Adaptive Minds: Machine Learning Approaches to Predict Online Learning Adaptability

Group 13

Vimukthi Ranathunga – S16058

Danuja Fernando – S16024

Lasandi Perera – S16388

Abstract

This study investigates factors influencing students' adaptability to online education using a dataset containing demographic, technological, and contextual variables such as gender, age, internet type, device used, and financial condition. Through data preprocessing, visualization, and predictive modeling, patterns are identified that reveal how access to technology, learning environment, and personal circumstances affect adaptability levels. The results highlight the critical role of digital readiness and resource availability in shaping students' success in virtual learning environments. These insights can support educators and policymakers in enhancing online education strategies for diverse student populations.

Table of Contents

Table of Figures

Introduction

This study examines students' adaptability to online education, focusing on the factors that influence their engagement and success in virtual learning. As digital learning becomes widespread, understanding how students respond to online education is crucial. Using a structured dataset of demographic, socioeconomic, and technological variables, we analyze the impact of internet connectivity, device availability, financial condition, and institutional support on adaptability. The findings provide insights for educators and policymakers to enhance online learning environments and support students in developing the resilience and flexibility needed for effective digital education.

Description of the question

This study investigates students' adaptability to online education, focusing on how demographic, socioeconomic, and technological factors influence their ability to engage and succeed in virtual learning environments. As online education becomes increasingly common across schools, colleges, and other learning institutions, understanding the determinants of student adaptability is crucial for designing effective digital learning experiences.

We focus on three main objectives:

1. **Analyze Patterns of Student Adaptability**
   We examine adaptability levels across different student groups based on factors such as age, gender, education level, institution type, and IT background. This analysis helps uncover trends in how students respond to online learning environments and identifies groups that may face challenges adapting.

2. **Identify Key Determinants Affecting Adaptability**
   We evaluate how technological and contextual factors such as internet type and network quality, device availability, class duration, self-learning platforms (LMS) usage, financial conditions, and load-shedding impact students' adaptability. These insights reveal the influence of both infrastructure and personal circumstances on online learning effectiveness.

3. **Build a Predictive Model for Adaptability Level**
   We develop machine learning models to classify students into adaptability categories (e.g., Low, Moderate, High) based on their demographic and contextual attributes. This approach identifies the most important factors affecting adaptability and provides a foundation for designing interventions, support programs, and policies to enhance digital readiness and equitable access to online education.

Description of the data set

The dataset, Students' Adaptability Level Prediction in Online Education, contains records describing student demographics, academic context, and technology-related conditions. The target variable is Adaptivity Level, which measures how well a student adjusts to online learning environments.

Features include:

- Demographics: Gender, Age, Financial Condition, Location in Town
- Educational Context: Education Level, Institution Type, IT Student, Self LMS availability
- Technology & Environment: Internet Type, Network Type, Load-shedding, Device used, Class Duration

Target Variable:

- Adaptivity Level: categorical measure of student adaptability (e.g., low, moderate, high).

Acknowledgement: his dataset is based on the research study Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches (DOI: 10.1109/ICCCNT51525.2021.9579741). Users are free to apply it for research and analysis, with proper citation (Hasan Suzan et al., 2021).

Descriptive Analysis

Initially in preprocessing stage we didn't find any missing values and outliers in this data set. However, we found 949 duplicates out of 1205 data entries. But we didn't remove them because this data set was gathered using a survey so in real world surveys, we can often encounter individuals who submit same details and also if we remove those duplicates that will cover the true patterns and our sample will not be representative of the population.

Cluster Analysis

The MCA method was first applied to reduce the dimensionality of the categorical dataset. The first two components explained approximately 25.5% of the total variance. The MCA plot showed no clear or well separated groups among data points instead overlapping patterns were visible across three Adaptivity Level Categories (Low, Medium, High). This illustrates that the student characteristics are mixed and distinct clusters are not visually separable in the lower dimensional space.

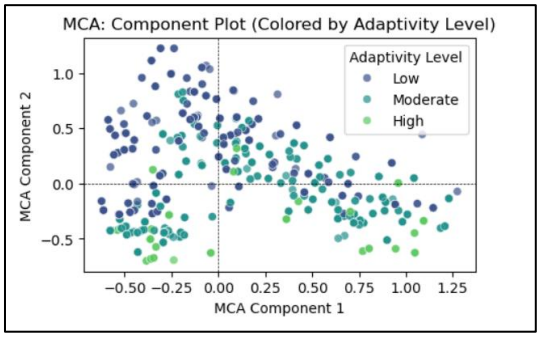


Figure 1: MCA component plot

The **K-Modes** method suggested two clusters, with Elbow method and Silhouette Score (~0.27) supporting this, though cluster separation was weak.

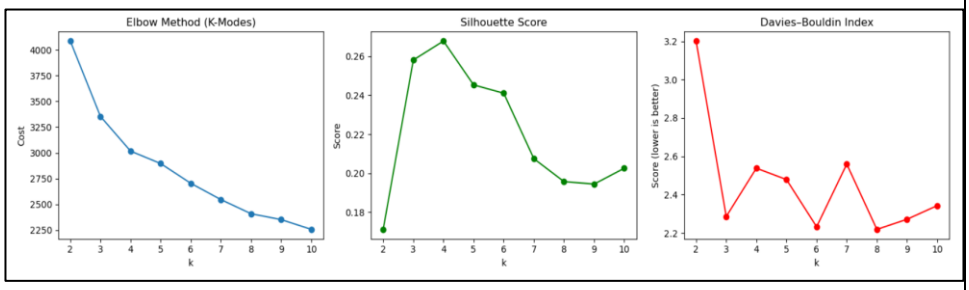


Figure2: Optimal cluster determination plots

Similarly, **Hierarchical Clustering** using Jaccard distance and complete linkage method. The resulting dendrogram showed gradual merging of clusters rather than the formation of distinct, well-separated groups. The silhouette Score of 0.2028 confirmed the weak cohesion among clusters.

The DBSCAN with Hamming Distance metric detected only one dense cluster, while the Latent Class Analysis using Gaussian Mixture Model identified three balanced clusters with 364, 306, and 294 observations in each group respectively. but a lower silhouette Score (0.18), showing weak cohesion. To better accommodate the mixed categorical nature of the data, Spectral Clustering was conducted using the Gower distance metric. This method converts pairwise dissimilarities into a similarity matrix before clustering. When set to form three clusters, the Spectral Clustering approach yielded a Silhouette Score of 0.2927, which was an improvement compared to the earlier methods. The t-SNE visualization provided a more structured distribution, showing partial separation among the three clusters, though overlaps remained visible. Cluster membership counts were 358, 320, and 286, respectively. This approach demonstrated that while clear separations did not exist, the Gower-based similarity captured subtle underlying patterns among students' adaptability levels.

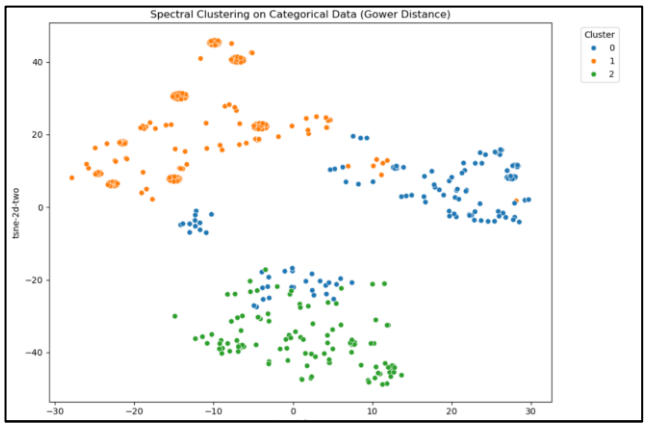
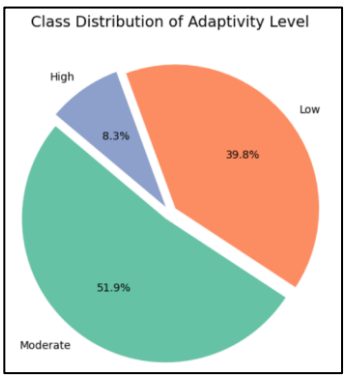


Figure3:Spectral Clustering plot

Overall, results showed that students' adaptability levels are not distinctly separable, as most clustering methods indicated overlapping characteristics.



The left plot illustrates the class distribution of Adaptivity Level among students. It can be observed that the majority of students fall under the Moderate adaptivity level (51.9%), followed by Low adaptivity (39.8%), while only 8.3% of students exhibit a High adaptivity level. This imbalance highlights that most students show moderate adaptability toward online education, with fewer demonstrating strong adaptability skills.

Figure 4: Distribution of adaptivity levels

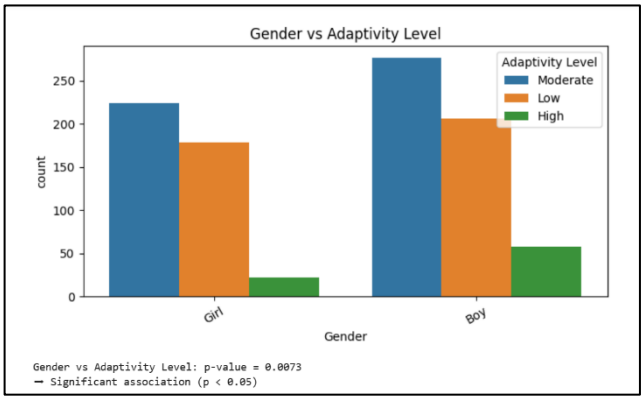


Figure 5: Gender vs Adaptivity Level

This grouped bar chart reveals the distribution of adaptivity levels across gender, with a chi-square test confirming a statistically significant association. While girls having lower adaptivity than in Moderate and in Low adaptivity category both genders have same adaptivity, a striking disparity emerges in High adaptivity where boys (~50) outnumber girls (~20) by 2.5 times. This gender gap suggests barriers such as unequal technology access, learning environment differences, or socio-cultural factors that hinder girls from achieving high adaptivity in online education. The findings highlight gender as a critical predictor variable for machine learning models and underscore the need for targeted interventions including gender-sensitive support programs, infrastructure improvements, mentorship initiatives, and policy measures to ensure equitable online education outcomes and prevent widening of existing gender gaps in educational achievement.

This grouped bar chart shows adaptivity level distribution across six age groups with a statistically significant association. The 11-15 and 21-25 age groups represent the largest cohorts, while all age groups having different distributions for adaptivity levels. A critical finding is that the 16-20 age group, despite moderate participation, shows predominantly Low (~120) and Moderate (~100) adaptivity with very low High adaptivity (~20), suggesting significant struggles during this transitional educational phase. Younger groups (1-5, 6-10) have minimal representation, while the 26-30 group shows lower participation and lower adaptivity when engaged. These patterns establish age as a crucial predictor variable and highlight the need for age-specific interventions, particularly targeted support for the 16-20 cohort through enhanced technological training, personalized learning resources, and adaptive pedagogical strategies to address their unique challenges in online education environments.

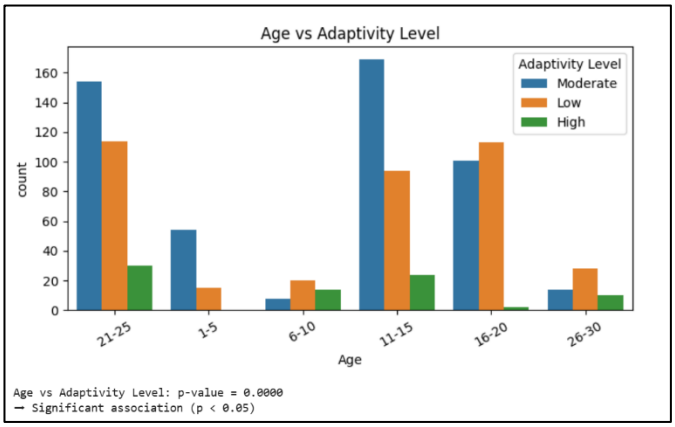


Figure 6 :Age vs Adaptivity Level

Below grouped bar chart shows adaptivity level distribution across three education levels with a statistically significant association. University and School students demonstrate relatively have higher adaptivity. In stark contrast, College students show dramatically lower participation (~75 Moderate, ~80 Low) with virtually zero High adaptivity, revealing severe adaptation challenges at this level. This disparity likely stems from inadequate institutional infrastructure, curriculum unsuited for online delivery, lack of support services, or demographic factors such as working students with limited resources. Education level emerges as a critical predictor variable, and the findings highlight urgent need for college-specific interventions including infrastructure upgrades, faculty training in online pedagogy, enhanced student support, flexible schedules, and policy reforms to ensure educational equity across all academic levels.

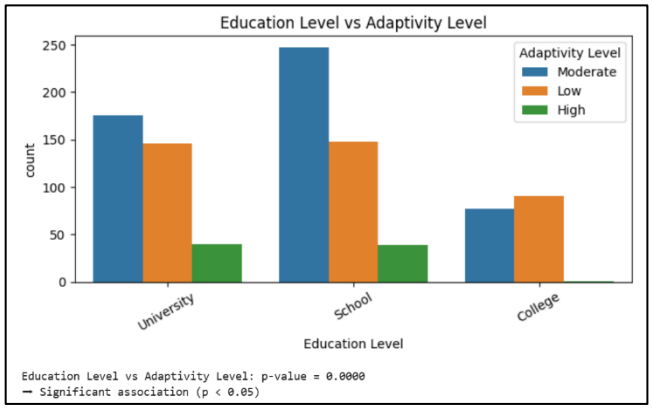


Figure 7: Education Level vs Adaptivity Level

Below grouped bar chart reveals adaptivity level distribution across institution types with a statistically significant association. Non-Government institutions demonstrate substantially higher participation and balanced distribution than others (~400 Moderate, ~200 Low, ~60 High), indicating strong online learning infrastructure and preparedness. Conversely, Government institutions show significantly lower engagement with Low adaptivity (~200) dominating at more than double the rate of Moderate (~95) and High (~10), exposing a critical public-private equity gap. This disparity likely stems from inadequate technological infrastructure, limited internet access, insufficient faculty training, and resource constraints in government institutions. Institution type emerges as a vital predictor variable, underscoring urgent need for targeted public sector interventions including increased digital infrastructure funding, student device and internet subsidies, faculty development programs, robust LMS implementation, and policy reforms to ensure equitable online education quality across all institutional types.

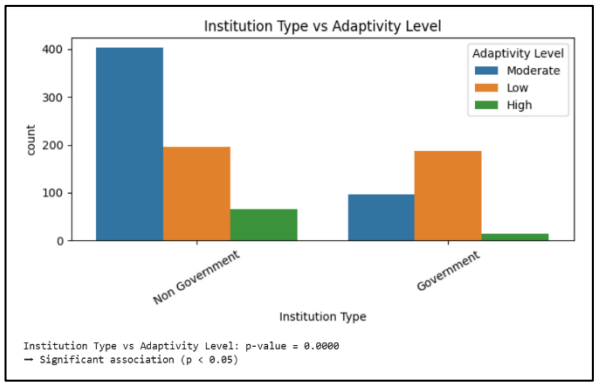


Figure 8: Institution vs Adaptivity Level

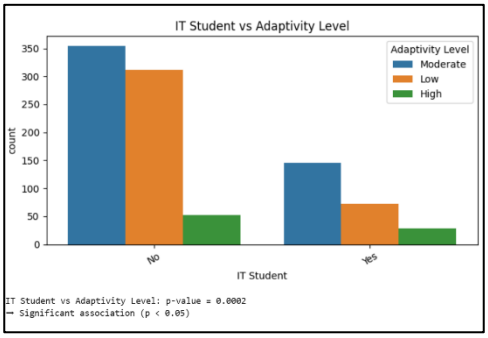


Figure 9: IT student vs Adaptivity Level

This grouped bar chart shows adaptivity level distribution between IT and non-IT students with a statistically significant association. Non-IT students, representing the majority, demonstrate a problematic pattern with Low adaptivity (~300) little less than Moderate (~350) and High (~50), indicating substantial struggles with online learning platforms. Conversely, IT students show superior proportion of adaptation with moderate adaptivity (~140) as the predominant category, exceeding low (~85) and high (~40), revealing that technological literacy provides significant advantages in navigating digital education tools and LMS platforms. IT student status emerges as a powerful predictor variable, and findings underscore urgent need for digital literacy training programs, technology orientation workshops, user-friendly platform interfaces, dedicated technical support, and foundational ICT skill development for non-IT students to bridge the adaptivity gap across academic disciplines.

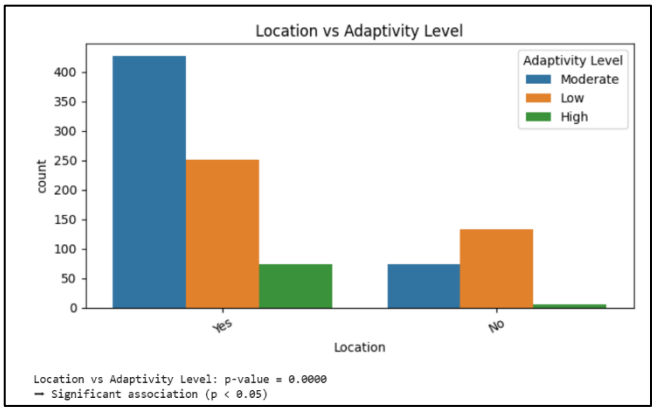


Figure 10: Financial Condition vs Adaptivity Level

This grouped bar chart reveals adaptivity level distribution based on location (town vs rural) with a statistically significant association. Town/urban students ("Yes") demonstrate substantially higher participation with Moderate adaptivity (~450) predominating over High (~50) and Low (~250), indicating access to superior infrastructure and conducive learning environments. Rural students ("No") show drastically limited engagement with Low adaptivity (~150) dominating at over double the rates of Moderate (~75) and High (~10), exposing a severe urban-rural digital divide. This disparity stems from poor internet connectivity, power outages, device scarcity, and inadequate learning spaces in rural areas. Location emerges as a critical predictor variable, underscoring urgent need for rural interventions including broadband expansion, internet subsidies, community learning centers, offline resources, device distribution programs, and policy reforms to ensure geographic equity in online education accessibility.

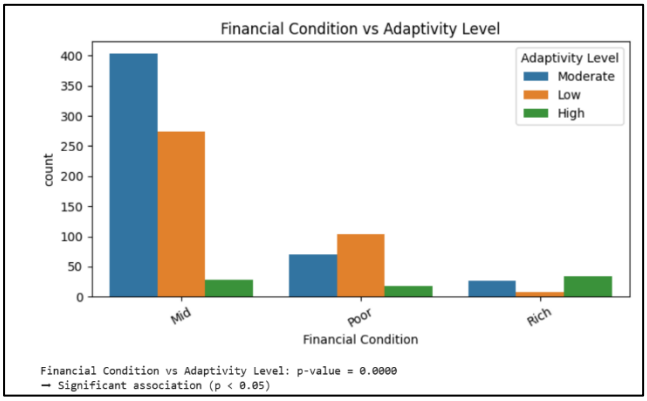


Figure 11: Financial Condition vs Adaptivity Level

This grouped bar chart shows adaptivity level distribution across financial conditions (Mid, Poor, Rich) with a statistically significant association. Mid-income students form the largest group with predominantly Moderate (~400) and Low (~270) adaptivity versus lower High (~40), indicating financial constraints limit optimal learning resources. Poor students show concerning patterns with Low adaptivity (~100) being most common among Moderate

(~70) and High (~20), reflecting barriers from inadequate devices, unstable internet, and unsuitable learning environments. Rich students demonstrate dramatically superior adaptation with High adaptivity (~45) vastly exceeding Moderate (~30) and Low (~5), revealing that financial privilege directly enables better online learning through premium devices, high-speed connectivity, and supplementary resources.

Advanced Analysis

After the exploratory data analysis, we did SMOTE resampling before move on to the advanced analysis. Since our response variable has high class imbalance. After balancing data classes, we did our advanced analysis.

Modeling: Neural Networks and Support Vector Machines

Neural Network Models

1. Simple MLP (Multi-Layer Perceptron)

Architecture: Single hidden layer with 64 neurons and ReLU activation, followed by a 3-class softmax output layer.

Performance: Achieved 73% accuracy with balanced performance across classes. Precision ranged from 0.70-0.86, with Class 2 (High adaptivity) showing lower recall (0.30), indicating difficulty identifying this minority class. The model demonstrates baseline capability but struggles with class imbalance.

2. Deep MLP

Architecture: Three hidden layers (128, 64, 32 neurons) with ReLU activation, creating deeper feature representations.

Performance: Achieved 90% accuracy, representing the best neural network performance. All classes showed excellent metrics precision (0.87-0.94), recall (0.80-0.94), and F1-scores (0.86-0.90). The deeper architecture successfully captured complex patterns and handled class imbalance effectively, making it the optimal neural network choice for this task.

3. Regularized MLP

Architecture: Two hidden layers (128, 64 neurons) with L2 regularization (0.001) and dropout (0.4, 0.3) to prevent overfitting.

Performance: Achieved 77% accuracy, a moderate improvement over Simple MLP but underperforming compared to Deep MLP. While regularization prevented overfitting, it may have been too aggressive, limiting the model's capacity to learn complex patterns. Class 2 recall (0.55) remained problematic.

Support Vector Machine Models

4. Linear SVM

Configuration: Linear kernel with default parameters.

Performance: Achieved 63% accuracy with poor performance. Failed completely on Class 2 (precision and recall both 0.00), indicating linear separability assumptions were violated. The model cannot capture non-linear relationships critical for this dataset.

5. RBF SVM

Configuration: Radial Basis Function kernel with default gamma='scale'.

Performance: Achieved 74% accuracy, substantial improvement over Linear SVM. However, Class 2 showed extremely low recall (0.15) despite perfect precision (1.00), indicating the model was overly conservative in predicting the minority class.

6. Tuned RBF SVM (GridSearchCV)

Configuration: Optimized hyperparameters (C=10, gamma=1) through 5-fold cross-validation.

Performance: Achieved 92% accuracy, the best overall model performance. Demonstrated excellent metrics across all classes' precision (0.81-0.93), recall (0.85-0.94), and F1-scores (0.83-0.93). The hyperparameter tuning successfully balanced model complexity and generalization, effectively handling the minority class challenge.

7. Polynomial SVM

Configuration: Polynomial kernel (degree=3, C=1).

Performance: Achieved 77% accuracy with moderate performance. While better than linear kernel, it underperformed compared to optimized RBF, suggesting polynomial relationships don't capture the data structure as effectively as RBF transformations.

Random Forest Model

Initially, without hyperparameter tunning we achieved Training Accuracy: 0.9413, Test Accuracy: 0.9004 from random forest model. After hyper parameter tunning we got 0.9044 test accuracy. That means model fit data well.

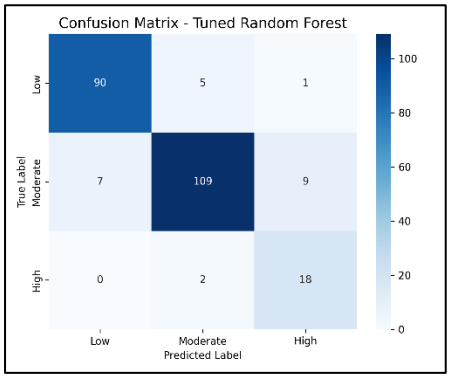


Figure 12: Confusion Matrix RF

Boosting Models

The predictive performance of four different tree-based and ensemble models XGBoost, LightGBM, CatBoost, and AdaBoost was evaluated to classify students' adaptability levels to online education into three categories: Low, Moderate, and High. The

models were trained using a dataset containing 12 demographic and contextual features, including gender, age, education level, IT background, internet type, network quality, class duration, LMS usage, and device availability.

1. XGBoost

XGBoost achieved a **training accuracy of 94.13%** and a **test accuracy of 90.04%**, indicating strong generalization without overfitting.

- **Strengths:** High precision and recall for the *Low* and *Moderate* adaptability classes (F1-scores of 0.93 and 0.90).
- **Limitations:** The *High* adaptability group showed lower precision (0.64) but relatively strong recall (0.90), meaning the model successfully identified most highly adaptive students but with a few false positives.

Overall, XGBoost effectively captured underlying adaptability patterns, performing robustly across major classes.

2. LightGBM

LightGBM achieved the **highest test accuracy of 90.87%**, slightly outperforming the other ensemble models.

- **Strengths:** Consistently high precision and recall across all categories, including *High* adaptability (precision 0.69, recall 0.90).
- **Interpretation:** This demonstrates LightGBM's strong capability to manage class imbalance while preserving overall predictive performance.

LightGBM proved to be the most balanced and accurate model for predicting students' adaptability levels.

3. CatBoost

CatBoost achieved a **test accuracy of 90.87%**, closely matching LightGBM's performance.

- **Strengths:** Comparable accuracy and F1-scores to XGBoost and LightGBM across *Low* and *Moderate* adaptability groups.
- **Interpretation:** CatBoost handled categorical features efficiently and maintained strong recall for all classes, especially in smaller categories.

CatBoost is a reliable model choice for multi-class adaptability prediction with mixed feature types.

4. AdaBoost

AdaBoost performed noticeably lower, with a **test accuracy of 63.9%**.

- **Performance:** The *Moderate* adaptability group showed moderate predictive strength (F1-score 0.72), but the *Low* and *High* categories had poor recall and precision.
- **Limitation:** AdaBoost struggled to capture complex, nonlinear relationships and handle the multi-class imbalance present in the dataset.

Thus, AdaBoost is less suitable for this problem compared to boosting algorithms like LightGBM, XGBoost, and CatBoost.

Conclusion

Among all evaluated models, the **Tuned RBF SVM** achieved the **highest overall accuracy (92%)**, demonstrating excellent precision, recall, and F1-scores across all adaptability levels. It effectively captured complex nonlinear relationships and handled class imbalance after hyperparameter tuning, making it the strongest individual performer.

However, **LightGBM** also delivered exceptional results with a **test accuracy of 90.87%**, maintaining consistently high precision and recall across all classes including the minority *High* adaptability group. Unlike SVM, LightGBM offers **greater interpretability**, computational efficiency, and scalability to larger datasets, while providing **feature importance insights** that enhance educational understanding and decision-making.

Considering both predictive power and interpretability, **LightGBM** stands out as the **most practical and explainable model** for predicting students' adaptability to online education, whereas the **tuned RBF SVM** represents the **best pure predictive performer**.

Appendix

Python code : You may access python codes here.