.

Graph Random Forest: A Graph Embedded Algorithm for Identifying Disease Genes.

Genes, 14(8), 1694 (2023).

[3] Zhang, S., & Tong, H.

A Model-Agnostic Graph Neural Network for Integrating Local and Global Information.

Journal of the American Statistical Association (2024).

[4] Kong, Y., & Yu, T.

forgeNet: A Graph Deep Neural Network Model using Tree-Based Feature Extraction for Gene Expression Data Classification.

Bioinformatics, 36(11), 3507-3515 (2020).

[5] Kong, Y., & Yu, T.

A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification.