IS-804 Project Final Report

Name: Uzma Hasan

Campus ID: KT98095

Title: Analyzing the E-learning System of Bangladesh in R

Table of Contents

1. Introduction

The emergence of COVID-19 in the first quarter of the year 2020 gave rise to a sudden crisis in the education system around the globe. The health concerns arising because of the deadly Coronavirus let the global communities decide to shut down in-person classes and shift to online classes. As a result of which over 1.2 billion students were out of classrooms for the longest period of time [1]. By the end of March, around 75 countries implemented or announced the closure of educational institutions [2]. Such closure led to the wide adoption of online classes in order to prevent any sort of disruption in the education sector. Since then, online education has changed the lives of students and teachers globally due to the fact that no one was prepared to shift entirely to online mode of study, dropping the age-old habit of in-person classes. Online learning vs traditional classroom environment varies profoundly in case of students motivation, satisfaction and interaction [3]. Mostly the interaction in online versus offline classes varies to a great extent.

Online classes have their own advantages and disadvantages. Authors in [4] state that some of the advantages of online classes are remote learning, comfort, accessibility, while the limitations are inefficiency and difficulty in maintaining academic integrity. Due to e-learning, a lot of time and effort is saved especially for students living in distant places from their educational institutes [5]. Also, the travel expenses and other expenses that accompany traditional learning are reduced to a great extent. However, the in-person hands on learning experience is often missing in online classes particularly for the subjects that require practical skills for learning. Often it becomes difficult for the students to perceive the knowledge or understand the practical implication of a certain theory without hands-on experience. Infact, instructors also face a lot of difficulties to deliver such topics online that must require practical or lab classes. Often it is really difficult for an instructor to interact properly with the students online. [6] proposes that E-learning should not only focus on the delivery of content, but also ensure that the students are able to work with the materials and receive appropriate and quick feedback.

Bangladesh, one of the densely populated countries in South Asia also adapted to the online mode of education to prevent its education system from collapsing due to the pandemic. The platforms such as Google Meet, Zoom, Microsoft teams, and other customized softwares are widely used for online classes in Bangladesh. These platforms have some useful features such as whiteboard, class recording, file/slide sharing, annotation, attendance count, chatbox, etc. Based on the online education system of Bangladesh, an online survey [7] was conducted where the participants were Bangladeshi students studying in higher secondary level or above. Education is a vital part of civilization. Hence, analyzing this major change in the education system and its effect upon the students is of great importance. The focus of this study is to understand *Bangladeshi students' perception and preference* towards online learning through the analysis of their survey responses. For performing such analysis, we will use appropriate statistical and machine learning techniques. Online classes can be a perfect substitute for the in-person classes only if they are designed in a structured way considering students' opinion and addressing their concerns. For effective and productive learning, it is important to consider the preferences of the students while designing the online courses. Investigating Bangladeshi online education and its impact on the students will allow us to have a better understanding of how *impactful* online learning is and determine which *aspects* need to be *improved or rectified* for forming *a better* virtual learning environment.

2. Objectives

The primary goals/objectives of this project are narrated below:

- Designing a classifier to predict if students are content with online learning or not, i.e. to analyze if students are preferring online classes more or less
- Study the impact of online education upon students grades, class performance and knowledge enhancement
- Analyze which category of students (male/female, in higher-secondary level/above, etc.) prefer online learning more
- Study the influence of various factors such as educational institutions type (public or private), location (urban/rural), etc. upon online learning
- Study the access to the required support and devices for online education based on gender

3. Motivation

Online education has become a vital part of a student's life since the emergence of the COVID-19 pandemic. Although the trend of online education has been there for a long time, but, it received a greater attention after the emergence of the global pandemic. Education like any other daily activity can't be off for a very long time. This led to the rise of online education as a replacement of in-person classes in the schools, colleges and universities. Even now when the pandemic seems to be almost over, many educational institutions still provide online learning besides in-person classes. Hence, it is very important to ensure the *quality* of online education and analyze how *useful* or *impactful* it is for the students. It needs to be studied how to *constantly improve* the quality of the online classes and identify what are the problems that are faced by the students during the online classes. These identified problems will help in *formulating* a *better online education system* that is more acceptable among the mass students.

4. Dataset

Data availability: The dataset chosen for conducting this study is publicly available at: https://www.kaggle.com/jehanbhathena/online-survey-data-of-bangladeshi-students.
Also available at : https://doi.org/10.7910/DVN/PLN7GM

Data size: It has a total of **17 columns and 8784 rows**. Dataset details:

Table 1

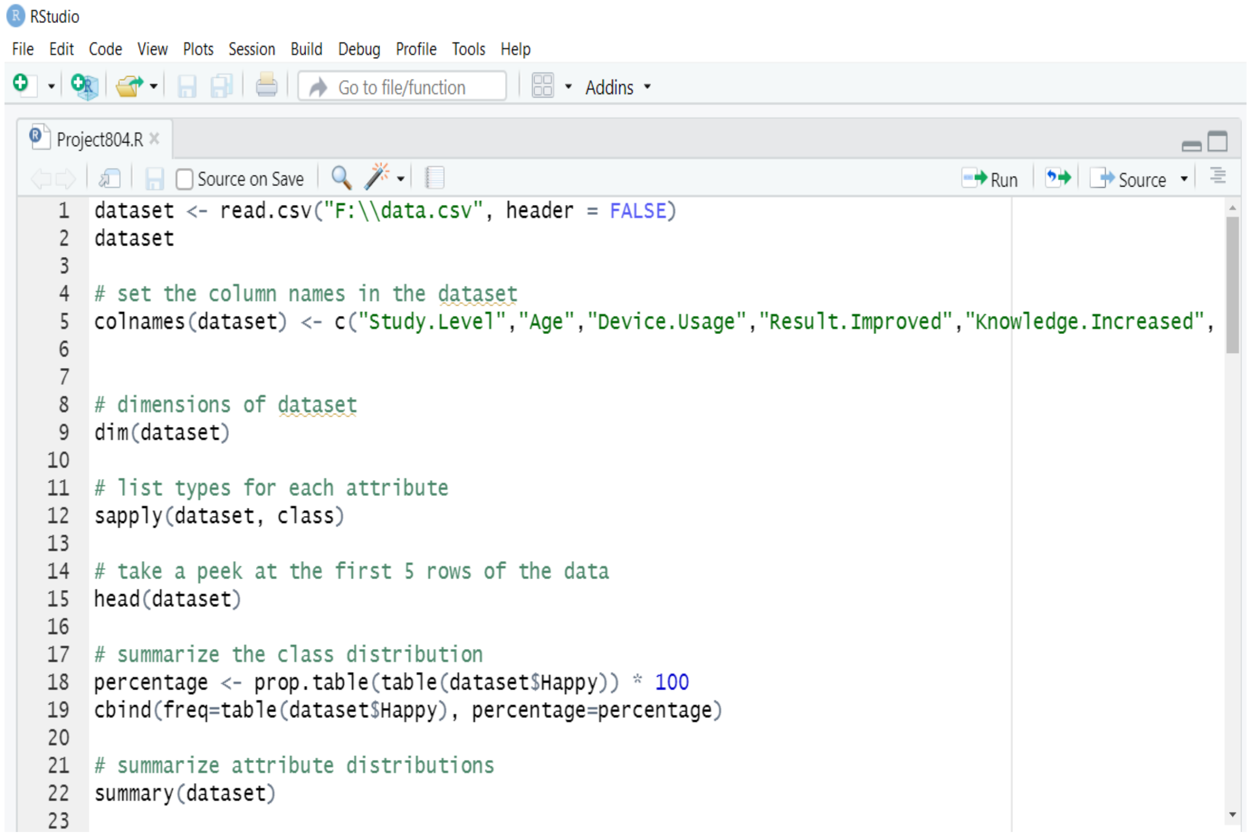
Variable Name	Possible values
1. Level of study	a. Upto Higher Secondary level (HSC) b. Honors or greater level
2. Age	Students of 16 to 26 years of age
3. Used smartphone/computer/laptop previously before online class?	Yes/No
4. Result increased after online education (comparatively)?	Yes/No
5. Knowledge increased after online education (comparatively)?	Yes/No
6. Education Institute Area	Urban/Rural
7. Have Internet availability?	Yes/No
8. Use Broadband / Mobile internet?	Broadband/Mobile Internet
9. Total hours of study before online education?	Any positive integer
10. Total hours of study after online education?	Any positive integer
11. Class performance increased in online education?	Yes/No
12. Institute Type	Public/Private
13. Current location (During Study) ?	Rural/Urban
14. Gender	Male/Female

15. Faced any issue with online class?	Yes/No
16. Preferred device for an online course	Mobile/Computer
17. Happy with online education?	Yes/No

Predictors/Input features: Age, gender, highest education level, access to devices, access to internet, internet type, school type, school location, issues faced, etc. are some of the notable input variables in the dataset. The variables 1 -16 listed in Table 1 are the input variables/features used for model training.

Outcome/Predicted variable: Task of the classifier is to predict if the students are *happy or content with the current online education system or not*, i.e. to predict the outcome of the variable **"Happy with online education?"** which is the 17th variable in the above Table-1.

Data-Preprocessing: The categorical variables are pre-processed before training any model. Label encoding is used to transform the categorical values into numerical ones. To handle the missing data, out of the 8784 samples, 5715 samples were used during model training. Rest of the 3069 samples had missing values, hence were not considered for experimentation. The **R code** for loading data and summarizing the attributes and class distribution is as follows:



```r
dataset <- read.csv("F:\\data.csv", header = FALSE)
dataset

# set the column names in the dataset
colnames(dataset) <- c("Study.Level","Age","Device.Usage","Result.Improved","Knowledge.Increased",


# dimensions of dataset
dim(dataset)

# list types for each attribute
sapply(dataset, class)

# take a peek at the first 5 rows of the data
head(dataset)

# summarize the class distribution
percentage <- prop.table(table(dataset$Happy)) * 100
cbind(freq=table(dataset$Happy), percentage=percentage)

# summarize attribute distributions
summary(dataset)

```

5. Main Research Questions

This study aims to answer the following vital research questions:

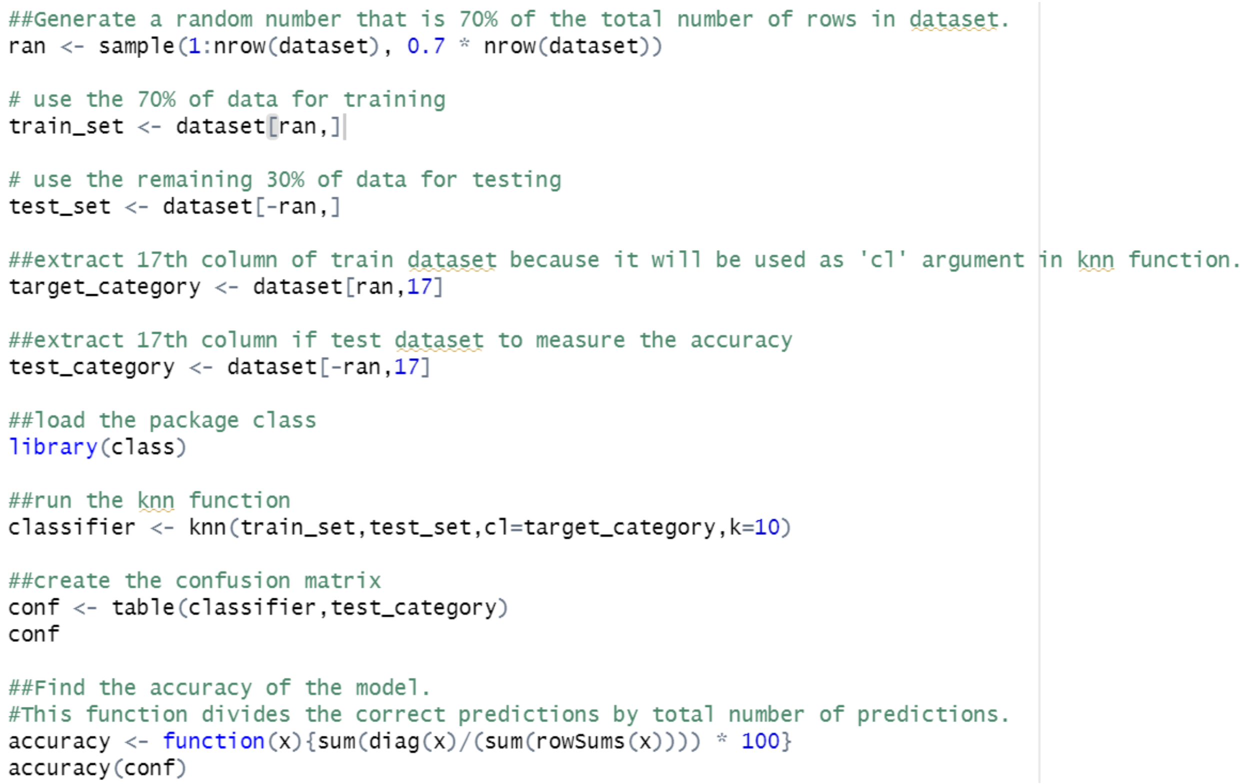
1. Are students **satisfied** with the **online education** structure of Bangladesh?
2. Is online learning actually **beneficial** to the **students** (in terms of both result and knowledge enhancement)?
3. What type of **factors impact** online learning the most?
4. How to **formulate** a **better e-learning** system?
5. Do male and female students **have equal access** to the support required for online education?

Each of the above research questions are significant to understand the effectiveness of the online education system of Bangladesh. The first question aims to understand how happy or satisfied the students are with the current structure of the online classes. Since the purpose of online education is to benefit the students, hence it is very important to make sure that the students are happy with it. If not, then immediately effective measures should be taken to improve the system and make learning enjoyable. The second question is very important as it will reveal if the current online education system is truly beneficial or not. This will be measured with respect to the improvement in their knowledge and results. The third question is to analyze what factors impact online learning the most. These factors are the ones that should be taken into account when re-structuring the system to improve its effectiveness. Some of these factors are the availability of enough support for online education, minimum knowledge of technology and devices for online classes, level of comfort with this mode of learning, etc. The fourth factor is to realize the necessary actions that need to be taken for forming a better online learning system. Without constant improvement of the present system by taking into account the students' concerns, it is not possible to make it acceptable among the students and retain it for a long time. Lastly, it is needed to make sure that there's no gap between the students irrespective of their genders in terms of the support received to carry on their online classes smoothly.

6. Analyses Done

K-Nearest Neighbors (KNN) - Here the **goal** of this analysis is to develop a *KNN classifier* that can predict if the students are Happy or Unhappy with the current online education system.

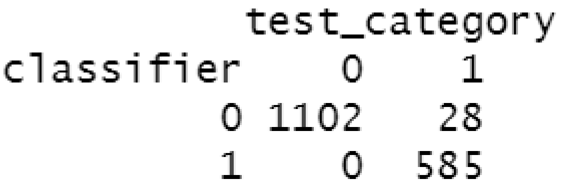
R code



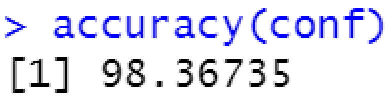
```
##Generate a random number that is 70% of the total number of rows in dataset.
ran <- sample(1:nrow(dataset), 0.7 * nrow(dataset))

# use the 70% of data for training
train_set <- dataset[ran,]

# use the remaining 30% of data for testing
test_set <- dataset[-ran,]

##extract 17th column of train dataset because it will be used as 'cl' argument in knn function.
target_category <- dataset[ran,17]

##extract 17th column if test dataset to measure the accuracy
test_category <- dataset[-ran,17]

##load the package class
library(class)

##run the knn function
classifier <- knn(train_set,test_set,cl=target_category,k=10)

##create the confusion matrix
conf <- table(classifier,test_category)
conf

##Find the accuracy of the model.
#This function divides the correct predictions by total number of predictions.
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(conf)
```

Results

The **accuracy** of the classifier is **98.36735%** and from the confusion matrix it is found that, out of 585 test instances of 'Happy', it predicts correctly all of those. But, out of the 1130 instances of not Happy, it predicts 1102 instances correctly and mis-classifies the rest 28 instances. Below are the attached screenshots of the results:



```
          test_category
classifier    0    1
         0 1102   28
         1    0  585
```



```
> accuracy(conf)
[1] 98.36735
```

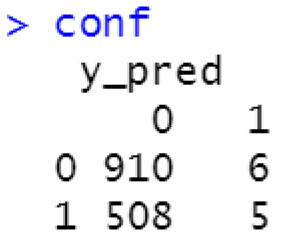
Support Vector Machines (SVM) - Here the **goal** of this analysis is to train a *SVM classifier* that can predict if the students are 'Happy' or 'Unhappy' with the current online education system.

R code

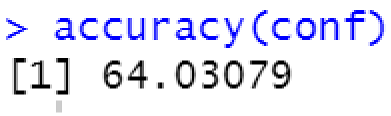
```
##Generate a random number that is 75% of the total number of rows in dataset
ran <- sample(1:nrow(dataset), 0.75 * nrow(dataset))

# use the 75% of data for training
train_set <- dataset[ran,]

# use the remaining 25% of data for testing
test_set <- dataset[-ran,]

# Fitting SVM to the Training set
install.packages('e1071')
library(e1071)

classifier = svm(formula = Happy ~ .,data = train_set, scale = TRUE,
                 type = 'C-classification',kernel = 'radial')

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set)

# Making the Confusion Matrix
conf = table(test_set[, 17], y_pred)
conf

##Find the accuracy of the model.
#This function divides the correct predictions by total number of predictions
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(conf)
```

Results

The **accuracy** of the classifier is **64.03079%** and from the confusion matrix it is found that, out of 513 test instances of 'Happy', it predicts *only* 5 *instances correctly*. While out of the 916 instances of not Happy, it predicts 910 instances correctly and mis-classifies the rest 6 instances. Here, it seems that *the classifier predicts the more frequent class well and* suffers due to class imbalance. Below are the attached screenshots of the results:



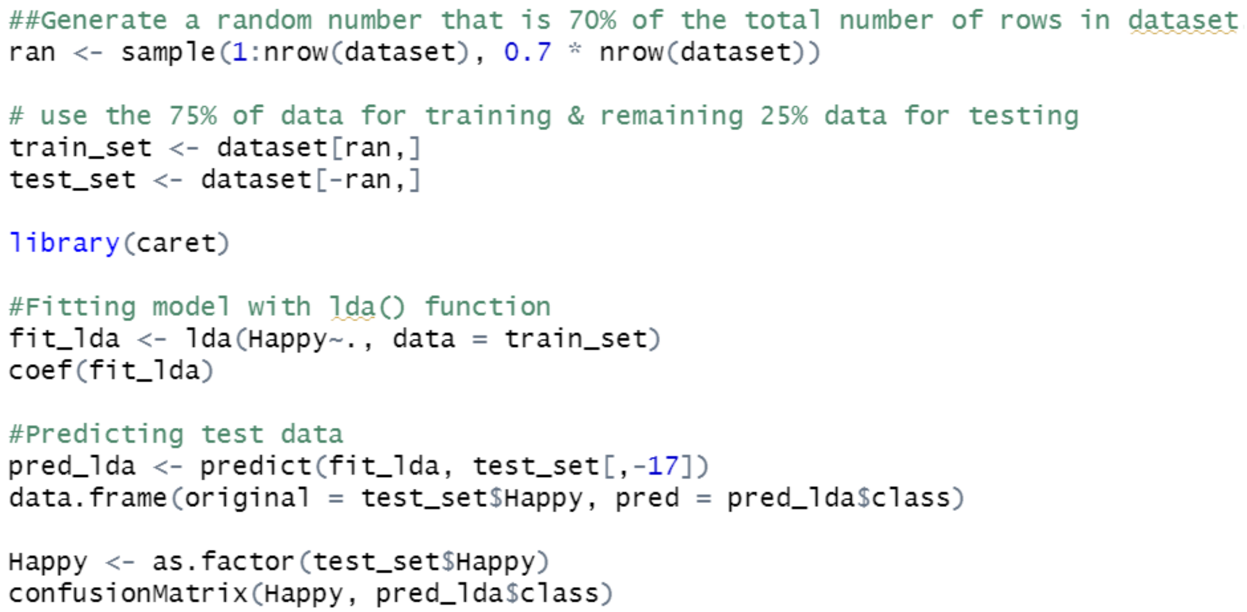
```
> conf
   y_pred
      0   1
  0 910   6
  1 508   5
```



```
> accuracy(conf)
[1] 64.03079
```

Linear Discriminant Analysis (LDA) - Here the **goal** of this analysis is to train a *LDA classifier* that can predict if the students are 'Happy' or unhappy with the online education system.

R code

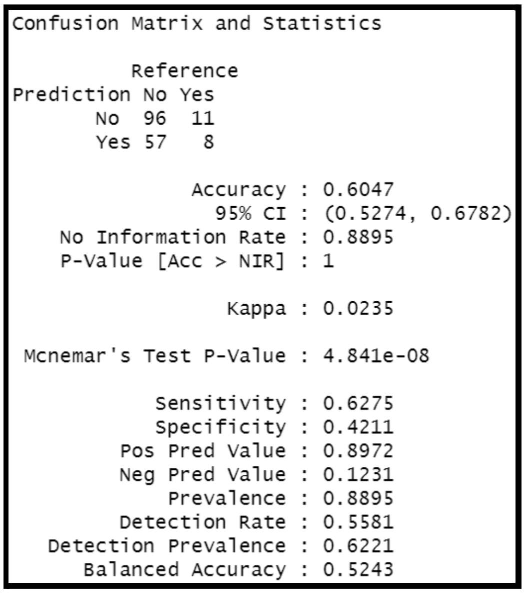


```
##Generate a random number that is 70% of the total number of rows in dataset
ran <- sample(1:nrow(dataset), 0.7 * nrow(dataset))

# use the 75% of data for training & remaining 25% data for testing
train_set <- dataset[ran,]
test_set <- dataset[-ran,]

library(caret)

#Fitting model with lda() function
fit_lda <- lda(Happy~., data = train_set)
coef(fit_lda)

#Predicting test data
pred_lda <- predict(fit_lda, test_set[,-17])
data.frame(original = test_set$Happy, pred = pred_lda$class)

Happy <- as.factor(test_set$Happy)
confusionMatrix(Happy, pred_lda$class)
```

Results

In the LDA analysis, it is observed that *using the entire dataset* results in a *very poor model performance* (predicts all 'Yes' as 'No'). Hence, a *randomly sampled* 10% of the dataset is being used for model training due to which the performance is raised a bit. The results are summarized below:

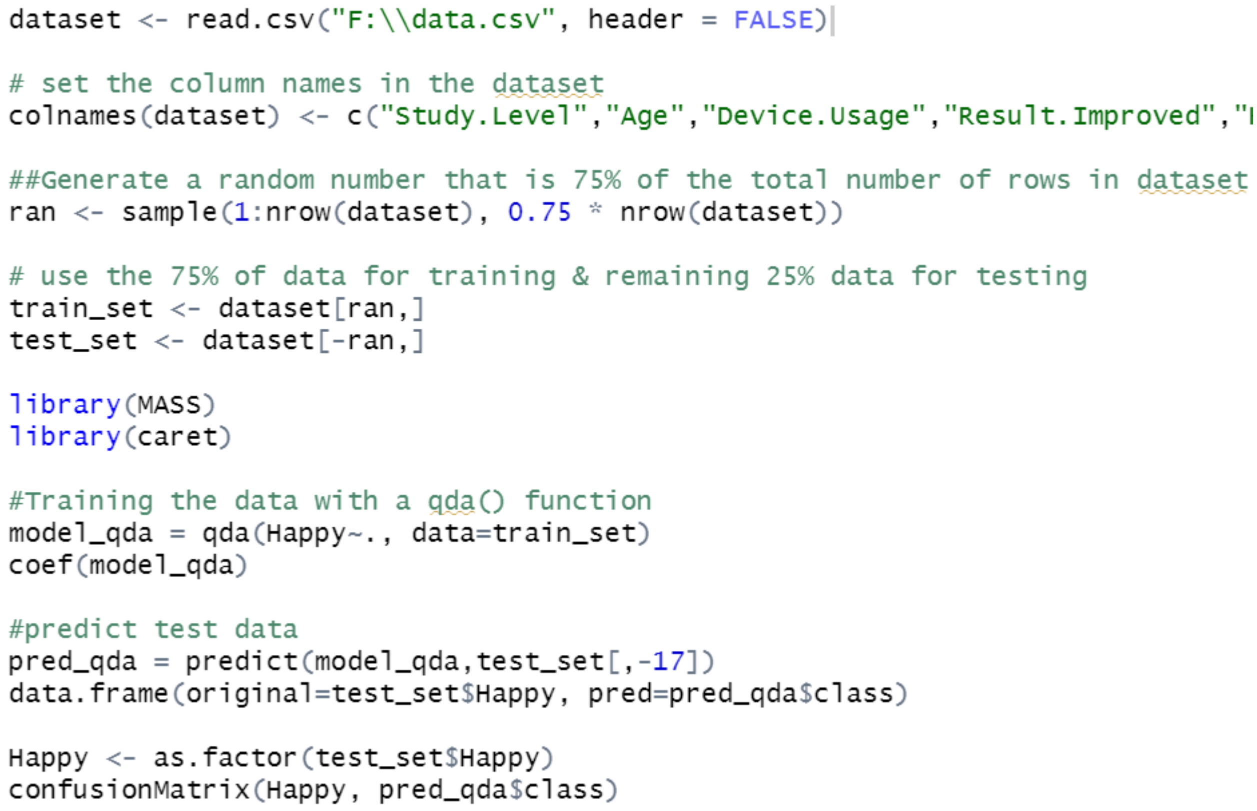
- The **accuracy** of the classifier is **60.47%**
- From the confusion matrix it can be seen that, out of the 65 test instances of 'Happy', it predicts *only* 8 *instances correctly*.
- While out of the 107 instances of not Happy, it predicts 96 instances correctly and mis-classifies the rest 11 instances.
- **Remarks**: Here also, it seems that *the classifier predicts the more frequent class well*. Class *imbalance* could be a reason for the low accuracy.



```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  96  11
       Yes 57   8

               Accuracy : 0.6047
                 95% CI : (0.5274, 0.6782)
    No Information Rate : 0.8895
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0235

 Mcnemar's Test P-Value : 4.841e-08

            Sensitivity : 0.6275
            Specificity : 0.4211
         Pos Pred Value : 0.8972
         Neg Pred Value : 0.1231
             Prevalence : 0.8895
         Detection Rate : 0.5581
   Detection Prevalence : 0.6221
      Balanced Accuracy : 0.5243
```

Quadratic Discriminant Analysis (QDA) - Here the **goal** of this analysis is to develop a *QDA classifier* that can predict if the students are 'Happy' or unhappy with the online education system.

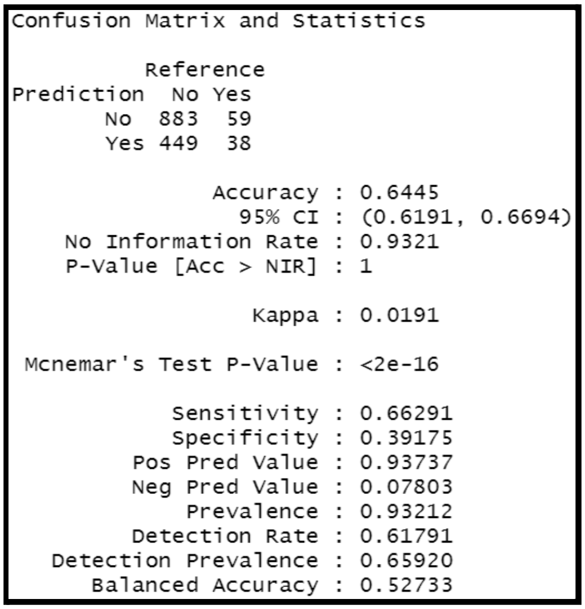
R code



```
dataset <- read.csv("F:\\data.csv", header = FALSE)

# set the column names in the dataset
colnames(dataset) <- c("Study.Level","Age","Device.Usage","Result.Improved","I

##Generate a random number that is 75% of the total number of rows in dataset
ran <- sample(1:nrow(dataset), 0.75 * nrow(dataset))

# use the 75% of data for training & remaining 25% data for testing
train_set <- dataset[ran,]
test_set <- dataset[-ran,]

library(MASS)
library(caret)

#Training the data with a qda() function
model_qda = qda(Happy~., data=train_set)
coef(model_qda)

#predict test data
pred_qda = predict(model_qda,test_set[,-17])
data.frame(original=test_set$Happy, pred=pred_qda$class)

Happy <- as.factor(test_set$Happy)
confusionMatrix(Happy, pred_qda$class)
```

Results

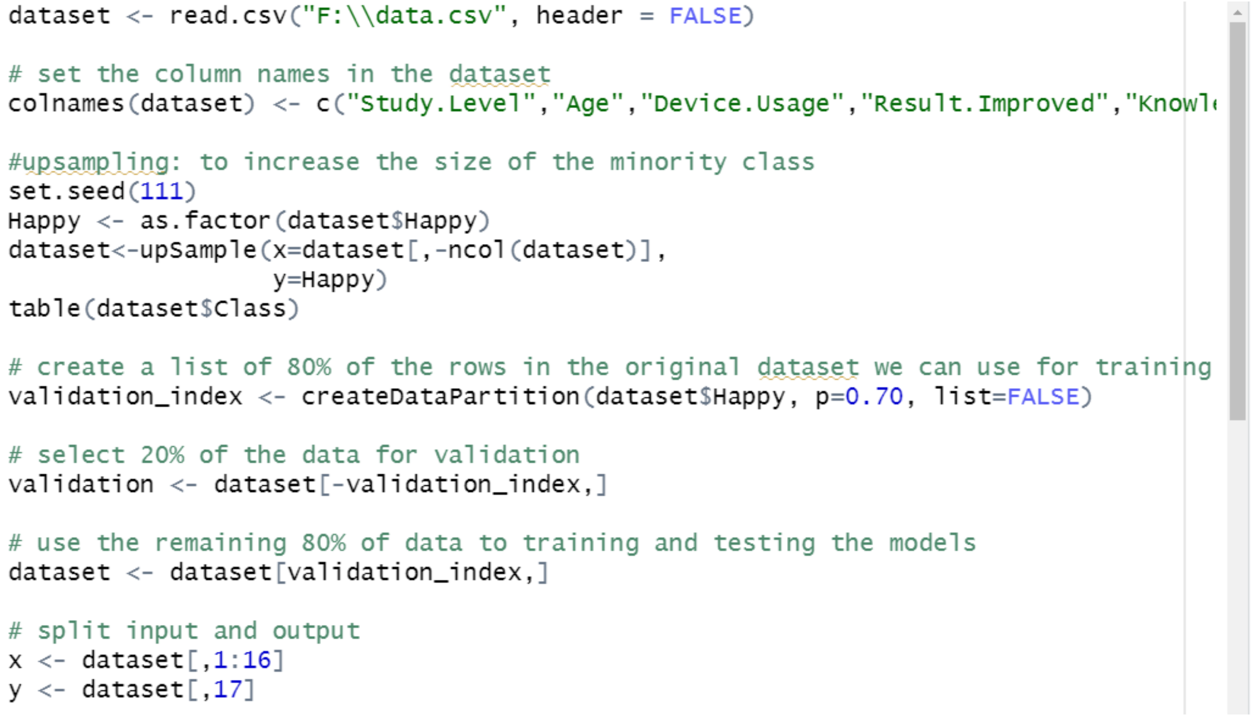
- The **accuracy** of the classifier is **64.45%**
- From the confusion matrix it can be seen that, out of 487 test instances of 'Happy', it predicts *only 38 instances correctly*.
- While out of the 942 instances of not Happy, it predicts 883 instances correctly and mis-classifies the rest 59 instances.
- Here also, it seems that *the classifier predicts the more frequent class well.*



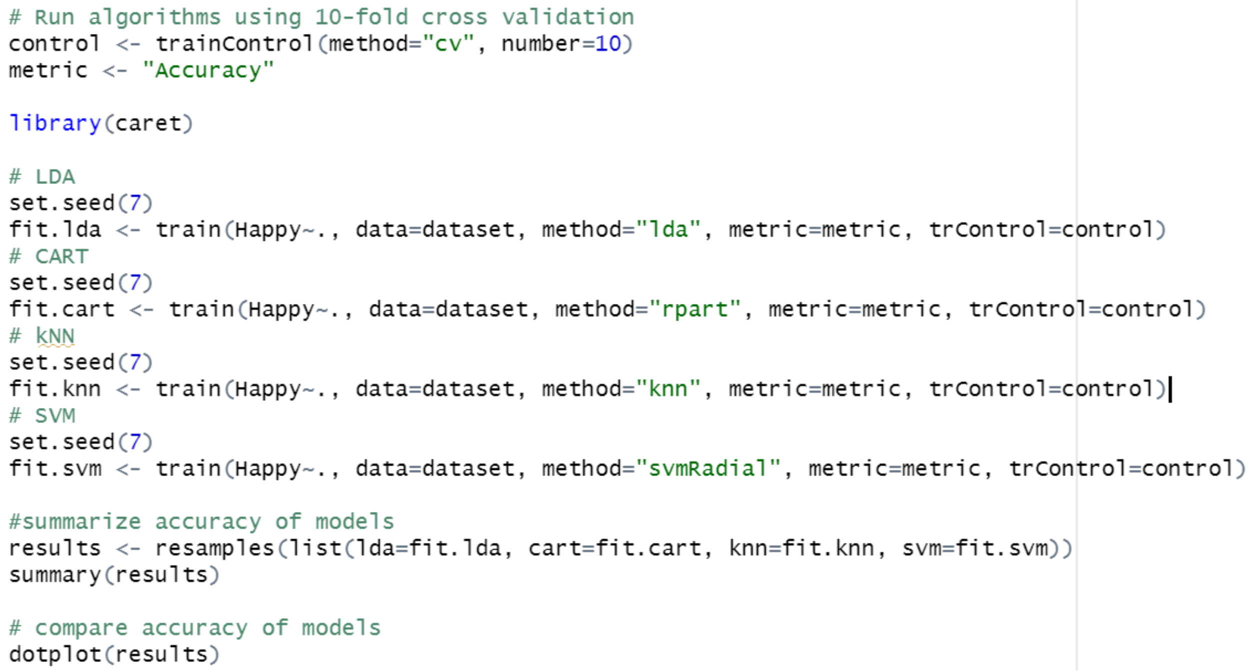
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  883  59
       Yes 449  38

               Accuracy : 0.6445
                 95% CI : (0.6191, 0.6694)
    No Information Rate : 0.9321
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0191

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.66291
            Specificity : 0.39175
         Pos Pred Value : 0.93737
         Neg Pred Value : 0.07803
             Prevalence : 0.93212
         Detection Rate : 0.61791
   Detection Prevalence : 0.65920
      Balanced Accuracy : 0.52733
```

<u>k-fold Cross Validation (K-fold CV)</u> - Here the **goal** of this analysis is to train LDA, KNN, SVM and CART (Classification and Regression Trees) classifiers using **10-fold CV** that can predict if the students are 'Happy' or unhappy with the online education system.

<u>R code</u>

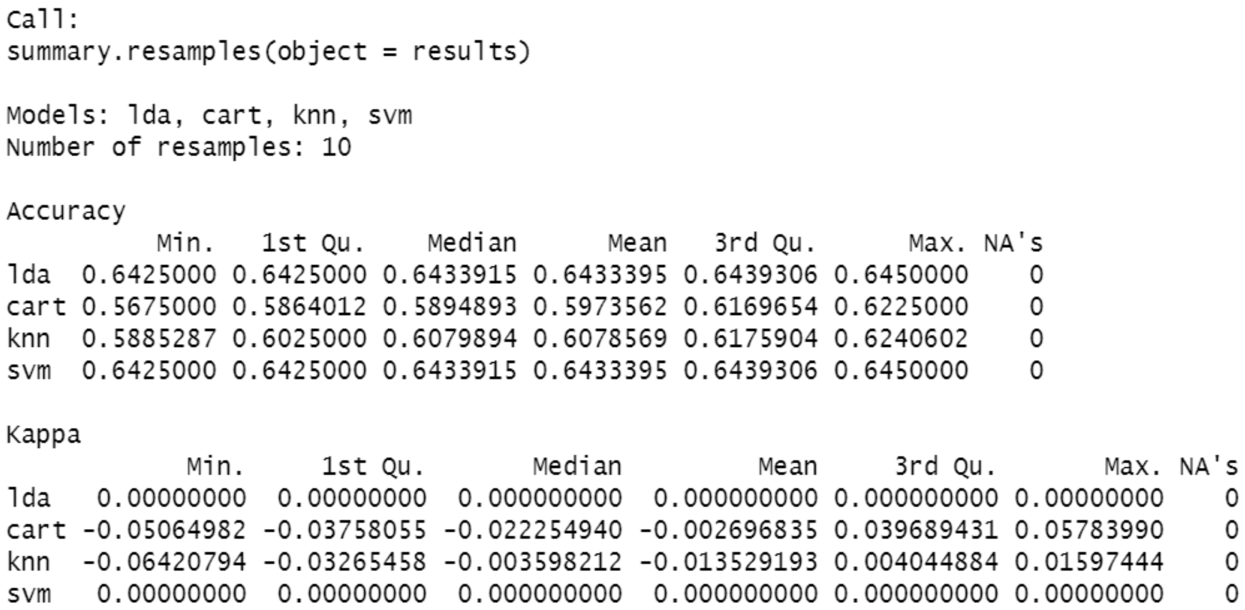


```
dataset <- read.csv("F:\\data.csv", header = FALSE)

# set the column names in the dataset
colnames(dataset) <- c("Study.Level","Age","Device.Usage","Result.Improved","Knowl

#upsampling: to increase the size of the minority class
set.seed(111)
Happy <- as.factor(dataset$Happy)
dataset<-upSample(x=dataset[,-ncol(dataset)],
                  y=Happy)
table(dataset$Class)

# create a list of 80% of the rows in the original dataset we can use for training
validation_index <- createDataPartition(dataset$Happy, p=0.70, list=FALSE)

# select 20% of the data for validation
validation <- dataset[-validation_index,]

# use the remaining 80% of data to training and testing the models
dataset <- dataset[validation_index,]

# split input and output
x <- dataset[,1:16]
y <- dataset[,17]
```



```
# Run algorithms using 10-fold cross validation
control <- trainControl(method="cv", number=10)
metric <- "Accuracy"

library(caret)

# LDA
set.seed(7)
fit.lda <- train(Happy~., data=dataset, method="lda", metric=metric, trControl=control)
# CART
set.seed(7)
fit.cart <- train(Happy~., data=dataset, method="rpart", metric=metric, trControl=control)
# kNN
set.seed(7)
fit.knn <- train(Happy~., data=dataset, method="knn", metric=metric, trControl=control)
# SVM
set.seed(7)
fit.svm <- train(Happy~., data=dataset, method="svmRadial", metric=metric, trControl=control)

#summarize accuracy of models
results <- resamples(list(lda=fit.lda, cart=fit.cart, knn=fit.knn, svm=fit.svm))
summary(results)

# compare accuracy of models
dotplot(results)
```

Results



```
Call:
summary.resamples(object = results)

Models: lda, cart, knn, svm
Number of resamples: 10

Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
lda  0.6425000 0.6425000 0.6433915 0.6433395 0.6439306 0.6450000    0
cart 0.5675000 0.5864012 0.5894893 0.5973562 0.6169654 0.6225000    0
knn  0.5885287 0.6025000 0.6079894 0.6078569 0.6175904 0.6240602    0
svm  0.6425000 0.6425000 0.6433915 0.6433395 0.6439306 0.6450000    0

Kappa
            Min.     1st Qu.       Median         Mean     3rd Qu.       Max. NA's
lda   0.00000000  0.00000000  0.000000000  0.000000000 0.000000000 0.00000000    0
cart -0.05064982 -0.03758055 -0.022254940 -0.002696835 0.039689431 0.05783990    0
knn  -0.06420794 -0.03265458 -0.003598212 -0.013529193 0.004044884 0.01597444    0
svm   0.00000000  0.00000000  0.000000000  0.000000000 0.000000000 0.00000000    0
```

Here, *LDA and SVM* seem to be the *best performing models* with a mean accuracy of **64.33%**. KNN and CART have almost similar results with a slight difference in accuracy. From the Kappa values, it can be seen that SVM and LDA have a mean value of 0 indicating that the classification is as good as random values. The CART and KNN models have negative Kappa values which is actually an indicator of their poor performance.

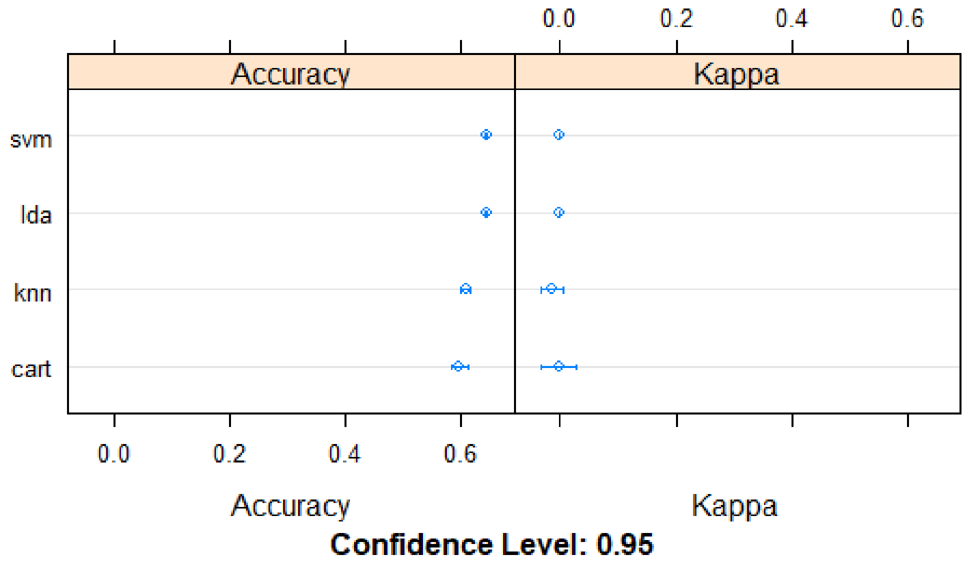
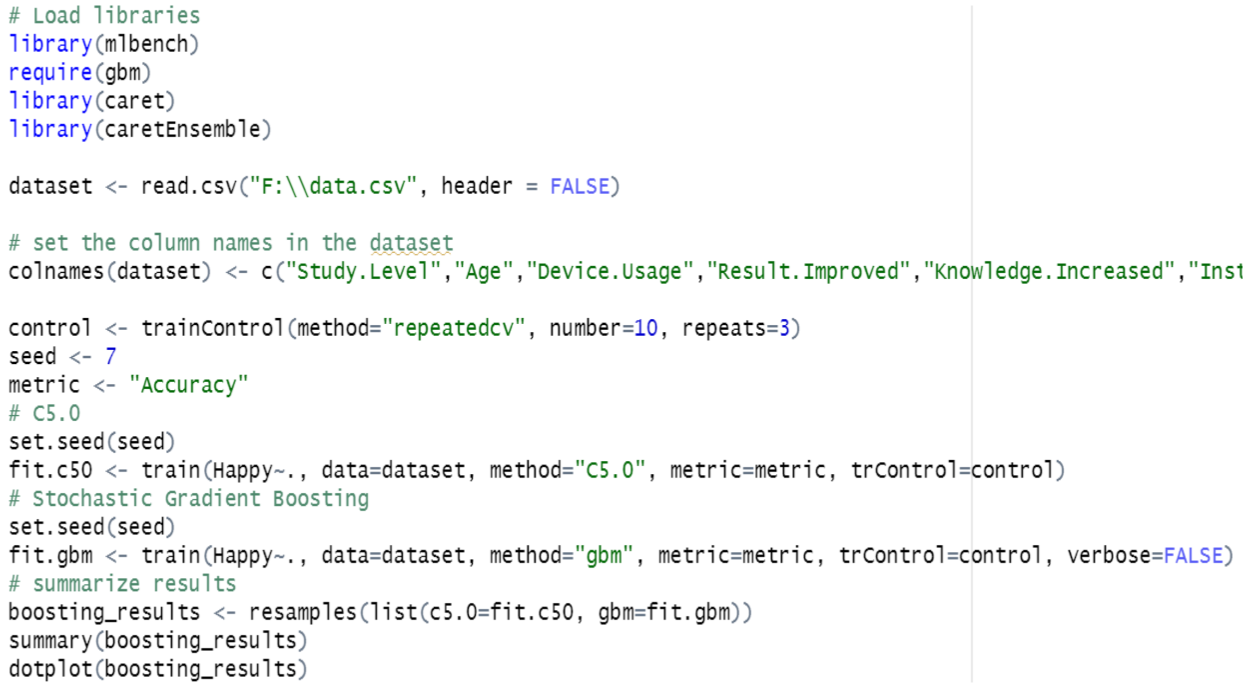


Figure: Evaluation Metrics : Accuracy and Kappa using 10-fold CV

Boosting - Here the **goal** of this analysis is to train the *boosting classifiers (C5.0 and gbm)* that can predict if the students are 'Happy' or unhappy with the online education system.

R code



```
# Load libraries
library(mlbench)
require(gbm)
library(caret)
library(caretEnsemble)

dataset <- read.csv("F:\\data.csv", header = FALSE)

# set the column names in the dataset
colnames(dataset) <- c("Study.Level","Age","Device.Usage","Result.Improved","Knowledge.Increased","Inst

control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7
metric <- "Accuracy"
# C5.0
set.seed(seed)
fit.c50 <- train(Happy~., data=dataset, method="C5.0", metric=metric, trControl=control)
# Stochastic Gradient Boosting
set.seed(seed)
fit.gbm <- train(Happy~., data=dataset, method="gbm", metric=metric, trControl=control, verbose=FALSE)
# summarize results
boosting_results <- resamples(list(c5.0=fit.c50, gbm=fit.gbm))
summary(boosting_results)
dotplot(boosting_results)
```

Results

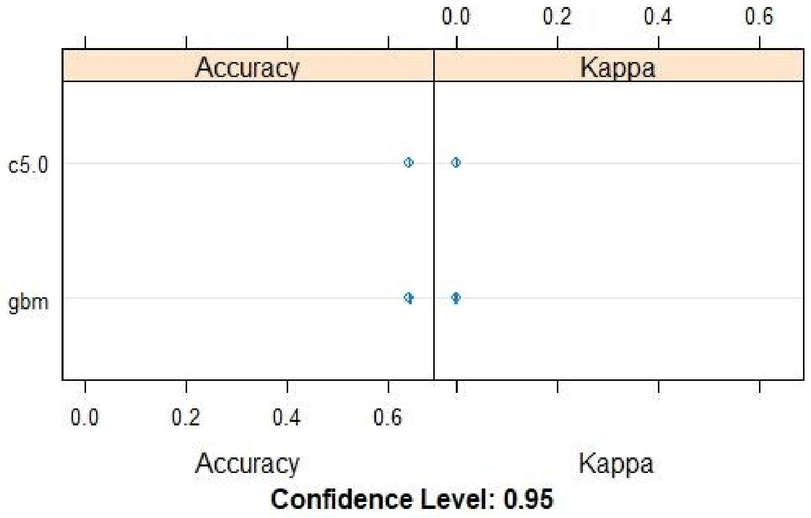
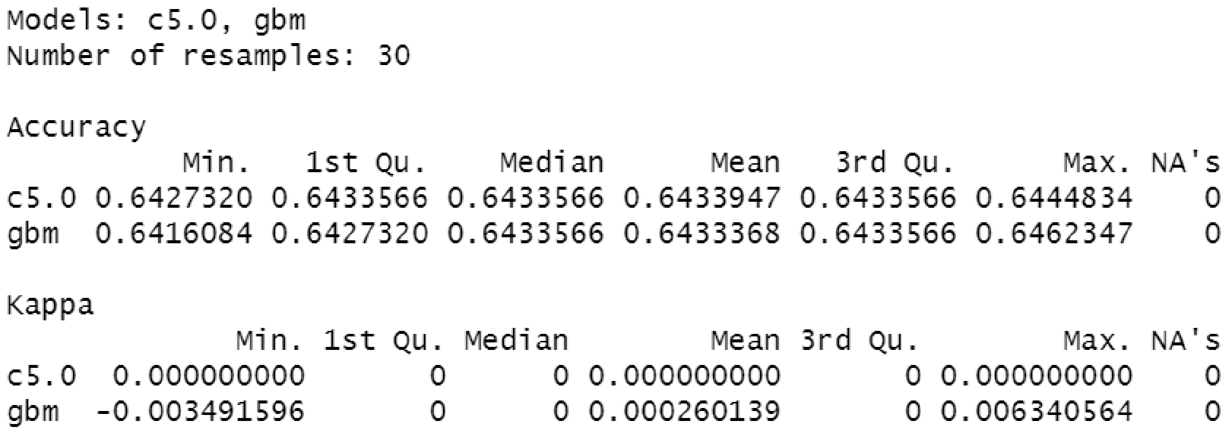


Figure: Accuracy and Kappa scores of the boosting models C5.0 and gbm

Both of the boosting models have almost the same mean accuracy and kappa scores. Detailed results given below:



```
Models: c5.0, gbm
Number of resamples: 30

Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
c5.0 0.6427320 0.6433566 0.6433566 0.6433947 0.6433566 0.6444834    0
gbm  0.6416084 0.6427320 0.6433566 0.6433368 0.6433566 0.6462347    0

Kappa
             Min. 1st Qu. Median        Mean 3rd Qu.        Max. NA's
c5.0  0.000000000       0      0 0.000000000       0 0.000000000    0
gbm  -0.003491596       0      0 0.000260139       0 0.006340564    0
```

The **analyses done** are **summarized** below for a better comparison:

Analysis/Models	**Accuracy**
K-Nearest Neighbors (KNN)	98.37%
Support Vector Machines (SVM)	64.03%
Linear Discriminant Analysis (LDA)	60.47%
Quadratic Discriminant Analysis (QDA)	64.45%
K-fold Cross Validation (10-fold CV) a. LDA b. CART c. KNN d. SVM	LDA - 64.33% CART - 59.74% KNN - 60.79% SVM- 64.33%
Boosting a. C5.0 b. gbm	C5.0 - 64.34% gbm - 64.33%

Here, KNN seems to be the best performing model with the highest accuracy. Except for KNN, all other algorithms suffer in performance with accuracy ranging mostly from around 60-65%. SVM, QDA and Boosting seem to have quite similar performance. A probable reason for the poor performance of most of the models could be due to the exhibition of a **higher class imbalance** in the dataset. In general, class imbalance can bias the loss function towards the majority class. Class imbalance can also affect the prediction performance through *oversmoothing,* as the converged representation will be more representative of the majority class.

The reasons for **not conducting** certain type of analyses are given below:

Analysis	**Reason**
Linear Regression	This is a classification problem. Hence applying linear regression that deals with continuous variables is unsuitable for this study.
Local Regression	This is a classification problem. Hence applying regression algorithms in this case is unsuitable.
Logistic Regression	Not quite suitable for this problem. Since it is often prone to noise and overfitting. Also, it is tough to obtain complex relationships using logistic regression.
Polynomial Logistic Regression	Since these models are significantly affected by outliers, it is not used for this study.
Ridge Regression and the Lasso	This is a classification problem. Hence applying regression algorithms in this case is unsuitable.
Regression Trees	This is a classification problem. Hence applying purely regression trees in this case is unsuitable.
Splines	Splines are not suitable for this project. This is a classification problem. Hence applying regression algorithms in this case is unsuitable.
Leave-One-Out Cross-Validation (LOOCV)	A k-fold CV is used in the project. k-fold CV is more appropriate and efficient in this case than LOOCV. Hence, I did not use LOOCV.
Best Subset Selection	It isn’t necessary to use this technique for this study.
GAMs	GAMs are usually not suitable for classification. They are mostly suitable for estimating non-linear functions. Hence not used here.

Exploratory Data Analysis / Generated Plots

1. Figure (a) *Pie chart* of the students who are either **'Happy' or 'Unhappy'** with the online education and **Figure (b)** *Pie chart* of the students whose **knowledge** either '*increased*' or '*not increased*' after the online education

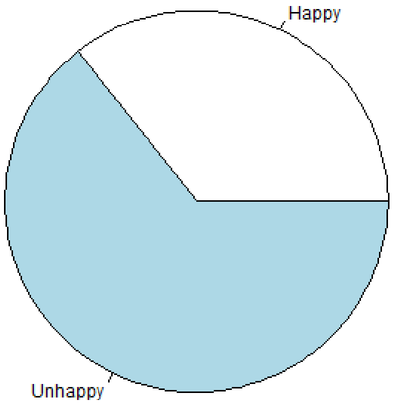


Fig: (a)

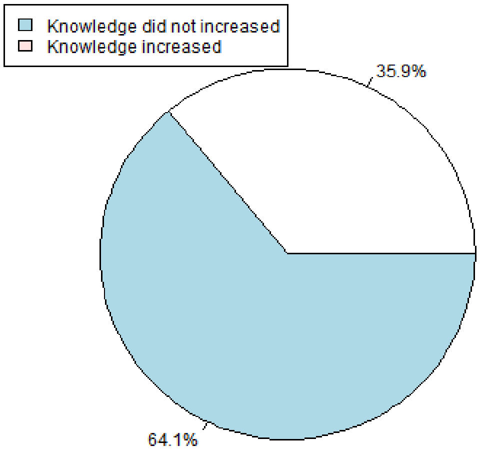


Fig: (b)

Out of the **5715** records, **2038** students are *happy with online education*. Rest of the **3677** students are *unhappy* with the online education. It seems that the majority of the students are in support of in-person learning and are not satisfied with the current form of online learning. From the perspective of the *knowledge enhancement* after online education, it is found that only **2051** students believe that online learning is helpful for them with respect to the increment of knowledge. Majority of the students (about 64%) don't find it useful w.r.t enhancement of knowledge.

2. Figure ***(a)*** *Pie chart* of the students whose **results** either '*increased*' or '*not increased*' after the online education and **Figure** ***(b)*** *Pie Chart* representing the % of *male and female* students who have *access to the internet*.

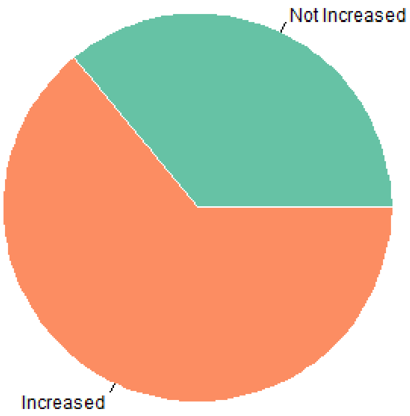


Figure: (a)

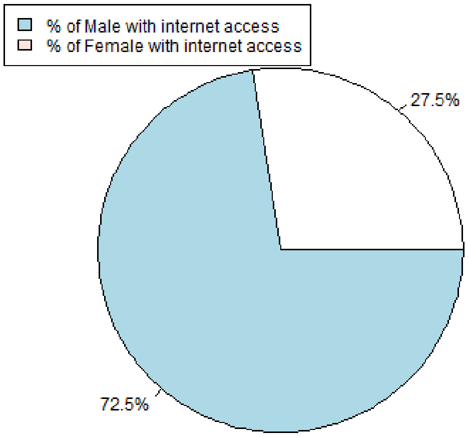


Figure: (b)

Majority of the students (about 3981) reported the improvement of their results during online education. That is about 70% of the students' results improved due to online education. Although, it contradicts with their opinion of being happy/satisfied with online education and also, the enhancement of knowledge factor. This proves that simply the improvement of results is not an indicator of a good education system. Improvement in results **does not** mean the improvement in knowledge or satisfaction with the learning method. Thus, results are *not a good indicator* of an efficient online learning system. From Figure (b), it can be seen that a gap exists between the male and female students in terms of their access to the internet facilities. This shows a picture of the discrimination towards female students in Bangladesh. This is infact a common scenario in the developing nations. Thus, while improving the quality of online education, it is also equally important to make sure that everyone has equal opportunities regardless of gender, age, race, ethnicities, etc.

3. *Stacked Barchart* of the *type of interne*t in rural and urban areas

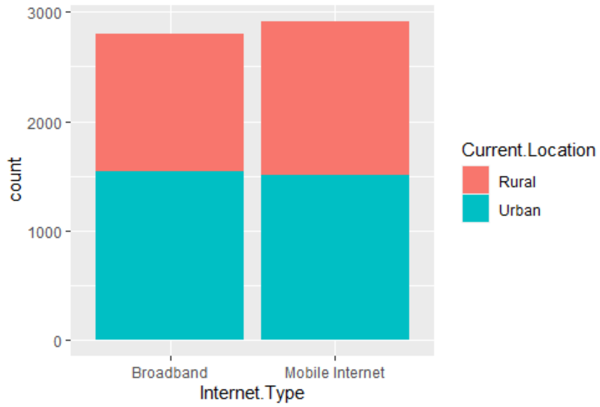
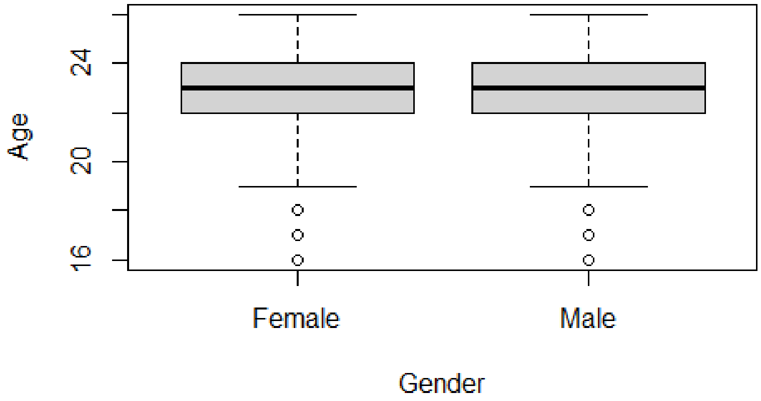


Figure (a)

The above Figure (a) depicts that compared to broadband connection, mobile internet is more frequently used in rural areas. While in urban areas, the usage of broadband and mobile internet is in almost similar ratio.

4. *Box plot* showing the age range of male and female students



5. *Stacked Barchart* of the *hours of study* spent by male and female students (a) before and (b) after shifting to the online mode of education

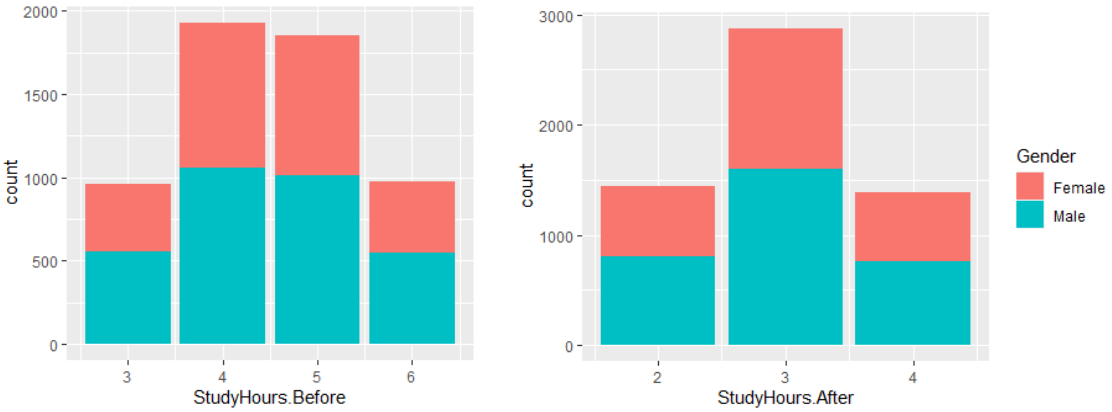


Figure: (a) **Figure: (b)**

The above charts highlight an important aspect of every student's life. There's a significant difference in the hours spent studying before and after the emergence of online classes. Before online classes students spent significantly more time studying. The minimum hours of study was 3 and maximum 6 hours. However, after shifting to the online mode of study, students spent a minimum of 2 and a maximum of 4 hours in their studies. This reveals a carelessness towards education due to the online classes.

7. Conclusion

This study presents the scenario, limitations and scopes for the improvement of the online education system in Bangladesh. It also reveals how effective online education is from the perspective of the students and studies students' satisfaction regarding online education. Analysis shows that the majority of the students (about 64%) are not satisfied or happy with online learning. This indicates their preference of in-person classes over the online ones. A surprising fact is that although the majority of the students reported that due to online classes there is an increase in their results. However, in terms of knowledge enhancement, only some of the students believe that online education has enhanced their knowledge. Majority believe the opposite regarding improvement in knowledge. This contradicts the improvement in results. It indicates that only the improvement of results is *not a good indicator* of an efficient online education system. Also, it means that an increase in results does not always mean an increase in knowledge too. Another factor that needs to be considered is to ensure that both male and female students have equal opportunities to pursue their online education smoothly. It is found that compared to the male students, a small percentage of the female students have access to the internet for attending online classes. This is creating a barrier for the female students for the smooth continuation of their studies.

The following measures should be taken to improve the current situation of online education in Bangladesh. It seems from the analysis that although results are improving, there's a lack of improvement in knowledge. Focus needs to be given on how to enhance the knowledge of the students via online classes. Also, all students should receive equal opportunities and support irrespective of their gender, age and residence area. This will remove the existing disparity and thus, make online education accessible for all.

In terms of the analyses done in this project, several models were applied to develop a classifier that perfectly predicts the students who are happy or unhappy with online classes. It is seen that KNN performs the best among all the algorithms/models. However, in future, a more robust classifier can be developed by feature engineering and applying different techniques to find out the most important features. Also, class imbalance needs to be handled in future to improve model performance.

References

1. Li, Lalani. The COVID-19 pandemic has changed education forever. This is how. *World Economic Forum*, https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/. Accessed 9 February 2022.
2. Muthuprasad, T., Aiswarya, S., Aditya, K.S. and Jha, G.K., 2021. Students' perception and preference for online education in India during COVID-19 pandemic. *Social Sciences & Humanities Open*, *3*(1), p.100101.
3. S. Bignoux, K.J. Sund. Tutoring executives online: What drives perceived quality? Behaviour & Information Technology, 37 (7) (2018), pp. 703-713
4. Mukhtar, K., Javed, K., Arooj, M. and Sethi, A., 2020. Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era. *Pakistan journal of medical sciences*, *36*(COVID19-S4), p.S27.
5. Maatuk, A.M., Elberkawi, E.K., Aljawarneh, S., Rashaideh, H. and Alharbi, H., 2021. The COVID-19 pandemic and E-learning: challenges and opportunities from the perspective of students and instructors. *Journal of Computing in Higher Education*, pp.1-18.
6. Bączek, M., Zagańczyk-Bączek, M., Szpringer, M., Jaroszyński, A. and Wożakowska-Kapłon, B., 2021. Students' perception of online learning during the COVID-19 pandemic: a survey study of Polish medical students. *Medicine*, *100*(7).
7. Ferdous, Jannatul, 2021. Online Survey Data of Bangladeshi Students. *Harvard Dataverse,* https://doi.org/10.7910/DVN/PLN7GM. Accessed 1 February 2022.