Exploratory Data Analysis and Machine Learning Model Evaluation: Understanding Student Flexibility in Online Learning.

Umanga Niroula
Email: umanganiroula83@gmail.com

***Abstract*—The COVID-19 pandemic has triggered major changes in education, which has caused a shift towards online learning. Further, understanding student adaptability in such context is very necessary for enhancing the online learning experience. As a result, this study utilizes exploratory data analysis (EDA) and machine learning model evaluation to investigate factors influencing student flexibility in online education. The dataset contains demographic and technological attributes, providing insights into adaptability levels. Machine learning algorithms are trained and assessed to predict flexibility levels, with the RandomForestClassifier becoming the most effective model. The findings highlight the complicated relationship between demographic trends, technological preferences, and model performance in optimizing student engagement.**

Keywords: COVID-19, demographic features, technological characteristic, student adaptability, exploratory data analysis, machine learning, RandomForestClassifier.

I. Introduction

In recent years, the landscape of education has had drastic change with the emergence of e-learning (Pulham Graham, 2018). COVID-19 pandemic is one of the major reasons which accelerated e-learning, as the educational institute required an alternative towards physical classes (Hodges et al., 2020). Institutes worldwide started viewing the e-learning platforms as the primary method of learning which brings forth a question: how can students thrive in this digital learning environment?

Understanding and predicting students' adaptability levels in online education is important for several reasons. Firstly, it allows educators and institutions to make learning experiences such to meet the diverse needs of students, ensuring that each student can engage effectively regardless of their curriculum (Ally, 2008). Moreover, by identifying students who may struggle with the e-learning platform, various strategies can be implemented to provide additional support and resources, ultimately providing productive and fair educational environment (Artino et al., 2011).

Machine learning techniques offer a promising approach for addressing this challenge. By using vast amounts of data on student demographics, behaviors, and learning outcomes, machine learning models can uncover hidden patterns and insights that traditional methods may overlook (Kotsiantis et al., 2007). These models have the potential to not only predict students' adaptability levels with a high degree of accuracy but also provide actionable recommendations for personalized interventions, thereby optimizing the online learning experience for all stakeholders (Hastie et al., 2009).

In this study, an exploration of machine learning approaches for predicting students' adaptability levels in online education is undertaken. Through a complex analysis of the dataset containing various student attributes and preferences, the aim is to highlight factors that influence students' ability to thrive in digital learning environments. By uncovering these insights, contributions to the ongoing efforts to uplift the effectiveness of online education for learners worldwide are sought.

II. Problem and Data Set Description

A. *Problem Statement*

In the wake of the COVID-19 pandemic, educational institutions worldwide have been compelled to swiftly transition to online learning platforms. This rapid shift has caused significant questions regarding how students are adapting to this new mode of education delivery. Likewise, there grows concern in need of understanding the perspective of students around such adaptation. Similarly, what changes is this going to bring in the mode of learning will it be positive or negative, what effects it will have in the shaping the future of the world, can it be enhanced to make the result more positive by identifying its weakness and working on them? Hence, understanding the verity of complexity for student's adaptability around online learning environments is important for educational institutions seeking to optimize their instructional strategies. By gaining insights into the various factors influencing student flexibility, schools can utilize various strategies to meet the needs of their student. From demographic characteristics (age, gender, education level, and location) to technological preferences (various tools, devices, and platforms), each aspect plays a crucial role in shaping students' experiences and outcomes in the online learning landscape.

The dataset analysis helps decision-making in online education. It helps institutions create fair strategies and courses based on students' backgrounds, ensuring equal accessibility for all. Additionally, it assists in resource allocation by understanding students' preferences for digital tools and internet access. Predictive models identify students needing extra support, like personalized tutors. Valuable insights are gained by spotting patterns between factors and students' adaptability levels, for example providing extra resources towards targeted students in regards to their need and convince can help dealing with hard times and not being deprived to quality education. Understanding these factors guides the design of interactive

and highly efficient online learning environments, benefiting all students. Moreover, the goal is to help schools improve online learning by understanding students' adaptability. With this insight, schools can create better learning experiences that empower all students to succeed in digital classrooms. By focusing on adaptability and efficiency, education can provide equal learning opportunities for everyone.

B. Data Set Description

The dataset used in this analysis, sourced from Kaggle, includes demographic details like education level, gender, and age, along with technological factors such as device and internet usage. It contains 1205 entries, each representing a different student profile. Further, upon analysis the dataset shows no missing values, indicating it's ready for analysis. Demographic variables like education level and gender display diversity, reflecting the student population's heterogeneity. Technological factors like device and internet usage offer insights into students' digital engagement with online learning platforms. Table I shows all the features and their type along with the categorical value range.

TABLE I: Features Description

No	Description	Type	Categorical Value Range
1	Education Level	Categorical	School, College, University
2	Institution Type	Categorical	Public, Private
3	Gender	Categorical	Male, Female
4	Age	Numerical	
5	Device	Categorical	Mobile, Tab, Computer
6	IT Student	Categorical	Yes, No
7	Location	Categorical	Town, Rural
8	Financial Condition	Categorical	Low, Mid, High
9	Internet Type	Categorical	Wifi, Mobile Data
10	Network Type	Categorical	3G, 4G, 2G
11	Flexibility Level	Categorical	Low, Moderate, High

To facilitate analysis, the age column was converted to a categorical variable, ensuring uniformity in data representation. Descriptive statistics reveal the distribution and frequency of categorical variables within the dataset, offering initial insights into the demographic and technological landscape of online learning participants.

Table II provides all the insights into demographics features such as the highest age group is 23, most of the student's educational level is school and most of them are private institutes out of three other types. Similarly, the technological features like mobile phone is the most wide used device type, who are not from IT background and are utilizing mobile data for internet connectivity.

Similarly, checking for the correlation of the features interesting insights into the factors influencing students' perspective on flexibility in online learning can be seen. Firstly, while being an IT student shows a slight positive correlation with flexibility level, suggesting IT students are more flexible as they have knowledge of online platforms, hence this insight can help institutes in incorporating basic IT skill classes to students in order to make them more flexible. Hence, such

TABLE II: Features Description

Feature	Count	Unique	Top	Frequency
Education Level	1205	3	School	530
Institution Type	1205	2	Private	823
Gender	1205	2	Male	663
Age	1205	6	23	374
Device	1205	3	Mobile	1013
IT Student	1205	2	No	901
Location	1205	2	Town	935
Financial Condition	1205	3	Mid	878
Internet Type	1205	2	Mobile Data	695
Network Type	1205	3	4G	775
Flexibility Level	1205	3	Moderate	625

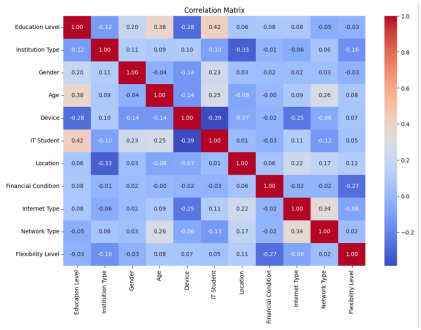

Fig. 1: Correlation of all Features.

kind of information can be derived from figure 1 of correlation between each of the 11 features.

III. METHODS

A. Data Pre-processing

In the data preprocessing stage, a technique known as Ordinal Encoding was applied to convert categorical variables into numerical representations. This process ensures compatibility with machine learning models (Pedregosa et al., 2011). For example, consider a categorical variable such as "Education Level" with categories like "High School," "Bachelor's Degree," and "Master's Degree." Through Ordinal Encoding, each category is assigned a unique numerical value. For instance, "High School" might be represented as 0, "Bachelor's Degree" as 1, and "Master's Degree" as 2. This encoding facilitates the interpretation of categorical data by machine learning algorithms, which typically require numerical input. By preprocessing the data in this manner, a dataset is created that can be effectively analyzed and utilized by machine learning models to extract valuable insights (Pedregosa et al., 2011). Following is the top 4 records of the dataset after the conversion of the data.

B. Model Exploration

In the phase of model exploration, five of classification algorithms was utilized to identify the most suitable approach for predicting student flexibility levels. The algorithms explored encompassed RandomForestClassifier, KNeighborsClassifier, Support Vector Classifier (SVC), Logistic Regression, and XGBoostClassifier. For each of these algorithms, the process

involved training the model using the training dataset, which comprises a subset of the available data. This training phase allows the model to learn the underlying patterns and relationships within the data (James et al., 2013). Following the training phase, the performance of each model was assessed using the testing dataset, which serves as an independent measure to evaluate the model's predictive capability. Metrics such as precision, recall, and F1-score were employed to gauge the model's effectiveness in correctly classifying instances across different flexibility levels (James et al., 2013). Basically, the dataset used might have had more data for certain flexibility levels compared to others. To make sure each of the model doesn't favor the more common flexibility levels and overlook the less common ones, technique called Synthetic Minority Over-sampling Technique (SMOTE) was used. SMOTE helps to balance out the dataset by creating more data points for the flexibility levels that are less represented. This way, each model gets trained on a more even distribution of data for each flexibility level, which should help it make more accurate predictions overall. (Chawla et al., 2002). Among the models evaluated, the RandomForestClassifier and XGBoostClassifier emerged as particularly promising, both achieving an accuracy of 86 percentage . These models demonstrated robust performance in classifying students into different flexibility levels, as evidenced by their precision, recall, and F1-score metrics (Saito and Rehmsmeier, 2015).

C. *Performance Assessment*

In assessing the performance of the models, a set of standard metrics was employed, including precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the models' effectiveness in accurately classifying instances across various flexibility levels (James et al., 2013).

TABLE III: Model Performance Detail

Model	Precision	Recall	F1-Score	Accuracy
RandomForestClassifier	0.90	0.98	0.94	0.86
KNeighborsClassifier	0.95	0.84	0.89	0.80
SVC	0.80	0.87	0.83	0.75
LogisticRegression	0.72	0.67	0.69	0.64
XGBClassifier	0.89	0.98	0.94	0.86

D. *Model Evaluation*

In analyzing the performance of the RandomForestClassifier, it achieved an accuracy rate of 86 percentage, indicating that it accurately classified 86 percentage of instances across all flexibility levels. Precision, which represents the proportion of true positive predictions out of all positive predictions made by the model, provides insights into the model's accuracy for each flexibility level. For example, a precision score of 0.90 for flexibility level 0.0 implies that 90 percentage of instances predicted as level 0.0 by the model were actually level 0.0. Similarly, recall, also known as sensitivity, indicates the proportion of true positives correctly identified by the model out of all actual positives. With a recall score of 0.98 for flexibility

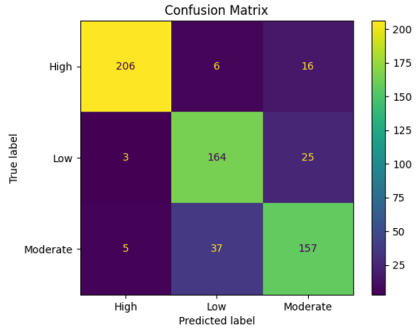

Fig. 2: Confusion Matrix of RandomForestClassifier.

level 0.0, the model captured 98 percentage of all instances actually at level 0.0. Additionally, the F1-score, a harmonic mean of precision and recall, offers a balanced measure of the model's performance. Moving on to the KNeighborsClassifier, it achieved an accuracy of 80 percentage, signifying its ability to correctly classify 80 percentage of instances across all flexibility levels. The Support Vector Classifier (SVC) demonstrated an accuracy of 75 percentage, showcasing its overall effectiveness in classifying instances. The Logistic Regression model achieved an accuracy of 64 percentage, underscoring its overall correctness in classifying instances. Lastly, the XGBClassifier, akin to the RandomForestClassifier, attained an accuracy of 86 percentage, demonstrating its effectiveness in classifying instances. Similar to the RandomForestClassifier, precision, recall, and F1-score metrics provide insights into the all the four model's performance for each flexibility level.

E. *Insights Visualization*

To delve deeper into model performance, a confusion matrix was generated specifically for the RandomForestClassifier. This visualization offers a clear depiction of the model's ability to correctly classify students into different flexibility levels. By visualizing the confusion matrix, we gain a deeper understanding of the model's strengths and weaknesses in accurately predicting flexibility levels.

F. *Overall Impact*

The incorporation of machine learning in education marks a significant step forward in predicting student adaptability to online learning. By using advanced algorithms, educators can delve into student data to uncover insights previously inaccessible. These insights help institutions tailor their online platforms to better meet student needs, fostering continuous improvement and innovation in online education. In essence, machine learning enables a more precise and effective approach to educational interventions, ultimately enhancing the learning experience for all students in the digital age. For example, machine learning algorithms can analyze students' online activity to predict who might struggle. Based on these predictions, personalized interventions like reminders or extra resources can be provided, improving student engagement and learning outcomes. This demonstrates how machine learning

enhances online education by offering targeted support to students.

IV. Experimental Setup

A. Feature Selection and Extraction

While all available features in the dataset were utilized for model training in this analysis, further exploration could involve feature selection techniques such as Recursive Feature Elimination (RFE) or feature importance analysis. Implementing these techniques could reveal the most influential features driving predictions of flexibility levels. By identifying and focusing on the most informative features, these methods streamline the model, potentially improving its performance and interpretability. This could lead to insights into which factors have the most significant impact on students' adaptability in online learning environments.

B. Classification Parameters

Each classification algorithm employed in the analysis came with its own set of parameters that could be fine-tuned to optimize model performance. For instance, in RandomForestClassifier, parameters like the number of trees in the forest (nEstimators), maximum depth of trees (maDepth), and minimum number of samples required to split an internal node (minSamplesSplit) could be adjusted. Similarly, in XGBoostClassifier, parameters like learning rate, maximum depth of trees, and regularization parameters could be optimized to achieve better predictive accuracy. Tuning these parameters ensures that the models are tailored to the specific characteristics of the dataset, potentially enhancing their ability to accurately predict flexibility levels for students in online learning settings.

V. Results

The analysis began with exploratory data analysis (EDA) to gain insights into the dataset's structure and distribution. The dataset comprised 1205 entries with 11 columns, including categorical features such as Education Level, Institution Type, Gender, Age, Device, IT Student, Location, Financial Condition, Internet Type, Network Type, and the target variable Flexibility Level. Upon inspection, there were no missing values in the dataset, indicating its completeness. Categorical variables were converted into numerical representations using Ordinal Encoding for machine learning compatibility. The distribution of features and the number of students across different flexibility levels were visualized through count plots, revealing patterns and trends within the data. For instance, the majority of students were male, attended private institutions, and used mobile devices. Machine learning models were then trained to predict students' flexibility levels based on their demographic and technological attributes. Five classifiers were employed: RandomForestClassifier, KNeighborsClassifier, SVC, LogisticRegression, and XGBClassifier. Each model's performance was evaluated using precision, recall, and F1-score metrics, with RandomForestClassifier achieving the highest accuracy of 86 percentage. Confusion matrices were also generated to visualize the models' predictive performance, highlighting areas of correct and incorrect classifications. Overall, the RandomForestClassifier emerged as the most effective model for predicting students' flexibility levels in online learning environments, achieving high precision and recall across all flexibility levels. This suggests that the RandomForestClassifier can accurately classify students into their respective flexibility categories, providing valuable insights for educational institutions to tailor online learning experiences according to students' adaptability.

VI. Discussion and Conclusions

The findings from this analysis provide valuable insights into factors influencing students' adaptability levels in online education. Several key observations can be drawn from the data and model predictions.The analysis revealed demographic trends such as gender distribution and educational institution types. Understanding these trends can help educators and policymakers design targeted interventions to support diverse student populations. Students' device usage, internet connectivity, and network types play significant roles in their adaptability to online learning. Institutions should ensure access to reliable technology infrastructure to facilitate effective learning experiences for all students. Among the machine learning models evaluated, the RandomForestClassifier exhibited the highest accuracy in predicting students' flexibility levels. This underscores the importance of employing robust classification algorithms to accurately identify students' adaptability to online learning. By leveraging insights from this analysis, educational institutions can implement tailored strategies to enhance students' online learning experiences. These strategies may include providing additional support for students with specific demographic characteristics or technological needs and optimizing online platforms for improved accessibility and usability. Further research could explore additional factors influencing students' adaptability to online education, such as socio-economic background, prior academic performance, and learning preferences. Additionally, longitudinal studies could investigate how students' adaptability levels evolve over time and the impact of interventions on improving adaptability and learning outcomes. In conclusion, this analysis sheds light on the complex interplay between demographic, technological, and educational factors in shaping students' adaptability to online learning. By understanding these dynamics and leveraging predictive modeling techniques, educators and policymakers can develop targeted interventions to support diverse student populations and enhance the overall effectiveness of online education.

References

Al Lily, A. E., Ismail, A. F., Abunasser, F. M., Alqahtani, R. H., Jayarajah, U. (2021). Distance education as a response to pandemics: Coronavirus and Arab culture. Technology in Society, 64, 101482.

Ally, M. (2008). Foundations of educational theory for online learning. In T. Anderson (Ed.), The theory and practice

of online learning (2nd ed., pp. 15–44). Athabasca University Press.

Artino, A. R., Brydges, R., Gruppen, L. D., Holmboe, E. S. (2011). Creating boundaries for electronic health records training: Educational and cognitive approaches. Academic Medicine, 86(6), 684–690.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

Dwivedi, Y.K., Hughes, D.L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J.S., Gupta, B., Lal, B. and Misra, S., 2021. Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. International Journal of Information Management, 57, p.102287.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hodges, C., Moore, S., Lockee, B., Trust, T., Bond, A. (2020). The difference between emergency remote teaching and online learning. Educause Review, 27.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160–166.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825-2830.

Pulham, E., Graham, C. R. (2018). Comparing K-12 online and blended teaching competencies: A literature review. Distance Education, 39(3), 411–432.

Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), p.e0118432.

UNESCO. (2020). Education: From disruption to recovery. Retrieved from https://en.unesco.org/covid19/educationresponse