Predicting a Student's Adaptability Level in Online Education

Rishika Sharma **Shubhaangi Verma** **Tanishk Arora** **Ayush Rawat**

1 Abstract

Since Covid-19 online education has become very crucial. It has been due to the fact that it has enabled people to continue their education even when the whole world had been hit by the pandemic. People pursued their studies, sitting in front of electronic devices in their homes worldwide. The increasing importance of online education led to the emergence of the need to determine the student's adaptability level in online education. The student's adaptability level in online education depends on various factors, which include the characteristics of the student itself, the student's location, access to technology and the internet etc. Different students with varying factors faced different difficulties in online education. So, it becomes extremely important for educational institutions to predict the adaptability level in online education for the student with given constraints w.r.t their location, internet and technology access, age etc. Using ML techniques, predicting the adaptability level beforehand is very beneficial for the students as it helps improve the adaptability level further to get an optimal level for that scenario.

2 Introduction

There has been a sudden substantial increase in the importance of online education with the advent of Covid-19 as it was the only medium left to pursue the education of an individual. But the major drawback of online education is that there is not an effective one-on-one interaction between teachers and students, so it becomes extremely important to predict a student's adaptability level. Different students have different adaptability levels depending on various factors related to a student, viz. Gender, Age, Education level, Institution type, location, internet access, etc. Prediction of adaptability level is done using different Machine Learning techniques (logistic regression, Naive Bayes, SVM, Decision Trees, Random Forests, KNN, ANN) and their performance is compared to find the best classifier.

3 Literature Survey

Technology allows for virtual or remote learning. Thanks to technological advancements, we can now create online education systems. Aspects of education will be held in digitization under current circumstances. Students must accept the challenge of adapting to online education to make these changes. In the following discussion, we will provide an overview of the findings from the analysis of related works on online education. In [1] and [2], the researchers have studied the improvement of the online education model. Online education decisions should be based on evidence of effectiveness rather than the assumption that face-to-face interaction is superior. A demonstration by Rojan et al. [1] helped us observe a significant difference in student performance and satisfaction and made us realize many benefits of online education for students. It has comparable off-campus and on-campus performances, offers, and student satisfaction, but communication has been difficult. An investigation was done by William et al. [2] on how to improve the Online Education Model by using Machine Learning and Data Analysis in a Learning Management System (LMS), He also focused on formative assessment for better learning and the exploratory results show that 85The researchers have examined how the ongoing pandemic is a worry for international education systems in [3] and [4]. During the pandemic, a vast majority of the countries shut down their schools. The research studies in [3] and [4] demonstrate the terrible impacts of the coronavirus on education and identify a number of obstacles that prevent interactions between students and instructors in online learning during the pandemic. They further took their research into depth and found out that as the rural areas didn't have adequate digital skills and had a lot of barriers

including technological barriers, domestic barriers, financial inadequacy, and poor electricity, hence, they faced even more challenges compared to the people in urban areas.

4 Dataset Analysis

4.1 Dataset Description

The dataset has 13 independent and 1 dependent feature. There are 1205 samples in the dataset. Some of the features have binary values like Yes/No and Low/High while others have multiple values. Since the data is categorical, no outlier is present. A brief description of the dataset attributes with their possible values has been provided the below Table1.

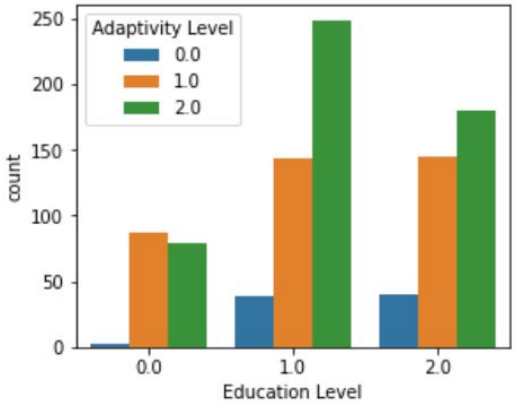

Figure 1: No. of Students vs Education Level

The above countplot depicts the distribution of different adaptivity levels as per the different types of education levels.

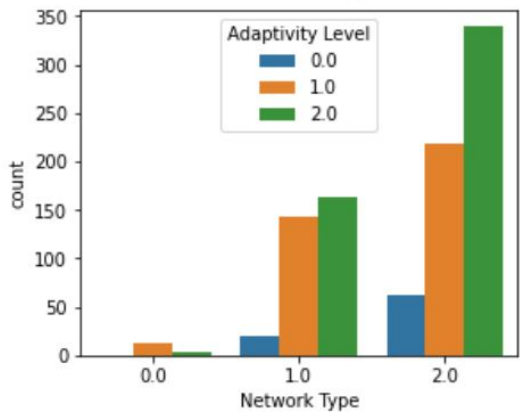

Figure 2: No. of Students vs Network type

The above countplot depicts the distribution of different adaptivity levels as per the different types of networks. The above scatterplot depicts the plot of class duration for different age groups.

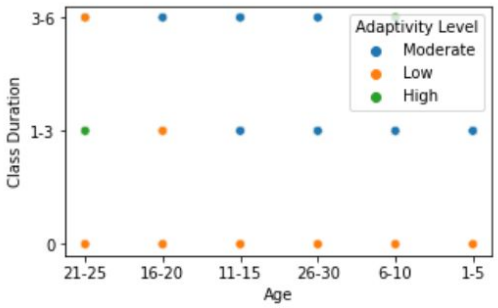

Figure 3: Scatterplot of class duration vs age groups

4.2 Data Preprocessing

Variable	Variable Type	Possible values
Gender	Independent	Girl(1), Boy(0)
Age	Independent	Around 1-5(5), 6-10(4) 11-15 (1), 16-20 (2), 21-25 (3),26-30 (0)
Education level	Independent	School(1), College(0), University(2)
Institution Type	Independent	Non Govt(1), Govt(0)
IT Student	Independent	No(0), Yes(1)
Location (Is town)	Independent	No(0), Yes(1)
Load Shedding	Independent	Low(0), High(1)
Financial Condition	Independent	Poor(1), Mid(0), Rich(2)
Internet Type	Independent	2G(0), 3G(1), 4G(2)
Network Type	Independent	Mobile Data(0),WiFi(1)
Class Duration	Independent	0 Hours(0), 1-3 Hours(1), 3-6 Hours(2)
Self lms	Independent	No(0), Yes(1)
Device	Independent	Tab(2), Mobile(1), Computer(0)
Adaptivity Level	Dependent	Low(1),Moderate(2), High(0)

Table1: Attribute Details

In the given dataset :
1. There are no null values present.
2. Feature Transformation: The data present is categorical, so the string values have been scaled to Integer for model prediction.
3. Feature selection: On the basis of Information Gain, and correlation matrix some of the attributes 'Load Shedding' and 'Self Lms' are dropped.

The below plot Fig 4. shows the information gain by different independent variables on 'Adaptivity Level'. It is clear that the features

like 'Load shedding' and 'Self Lms' have very low correlation w.r.t target variable and low information gain and hence are dropped while selecting features to improve model performance.

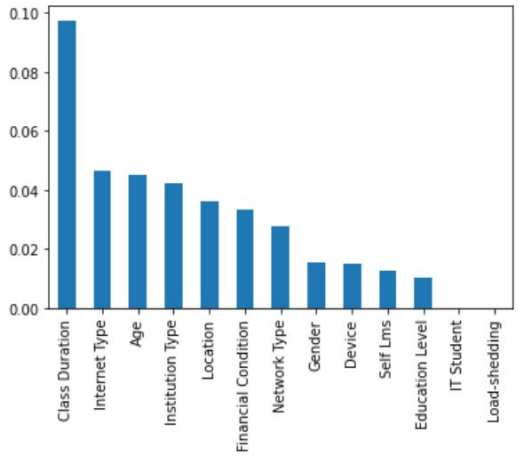

Figure 4: No. of Students vs Network type

5 Methodology

After preprocessing the data, we used several Machine Learning models to predict the Adaptivity Level based on the features in the sample. We have used the following classifiers:

1. **Logistic Regression:** It is a type of regression used in case of classification problems. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.
2. **Gaussian Naive Bayes:** It is a type of classification model which uses the Bayes algorithm. It is easy and fast in multiclass classification as it needs less training data. It is used to determine the benchmark performance of the models.
3. **Decision Tree Classifier:** A decision tree is a non-parametric supervised learning algorithm which provides interpretability while doing classification. At each level, a feature is chosen as per its information gain or entropy for classifying data and final classification is obtained at the leaf level.
4. **Random Forests Classifier:** It is ensemble learning of Decision Trees(which provides interpretability and is non-parametric in nature) where some weak classifiers are combined and the prediction is done by majority voting for classification problems.
5. **SVM:** A support vector machine (SVM) is a supervised learning algorithm to classify or predict data groups. The goal of the SVM is to determine the unique decision boundary known as Optimum Separating Hyperplane (OSH) that can segregate n-dimensional space into the required number of regions for classification.
6. **KNN:** The k-nearest neighbours' algorithm, also known as KNN or k-NN, is a supervised learning classifier that makes predictions or classifications about the clustering of a single data point based on proximity. It makes the assumption that the new case and the existing cases are similar, classifies the new case into the category that most closely resembles the existing categories, stores all the existing data, and then assigns a new data point based on the similarity.
7. **ANN:** Artificial Neural Network comprises several different layers viz. input layer, one or more hidden layer, and output layer. Each node connects to the node in the previous and next layer having a certain weight associated with it. The node can be activated or deactivated depending upon the output it generates.

6 Results and Analysis

Model	Class	Accuracy	Precision	Recall	F1-score
LR	Low	64.73%	0.80	0.22	0.35
	Mod	64.73%	0.66	0.56	0.60
	High	64.73%	0.64	0.79	0.70
NB	Low	63.07%	0.32	0.39	0.35
	Mod	70.12%	0.65	0.62	0.64
	High	70.12%	0.67	0.67	0.67
DT	Low	82.98%	0.72	0.72	0.72
	Mod	82.98%	0.86	0.84	0.85
	High	82.98%	0.82	0.84	0.83
RF	Low	86.72%	1.00	0.88	0.84
	Mod	86.72%	0.67	0.85	0.92
	High	86.72%	0.80	0.86	0.88
SVM	Low	85.89%	0.62	0.72	0.67
	Mod	85.89%	0.87	0.93	0.90
	High	85.89%	0.90	0.82	0.85
KNN	Low	81.74%	0.87	0.72	0.79
	Mod	81.74%	0.85	0.77	0.81
	High	81.74%	0.79	0.87	0.83
ANN	Low	82.57%	0.62	0.72	0.67
	Mod	82.57%	0.87	0.93	0.90
	High	82.57%	0.90	0.82	0.85

Table 2: Model Results

Fig 5. depicts the Loss vs Epochs curve of

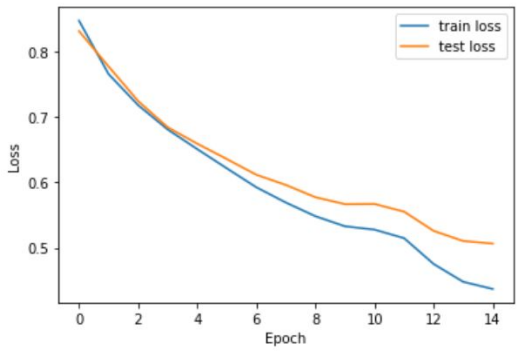

Figure 5: Loss vs epochs

Multilayer Perceptron model, the minima is achieved after 10 epochs showing that the model is training quickly as the dataset is not large enough.

The given table presents the performance metrics (accuracy, precision, recall, and f1-score) of different machine learning models on testing data after training on a given dataset. Among the models used, Random Forest outperformed all the others with an accuracy of approximately 87%. The reason behind this success is that Random Forest is an ensemble learning technique that combines several decision trees to achieve high accuracy in classification tasks. Decision trees work well with categorical data, and the combination of multiple trees reduces the overfitting problem.

On the other hand, Logistic Regression performs poorly with an accuracy of only 64.7%. This is because the data in the dataset is complex and non-linear, and Logistic Regression assumes a linear relationship between the features and the target variable. Naive Bayes also gives poor results with an accuracy of 63.1%, as it works well with probabilistic data, which is not the case in the given dataset.

Support Vector Machine (SVM) performs almost as well as Random Forests, with an accuracy score close to 87%. SVM is known for its ability to generalize well in high dimensional feature spaces, and it eliminates the need for feature selection. Therefore, SVM can work well with complex datasets like the one used in this study.

In summary, Random Forest and SVM are the best performing models for the given dataset, while Logistic Regression and Naive Bayes are not suitable due to the complexity of the data. However, it is essential to note that the choice of the model depends on the nature of the dataset and the specific task at hand. Therefore, a thorough analysis of the data and evaluation of multiple models is necessary to choose the most appropriate one.

7 Conclusion

The ability to select the most effective machine learning model for a given problem is crucial in achieving accurate predictions. It is important to consider various techniques since no single model is always the best choice. In this project, various machine learning models, such as Logistic Regression, Naive Bayes, Random Forest, KNN, SVM, and ANN, were applied to forecast the adaptability of students in online learning. The results showed that Random Forest outperformed all other models, emphasizing the importance of ensemble learning methods for categorical data.

The project also demonstrates the potential of machine learning techniques to assist educational decision-makers in designing effective learning plans. By predicting a student's level of adaptability in an online learning environment, personalized learning strategies can be created for each student. This approach can help bridge the gap between students' learning styles and preferences and the resources and techniques used in online learning.

Overall, this project highlights the value of machine learning in addressing complex educational issues and demonstrates the importance of considering multiple techniques to select the most effective model. The results have significant implications for educational decision-makers and the potential to improve students' learning outcomes in online learning environments.

8 Individual Contribution

Dataset description: Rishika Sharma
Model training: Rishika Sharma, Ayush Rawat
Analysis: Rishika, Tanishk, Shubhaangi, Ayush
Report and Literature Review: Rishika Sharma
Presentation: Shubhaangi Verma
Graphical User Interface: Tanishk Arora, Shubhaangi Verma

9 References

Dataset from Kaggle

[1] R. Afrouz and B. R. Crisp "Online education in social work, effectiveness, benefits, and challenges: A scoping review," Australian Social Work, vol. 74, no. 1, pp. 55–67, 2021.

[2] D. Wiliam "Assessment in Education: Principles, policy practice," Assessment in Education: Principles, Policy and Practice, vol. 15, no. 3, pp. 253–257, 2008.

[3] M. Onyema, N. C. Eucheria, F. A. Obafemi, S. Sen, F. G. Atonye,A. Sharma, and A. O. Alsayed, "Impact of coronavirus pandemic on education," Journal of Education and Practice, vol. 11, no. 13, pp. 108–121, 2020.

[4] R. E. Baticulon, J. J. Sy, N. R. I. Alberto, M. B. C. Baron, R. E. C.Mabulay, L. G. T. Rizada, C. J. S. Tiu, C. A. Clariational survey of medical students in the philippines," Medical scieon, and J. C. B. Reyes, "Barriers to online learning in the time of covid-19: A nnce educator, pp. 1–12, 2021.