automation

Scientific paper

Predicting Students' Adaptability to Online Education

Mohamad Slim

Abstract:

Since the COVID-19 epidemic, which drove many educational institutions throughout the globe to embrace total or partial online education, online learning has seen a tremendous increase and has been regarded as one of the most effective and efficient means of learning. However, many underdeveloped countries have taken a long time to introduce online education at all levels of the educational system. The capacity to deal with unexpected changes, such as COVID-19, is referred to as adaptation, and since the flexibility shown by students is a crucial determinant of the success of an online course, machine learning techniques were employed in this research that attempts to construct a data-centric system that predicts how well a student would adapt to online schooling. The dataset used in this study comes from a survey taken by students in Bangladesh between 2020 and 2021, and it focuses on features related to students such as their age, education level, financial conditions, network connections, and so much more across all three levels (School, College, and University). Using both online and in-person surveys we were able to predict adaptability based on three metric levels that include low, moderate, and high levels. After implementing a variety of machine learning algorithms on our dataset, such as Logistic Regression, Decision Tree (DT), Random Forest (RF), XGBoost, Support Vector Classification (SVC), and K-Nearest Neighbors (KNN), we were able to predict students' adaptation to online education (Suzan et al., 2021). We achieved a maximum accuracy of 92.5 percent with the Decision Tree classifier, 91.6 percent with XGBoost and 90.7 percent with Random Forest on an imbalanced dataset, whereas after balancing the data we achieved 92.5 percent with XGBoost, 92.1 percent with Random Forest and 92 percent with Decision Tree classifier, in comparison to the other approaches used.

Keywords: Data Analysis, Online Education, Adaptability Level, Artificial Intelligence, Machine Learning.

1. Introduction

Online education is a form of education that enables administrators to deliver classes virtually using the internet. Implementing online education platforms implies that every interaction between teachers and students would be over the internet. The process of online education has gained a lot of attention since the COVID-19 pandemic, and a lot of educational institutions have opted for a virtual learning platform for their students due to its benefits (Singh and Thurman, 2019). The significance of online education in the twenty-first century cannot be overemphasised. Online education allows professors and students to develop and access educational information at their own convenience and leisure; flexible access to educational information aids students' time management; moreover, for students engaged in work-study programs, online education allows them to effortlessly strike a balance between employment and study. Although an internet connection requires a monthly fee, it is little when compared to the expense of transportation and lodging, because the process of online education is not constrained by physical barriers. Any electronic mobile device, including tablets, PCs, and mobile phones, may connect to the internet and made education available digitally as well as accessible from anywhere.

However, because the COVID-19 pandemic was unexpected, several institutions encountered difficulties in transitioning to an online education system. There are several challenges faced by students as well as educational institutions attempting to adapt to online learning, and some of the contributing factors have been highlighted in the dataset.

While the benefits of online education have already been mentioned, the ease with which students may adjust to online education is the subject of this research employing machine learning algorithms.

2. Related Work

Despite the COVID-19 circumstance, many people have been able to continue teaching and studying via online learning. The global pandemic is generating the most widespread innovation in online education. Given its broad use and expansion, recognising the advantages and downsides of online learning, as well as investing in, preparing for, and delivering it, requires a methodical approach. (Vlachopoulos, D.). Certain nations have adopted various online learning modalities to circumvent the digital divide amid the present crisis. New Zealand, for example, used a hybrid approach that blended two television channels with Internet distribution and a hard-copy curriculum resource to convey educational information. Due to Australia's restricted Internet access, television has been utilised to engage parents in order for them to aid their children in learning. Because of a collaboration between schools and post office services in Portugal, hard-copy instructional materials were quickly sent to children's homes. (Drane, C. et al.).
(Yusuf, B.N.) suggests that universities provide improved e-learning platforms in order to improve Internet access and develop an interactive learning approach. Seminars or training for teachers and students to improve their technical and pedagogical skills in online learning are also necessary. The problem of inclusion is critical when it comes to emergency remote training. Inclusion varies from nation to country. (Thomas, M.S.C., Rogers, C.) found that school-provided IT solutions are often overpriced, clumsy, and swiftly out of date based on their experiences with online learning during the epidemic. They argue that students should be allowed to bring their own electronic gadgets to school. The authors also suggest that governments provide incentives to enterprises that develop engaging and effective educational games and virtual worlds. A child's interest and engagement in school may be boosted by using game mechanics. The researchers (D. Kucak et al.) looked at how well machine learning worked in educational environments. The primary objective of his study was to evaluate how practical it would be to use machine learning in the academic sector. Student evaluation comes first among the four main components. When evaluating students, machine learning can do it without introducing any of the inherent biases that humans have. The group's goal was to enhance how schools evaluate students' problem-solving skills. Enhancing the rate at which students are retained is the second goal. The subsequent stage is to make projections about student performance, and the last step is to provide assessments to students.
In our work, using machine learning methodologies, we attempted to determine students' adaptation to online education under this pandemic condition.

3. Materials and Methods

3.1. Data Exploration

Data exploration is the first step that is used to visualize, explore and understand the data to attain valuable as well as actionable insights (Asmael et al., 2017). Data exploration helps in the process of data pre-processing to optimize machine learning predictions. The data was collected from the Kaggle repository. The description of the dataset is as shown in Table 1. It consisted of 1205 records and 14 features. The features represent the information of students enrolled in colleges, universities, and schools. The record contains 13 predictor variables: age, gender, level, govt./non-govt. institution, location, IT student or not, educational background, load shedding level, internet quality, class-time, the economic condition of the family, device type used while attending classes, and the availability of the institution's own LMS to predict a target variable: adaptability level (Suzan et al., 2021). Since the data was collected from the survey, student adaptability was binned into three levels namely low, moderate, and high levels. It was observed that all the data type of all the fourteen columns in the data frame was of object type and therefore to make the data uniform, all the variables were converted into the category datatype. As well as the dataset was examined for the presence of missing values and we didn't found any (Figure 1).
Also, the integrity of the values in every row and column of the dataset has been verified using data visualization methods.

Table 1. Features names and Description

S/N	Column	Variable Definition
1	Gender	Gender of the Student
2	Age	Age range of the Student
3	Education Level	Education Institution Level (School, College, University)
4	Institution Type	Education Institution Type (Government, Non-Government)
5	IT Student	Studying as IT student or not
6	Location	Is student location in town or not
7	Load-shedding	Level of load shedding
8	Financial Condition	Financial condition of family
9	Internet Type	Internet type used mostly in device
10	Network Type	Network connectivity type
11	Class Duration	Daily class duration
12	Self LMS	Is Learning Management System (LMS) owned by Institution
13	Device	Device used mostly in class
14	Adaptivity Level	Adaptability level of the student

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1205 entries, 0 to 1204
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Gender               1205 non-null   object
 1   Age                  1205 non-null   object
 2   Education Level      1205 non-null   object
 3   Institution Type     1205 non-null   object
 4   IT Student           1205 non-null   object
 5   Location             1205 non-null   object
 6   Load-shedding        1205 non-null   object
 7   Financial Condition  1205 non-null   object
 8   Internet Type        1205 non-null   object
 9   Network Type         1205 non-null   object
 10  Class Duration       1205 non-null   object
 11  Self Lms             1205 non-null   object
 12  Device               1205 non-null   object
 13  Adaptivity Level     1205 non-null   object
dtypes: object(14)
memory usage: 131.9+ KB
```

Figure 1. Properties of the data

3.2 Exploratory Data Analysis

(EDA) is a term responsible for extracting and visualizing patterns in the dataset to develop machine learning algorithms.

3.2.1 Target Analysis (Adaptivity feature)

The target variable in this study was the column "Adaptivity Level" which defines the quality of the student being adjusted to the online learning environment (Figure 2). Since the target feature is categorical data with 3 classes, this machine learning project was modeled as a multi-class, classification task.

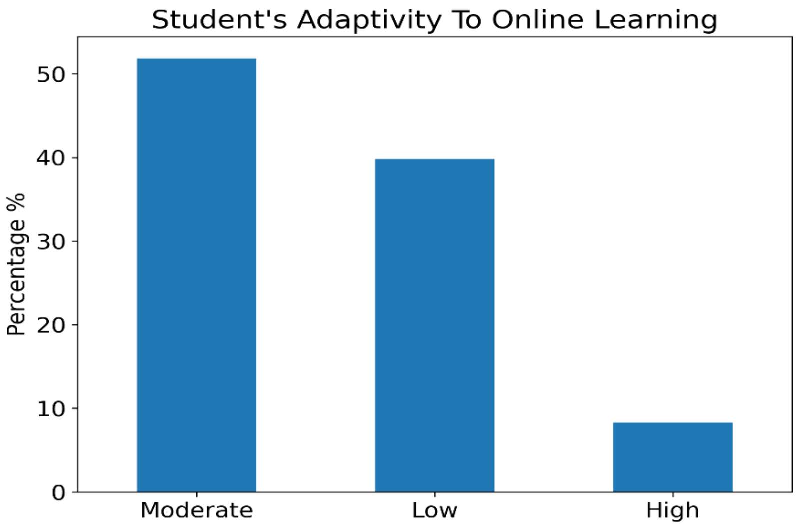

Figure 2. Bar Chart showing the distribution of Adaptivity levels

The dataset considered for this study consisted of imbalanced classes because there are significantly fewer samples (~10%) with a high level of adaptability, compared to moderate and low adaptive levels.

3.2.2 Relationship Between Age-group and Adaptivity Level

Visualization between the different age groups that have unique proportions and distributions concerning the adaptivity level was conducted to develop.

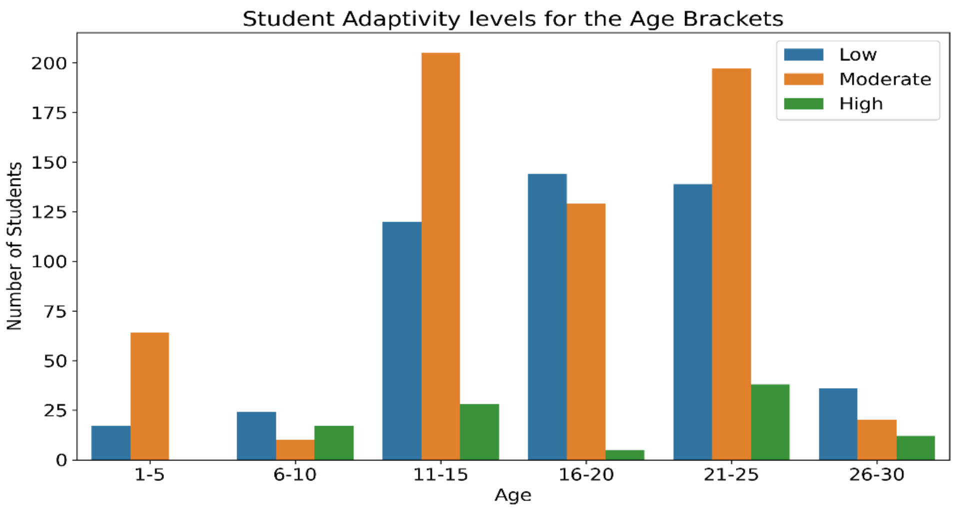

Figure 3. **Bar chart of adaptivity levels for the different age ranges.**

It was observed that the age group 16-20 contains the greatest number of students with low adaptability levels to online education, but also has a relatively high number of students with moderate levels of adaptability. None of the students in age brackets 1-5 were recorded to have a high adaptivity level to online learning.

3.2.3 Relationship Between Education Level and Adaptivity Level

It was observed that a good proportion of the respondents in the survey were currently at the school level. Comparatively few students with a high level of adaptability to online education were in college as compared to the total number of college students (Figure 4). Whereas the significant difference is the number of students with moderate adaptability levels. School and university education level students showed a decent number of students with a high adaptability level concerning each total number of students.

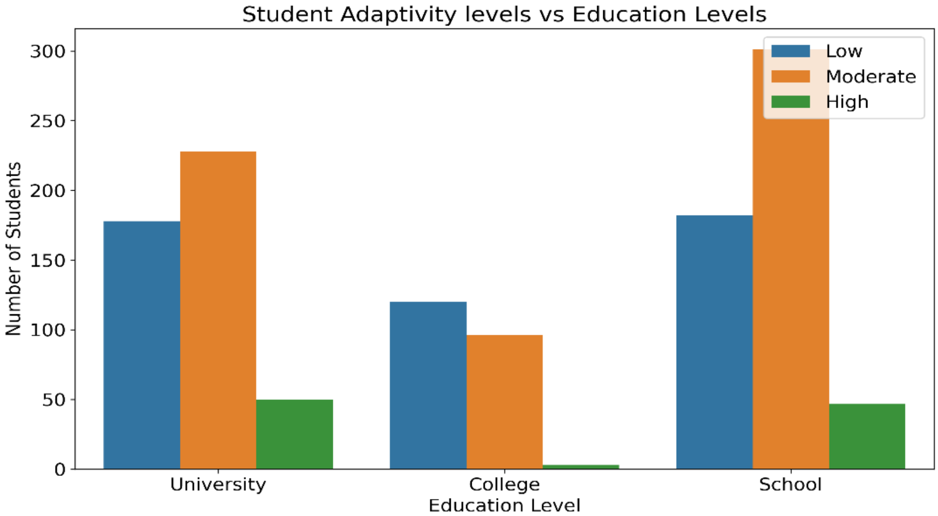

Figure 4. Bar chart of adaptivity levels for each education level

3.2.4 Relationship Between Network Type and Adaptivity Level

The mapping of the relationship between network type and adaptivity level was visualized to get an idea about the influence of network type on the online courses in terms of students' perspective.

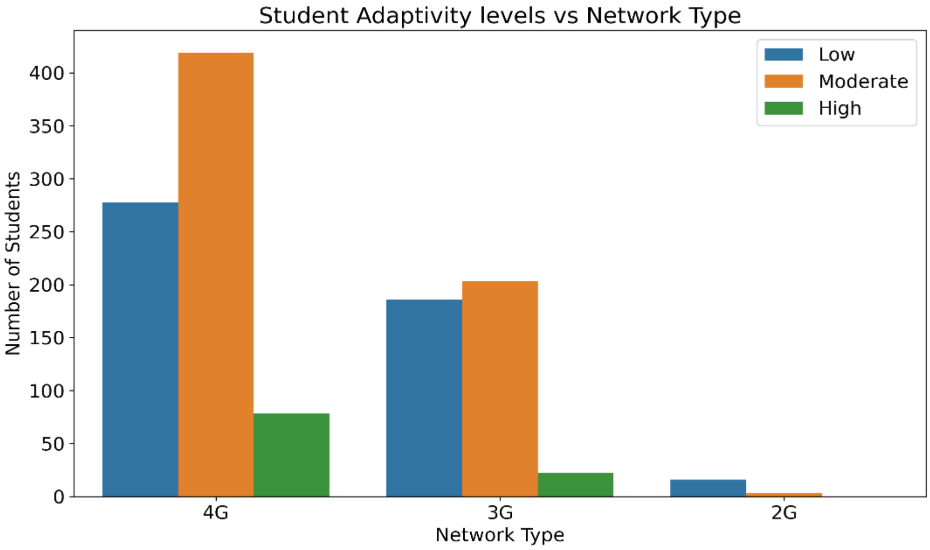

Figure 5. Bar chart of adaptivity levels of Students with different Networks

As shown in Figure 5, we can observe that the 4G network is the most used network amongst the students. Students that only have access to a 2G network are likely to show low adaptability to online education, this makes sense because the 2G network is the slowest amongst the network types. In compared to 2G, 3G and 4G appear to be superior possibilities for an online education system.

3.3 Data Preprocessing

Data preprocessing was done based on different variables included in this section.

3.3.1 Age Transforming

The age values are collected within a range, to use the age range values, a function is used to transform the age brackets and returns the average of age values within the bracket.

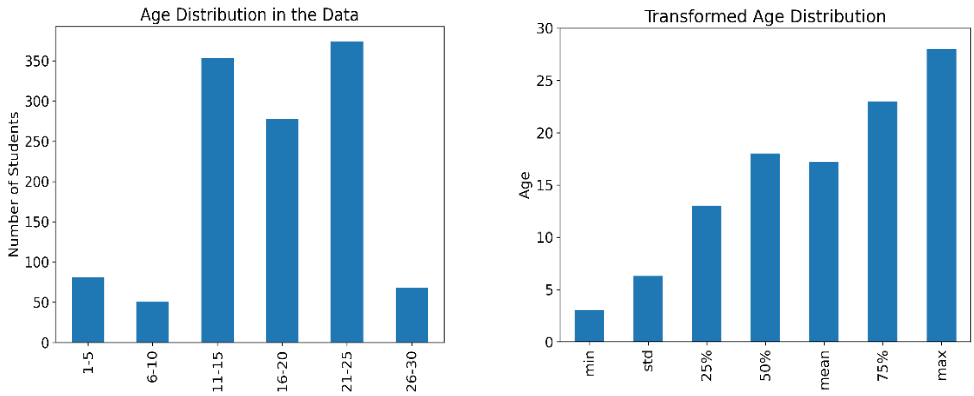

Figure 6. Tranformed age statistics

From Figure 6, after transforming the age brackets to numerical values, the average age of students is slightly less than the median (50% percentile), which means that the age distribution is slightly (left) skewed, this interprets that the ages of students in our dataset are slightly more concentrated on the right (older) side.

3.3.2 Labeling Features

To make the string attribute values more obvious and accessible to the models, we converted them to numeric values.

First, we mapped the ordinal features, where each category in the feature retains its order. A viable alternative for mapping is Label Encoder which converts values to numbers but it does not keep the order of the categories (Table 2).

Then, Gender, IT Student, Location, Load-shedding, and Self LMS are encoded using dummy encoding. dummy encoding avoids a dummy variable trap. (Table 3) below shows how Dummy Encoding (n-1) would transform the Gender feature. From the python code for dummy encoding, (drop-first = True) ensures that the Gender-Male column is not created.

The dependent variable which is the Adaptivity Level was encoded using the Label Encoder which transforms the adaptivity level by converting the values into numbers.

Table 2. Features and mapped values

Features	Mapped Values
Education Level	School: 1, College: 2, University: 3
Financial Condition	Poor: 1, Mid: 2, Rich: 3
Institution Type	Government: 0, Non-Government: 1
Internet Type	Mobile Data: 0, Wi-Fi: 1
Network Type	2G: 1, 3G: 2, 4G: 3
Class Duration (Hours)	0: 0, 1-3 : 1, 3-6 : 2
Device	Tab: 1, Mobile : 2, Computer : 3

Table 3. Dummy Encoding of Gender

Gender_Female
1
0
1

4. Model Implementation

Algorithms in machine learning are responsible for learning patterns in datasets in association with a target feature, once learning is completed, the algorithm is smart enough to identify patterns in an unseen dataset and make a prediction (Hyun and Kim, 2022), the following machine learning algorithms with their description are used to predict student adaptivity.

Logistic regression is a sort of regression that predicts the likelihood of occurrence of an event by fitting data to a logistic function. This type of regression is described in more detail in the following paragraph. In the same way as other kinds of regression analysis, logistic regression makes use of a few different predictor variables, each of which can be either numerical or categorical. (Feng, 2014)

Decision Tree is a representation of a classifier that is expressed as a recursive split of the instance space. The decision tree is made up of nodes that come together to form what is known as a root tree. This indicates that the decision tree is a distributed tree with a fundamental node known as root and no incoming edges. (Charbuty, 2021)

K-Nearest Neighbors (KNN) is among the most straightforward machine learning (ML) methods. The success of the KNN can be due to its Simple interpretation and short calculation time KNN selects the number k of neighbors and computes their distances. The distance function computes the distance between k neighbors the distance function most usually used to assign a class among its k closest neighbors. (Jawthari, 2021)

Random Forest algorithm is a form of supervised machine learning that can perform both classification and regression. According to what its name says, it is a collection of decision trees that can be used collectively. The fact that each tree protects each other from their own particular faults is one of the most significant benefits of RF (Abdulkareem, 2021).

Support Vector Machine (SVM) In the realm of machine learning, (SVM) is a supervised method that has applications in both classification and regression. However, its primary use is in resolving categorization issues. When using the SVM technique, each data point is represented by a single coordinate in an n-dimensional space (where n is the number of features). Then we classify by finding the hyperplane that most clearly separates the two groups (Sunil, 2017).

Extreme Gradient Boosting, more often known as XGBoost, is a prominent supervised learning technique that is used for regression and classification on big datasets. It avoids overfitting by utilising a training strategy that is extremely scalable and builds shallow decision trees in a sequential fashion in order to produce accurate results. (Asselman, 2021).

The purpose of using six models on the data is to get a credible model by comparing the performance of the various models.

The following classification metrics are used to assess the performance of the models.

- Accuracy: It is the proportion of correctly predicted data points out of all the data points.
- ROC AUC Score: ROC AUC score describes the predictive quality of the model across all thresholds for the probabilities gotten from the model. The higher the AUC score the better the model
- F1 Score: The F1 score is the harmonic mean of precision and recall, as it is a suitable metric to compare the performance of models for this project (Suzan et al., 2021).

To be able to evaluate the models on unseen data, the dataset was split into 80% train and 20% test splits. Then, all models were fitted on the training subset twice, once on imbalanced data, and once after balancing the dataset.

5. Evaluation and Results

This section is divided into three parts, results from imbalanced data, from balanced data, and comparison of results between them.

5.1. Using Imbalanced Dataset

5.1.1 Evaluating metrics

Below Table 4, you can observe the results of each metric corresponding to each model when inference is made using the test dataset. The highest accuracy for each metric was highlighted in red. Since the F1 score gives precisely the most accurate metric based on this study, the best model was found to be the Decision Tree Classifier with an F1 score of 0.925.

Table 4. Performing of ML Models (Imbalanced Data)

Model Name	**Accuracy**	**F1_Score**	**ROC_AUC**
Logistic Regression	0.722	0.690	0.789
Decision Tree Classifier	0.925	0.925	0.981
nearest Neighbors	0.759	0.756	0.928
Random Forest Classifier	0.909	0.907	0.986
Support Vector Classifier	0.643	0.595	0.837
XGBoost Classifier	0.917	0.916	0.988

5.1.2 Applying Hyperparameter Tuning

Hyperparameter Tuning where applied on all the models trying to achieve better accuracy. Tuning the hyperparameters of tree models such as Decision Tree, Random Forest and XGBoost does not yield a significant improvement in F1 score.

The model that improved the most after hyperparameter tuning was the SVC (Figure 7).

	Model_Name	F1_Score	F1_Score_tuned
0	Logistic Regression	0.690257	0.703179
1	Decision Tree	0.924919	0.924919
2	KNN	0.751760	0.870433
3	Random Forest	0.907475	0.885451
4	SVC	0.595238	0.789035
5	XGBoost	0.916268	0.916268

Figure 7. Comparing F1_score after applying hyperparameters tuning

5.1.3 Confusion Matrix

Confusion Matrix visualises the properly and erroneously predicted classes to provide understanding about the algorithm's performance. Since Decision Tree Classifier had the highest F1 score, its confusion matrix was visualized.

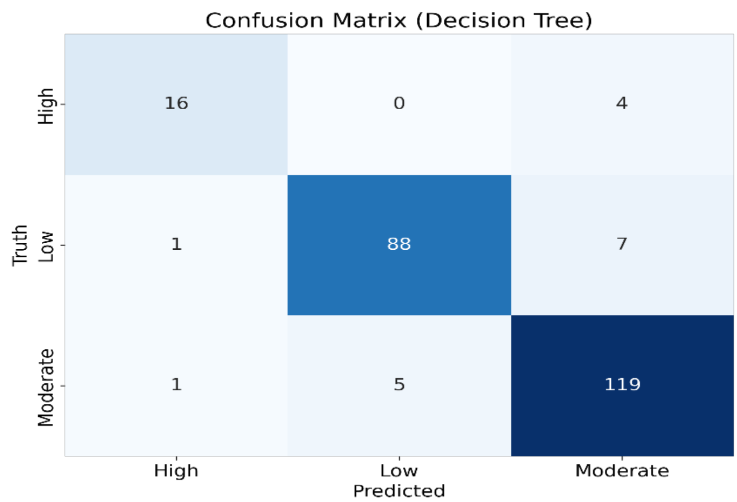

Figure 8. Confusion matrix of Decision Tree

Based on the metrics from the confusion metrix (Figure 8), it could be inferred that the decision tree classifier can make high level of accurate predictions for all 3 classes with small amounts of misclassifications.

5.1.4 Receiver Operating Characteristic (ROC AUC)

The ROC curve as known is a curve showing the effectiveness of models across categorization levels. It does so by plotting the TPR against the FPR at a variety of different classification levels. When the threshold for positive classification is lowered, more objects are counted as positive; this increases both false positives and true positives. The ROC curve for the decision tree model is displayed in the following (Figure 9)

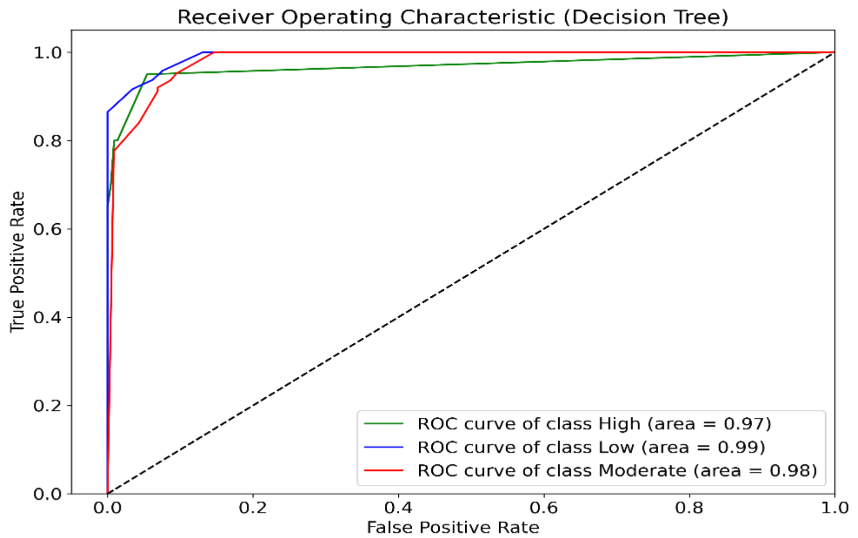

Figure 9. ROC curve for Decision Tree

5.1.5 Analyze of Feature importance of the Decision tree classifie

Feature importance is a score that shows the impact of the features that are used in training a model. It helps to determine the most significant feature in predicting the adaptivity level.

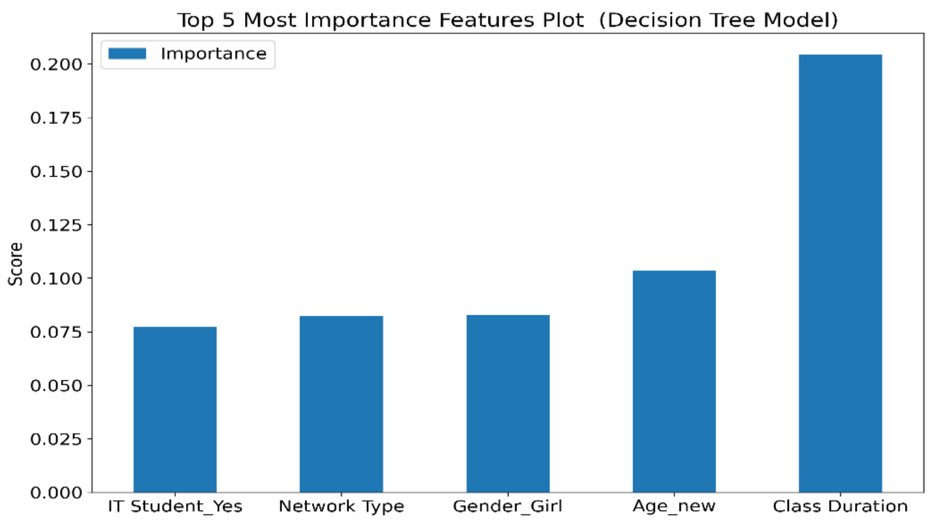

Figure 10. Feature Importance Plot for Decision Tree

From the bar plot above in (Figure 10), Class Duration Feature had the most impact on the Decision Tree model.

5.2 Using Balanced Dataset

The original dataset is imbalanced, and this will affect the performance of some machine learning algorithms, to discover the impact of using a balanced dataset, the random oversampling method was used in making the dataset balanced.

5.2.1 Random Oversampling

Random oversampling balances a dataset by randomly duplicating the records of a minority class until the number of records for the minority class equals the number of records of the majority class Figure 11. The result of implementing random oversampling is a balanced dataset with a 55.60% data point increase from the original data (Al-Shabi, 2019).

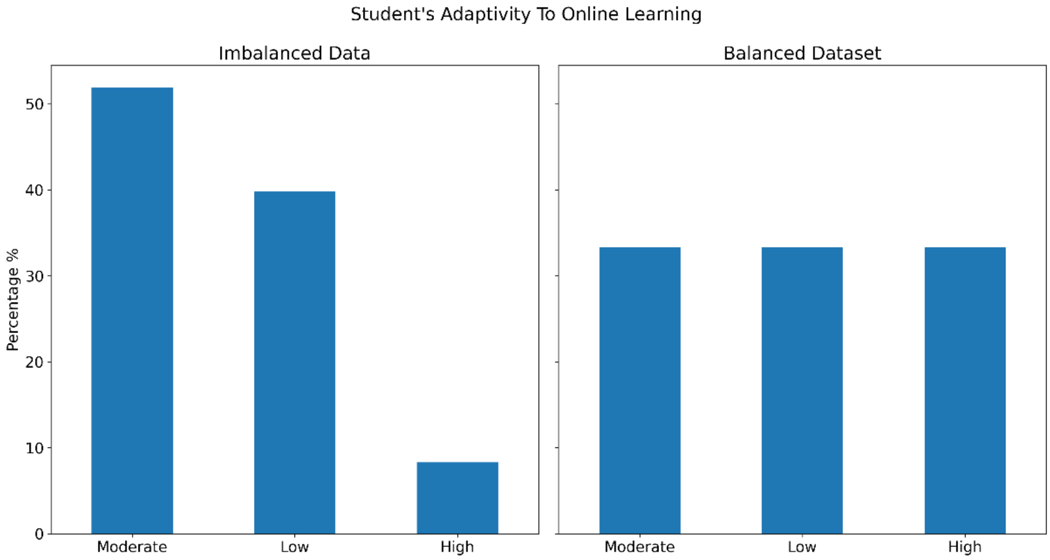

Figure 11. Comparison of balanced and imbalanced data

5.2.2 Revaluating models

Accuracy, ROC_AUC, and F1_score metrics were calculated for evaluating the models on the test dataset of the balanced data and the best scores for each metric were highlighted in red as shown below. The best model on the test dataset was found to be XG-Boost based on its F1 score.

Table 5. Results on balanced Dataset

Model Name	Accuracy	F1_Score	ROC_AUC
Logistic Regression	0.707	0.703	0.797
Decision Tree Classifier	0.920	0.919	0.983
nearest Neighbors	0.840	0.840	0.956
Random Forest Classifier	0.923	0.922	0.985
Support Vector Classifier	0.590	0.593	0.829
XGBoost Classifier	0.925	0.925	0.985

5.2.3 Results after Hyperparameter Tuning

Even after hyperparameter tuning was done again on all models and they were re-trained, XGBoost was still able to get the best accuracy, with 92.5 percent (Figure 12)

	Model_Name	F1_Score	F1_Score_tuned
0	Logistic Regression	0.702916	0.702916
1	Decision Tree	0.919249	0.919249
2	KNN	0.836890	0.866159
3	Random Forest	0.921964	0.921777
4	SVC	0.592858	0.799199
5	XGBoost	0.924656	0.924656

Figure 12. Comparing F1_score after applying hyperparameters tuning

5.2.4 Confusion matrix

The confusion matrix suggested that the XG-boost Classifier can make accurate predictions for the 3 classes without performing poorly in predicting any class with low levels of misclassifications (Figure 13).

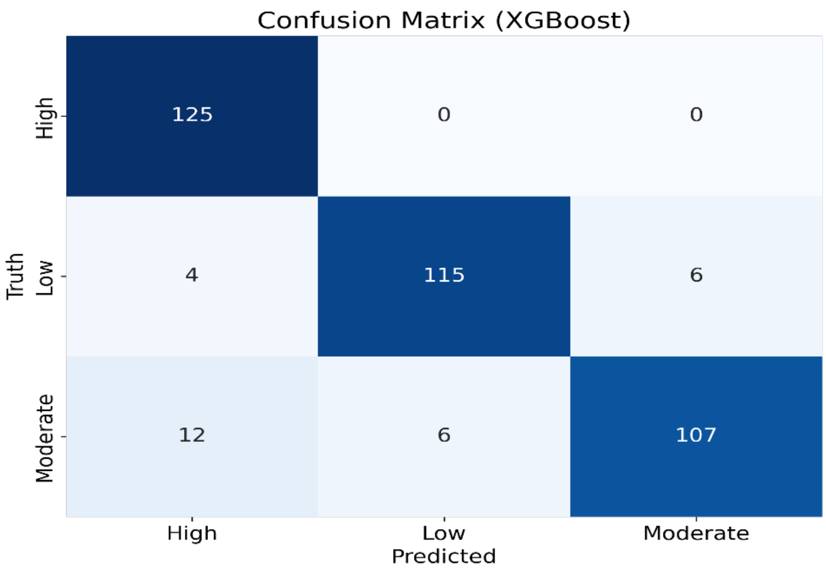

Figure 13. Confusion matrix of XG-Boost classifier

5.2.5 Cross-Validation

It is a method that helps verify how accurate a machine learning model would perform on different subsets of data. K-fold is a cross-validation technique that iterates the dataset K times and divides the dataset into training and testing subsets.

In this study, 5-Folds were used on the training dataset, and therefore for each iteration, the validation and training datasets were generated.

```
n = 1
test_preds_cv = []
for train_index, test_index in fold.split(X_train_bal, y_train_bal):
    print('{} fold'.format(n))
    train_set, val_set = X_train_bal.iloc[train_index], X_train_bal.iloc[test_index]
    y_train_set, y_val_set = y_train_bal.iloc[train_index], y_train_bal.iloc[test_index]

    model = XGBClassifier(random_state=20, n_estimators=1000, tree_method="hist",
                          enable_categorical=True,  early_stopping_rounds=50, eval_metric = 'auc')
    model.fit(train_set, y_train_set, eval_set=[(val_set, y_val_set)], verbose=100)

    test_preds=model.predict_proba(X_test_bal)
    test_preds_cv.append(test_preds)
    n+=1
    del model, test_preds
```

The held-out test set was used to check the implemented cross-validation technique for improvement since the cross-validation aims to achieve better training for our model. Achieving the average, the predictions on the held-out test set were obtained as probabilities during each iteration. The averaged predictions were then used to generate metric scores against the actual test labels. Training the XG-Boost classifier on the training dataset in one pass produced an F1 Score of 0.925, while the result of using a 5-Fold cross-validation on the training dataset produces the same F1 Score of 0.925. Since cold is more time-consuming to train and generate predictions and it does not give an improvement in F1_Score, k-Fold cross-validation was not found to be effective in predicting student adaptability with the XG-Boost classifier.

5.3 Comparison of Model Performance (Imbalanced and Balanced)

After training the 6 machine learning algorithms on the balanced dataset and imbalanced datasets. F1_score metric is used to determine if the models perform better on the imbalanced or the balanced dataset.

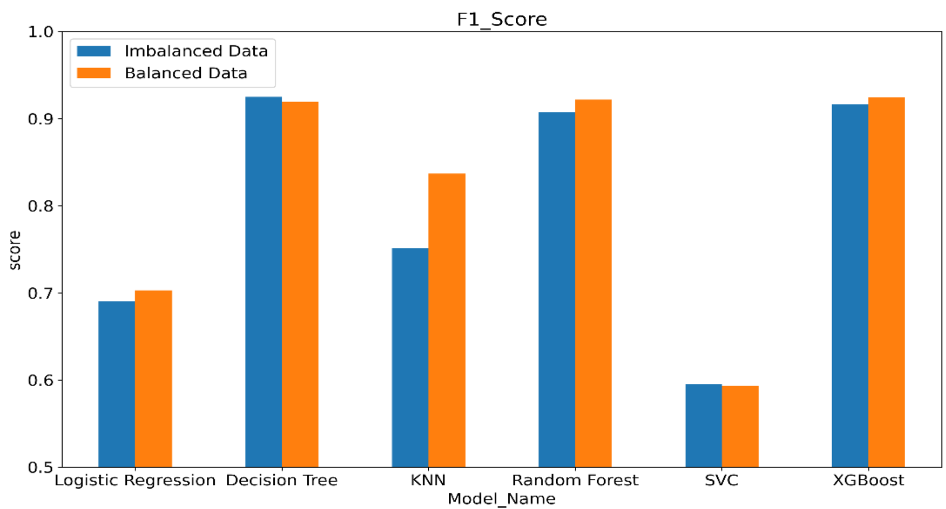

Figure 14: Comparison between model performance balanced and imbalanced

From Figure 14, the F1 score improved for most models when trained on the balanced dataset. However, there were exceptions like Decision Tree and Support Vector classifiers where there was a slight decrease in F1_score whereas KNN sees an 11% increase, the highest improvement in F1_score when trained on the balanced dataset.

6. Discussion

The influence of balanced and imbalanced data on the deployment of machine learning algorithms to predict adaptability was examined in this work. Oversampling seems to be a successful strategy for imbalanced data classification, as the F1 score increased for most models when retrained on the balanced dataset. However, this indicates the importance of balancing the data because imbalanced input affects the machine learning algorithms, causing it to anticipate the imbalanced distribution of classes for the majority of samples (Kaur and Gosain, 2018).

On the other hand, throughout the experiment, the Tree-based models, Decision trees, Random Forest, and XGboost obtained the best accuracies on balanced and unbalanced data, even after adjusting hyperparameters and applying cross validation. From my point of view, this demonstrates the resiliency of these models on our dataset, proving our proper decision to deploy the best model with the highest accuracy.

7. Future Work

Even though the study focused on the data relevant to the students, adaptability could be predicted based on factors related to the teaching style, which may limit the study. It was observed that "IT students" tend to have high adaptation to online education. Also, as "Class duration" was found to be the most impactful variable in predicting the adaptability levels, so, in my opinion, faculty play a significant role in creating an exciting online course, thereby shaping the attitudes of the students and helping them pass.

However, the study has contributed to finding patterns based on student variables. So, similar methods and performance metrics could be used in a future study to figure out the relationship between how well a student can adapt and the type of teaching method.

8. Conclusion

In this study, six machine learning algorithms were trained to predict students' adaptability level to online education. The decision Tree Classifier achieved the best F1 Score of 0.925 when trained on the original imbalanced dataset as the weights in the decision tree classifier enable the model to create a bias to compensate for the classes that have little representation in the overall dataset (Truica and Leordeanu, 2017). However, XG-Boost Classifier achieved the best F1 Score of 0.925 when trained on the oversampled balanced dataset as it creates individual tress through applying multiple cores along with organizing the data to minimize the time for lookup which in turn decreases the time for training a model that in turn increases the performance of XG-Boost (Ramraj et al., 2016).

The output of this project can serve as a baseline to help enhance the decision of administrators in educational institutions as they try adopting online education.

Funding: This research received no external funding

Data Availability Statement: Mahmudul Hasan Suzan. Students Adaptability Level in Online Education Dataset. [Kaggle]

Conflicts of Interest: I declare no conflict of interest

References

1. SUZAN, M. H., SAMRIN, N. A., BISWAS, A. A. & PRAMANIK, A. Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021. IEEE, 1-7.
2. SINGH, V. & THURMAN, A. 2019. How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018). *American Journal of Distance Education,* 33**,** 289-306.
3. Vlachopoulos, D. (2020). COVID-19: Threat or Opportunity for Online Education? *Higher Learning Research Communications, 10* (1).DOI:10.18870/hlrc.v10i1.1179
4. Drane, C.; Vernon, L.; O'Shea, S. The Impact of 'Learning at Home' on the Educational Outcomes of VulnerableChildren in Australia during the COVID-19 Pandemic; Literature Review Prepared by the National Centre forStudent Equity in Higher Education; Curtin University: Bentley, Australia, 2020.
5. Yusuf, B.N. Are we prepared enough? A case study of challenges in online learning in a private higher learning institution during the Covid-19 outbreaks. Adv. Soc. Sci. Res. J. 2020, 7, 205–212.
6. Thomas, M.S.C., Rogers, C. Education, the science of learning, and the COVID-19 crisis. Prospects 49, 87–90 (2020). https://doi.org/10.1007/s11125-020-09468-z
7. D. Kucak, V. Jurici´c, and G. ambi´c, "Machine learning in education-a survey of current research trends." Annals of DAAAM & Proceedings, vol. 29, 2018
8. ASMAEL, N. M., EMBABY, A. K., MOUSA, B., ASARE, M. T., MISSAH, Y. M., AHMAD, S. A., JAVED, M. N., SAEED, M. Z., SYED, H. & ASLAM, M. A. 2017. International Research Journal of Engineering and Technology (IRJET).
9. HYUN, Y. & KIM, D. 2022. Development of Deep-Learning-Based Single-Molecule Localization Image Analysis. *International Journal of Molecular Sciences,* 23**,** 6896.
10. Feng, J. X. H. M. S. a. Y. S., 2014. Robust logistic regression and classification. *Advances in Neural Information Processing Systems*, Volume 27.
11. Charbuty, B. a. A. A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1), pp. 20-28.
12. Jawthari, M. a. S. V., 2021. Predicting students' academic performance using a modified kNN algorithm. *Pollack Periodica*, 16(3), pp. 20-26.
13. Abdulkareem, N. a. A. A., 2021. Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business,* 5(2), pp. 128-142.
14. Sunil, R. 2017. Understanding Support Vector Machine(SVM) algorithm from examples (along with code). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code.
15. Asselman, A. K. M. a. A. S., 2021. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments,* pp. 1-20.
16. AL-SHABI, M. 2019. Credit card fraud detection using autoencoder model in unbalanced datasets. *Journal of Advances in Mathematics and Computer Science,* 33**,** 1-16

17. KAUR, P. & GOSAIN, A. 2018. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. *ICT Based Innovations.* Springer.
18. TRUICA, C.-O. & LEORDEANU, C. A. 2017. Classification of an imbalanced data set using decision tree algorithms. *Univ. Politech. Bucharest Sci. Bull. Ser. C Electr. Eng. Comput. Sci,* 79**,** 69-84.
19. RAMRAJ, S., UZIR, N., SUNIL, R. & BANERJEE, S. 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications,* 9.