Class Survey Project

Data Dictionary

Field Name	Data Type	Description	Possible Values for Categorical
Week			
Whatsapp	Numeric	In hours, how much time do you spend on Whatsapp this week.	-
Instagram	Numeric	In hours, how much time do you spend on Instagram this week.	-
Snapchat	Numeric	In hours, how much time do you spend on Snap this week.	-
Telegram	Numeric	In hours, how much time do you spend on Telegram this week.	-
Facebook	Numeric	In hours, how much time do you spend on Facebook/Messenger this week.	-
BeReal	Numeric	In hours, how much time do you spend on BeReal this week.	-
TikTok	Numeric	In hours, how much time do you spend on TikTok this week.	-
Wechat	Numeric	In hours, how much time do you spend on Wechat this week.	-
Twitter	Numeric	In hours, how much time do you spend on Twitter this week.	-
Linkedin	Numeric	In hours, how much time do you spend on Linkedin this week.	-
Messages	Numeric	In hours, how much time do you spend on Messages this week.	-
Total Social Media Screen Time	Numeric	The sum of all the above times in hours.	-
Number of times opened (hourly intervals)	Numeric	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

This project is used to determine whether a person is addicted to social media or not. There multiple apps which are taken into consideration. The amount of time (hours) spend on each app, the total screentime of each person and the number of times the app was opened are few factors which contribute in determining addiction.

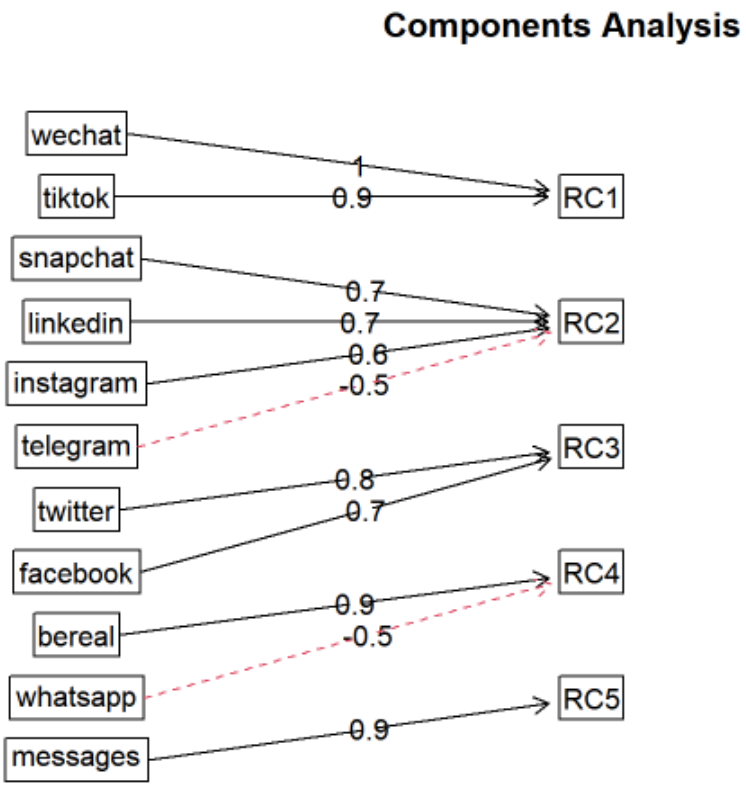
In this project I am using Principal Component Analysis to determine the significant features in the data. I am using cluster analysis to find the commonalities in the data which can be grouped together. Also, I have checked if the data contains any underlying factors using Exploratory Factor Analysis. For prediction I am using Logistic Regression.

The data is imported and data is prepared by removing student, week, total screen time and number of times opened fields for yielding better results.

The following questions are answered with the help of this project.

Q1. Are there any underlying factors within the data?

To answer this question,we use Exploratory Factor Analysis.



From this diagram we can consider RC1, RC3 and R4

RC1 contains WeChat, TikTok.

RC3 contains Twitter and Facebook.

R4 contains Breal and WhatsApp.

fit.pc$scores

```
##              RC1        RC2        RC3        RC4        RC5
## [1,] -0.15249728 0.2694719 -0.3467212 -0.1113600 -0.1380693
## [2,] -0.20912237 0.8189327 -0.1902031 -0.3947255 -0.3995048
## [3,] -0.13314999 1.7668179 -0.0301857 -0.5580449 -0.4714241
## [4,] -0.20803139 1.3778233 -0.2661032 -0.4451929 -0.4063803
## [5,] -0.08200773 0.8264971 -0.2212293 -0.2685068 -0.2272910
## [6,] -0.14132275 0.6393914 -0.3196369 -0.4046006 -0.4439738
```

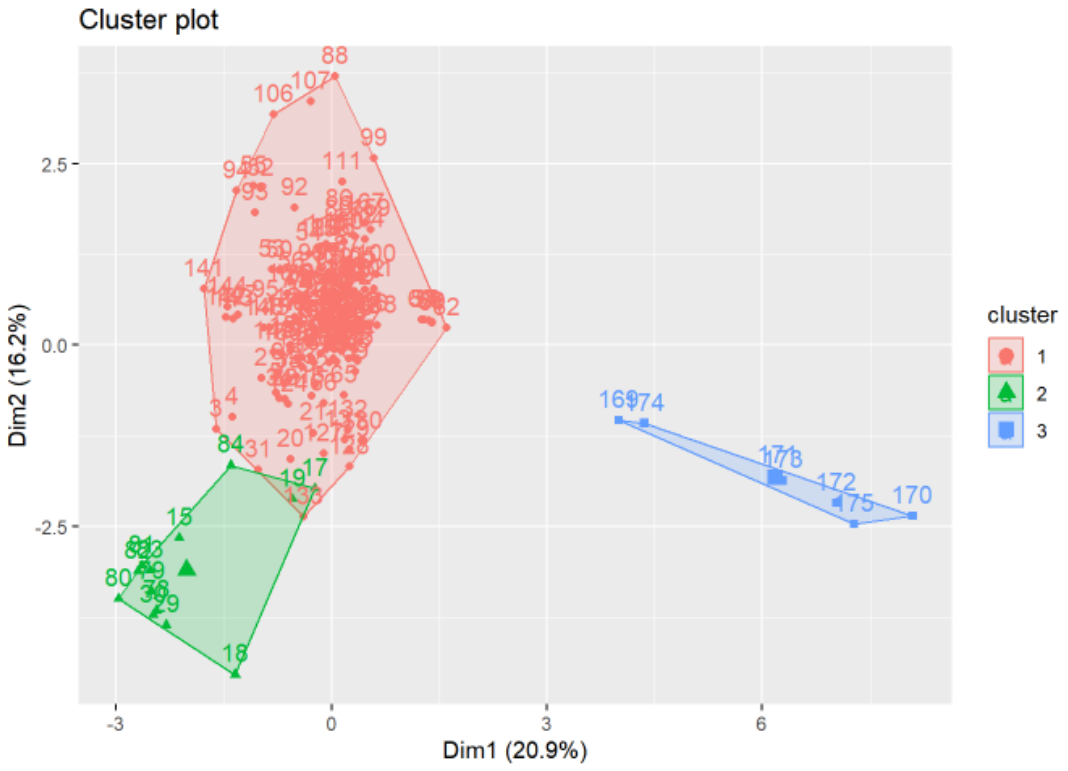
The pc$scores tells us the following:

RC2 has the highest factor loading of 0.269, followed by RC5 −0.138 and RC1 −0.152

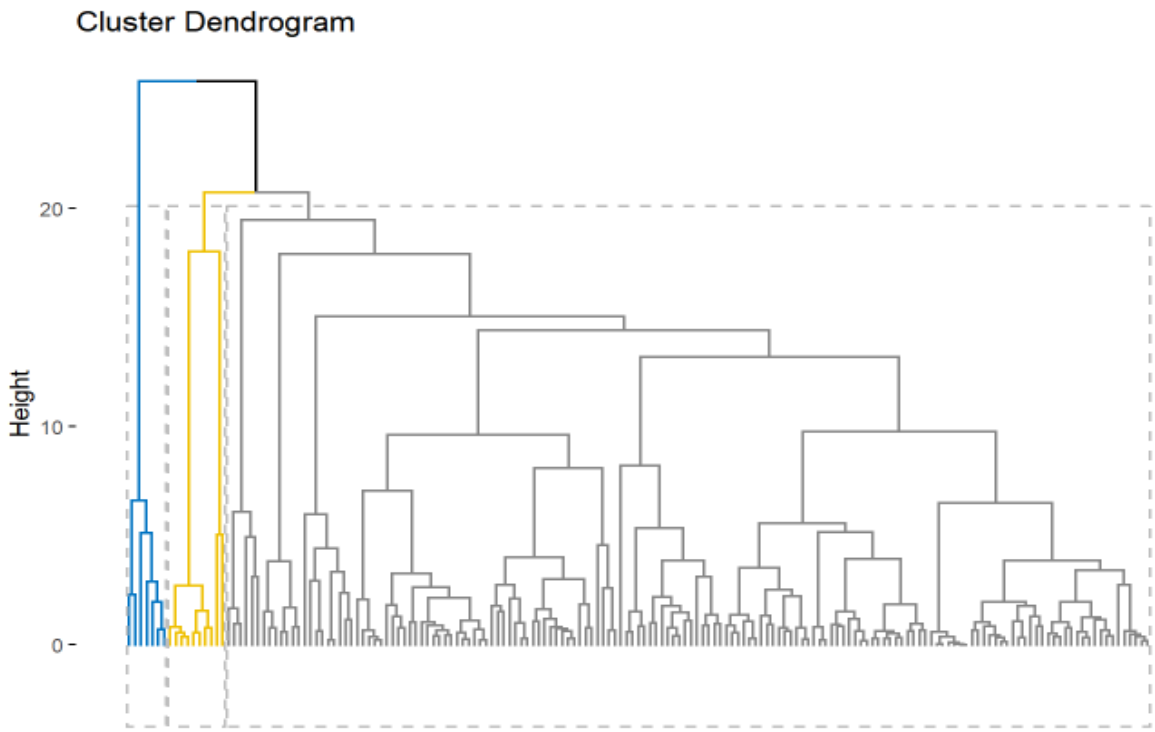
RC3 and RC4 have negative loadings. This suggests that the variables associated with RC2 may be more strongly related to the underlying factor than the variables associated with RC3 and RC4.

Q2. Are there any commonalities in the data

To answer this question we did Cluster Analysis in the data



Here it is observed that 3 clusters have better visualization. There is significantly less overlapping between the clusters.




```
fviz_silhouette(res.hc)
```

```
##   cluster size ave.sil.width
## 1       1  158          0.49
## 2       2   10          0.35
## 3       3    7          0.51
```

From above data we can conclude:

Cluster 1 indicates that observations here are well matched.

Cluster 2 has the average sil widh less than cluster 1

Cluster 3 has highest average sil width, indicating that observations are well matched.

Q3. Using the Class dataset predict whether the students are addicted to social media or not and determine if the model's accuracy.

To answer the above question we use Logistic Regression model, if the model has high accuracy, it will be a good fit for predicting whether the student is addicted or not.

For this the data needs to be split into training and testing set. For this, we divide 70% of the data for training set and 30% for testing set.

```
#Split the data in training and testing set
# Set the seed for reproducibility
set.seed(123)
trainIndex <- createDataPartition(class$addiction, p = 0.7, list = FALSE) # Split the data into 70% training and 30% testing
training <- class[trainIndex, ]
testing <- class[-trainIndex, ]
nrow(class)
nrow(training)
nrow(testing)
table(training$addiction)
```

Now we fit the Logistic Regression model

```
#Logistic Regression
model <- glm(addiction~., data= training, family = binomial)
summary(model)
PredictTrain <- predict(model, newdata= testing, type = "response")
summary(PredictTrain)
```

```
##
## Call:
## glm(formula = addiction ~ ., family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5326  -0.8421   0.4701   0.7446   1.5334
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.779e-01  7.521e-01  -0.237  0.81298
## whatsapp    -6.421e-02  6.390e-02  -1.005  0.31495
## instagram    2.028e-01  6.232e-02   3.254  0.00114 **
## snapchat    -1.774e-01  2.021e-01  -0.877  0.38032
## telegram     2.008e-01  7.157e-01   0.281  0.77905
## facebook    -1.516e-01  7.561e-01  -0.200  0.84110
## bereal      -4.073e-01  7.210e-01  -0.565  0.57210
## tiktok      -4.252e+01  1.173e+04  -0.004  0.99711
## wechat      -5.127e+01  4.702e+03  -0.011  0.99130
## twitter     -8.385e-01  7.737e-01  -1.084  0.27849
## linkedin     7.327e-02  7.438e-02   0.985  0.32455
## messages    -4.438e-01  2.304e-01  -1.926  0.05410 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 168.16  on 122  degrees of freedom
## Residual deviance: 114.74  on 111  degrees of freedom
## AIC: 138.74
##
## Number of Fisher Scoring iterations: 21
```

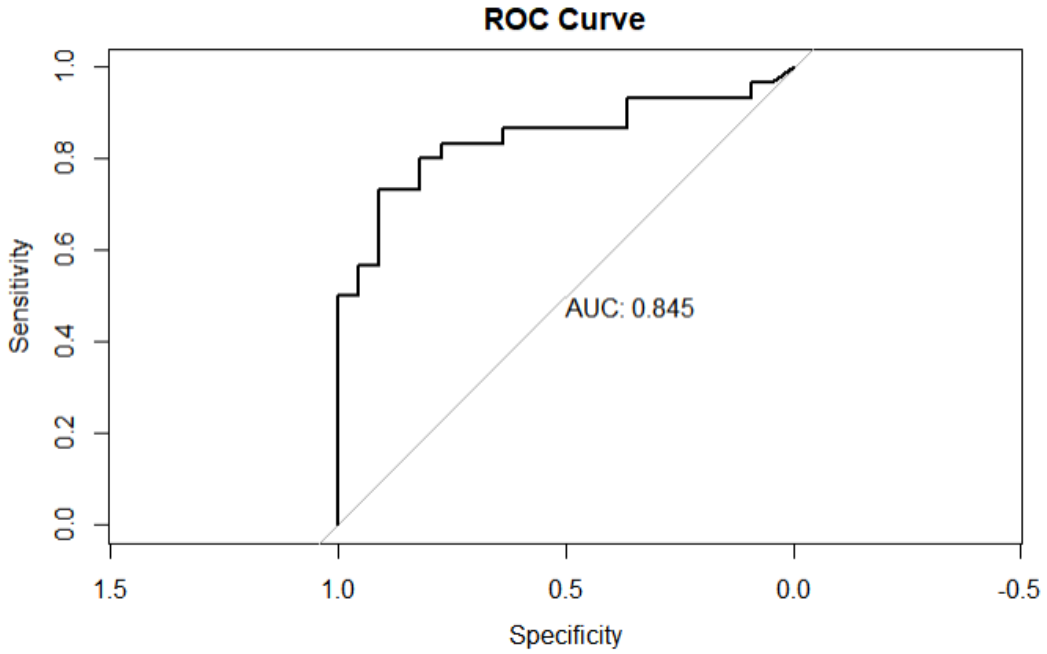
Confusion Matrix

```
##                 actual_class
## predicted_class  0  1
##               0 18  6
##               1  4 24
```

The confusion matrix indicates the accuracy to be 82%

(18 + 24) / (18 + 6 + 4 + 24) = 0.82, or 82%

ROC AUC score



The ROC AUC score for this model is 84.5%, which determines that the model is a good fit for predicting if the student is addicted to social media or not.