Predicting Student's Adaptability to Online Learning

Kayde Varona

2022-06-16

I. Introduction

Many drastic changes have happened to us every since COVID hit. One of the many changes that education institutions adopt is online learning. Most schools worldwide shifted from the traditional classroom setting to an online environment. Nearly every student in every grade level has some form of an online class that they took. It is essential to know how students adapt to change and what challenges students face in an online environment.

Objectives

- The main objective of this paper is to predict the level of adaptability of a student using different machine learning techniques.
- Second is to identify a group of students more likely to adapt seamlessly to an online environment.
- Lastly is to point out factors the student needs to succeed in online learning.

Results

- The best model identified by the accuracy and kappa is the Bagged Model
- Groups that are more lily to adapt to online environment are students ages 21-25 at University Level. Students with better resources such as better network/internet connectivity and financial condition are more likely to succeed.
- Age, class duration, location (urban or rural area), network type, and educational level are the most important factors that should be taken into consideration for students to succeed in an online environment.

II. Data Sources and Variable Definitions

The data utilized in this project is from the "Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches" paper published in 2021. This dataset is imported from Kaggle. The authors of the said paper used online and offline surveys collected from December 10, 2020, to February 5, 2021, in Bangladesh. For more information regarding this paper, please click here https://ieeexplore.ieee.org/document/9579741

One limitation of the data is that it is imbalanced. Out of 1205 observations, there are more students with low and moderate adaptability at 480 and 625 respectively than the positive outcome, which is high at 100 observations

```
High Moderate      Low
 100      625      480
```

The variables used in this study are

(1) Gender: Girl (0), Boy (1)
(2) Age: Age group Around 1 to 5 (0), 6 to 10 (1), 11 to 15 (2),16 to 20 (3), 21 to 25 (4), 26 to 30 (5)

(3) Education Level: School (0), College (1), University (2)
(4) Institutional Type: Non Government (0), Government(1)

(5) IT Student : No (0), Yes (1)

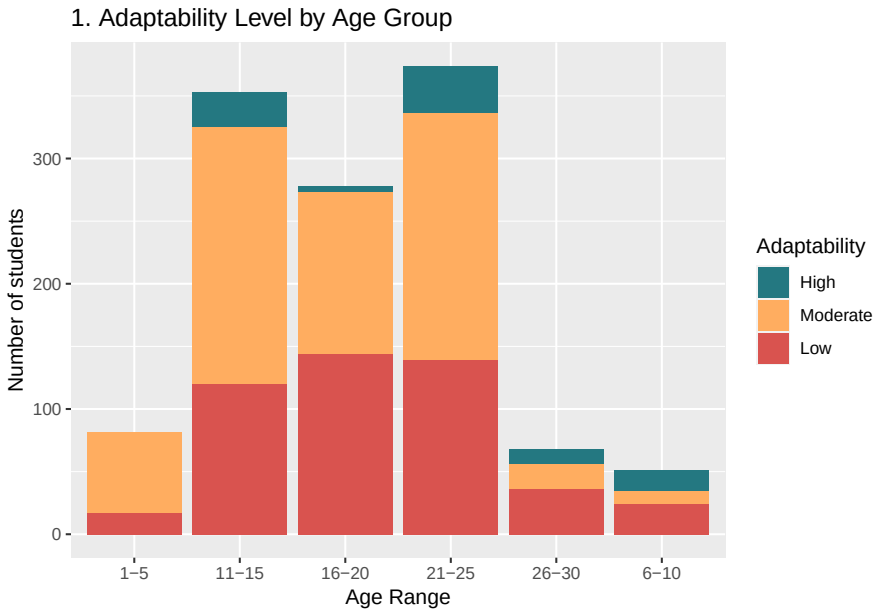
(6) Location: Is the educational institution located in the city/town No (0), Yes (1)
(7) Load-Shedding: Low (0), High (1)
(8) Financial Condition: Poor (0), Mid (1), Rich (2)
(9) Internet Type: Mobile Data (0), Wifi (1)
(10) Network Type: 2G (0), 3G (1), 4G (2)
(11) Class Duration: 0 (0), 1 to 3 Hours (1),3 to 6 Hours (2)
(12) Self LMS: No (0), Yes (1)
(13) Device: Most used device Tab (0), Mobile (1), Computer (2)
(14) Adaptability Level: Low (0), Moderate (1), High (2)

III. Data Exploration

To get more insight regarding this dataset, exploratory data analysis is utilized using ggplot2.

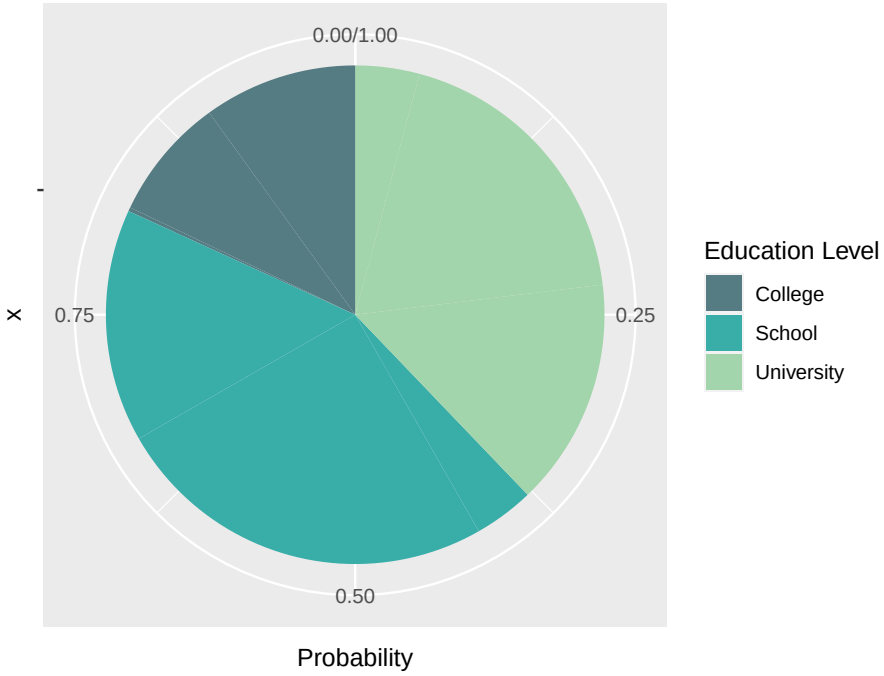
This paper focused on visualizing the adaptability level using different groups of students.

The first graph shows the adaptability level of different age groups. Students ages 1 to 5 do not adapt to online learning. Students start having high adapting at the age of 6-10. Most of the students in this study are ages 21-25, most of whom have moderate adaptability.



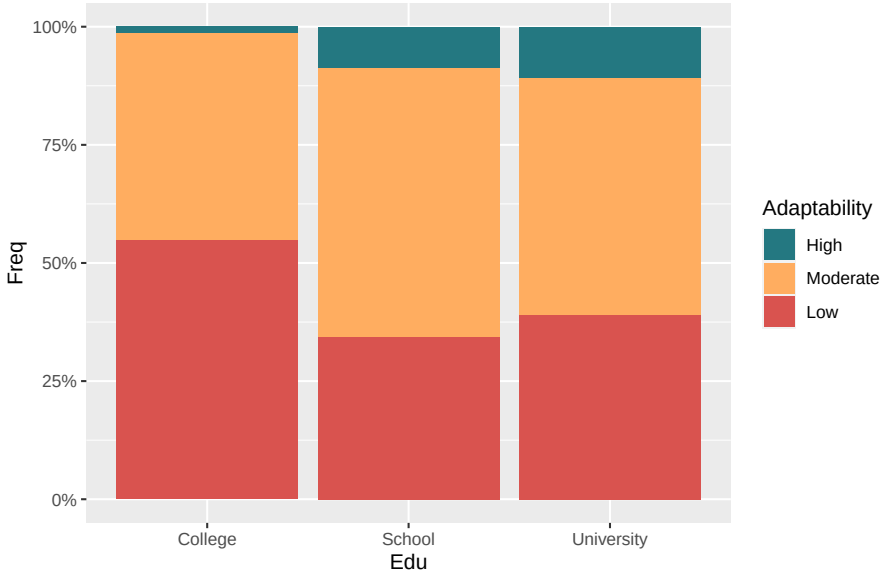
The second graph shows the proportion of educational institution level. Most of the students in the dataset are either in school or university.

2.a. Pie Chart of Eduction Level

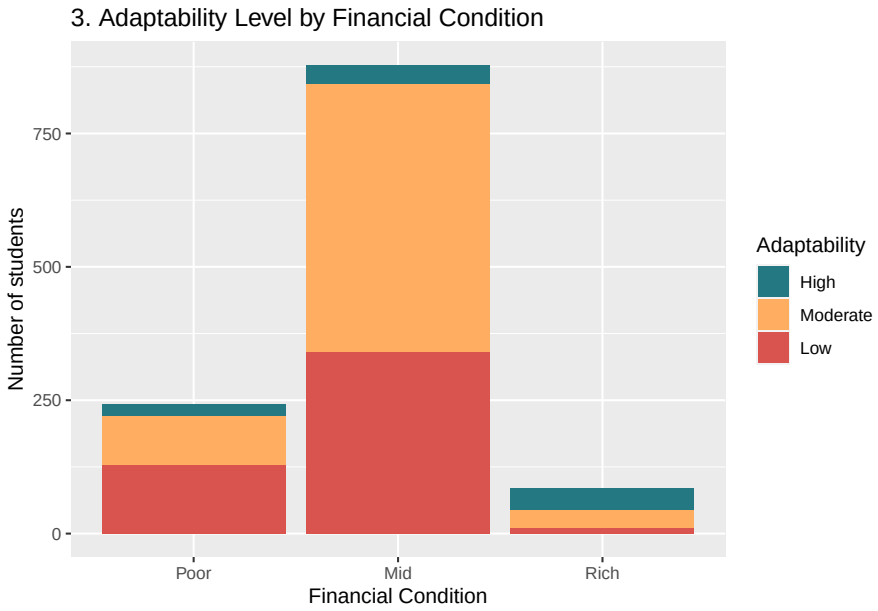


Looking more at the proportion of educational level, on the third graph, students on the college level have overall the lowest adaptability level. The groups with the highest adaptability scores are groups of students in University. Even though university students have the highest adaptability overall, 40% are still considered with low adaptability scores.

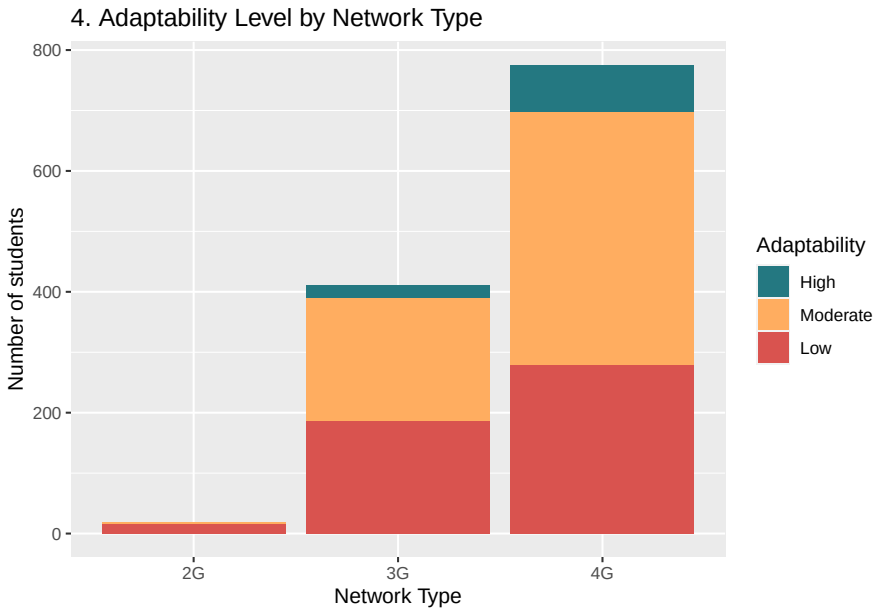
2.b. Adaptability per Educational Level



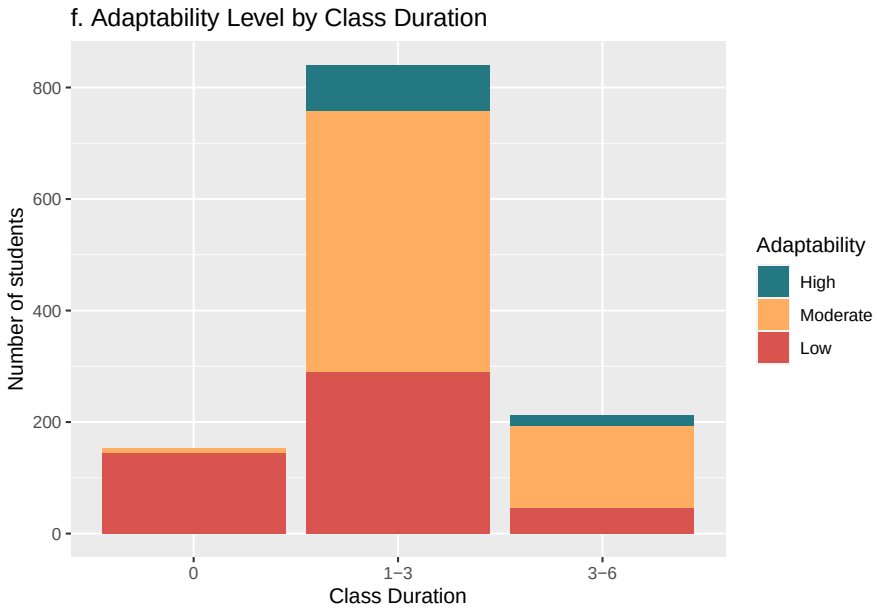
The graph number 3 shows adaptability levels by financial condition; financial conditions are divided into three groups, poor, middle class, and rich. Most of the students in the dataset are considered middle class, followed by poor, and lastly, rich. Students with better financial conditions are more likely to have higher adaptability.



The fourth graph shows the network types available in Bangladesh, 4G being the fastest. Despite having internet connection issues in the rural areas. The majority of the students either have 3G or 4G. Students with higher network connectivity tend to be more adaptive than those only having 2G connectivity.



The last graph shows the different class duration; the optimum number of hours when students have a high adaptability level is a class with a one to three-hour period. Below that, they mostly have low adaptability.



III. Data Preparation

The dataset is converted to a data frame to prepare the data for analysis. The values are also binarized and converted into factors. All the variables in this study are factors. The data set is divided into training and testing sets, 80% used for training and 20% for testing. The outcome variable is also removed from the training set.

There are three different Machine Learning algorithms used to predict the student's adaptability level. The first model used to indicate adaptability is a **Decision Tree Model**. This model is used because it is effective in categorizing the dataset. The second model used is the **Bagged Model**. This model is utilized to improve the model accuracy and reduce overfitting. The final model that is used is a **Random Forest Model**. This model is used because it is good with unbalanced data like the adaptability scores. Utilizing the random forest model will reduce the variance of the decision trees. The accuracy and kappa are analyzed to determine the best model.

Summary of the Data

```
 Gender  Age     Education_Level Institution_Type IT_Student Location
 0:542   0: 81   0:530           0:823            0:901      0:270
 1:663   1: 51   1:219           1:382            1:304      1:935
         2:353   2:456
         3:278
         4:374
         5: 68
 Load_shedding Financial_Condition Internet_Type Network_Type Class_Duration
 0:1004        0:242               0:695         0: 19        0:154
 1: 201        1:878               1:510         1:411        1:840
               2: 85                             2:775        2:211


 Self_Lms Device   Adaptability_Level
 0:995    0:  30   0:480
 1:210    1:1013   1:625
```

```
          2: 162   2:100
```

IV. Results

Decision Tree Model Analysis

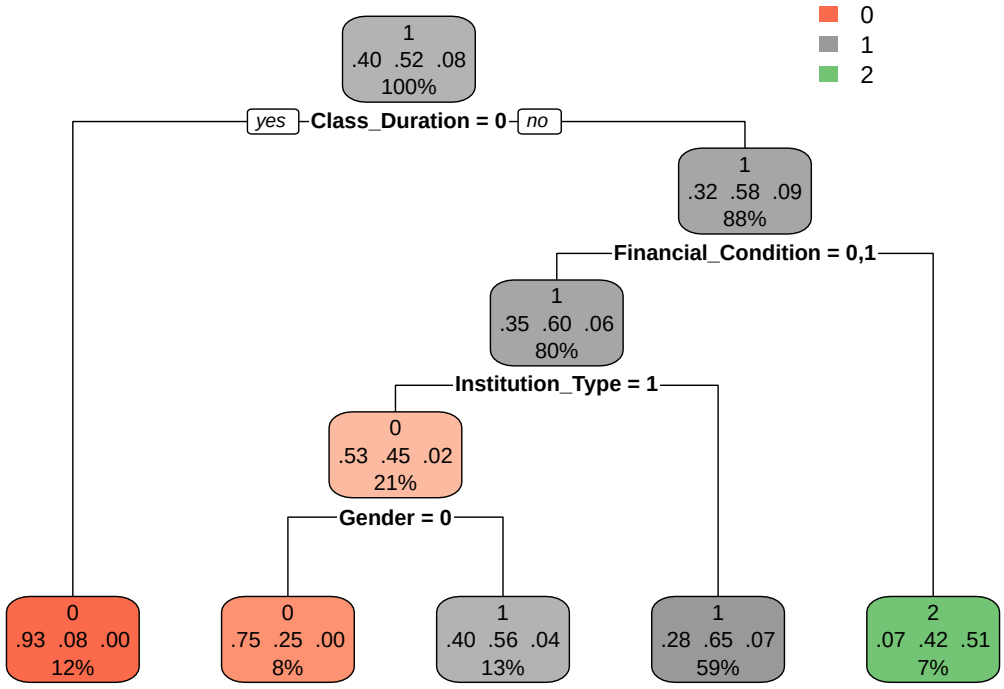
a. CART Results

```
CART

964 samples
 13 predictor
  3 classes: '0', '1', '2'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 964, 964, 964, 964, 964, 964, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.01831897  0.6642858  0.3773188
  0.03089080  0.6410012  0.3171650
  0.21982759  0.5656827  0.1131574

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01831897.


n= 964

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 964 464 1 (0.39834025 0.51867220 0.08298755)
   2) Class_Duration=0 120   9 0 (0.92500000 0.07500000 0.00000000) *
   3) Class_Duration=1,2 844 353 1 (0.32345972 0.58175355 0.09478673)
     6) Financial_Condition=0,1 773 312 1 (0.34670116 0.59637775 0.05692109)
      12) Institution_Type=1 203  96 0 (0.52709360 0.44827586 0.02463054)
        24) Gender=0 73  18 0 (0.75342466 0.24657534 0.00000000) *
        25) Gender=1 130  57 1 (0.40000000 0.56153846 0.03846154) *
      13) Institution_Type=0 570 200 1 (0.28245614 0.64912281 0.06842105) *
     7) Financial_Condition=2 71  35 2 (0.07042254 0.42253521 0.50704225) *
```

b. Decision Tree Plot

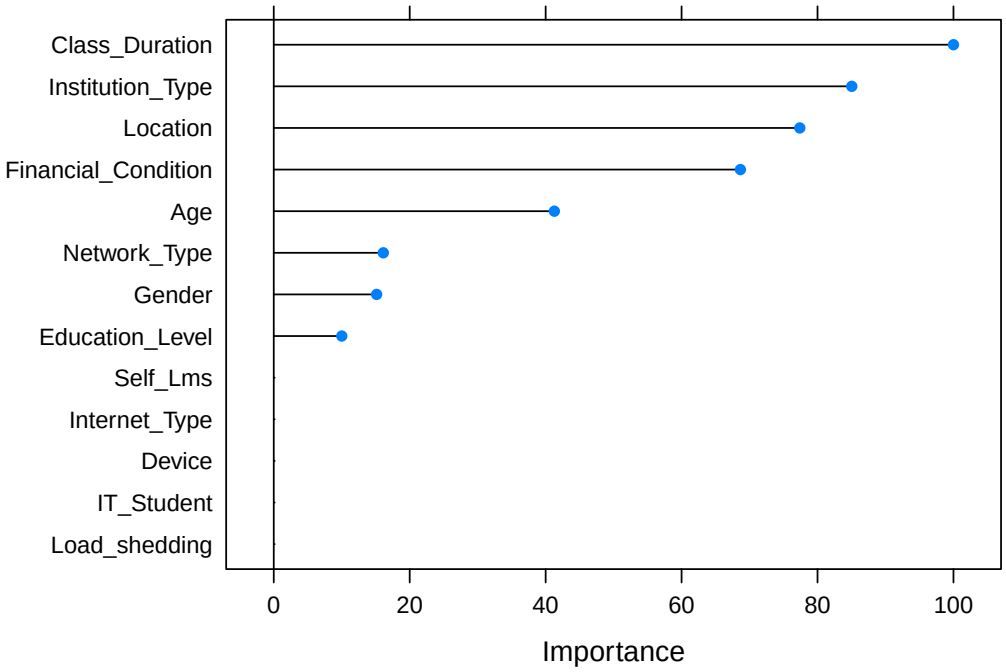
The first leaf of the decision tree shows that when class duration is less than an hour (class duration = 0), 12% of that data is composed of low adapters; otherwise, they are moredate level of adaptability. The following boundary/ leaf is set: if the class is more than 1 hour, the financial condition is considered. If the student is rich (financial condition = 2), they are more likely to be very adaptive to an online setting. This may be the case because they have more resources such as better internet connectivity, faster computers, and alike. If they are part of a low median income, the institution type, whether they are in a private school

(0) or public school (1), is taken into account. When the institution is private, 21% of our students with low adaptability level, and the following boundary is distinction after that is the student's gender. 59% of the students from public schools are moderately adaptive to online learning.



c. Decision Tree Important Variables:

The important variables identified by the decision tree are class duration, institution type, location financial condition, age network type, gender and education level.



d. Decision Tree Confusion Matrix

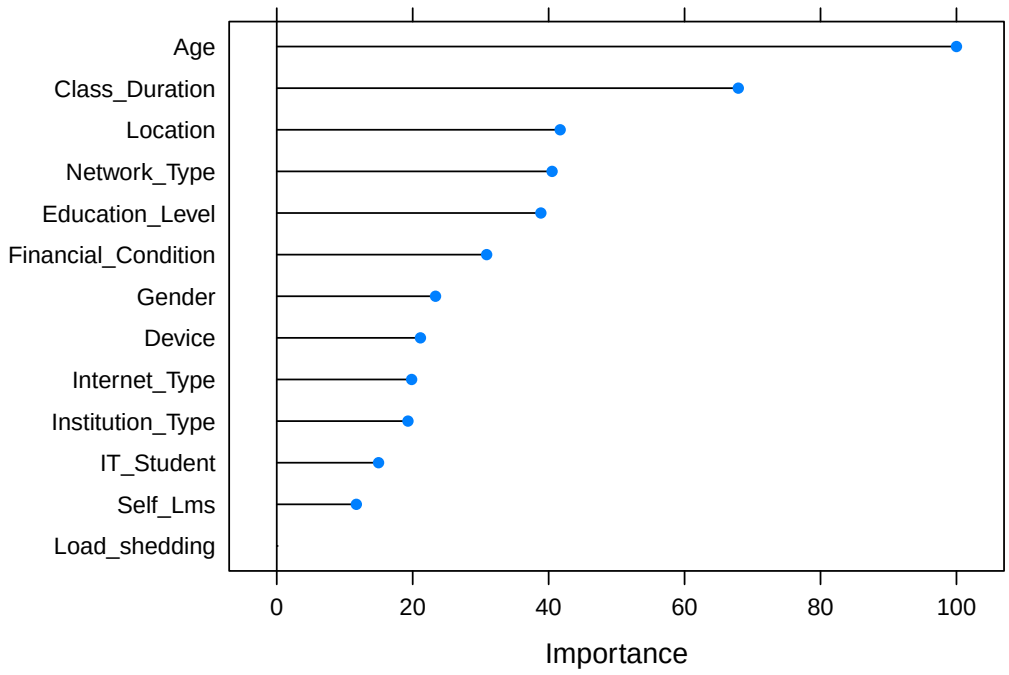
The Decision Tree's accuracy is at 70.12% at 0.4315 kappa.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0  43   2   0
         1  50 120  14
         2   3   3   6

Overall Statistics

               Accuracy : 0.7012
                 95% CI : (0.6392, 0.7583)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : 6.107e-09

                  Kappa : 0.4315

 Mcnemar's Test P-Value : 9.105e-12

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity            0.4479   0.9600  0.30000
Specificity            0.9862   0.4483  0.97285
Pos Pred Value         0.9556   0.6522  0.50000
Neg Pred Value         0.7296   0.9123  0.93886
Prevalence             0.3983   0.5187  0.08299
Detection Rate         0.1784   0.4979  0.02490
Detection Prevalence   0.1867   0.7635  0.04979
Balanced Accuracy      0.7171   0.7041  0.63643
```

Bagged Model Analysis

Bagged Model Results The bagged models show that ***age, class duration, location, network type, and educational level*** are the top five determining factors of adaptability level. This model has 92.53% accuracy with 0.87 kappas. The bagged model is better at predicting if the student has low adaptability with an overall higher specificity ranging from 92% to 99%. The sensitivity for the model is 90% for Low, 94% for Moderate, and 90% for High.

a. Bagged Model Important Variables



b. Bagged Model Confusion Matrix

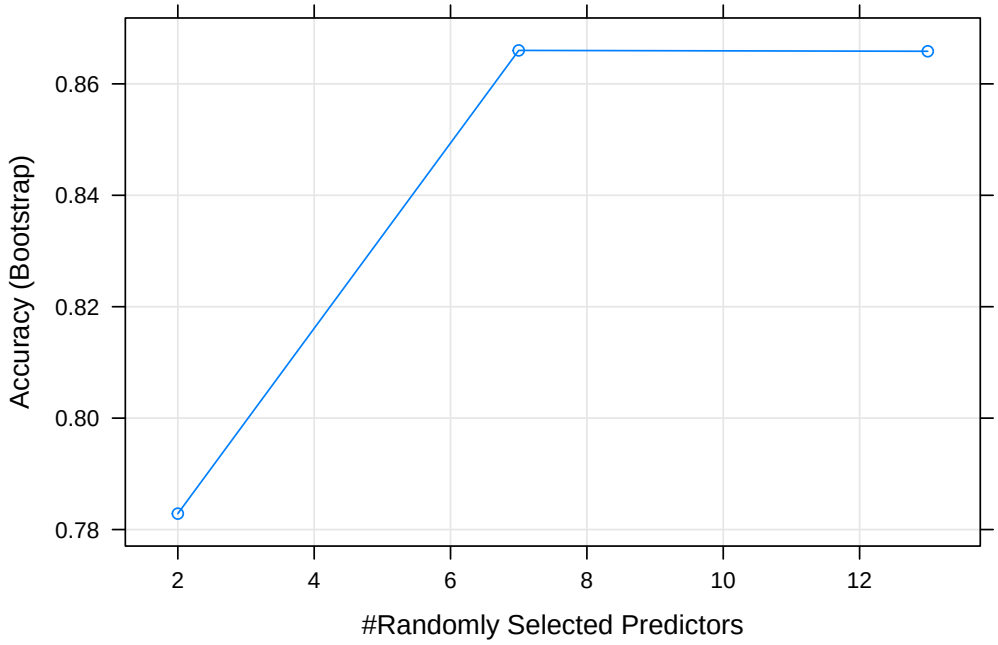
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0  87   7   0
         1   7 118   2
         2   2   0  18

Overall Statistics

               Accuracy : 0.9253
                 95% CI : (0.8845, 0.9551)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8677

 Mcnemar's Test P-Value : 0.2615

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity            0.9062   0.9440  0.90000
Specificity            0.9517   0.9224  0.99095
Pos Pred Value         0.9255   0.9291  0.90000
Neg Pred Value         0.9388   0.9386  0.99095
Prevalence             0.3983   0.5187  0.08299
Detection Rate         0.3610   0.4896  0.07469
Detection Prevalence   0.3900   0.5270  0.08299
Balanced Accuracy      0.9290   0.9332  0.94548
```

Random Forest Model Anlysis

Random Forest Model Results The random forest model shows that the ***duration of the class, age group, financial condition, gender, and network type*** are among the most important variables for this model. The number of predictors used for the model is 13. Looking at the confusion matrix, the overall accuracy of the model is 91.29% with a kappa of 0.84. Like the bagged model, the specificity range is greater than the sensitivity.

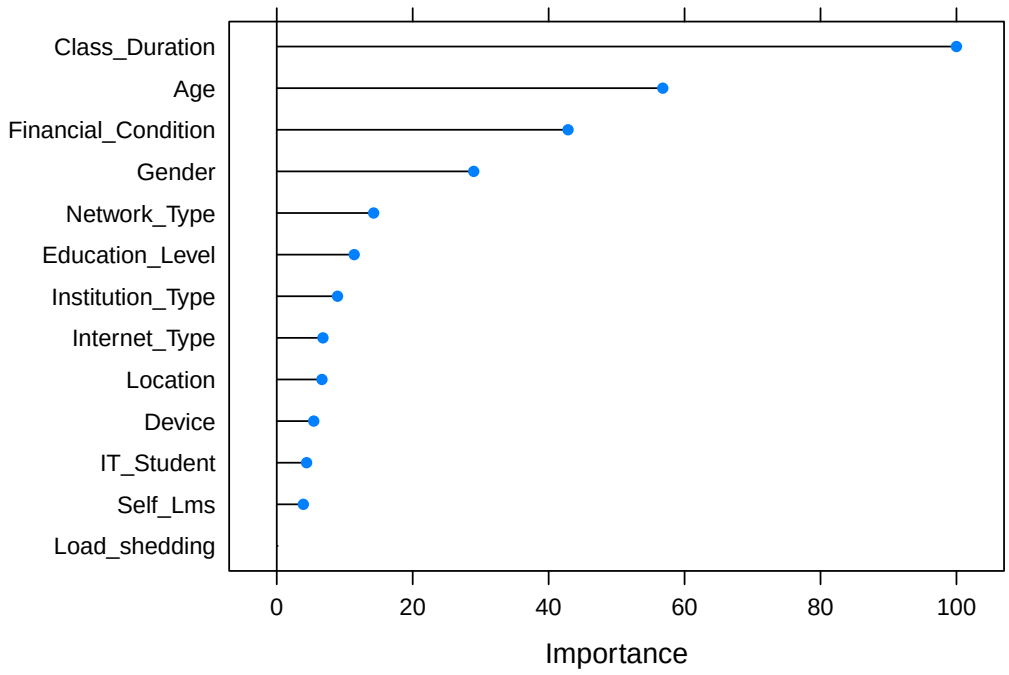
a. Random Forest Optimal Number of Predictors

The optimal number of predictors is 13. This is when the accuracy is at the highest.




```
Random Forest

964 samples
 13 predictor
  3 classes: '0', '1', '2'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 964, 964, 964, 964, 964, 964, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7828566  0.6008168
   7    0.8660080  0.7599857
  13    0.8658520  0.7605260

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 7.
```

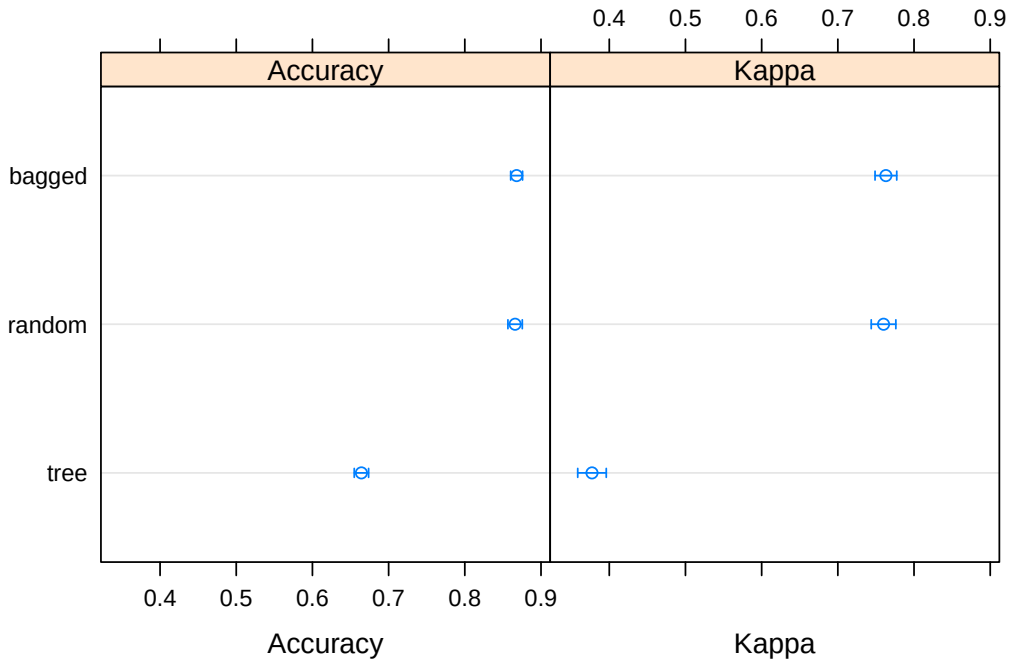
b. Random Forest Model Important Variables



c. Random Forest Model Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0  86   7   0
         1   8 118   4
         2   2   0  16

Overall Statistics

               Accuracy : 0.9129
                 95% CI : (0.8699, 0.9453)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8445

 Mcnemar's Test P-Value : 0.1084

Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity            0.8958   0.9440  0.80000
Specificity            0.9517   0.8966  0.99095
Pos Pred Value         0.9247   0.9077  0.88889
Neg Pred Value         0.9324   0.9369  0.98206
Prevalence             0.3983   0.5187  0.08299
Detection Rate         0.3568   0.4896  0.06639
Detection Prevalence   0.3859   0.5394  0.07469
Balanced Accuracy      0.9238   0.9203  0.89548
```

V. Conclusions and Recommendations



The best model is the **Bagged Model** with the highest accuracy at **92.53%**, compared to only 91.29% for random forest and 70.12% for Simple Tree Model. The students with a higher chance of having high adaptability levels are those aged 21 to 25 and are at University. Students with better resources, such as having a better network type and financial income, are more likely to have better adaptability. The decision-makers of the educational institutions consider the students' age group when it comes to implementing an online learning environment. Younger students need one-on-one physical interaction with their teachers or instructors to be effective in school as they do not adapt online. They should also consider the class's length or duration since it has been proven that this is a critical factor across different models. They should also ensure that the students have access to a reliable network.

Citation & Reference

M. Hasan Suzan, N. A. Samrin, A. A. Biswas and A. Pramanik, "Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-7, doi: 10.1109/ICCCNT51525.2021.9579741.