MVA_Class_Survey_aa2569

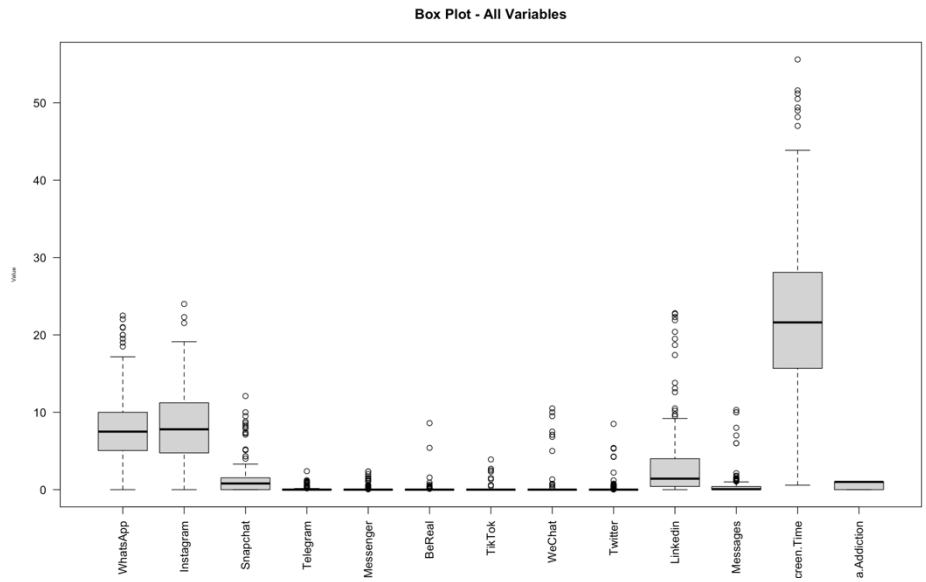
GitHub Link: https://github.com/ajeyvishnu/SocialMedia_Addiction

Defining the Data:

Data Dictionary

Field Name	Data Type	Units	Description	Possible Values for Categorical
Week	Alphanumeric	-	Week start and end date	-
Whatsapp	Numeric	Hours	Time spent on Whatsapp per week	-
Instagram	Numeric	Hours	Time spent on Instagram per week	-
Snapchat	Numeric	Hours	Time spent on Snapchat per week	-
Telegram	Numeric	Hours	Time spent on Telegram per week	-
Facebook/Messenger	Numeric	Hours	Time spent on Facebook/Messenger per week	-
BeReal	Numeric	Hours	Time spent on BeReal per week	-
TikTok	Numeric	Hours	Time spent on Tiktok per week	-
Wechat	Numeric	Hours	Time spent on WeChat per week	-
Twitter	Numeric	Hours	Time spent on Twitter per week	-
Linkedin	Numeric	Hours	Time spent on LinkedIn per week	-
Messages	Numeric	Hours	Time spent on Messages per week	-
Total Social Media Screen Time	Numeric	Hours	Total time spent on social media per week	
Number of times opened (hourly intervals)	Numeric	Nos	Considering the 24-hour slots in a day, how many hour slots did the user open social media apps. This is for one day. Consider the above count and add the daily counts over the week and input that data	-
Social Media Addiction Level	Categorical	-	Is the person addicted to social media or not?	Times opened >= 105 - Addicted Times opened < 105 - Not Addicted
LEGEND				
Green Label	Entertainment			
Blue Label	Career/Job/Work			
Red Label	Data used to determine Categorical variable (Will not be used for analysis)			
Yellow Label	Categorical variables			
***Note: A week is considered from Sunday to Saturday**				

- The data is collected from a class of 25 students who have reported their usage of social media apps for seven weeks.
- The data has 15 columns and 175 rows in total.
- The data is not cumulated for each student. Instead, we have considered each row as an individual data entry for better analysis.



- For the analysis, we have excluded columns 1, 2 and 14. Columns 1 & 2 are the Name and Week numbers we can exclude. We exclude column 14 as well because it is just the summation of all the other social media apps.
- The boxplot shows that the columns Telegram, Facebook, TikTok, WeChat, and Twitter have many outliers because we have very few students who use these chats.

Two Approaches

- As we have many outliers, we can analyse the data in two methods and check for any differences.
- We can consider two approaches - One considering all the Data, One considering only the columns WhatsApp, Instagram, Snapchat, LinkedIn, Total Social Media hours, and Addiction.

QUESTIONS and HYPOTHESIS

Questions

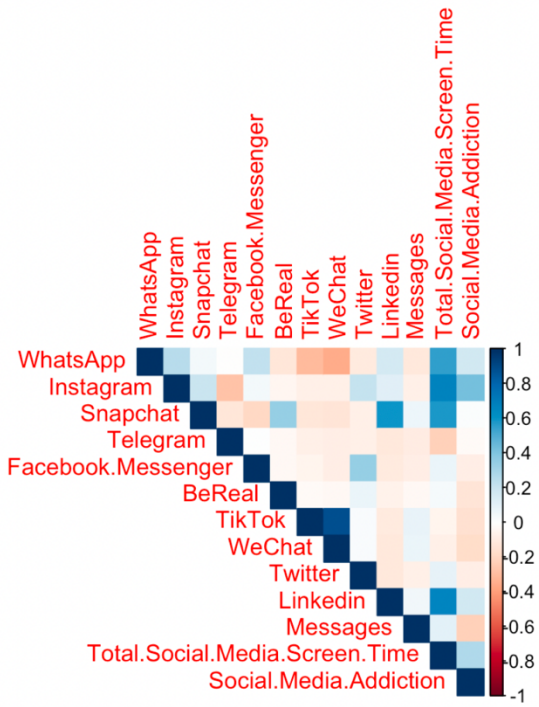
1. Based on the given variables, can we classify if the student is addicted to Social Media or not?
2. Based on the given variables, can we predict if the student is addicted to Social Media or not?

Hypothesis

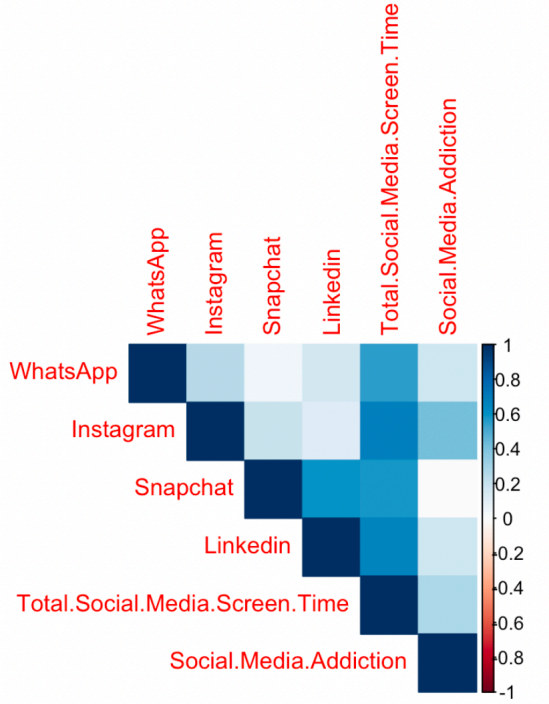
We can predict if the student is addicted to social media based on the time they have spent on the individual social media apps.

Correlation Plots

Approach 1

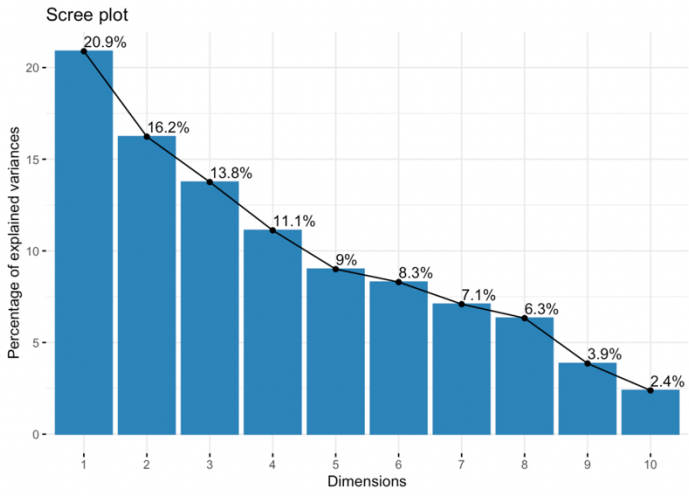


Approach 2



- The correlation matrix shows us a correlation between the columns in both cases.
- Hence, Principal Component Analysis (PCA) can be used to reduce the number of columns for the analysis.

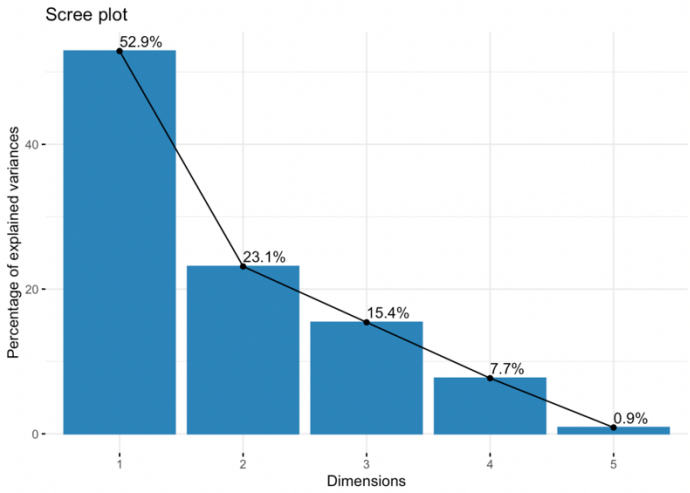
Principal Component Analysis (PCA)
Approach 1



- The scree diagram shows us that sum of the first 2 principal components is less than 70%.

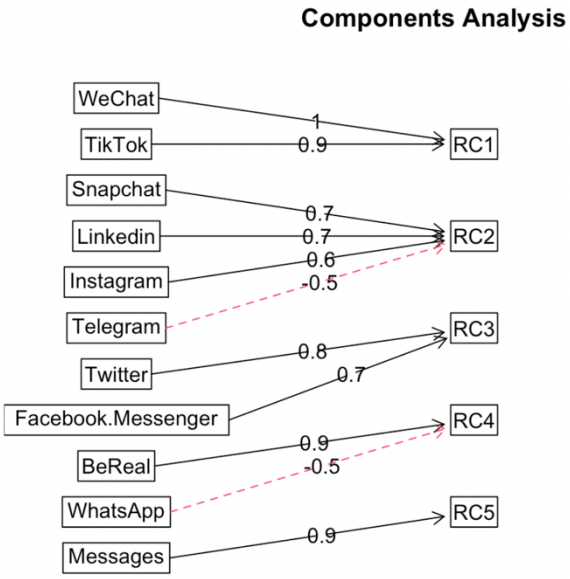
- So, we cannot move forward using PCA for column reduction.
- We now move on to check EFA for this main dataset.

Approach 2



- The scree diagram shows us that sum of the first 2 principal components is 76%.
- So, we can use PCA for column reduction.
- No need to consider EFA for this approach as there are already less columns (5) in this dataset.

Exploratory Factor Analysis (EFA)



Defining the factors obtained

RC1

- Both WeChat and Tiktok are popular apps in Asia region specifically.
- WeChat has multiple uses for chatting, payments, whereas Tiktok is used only for social media purpose to share and view videos.

RC2

- Snapchat, LinkedIn, Instagram, Telegram are popular all over the world.
- LinkedIn is used for professional purposes. Snapchat, Telegram are used for chatting and Instagram is used for posting photos and videos.

RC3

- Twitter & Facebook are popular all over the world.
- Both are used for posting photos and videos.

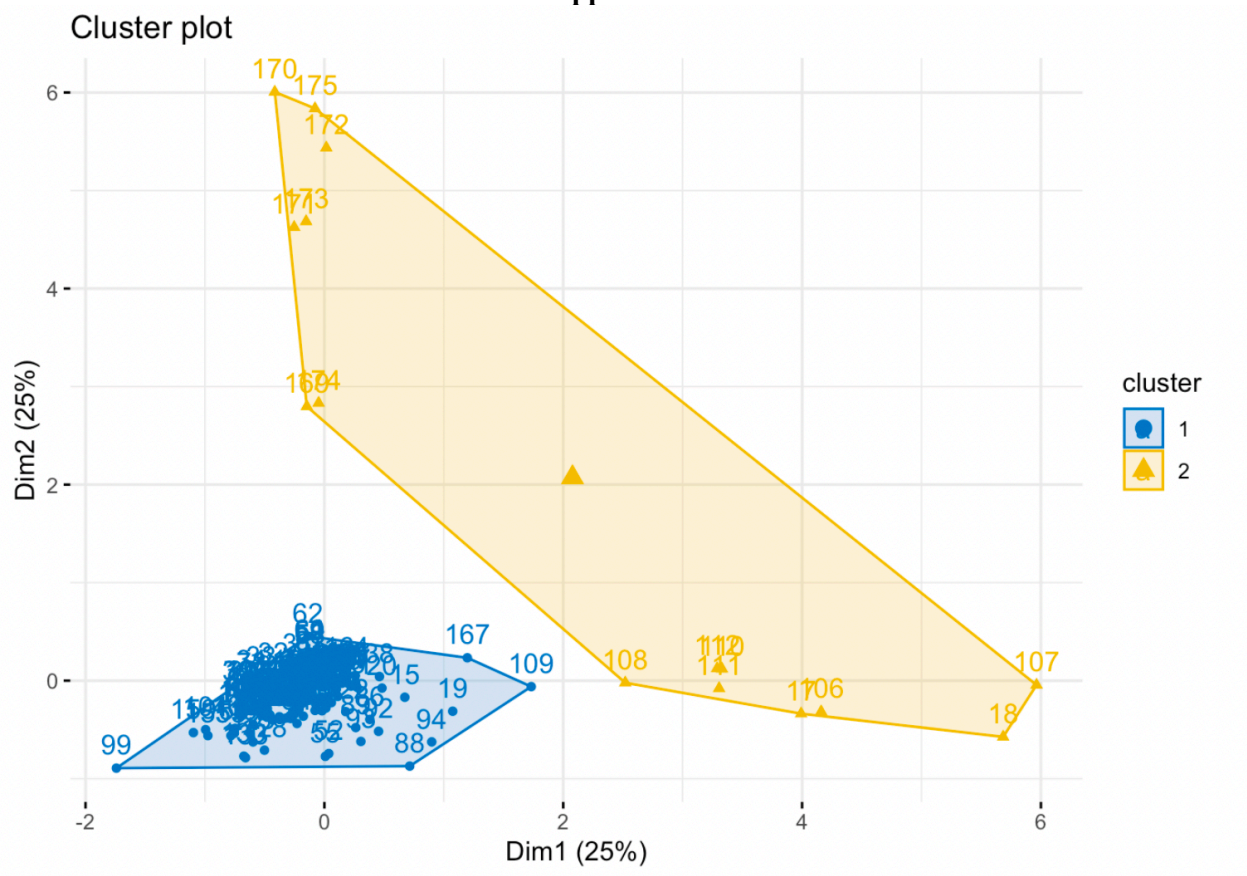
RC4

- WhatsApp is popular world wide and BeReal is a new app that has entered the social media market.

RC5

- RC5 has only one variable pertaining to it.
- So, we excluding the RC5 and consider the Messages column directly.

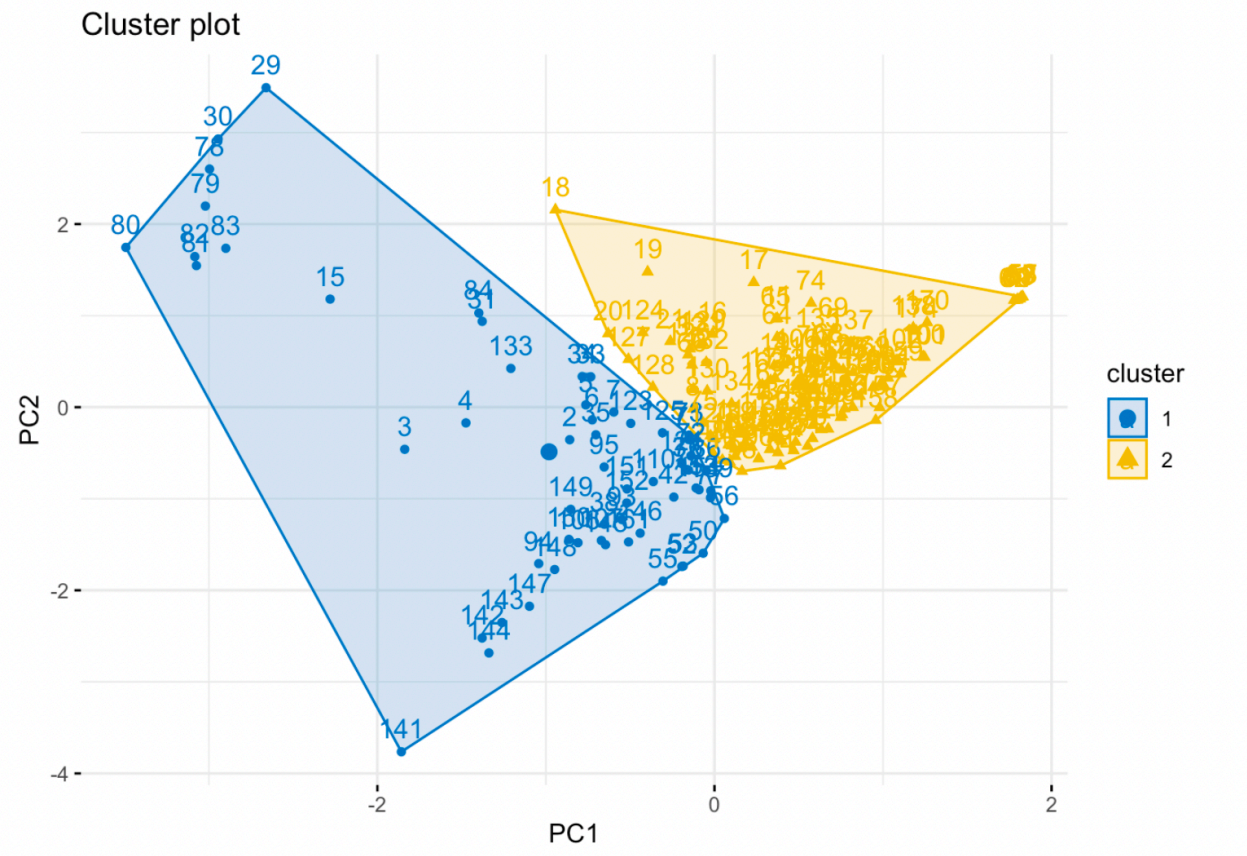
Cluster Analysis
Approach 1



```
##                Actual
## Clustered      Addicted Not Addicted
##   Addicted           97            63
##   Not Addicted        3            12
```

- Although we have a recall of 1, we can see that the confusion matrix shows the clustering is done in a way where almost all the users are Addicted.
- This shows that we cannot classify our data into Addicted and Not addicted based on the variables given.
- We can now check the clustering using the second approach.

Approach 2



```
##                Actual
## Clustered1     Addicted Not Addicted
##   Addicted            45           15
##   Not Addicted        55           60
```

- The precision is obtained to be 80% which is not so bad.
- Yet, we see that 55 inputs who are addicted to social media have been clustered to not addicted.
- This shows that we cannot classify our data into Addicted and Not addicted based on the variables given.

Classification Summary

Considering both approaches, we can see that we cannot classify the data into addicted and not addicted.

Answer to our question 1: No, we cannot classify.

Logistic Regression

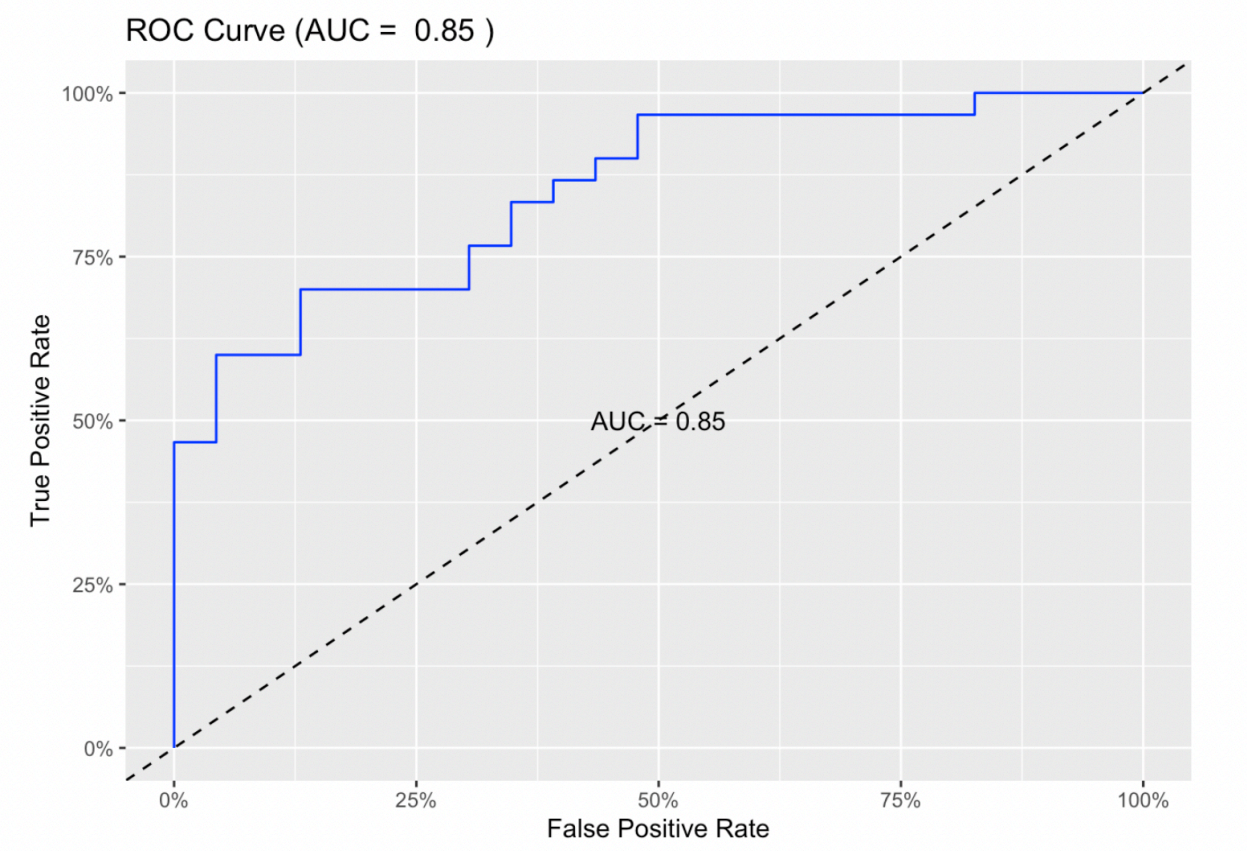
Approach 1

```
##
## Call:
## glm(formula = Ytrain_sm ~ ., family = "binomial", data = x_sm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9232  -0.8199   0.2994   0.7769   1.7039
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.98716    0.63774  -1.548   0.1216
## WhatsApp           -0.01881    0.05413  -0.348   0.7282
## Instagram           0.27012    0.06763   3.994  6.5e-05 ***
## Snapchat           -0.36162    0.17298  -2.090   0.0366 *
## Telegram            0.37880    0.68887   0.550   0.5824
## Facebook.Messenger -0.82205    0.55113  -1.492   0.1358
## BeReal             -0.21339    0.78789  -0.271   0.7865
## TikTok              0.26279    1.02170   0.257   0.7970
## WeChat             -0.23816    0.31745  -0.750   0.4531
## Twitter            -0.41172    0.20916  -1.968   0.0490 *
## Linkedin            0.16061    0.07702   2.085   0.0370 *
## Messages           -0.46549    0.20845  -2.233   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.46  on 121  degrees of freedom
## Residual deviance: 120.46  on 110  degrees of freedom
## AIC: 144.46
##
## Number of Fisher Scoring iterations: 5
```

- The regression summary shows that we have significant variables that affect the output variable.
- We check the confusion matrix, precision and recall of our regression below.

```
##                actual_sm
## predicted_sm2 No Yes
##            No  14   4
##            Yes  9  26
```

- Precision we got for the first approach is good with 86.7%



- AUC of the ROC curve = 85%
- Considering both AUC of the ROC curve of 85% and a precision of 86.7% we can say that the regression model works well and we will be able to predict the if the student is addicted to social media or not based on the variables provided.
- We can now check how the logistic regression for the second approach works.

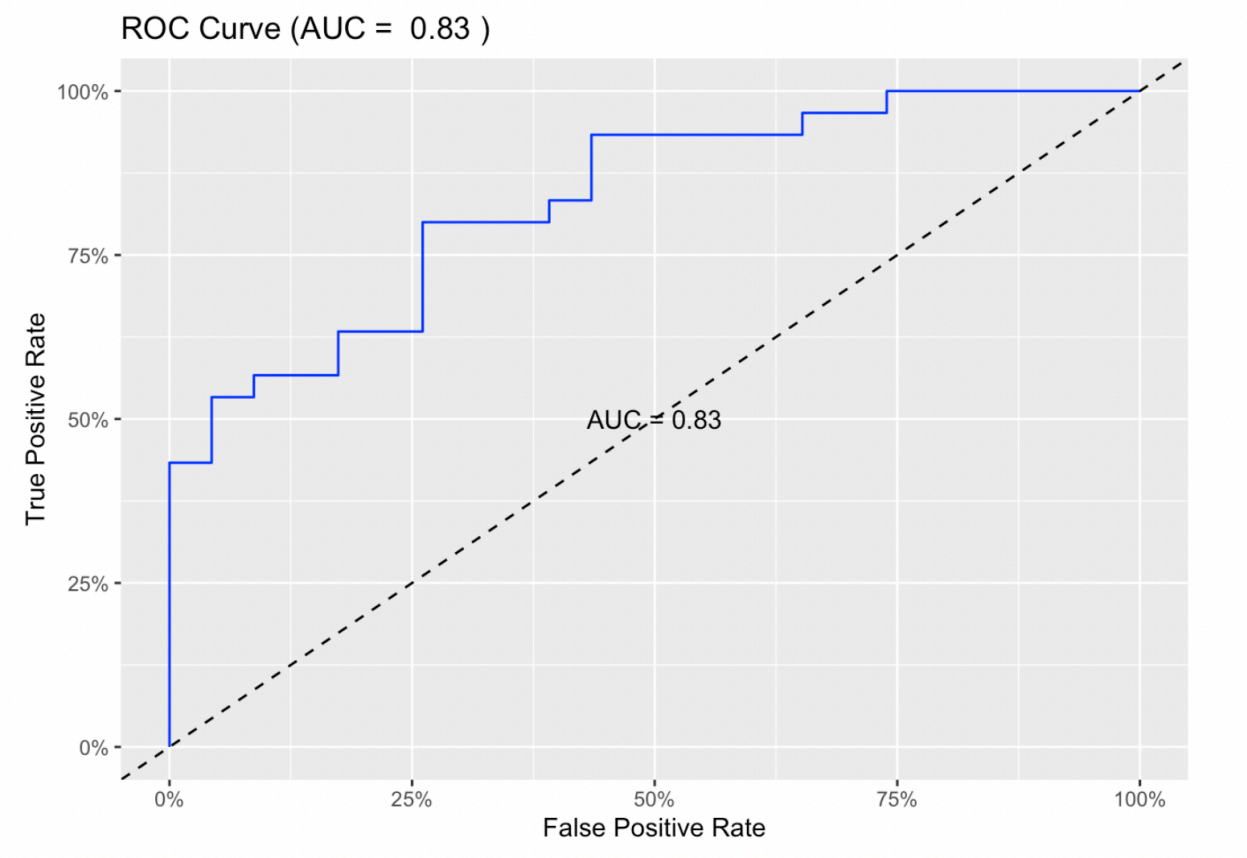
Approach 2

```
##
## Call:
## glm(formula = Ytrain_sm_new ~ ., family = "binomial", data = x_sm_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7926  -0.8728   0.3926   0.7931   2.3579
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.598714   0.543588  -1.101 0.270718
## WhatsApp                        0.291979   0.098133   2.975 0.002927 **
## Instagram                       0.555039   0.126742   4.379 1.19e-05 ***
## Snapchat                        0.008102   0.174537   0.046 0.962973
## Linkedin                        0.481654   0.127204   3.786 0.000153 ***
## Total.Social.Media.Screen.Time -0.329884   0.099419  -3.318 0.000906 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.46  on 121  degrees of freedom
## Residual deviance: 125.24  on 116  degrees of freedom
## AIC: 137.24
##
## Number of Fisher Scoring iterations: 5
```

- The regression summary shows that we have significant variables that affect the output variable.
- We check the confusion matrix, precision and recall of our regression below.

```
##                 actual_sm
## predicted_sm3 No Yes
##            No  14   5
##            Yes  9  25
```

- Precision we got for the first approach is good with 83.3%.



- AUC of the ROC curve = 83%
- Considering both AUC of the ROC curve of 83% and a precision of 83.3% we can say that the regression model works well and we will be able to predict the if the student is addicted to social media or not based on the variables provided (Only 5 columns in this approach compared to 12 in the first approach)

Regression Summary

Considering both approaches, we can see that we can predict if the student is addicted to social media based on the input variables.

Answer to our question 2: Yes, we can predict.

Our hypothesis that we can predict addiction can be proved right.