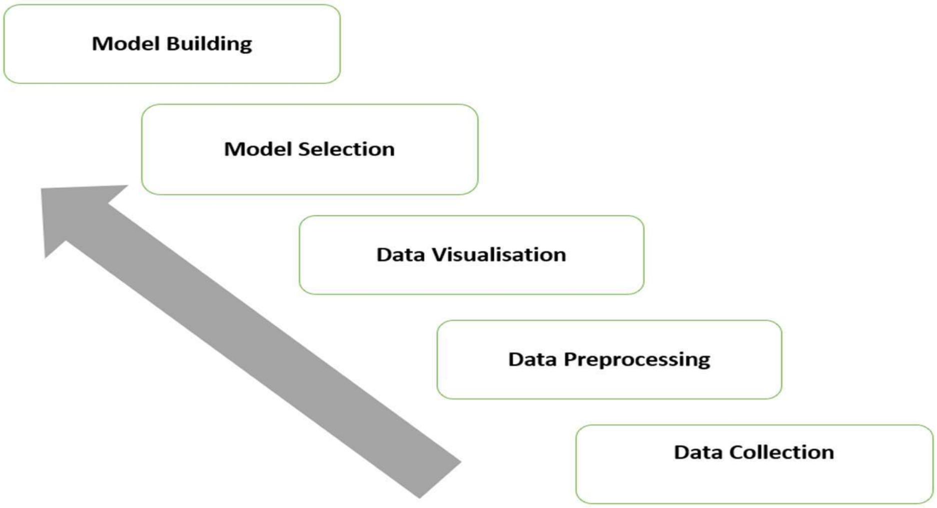
Data Set and Analysis:

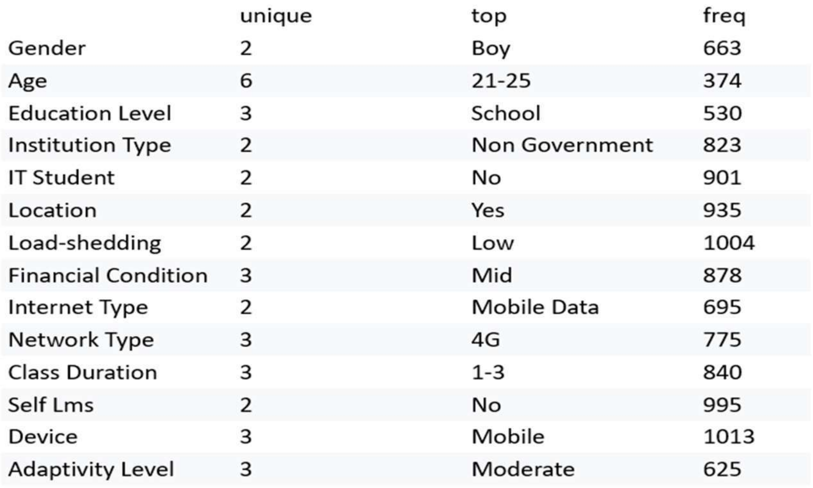
Methodology used :

The current technological advances in the field of education is a boon for the society especially in the COVID era where all the educational institution has been closed to curb the advances of virus outbreak. The methodology which we have used to analyse the dataset using a machine learning model is described by the flowchart



DATA COLLECTION

The dataset used in this dataset is from Kaggle. The Data has been taken from the online and offline survey analysis in Bangladesh from December 10, 2020 to February 5, 2021. The survey is taken from various educational levels (School , University, and colleges). The Dataset contains several features for determining the adaptability of students towards the online education. The salient features of the dataset are described by the following table:



	unique	top	freq
Gender	2	Boy	663
Age	6	21-25	374
Education Level	3	School	530
Institution Type	2	Non Government	823
IT Student	2	No	901
Location	2	Yes	935
Load-shedding	2	Low	1004
Financial Condition	3	Mid	878
Internet Type	2	Mobile Data	695
Network Type	3	4G	775
Class Duration	3	1-3	840
Self Lms	2	No	995
Device	3	Mobile	1013
Adaptivity Level	3	Moderate	625

According to the dataset we found out the adaptivity level of students of different age groups by the Bar chart shown in the figure. The Bar chart is a data visualisation technique to analyse the count of several parameters of the dataset. Here in this case, The adaptivity level of the students are shown among various age groups. The Comparison between the levels of adaptability among various age groups gives us a clear idea that the students with lower age groups are comparatively less adaptable to the digital learning.

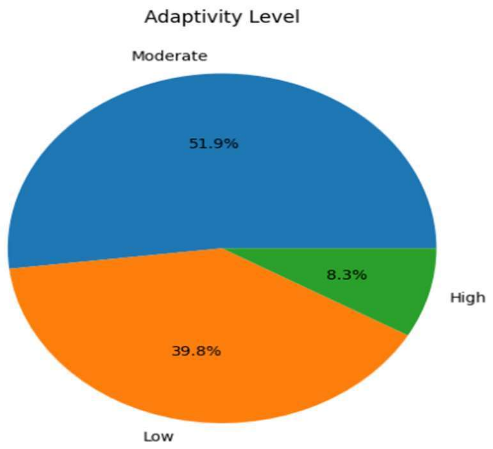
DATA PREPROCESSING

The data pre-processing involves converting the raw data into a structured form to provide it to the machine learning models for the further predictions. In this phase we have pre-processed the data into an encoded numerical format. The unique values present in each attribute has been converted to integer to provide it to machine learning model for training and testing. The libraries used for this purpose are numpy, pandas, scikit-learn.

DATA VISUALISATION

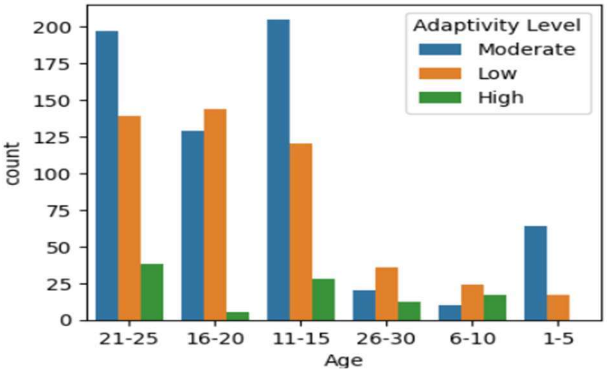
The data visualisation technique lets us to represent the data features graphically. The attributes and their dependency upon each other can be determined by using the visual elements and to understand the trend and outliers present in the data visually using graphs, charts, maps etc.

The following pie charts are shown to get the Visual analysis of the overall adaptability of the students



There are several key factors affecting this dependent variable, For better understanding we have taken all the key points under consideration.

The adaptivity level of different age groups towards online learning platform is shown by the bar graph as :



This clearly shows that students with less age groups are quite resistant towards the online learning while the students with higher age groups are productive and shows an adaptive behaviour towards the digital learning.

MODEL SELECTION

We have used many machine learning models to train our data. Selecting a machine learning model is a crucial task for achieving the best approaching result. The data is first divide into training and testing data. The model then trained using the training data and finally the testing of the formed model is performed on the testing data. Then according to the testing result we conclude the model deployment for further prediction. We have used several machine learning algorithms like Naïve Bayes, SVM, RandomForest, KNN, XGBoost for training the model. We, then, compared the results and accuracy of each models and found out the best models for the further data analysis.

Naïve Bayes Classifier:

This algorithms works on the probability theory under Bayes' theorem. The bayes theorem gives us the posterior probability based under given condition. The Bayes theorem formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) gives us the probability of occurrence of event A given that B has already occurred.

P(B|A) gives us the probability of occurrence of event B given that A has already occurred.

P(A) gives us the probability of occurrence of event A.

P(B) gives us the probability of occurrence of event B.

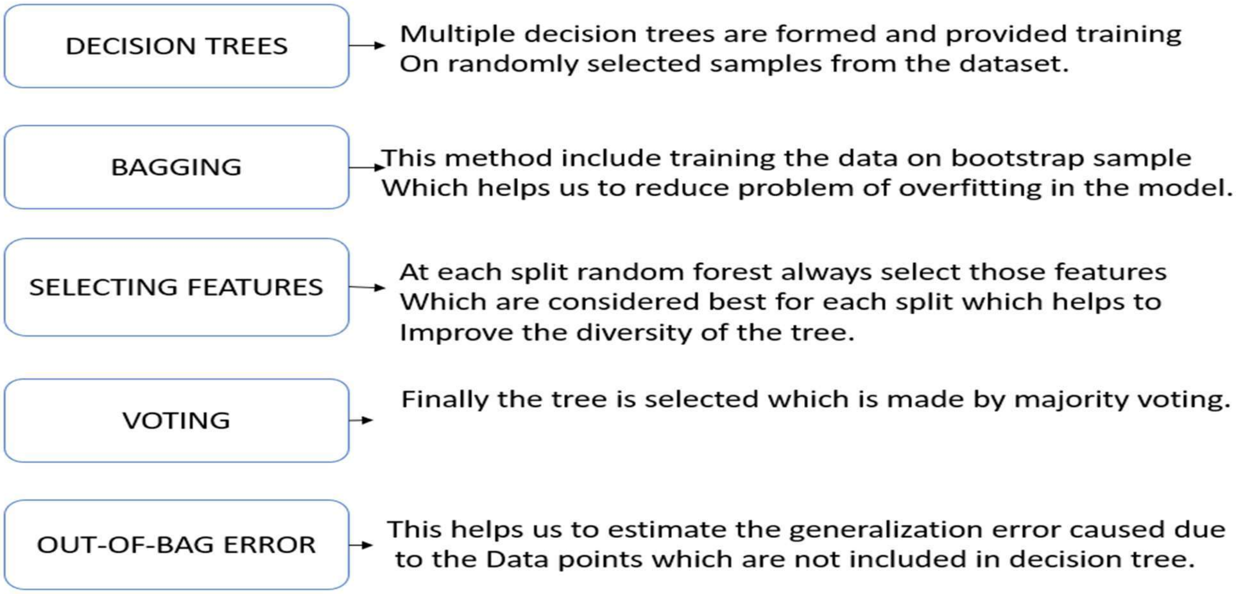
This algorithm gives us the naïve assumptions about the classification of an attribute under several correlated features. The theorem is most preferred because of its simplicity and speed. It results in an accurate predictions and the further classification of new data based on the model.

Support Vector Machines(SVM)

It is a powerful machine learning tool which is used to classify the model under different categories based on a decision boundary called hyperplane. The support vectors are made closest to the hyperplane and support vectors helps the model to classify the newly generated data points and results in a better analysed result.

Random Forest

Random forest works by creating several trees and each tree is trained by the subsets of the dataset created randomly. This algorithm uses following:



DECISION TREES → Multiple decision trees are formed and provided training On randomly selected samples from the dataset.

BAGGING → This method include training the data on bootstrap sample Which helps us to reduce problem of overfitting in the model.

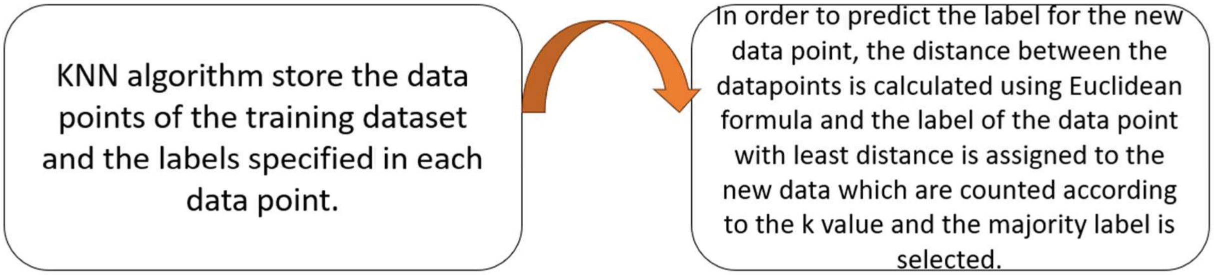
SELECTING FEATURES → At each split random forest always select those features Which are considered best for each split which helps to Improve the diversity of the tree.

VOTING → Finally the tree is selected which is made by majority voting.

OUT-OF-BAG ERROR → This helps us to estimate the generalization error caused due to the Data points which are not included in decision tree.

KNN (K Nearest neighbour):

This algorithm is also known as lazy learning algorithm. It does not involve training the model instead it memorizes training data.



KNN algorithm store the data points of the training dataset and the labels specified in each data point.

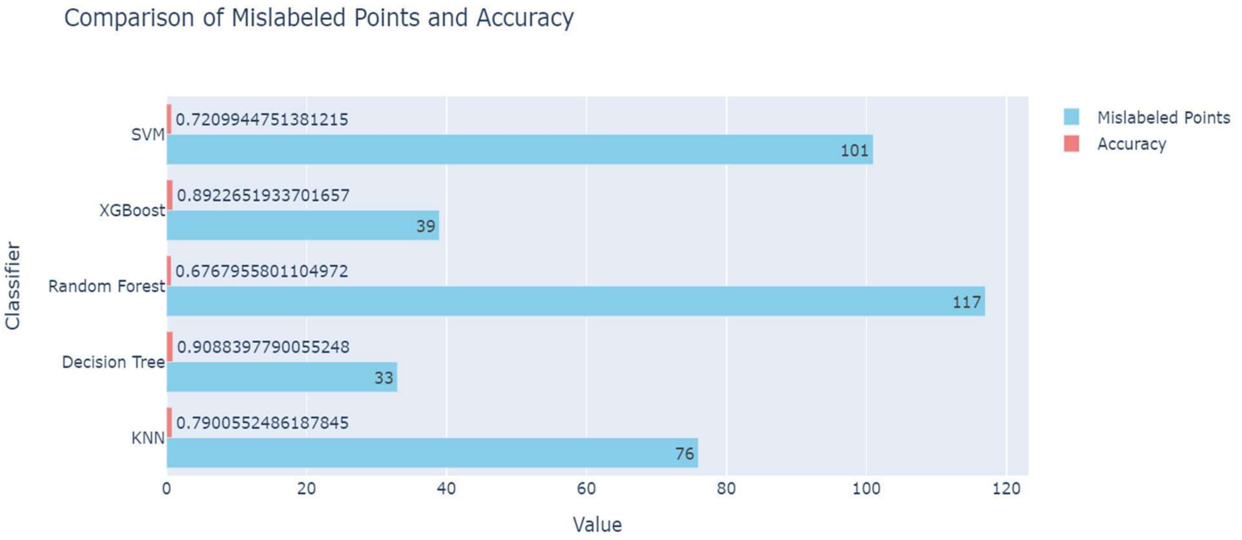
In order to predict the label for the new data point, the distance between the datapoints is calculated using Euclidean formula and the label of the data point with least distance is assigned to the new data which are counted according to the k value and the majority label is selected.

XGBoost:

It is also called Extreme gradient Boosting. It is used for the supervised learning algorithms like regression and classification. It is based on the gradient boosting technique which also helps in handling the missing values in turn reducing the time for data pre-processing.

MODEL BUILDING:

After building the models based on the dataset we conclusively predicted the data points and checked the accuracy and compared the accuracy of each models. Here we got the highest accuracy for Decision Tree classifier model (90.8%) and the lowest accuracy rate was shown by Random Forest Model (about 67.67%)


Comparison of Mislabeled Points and Accuracy
Mislabeled Points
Accuracy
Classifier
SVM
XGBoost
Random Forest
Decision Tree
KNN
Value