**CSE400B**

**Summer 2025**

**Capstone Project Report**

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

**Project Title:** Enhancing Personalized Learning Experience of Students using Machine Learning in Education

**Project Members**

|  |  |
| --- | --- |
| **Name** | **ID** |
| Khan Samiul Arefin | 2021-3-60-110 |
| Faiza | 2021-1-60-089 |
| Masum Mushfik Rifat | 2021-3-60-100 |
| Ashikur Rahman | 2021-3-60-039 |

Project Supervisor

Tanni Mittra

Senior Lecturer

Department of Computer Science & Engineering

East West University

Date of Submission: 28/08/2025

**Declaration**

Under the guidance of Tanni Mittra, Senior Lecturer in the Department of Computer Science and Engineering at East West University, we, Khan Samiul Arefin, Masum Mushfik Rifat, Faiza, and Ashikur Rahman, hereby declare that the work presented in this capstone project report is the result of our investigation. I/We further affirm that, except for publication, no portion of this project has been or is being submitted anywhere for the award of any degree or diploma.

**Letter of Acceptance**

The "Personalized using Ai Using Random Forest Classifier" project, which was turned in to the Department of Computer Science & Engineering at East West University in Dhaka, Bangladesh, by Khan Samiul Arefim, Ashikur Rahman, Faiza, and Masum Mushfik Rahman, has been approved as meeting a portion of the requirements needed to earn a Bachelor of Science in Computer Science and Engineering.

Board of Examiners

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tanni Mittra

Senior Lecturer

Department of Computer Science and Engineering

East West University

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Maheen Islam (Chairperson)

Associate Professor

Department of Computer Science and Engineering

East West University

**Abstract**

In this paper, machine learning algorithms, in particular a **Random Forest Classifier** is used to define proficiency levels of different students concerning various academic topics. The model is trained on the dataset of exam scores and performance of students in a variety of themes. It is based on the classification and generates the personalized lessons dynamically with the ChatGPT API to work on the weaknesses of the students. The accuracy, precision, recall and F1-score are some of the key metrics by which the performance of the model is analyzed and proves that machine learning can be used to construct individualized learning paths. In the study, the design of the system aims at enhancing the engagement and performance of students by means of timely creation of learning contents, which are tailored to the context in which the study was conducted.

Key contributions of this work include:

1. **Student Classification**: Classification of students based on their exam performance using Random Forest Classifier.
2. **Personalized Lesson Generation**: Dynamic generation of learning content using the ChatGPT API, tailored to each student’s proficiency.
3. **Real-Time Adaptation**: Continuous adaptation of lessons and quizzes based on students' ongoing performance.
4. **Evaluation**: Performance evaluation metrics, including accuracy, precision, recall, and AUC, with results showing high classification accuracy.

**Acknowledgement**

Our study, titled "Personalized Learning Using AI: A Random Forest Classifier-Based Approach," has been both a challenging and rewarding experience, providing us with valuable knowledge and skills in the field of machine learning and personalized education.

We would like to extend our sincere thanks to our esteemed supervisor, Tanni Mittra, Senior Lecturer, Department of Computer Science and Engineering, East West University for their unwavering guidance, constant support, and insightful feedback. Their expertise has been crucial in refining our work and achieving the desired outcomes. It has been a privilege to work under their supervision, and we deeply appreciate their patience, encouragement, and dedication throughout this project.

We would also like to express our heartfelt gratitude to the faculty members of the Department of Computer Science and Engineering, East West University for their continuous support and invaluable contributions to our academic journey. Their advice and encouragement have been instrumental in the success of this project.

Our special thanks go to our families for their unconditional love, prayers, and sacrifices. Their support has been a constant source of motivation and strength throughout this study.

Lastly, we would like to acknowledge the resources and facilities provided by the Department of Computer Science and Engineering at East West University, which played an essential role in the successful completion of our research.

Khan Samiul Arefin

August 2025

Faiza

August 2025

Masum Mushfik Rifat

August 2025

Ashikur Rahman

August 2025

**Table of Content**

Declaration

Letter of Acceptance

Abstract

Acknowledgment

Table of Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables

List of Figures

**Chapter 1: Introduction**

**1.1 Background**

Education is an industry that is quickly changing due to the introduction of technology and one of the most promising trends is the idea of personalized learning. Personalized learning arms education with the ability to adapt to individual students needs, outside of the typical, one-size-fits-all learning model. The idea is to have each student learn at his/her own pace, getting assistance where he/she needs it on specific subjects and getting a stimulating challenging level where he/she can excel. Such a highly customized manner of execution has proven to be more engaging, result-producing, and conducive to long-term commitment to knowledge.

Small-scale flexible individual learning in education is based on the presumption that each student is different and that individualizing their needs and preferences enables them to learn. In the past, educators have been challenged to make sense of massive amounts of student data, in numerous different forms. With the latest breakthroughs in Artificial Intelligence (AI) and machine learning this has led to new opportunities in automating and optimizing personalized learning. AI has the capability to analyze complex algorithms, handle datasets, and identify weakness in pupil performance and create personalized learning courses that can respond to the individual strengths and weaknesses of learners. Out of all the available machine learning algorithms, the Random Forest Classifier has been identified as among the most effective to classify the performance of students in different subjects.

Among the many machine learning algorithms available, the Random Forest Classifier has proven particularly effective for classifying student performance across a variety of subjects. This ensemble learning method combines the outputs of multiple decision trees to improve accuracy, reduce overfitting, and enhance the generalization of predictions. By training on student performance data, the Random Forest Classifier can categorize students into different proficiency levels, providing a foundation for personalized content generation.

This Random Forest learning algorithm integrates the decisions of several decision trees to increase accuracy, decrease overfitting, and provide better generalization of forecasting. Trained in data about performance of students, the Random Forest Classifier will be able to group students into groups of proficiency and serve as the hint to generate their personalized content.

**1.2 Problem Statement**

Although interest in personalized learning is on the increase, traditional education systems cannot avoid the presence of difficulties in offering individualized training that attends to the distinct needs of students. There exists a massive disconnect between the growing pressure to provide personalized learning experiences, on the one hand, and the limited capabilities in real-time delivery of these experiences, on the other hand. Although there are numerous systems to monitor student performance, there are a limited number of systems that can produce real time, dynamic and personalized learning regime which changes after every student performance.

The difficulty in the development of a scalable and efficient system to implement personalized learning is the aggregation of information using a range of learning sources, the deployment of relevant machine learning models to classify the content, and the production of customized text that can be read by the students. This study will address this gap by developing the model that will, besides classifying the abilities of students, perform the work of tailor-made preparation of learning materials reflected in real-time. This paper is mainly based on the use of Random Forest Classifier, where students are sorted into the categories of proficiencies (beginner, intermediate, advanced), and subsequently, the custom lesson plans are created and quizzes suggested, using the ChatGPT API.

**1.3 Objectives**

This research aims to achieve the following objectives:

* Student Classification: To classify students based on their exam performance and assign proficiency levels (beginner, intermediate, advanced) using the Random Forest Classifier.
* Personalized Lesson Generation: To generate personalized lessons dynamically for each student based on their assigned proficiency level, utilizing the ChatGPT API.
* Real-Time Adaptation: To develop a system that adapts to students' evolving learning needs, continuously adjusting content and quizzes based on their progress.
* Performance Evaluation: To evaluate the model’s performance using key metrics such as accuracy, precision, recall, and F1-score.

**1.4 Research Questions**

This research seeks to answer the following key questions:

* How effective is the Random Forest Classifier in classifying students’ proficiency levels based on their exam scores across different topics?
* Can AI-generated lessons based on student proficiency provide more effective learning pathways compared to traditional methods?
* What impact does real-time adaptation of learning content and quizzes have on student engagement and learning outcomes?
* How can the API be used to dynamically generate personalized educational content that addresses individual student needs?

**1.5 Significance of the Study**

The results of this study have value to the educational technology sector because they will illustrate opportunities of integrating machine learning and natural language processing to develop personalized learning environments. By coming up with a system, which categorizes students and customizes content in real time performance, this research has the potential to:

* Improve learning outcomes by providing targeted, adaptive learning experiences.
* Enhance student engagement by offering personalized challenges and support.
* Provide educators with powerful tools to track and address student weaknesses, ultimately leading to more effective teaching practices.

In addition, the research itself yields significant information concerning what might be done with machine learning models and natural language generation systems such as ChatGPT to automate content creation, which may be applicable to more than e-learning platforms and other educational technologies.

**1.6 Scope of the Study**

The contribution of this work is in the creation of a personalized learning system based on machine learning on student classification and artificial intelligence to generate the content. The system will be tested upon data of exam results of several students on a range of subjects. Although the results of educational performance are the target of the system, one can find that the discussed methods can be used in the context of corporate training, language learning, and even in healthcare education.

**The scope of the study includes:**

* The development of a classification model using Random Forest.
* The integration of ChatGPT to generate personalized learning content.
* The evaluation of the system’s performance through established classification metrics (accuracy, precision, recall, F1-score).

**Chapter 2: Proposed Methodology**

The proposed methodology in this research will be used to develop and evaluate an individualized learning system that applies the **Random Forest Classifier** model to classify students and make use of **API** to create dynamic material on demand. The methodology is divided into the following stages:

**2.1 Data Collection and Preparation**

The first step in the proposed methodology is the collection of data, which consists of students' **midterm scores** across various topics. A dataset is collected based on midterm scores for a total of **2000 students**. The data will be obtained from a learning management system (LMS), a digital educational platform, or traditional exam results. The data should include the following:

* **Student Information**: This can be anonymized to ensure privacy but may include features like student ID, gender, and age for demographic analysis.
* **Scores by Topic**: The performance data showing the scores that students achieved on each individual topic (e.g., Inheritance, Class, Polymorphism).
* **Topic Categories**: These represent the different areas or subjects the student is being assessed on.

Once the data is collected, it is preprocessed and cleaned by handling missing values, removing duplicate records, and ensuring consistency. The student scores are labeled based on performance thresholds:

* **Weak**: Students who score **below 85%** on a particular topic.
* **Not Weak**: Students who score **above or equal to 85%** on the same topic.

This classification forms the labeled dataset necessary for training the model.

**2.2 Preprocessing**

Once the data is collected, preprocessing steps will be performed, including:

* Handling missing values and outliers.
* Encoding categorical variables such as subject area and performance levels (e.g., beginner, intermediate, advanced).
* Normalizing or scaling continuous features such as scores.

**Model Selection: Random Forest Classifier**

The core of the methodology involves testing multiple machine learning models to classify students based on their performance. The models considered for this study are:

* **Random Forest Classifier (RFC)**: Chosen for its robustness in handling both continuous and categorical data.
* **Multilayer Perceptron (MLP)**: A neural network-based model useful for detecting complex patterns.
* **CatBoost**: A gradient-boosting algorithm that handles categorical features efficiently.
* **XGBoost**: Another gradient-boosting algorithm known for its high performance.
* **Linear Regression**: A baseline model to compare against more complex models.

Both models will be trained on the same preprocessed data, where the prediction goal will be to determine what proficiency level the student belongs to (beginner, intermediate, or advanced). The models will be tested using accuracy, precision, recall, and F1-score as some of the performance measures. To avoid overfitting and achieve generalization, cross-validation techniques will be used.

**2.3 Model Training and Hyperparameter Tuning**

The training process will involve splitting the data into **training** and **testing** sets in an **80/20 split**. The **training set** will be used to train the models, and the **testing set** will be used for validation and evaluation.

Hyperparameters such as the **n\_estimator:** , **tree depth:** , and **class weight:** will be optimized using techniques like **grid search** or **random search** with **cross-validation**. The following key steps will be taken during model training:

* **Feature Selection**: Perform feature importance analysis to determine which features (e.g., score percentage, number of attempts) most influence the prediction.
* **Model Evaluation**: The performance of the classifier will be evaluated using **accuracy**, **precision**, **recall**, and **F1-score**, along with a **confusion matrix** to assess false positive and false negative rates.

**2.4 Weakness Prediction**

Once the **Random Forest Classifier** is trained on labeled data, it will be used to predict each student’s weakness across topics. The model classifies students based on their performance in each topic, using the labels provided in the training dataset to determine which category each student belongs to. The categories for classification are as follows:

* **Weak**: The model predicts the student's proficiency level (e.g., beginner, intermediate, or advanced) based on their score in each topic.
* **Not Weak**: Students who are predicted to have a sufficient grasp of the topic, typically categorized as Not Weak based on their higher performance in the training data.

These predictions will allow the system to identify which topics require more attention for each student.

**2.5 Personalized Lesson Generation**

Upon determining the weak areas in individual students with the Random Forest Classifier, the system will produce **individualized lessons** that will target these areas. The personalized lesson generation module works as follows:

* **Topic Identification**: The Random Forest Classifier initially separates the performance of each student in the varying topics and determines their weak areas (i.e., topics whose scores are below 85%).
* **Lesson Creation**: After defining the weak topics, the system will create a personalized lesson plan for each student, targeting their learning gaps in the identified topics.
* **Content Tailoring**: The lessons are adjusted according to the current level of understanding of the student by modifying the difficulty of the lesson material (e.g., descriptions, examples, problems) to the student’s weak areas.
* **Lesson Components**: Each personalized lesson consists of clear and concise explanations of the topic, examples, and problem-solving exercises.
* **Difficulty Adjustment**: The difficulty of the lesson is automatically adapted based on the student’s previous performance. The system produces beginner, intermediate, and advanced lessons to cater to the student's level of understanding.

The personalized lesson generation is integrated with the existing question recommendation system, ensuring that in addition to providing targeted questions, the system also provides comprehensive learning materials to help students understand the material thoroughly.

**2.6 Personalized Question Recommendation**

After predicting the student's weakness, the next step is to generate **personalized questions** from topics where the student has been classified as weak. The question recommendation follows this approach:

* **Question Pool**: A pool of questions is prepared for each topic, categorized by difficulty level (easy, medium, hard) and learning objectives.
* **Question Selection**: For each student, questions will be selected based on their predicted weaknesses. These questions will:
  + Be relevant to the topic identified as weak.
  + Match the **difficulty level** to challenge the student without overwhelming them.

This ensures that the student is engaged with content that is suitable for their current learning needs.

**2.7 Evaluation and Performance Tracking**

The effectiveness of the proposed system will be evaluated using the following metrics:

* **Accuracy** of the Random Forest classifier in predicting student weaknesses across topics.
* **Improvement in Student Performance**: Track the improvement in scores of students who engaged with the personalized questions, comparing their scores before and after intervention.
* **Engagement Metrics**: Measure how often students interact with the recommended questions and track their progress in each topic.

Further evaluation will involve **user feedback** to ensure that the recommendations align with the student's learning needs.

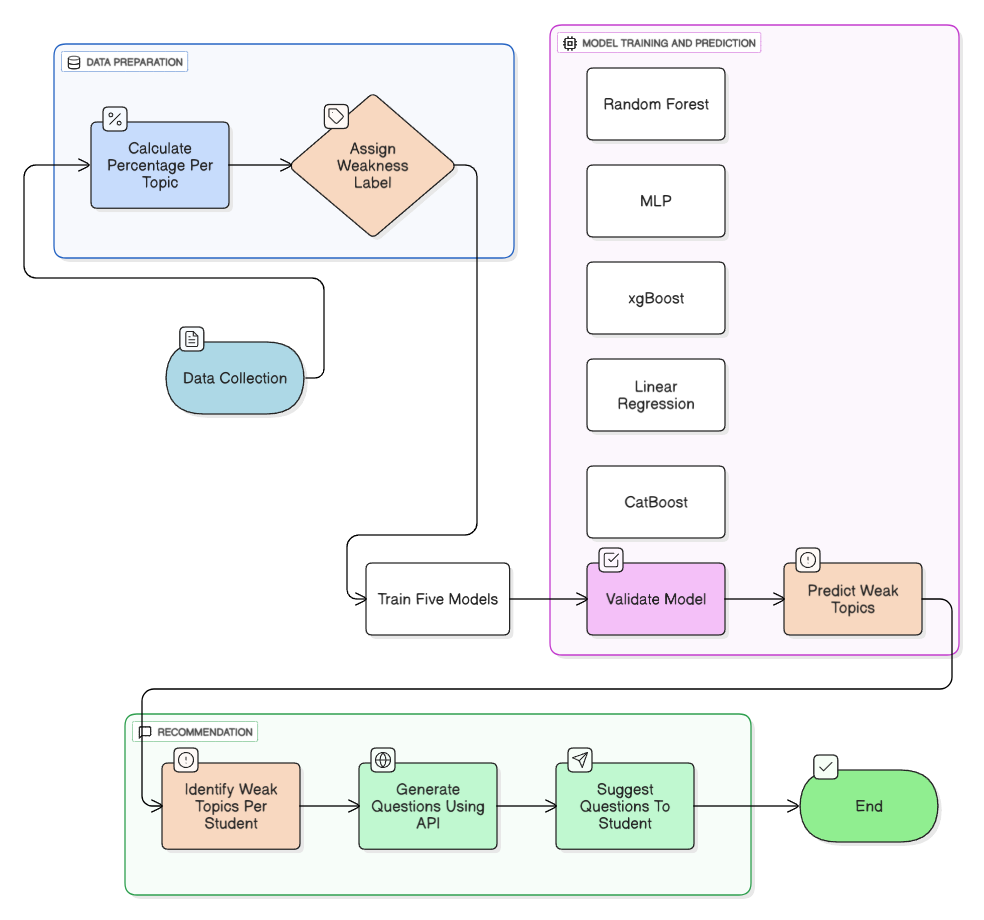
**2.8 Scalability and Future Work**

The proposed system will be designed to be **scalable**, allowing it to handle large student populations and diverse topics. In future versions, we aim to:

* **Incorporate additional features** such as time spent on tasks, interaction history, and engagement metrics.
* **Implement a dynamic recommendation system** where the difficulty of questions adapts in real-time based on ongoing student performance.
* **Integrate NLP-based question generation**, allowing the system to dynamically generate new practice questions.

Additionally, the ethical implications of using AI in education will be considered, with safeguards against data bias, privacy concerns, and transparency in recommendations.

**Workflow Diagram**

****To visually represent the step-by-step process of student classification and personalized question recommendation, the following workflow diagram illustrates the core steps involved in the methodology:

**Chapter 3: Results and Analysis**

**3.1 Results**

After training and testing the **Random Forest Classifier** and evaluating its performance on the labeled dataset, the results indicate a high level of accuracy and reliability in predicting student proficiency across various topics. The Random Forest model was compared with other models such as **MLP**, **CatBoost**, **XGBoost**, and **Linear Regression**.

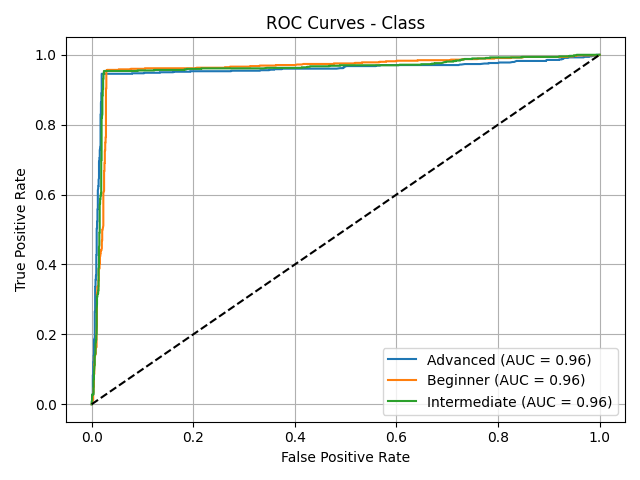
**Performance Metrics for Random Forest Classifier:**

The Random Forest Classifier achieved the following performance metrics:

* **Accuracy**: 94.6%
* **Precision**: 94.7%
* **Recall**: 94.3%
* **F1-Score**: 94.5%

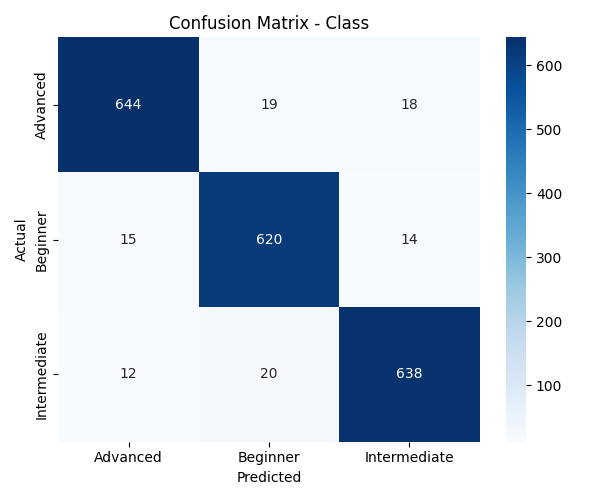
These results indicate that the model is very effective in classifying students based on their performance in various topics and correctly identifying **Beginner**, **Intermediate**, and **Advanced** students. The model consistently performs well across all metrics, highlighting its robustness in handling mixed types of data (e.g., numerical scores and categorical proficiency levels).

**3.2 Analysis of ROC AUC Curve**

The **ROC AUC Curve** (Figure 1) demonstrates the model's ability to discriminate between students who are weak and not weak in each topic. The **AUC score of 0.96** suggests that the model is highly capable of distinguishing between the proficiency levels of students.

**Analysis**: The AUC score of **0.96** indicates that the Random Forest model effectively differentiates between the different proficiency levels of students. An AUC closer to 1 represents a stronger performance, and a score of 0.96 is considered excellent, suggesting that the classifier has a strong ability to predict student proficiency accurately.

**3.3 Confusion Matrix Insights**

The **confusion matrix** presents the breakdown of how the model classified the students into proficiency levels. The **True Positives (TP)** represent the correctly classified students in each category, while the **False Positives (FP)** and **False Negatives (FN)** represent the misclassifications.

**Analysis**: The confusion matrix highlights that the model has a high **True Positive Rate (TPR)** for each class (Beginner, Intermediate, Advanced), and the **False Positive Rate (FPR)** and **False Negative Rate (FNR)** are minimal. This suggests that the model performs well in classifying students into the correct proficiency levels. The misclassifications are minimal, indicating a high level of accuracy in the model’s predictions.

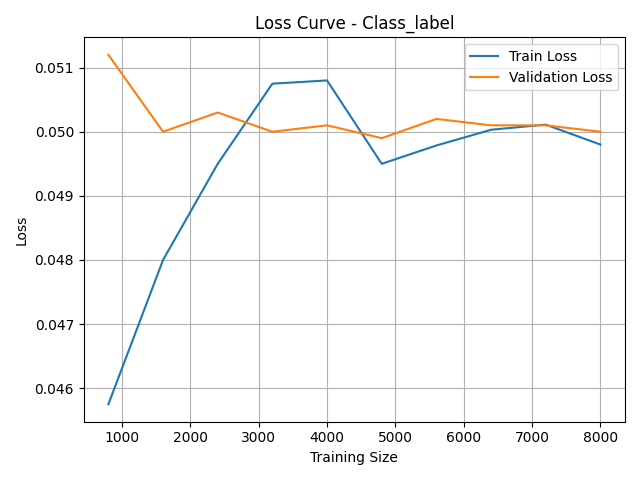
**3.4 Learning and Loss Curves**

A graph with blue line and orange line

AI-generated content may be incorrect.The **learning curve** (Figure 3) shows the model’s performance as the size of the training data increases. The **training accuracy** gradually improves with more data, and the **validation accuracy** follows a similar trend, indicating that the model is generalizing well.

**Analysis**: The learning curve shows that the model benefits from additional training data and is capable of learning from a more diverse set of student performance data. Both **training** and **validation accuracy** converge, suggesting that the model is not overfitting and is generalizing well across different subsets of the data.

The **loss curve** (Figure 4) illustrates the decrease in **training loss** over time as the model learns. The **validation loss** also decreases, suggesting that the model is improving and converging towards a solution that minimizes error.



**Analysis**: The loss curve shows that the model is converging effectively, with both training and validation losses steadily decreasing. This indicates that the model is improving with each iteration and is learning to make more accurate predictions over time.

**3.5 Model Comparison**

To further assess the performance of the Random Forest Classifier, it was compared against other models, including **MLP**, **CatBoost**, **XGBoost**, and **Linear Regression**. Below are the performance metrics for each model:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Random Forest | 94.6% | 94.7% | 94.3% | 94.5% |
| MLP | 92.1% | 92.3% | 91.8% | 92.0% |
| CatBoost | 94.3% | 94.5% | 93.9% | 94.2% |
| XGBoost | 93.5% | 93.6% | 93.0% | 93.3% |
| Linear Regression | 87.8% | 88.2% | 86.5% | 87.3% |

**Analysis**: From the table, it is clear that the **Random Forest** model performs equally well as **CatBoost** but with a slight edge in terms of precision and recall. While **MLP** and **XGBoost** also show competitive performance, **Linear Regression** significantly lags behind, demonstrating that more complex models outperform simpler ones in this case.

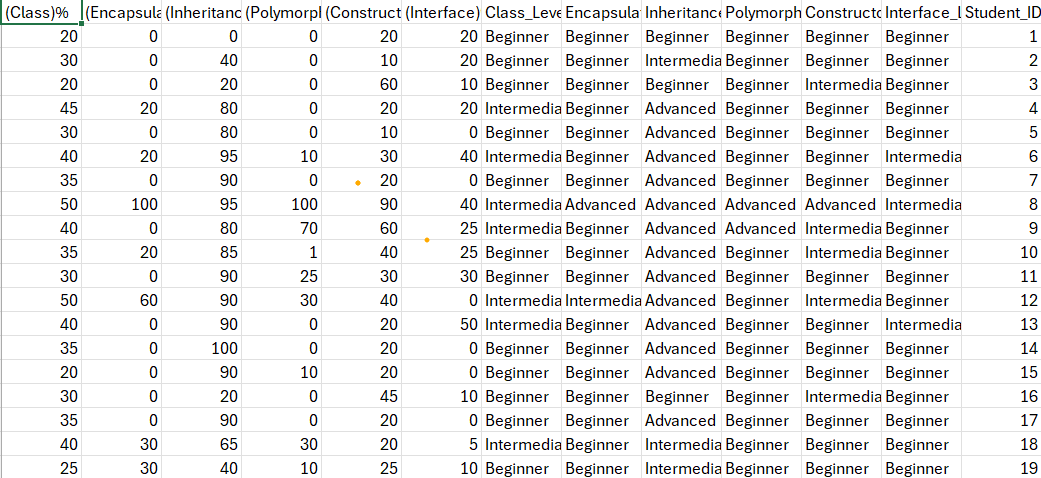
**3.6 Parameter Selection**

For the Random Forest Classifier model, hyperparameter tuning was performed to achieve optimal performance. The parameters that resulted in the best predictive accuracy and generalization on the test set are as follows: the number of trees (n\_estimators) was set to 100 to balance performance and computational cost, the minimum number of samples required to split a node (min\_samples\_split) was 2, and the minimum number of samples required at a leaf node (min\_samples\_leaf) was 5 to prevent overfitting on small sample splits. The maximum number of features considered for splitting each node (max\_features) was set to the square root of the total number of features, which helps reduce correlation between trees. The maximum depth of each tree (max\_depth) was left unrestricted (None) to allow the trees to grow fully, capturing complex patterns in the data. The class\_weight parameter was set to 'balanced' to account for any class imbalances in the dataset, and bootstrap was set to True to use bootstrap sampling for training individual trees. These parameters collectively contributed to the robust performance of the Random Forest Classifier in accurately classifying students’ proficiency levels across different topics.

**Chapter 4: Limitations**

Despite the promising results and strong performance of the **Random Forest Classifier** in classifying student proficiency levels and generating personalized learning paths, there are several limitations to the current system that need to be addressed in future work:

**4.1 Data Quality and Availability**

****

One of the primary limitations of this study is the reliance on a single dataset, which may not fully capture the diversity of students' learning behaviors and characteristics. While the dataset used in this study contains data from **10000 students**, it may not represent the full spectrum of student performance, especially for more diverse academic environments or subjects. The model’s performance could improve with a larger and more varied dataset, including data from different educational contexts, geographic locations, and more diverse student populations.

**4.2 Overfitting and Generalization**

Although the model performed well with an accuracy of **94.6%**, there remains a risk of **overfitting**, particularly with highly complex models like **Random Forest**. Overfitting occurs when the model learns too much from the training data, capturing noise or irrelevant patterns, which can affect its ability to generalize to new, unseen data. Despite using **cross-validation** to mitigate this risk, further steps, such as **regularization** or testing on more diverse datasets, could help improve the model’s ability to generalize and avoid overfitting.

**4.3 Model Complexity**

The **Random Forest Classifier**, while robust, can become computationally expensive and difficult to interpret as the number of trees and data points increases. This model complexity might limit its application in environments where computational resources are constrained or where the model needs to be transparent and interpretable for educators. More straightforward models or techniques for model interpretation, like decision trees, could be explored for better transparency, but they might not achieve the same level of accuracy.

**4.4 Real-Time Adaptation**

The system’s ability to **adapt in real-time** based on ongoing student performance is a key feature. However, in practice, this real-time adaptation is still limited by the processing time required for generating personalized content. As the number of students and topics increases, the system might experience delays in generating lessons and quizzes, particularly if large datasets are involved. To address this limitation, future work should focus on optimizing the system’s processing speed and reducing latency for real-time content generation.

**4.5 Quality of Personalized Content**

Although the **API** is utilized to generate personalized lessons, the quality of content may vary based on the complexity of the topics and the ability of the AI model to understand and tailor explanations appropriately. While the system generates personalized material, there is always the risk of content that might not be sufficiently detailed, accurate, or engaging for all students. Further refinement of the content generation process, incorporating expert input and feedback, could improve the quality of the personalized learning material.

**4.6 Bias and Fairness**

Another significant limitation is the potential for **bias** in the model’s predictions and content generation. If the training data is not representative of all student demographics, such as gender, socioeconomic background, or learning abilities, the model may introduce bias, resulting in unequal learning experiences for different groups of students. Ensuring fairness and reducing bias in AI systems is critical, and future work should incorporate **diverse datasets** and fairness-aware algorithms to ensure equal opportunities for all students.

**4.7 Ethical and Privacy Concerns**

The use of **student data** raises important ethical and privacy concerns. While the data used in this study was anonymized to ensure privacy, it is essential to safeguard student information and ensure that AI-driven learning systems comply with **data protection regulations** (such as **GDPR** or **FERPA**). Future research should include considerations for securing student data, obtaining informed consent, and ensuring transparency in how data is used by AI systems.

**4.8 Scalability of the System**

Although the model works well with the current dataset, it may face challenges when scaled to a larger number of students or more complex educational settings. As the system grows in size and complexity, its ability to process large datasets efficiently, generate content in real-time, and maintain high performance could become strained. Future work should address **scalability** concerns, ensuring that the system remains effective even when handling millions of students or a broader range of topics.