



University of Mumbai

Department of Statistics

Vidyanagari Mumbai – 400098

CERTIFICATE

This is to certify that

**Ms. Aparna Sunil Raut**

**Ms. Dipali Kishor Pingle**

**Ms. Maseera Sayyed Riyaz**

**Ms. Mehvash Atique Shaikh**

of M.Sc. Part-II have successfully completed the project entitled

**“IMPACT OF COVID-19 ON STUDENT’s**

**HEALTH, EDUCATION & SOCIAL LIFE.”**

During the academic year **2021-2022**.

This work to the best of our knowledge and belief is original.

Mrs.Ashwini.M.Ghanekar

(Project Mentor)

**ACKNOWLEDGEMENT**

We are extremely thankful to our guide ***Mrs.Ashwini Ghanekar*** for her guidance and support. We are extremely glad to have her as our project guide. We are also grateful to our HOD **Dr. *Vaijayanti Dixit*** for extending her help and support. We also thank other professors of the department ***Dr. Santosh Gite, Dr. Alok Dabade ,Prof.Sheetal Chabukswar, Prof. Priyesh Tiwari*** and the entire staff of Statistics Department, Mumbai University for providing us with all the necessary infrastructure and co-operation with us at every stage of our project. Last but not the least, our special thanks to all the people for providing us with their vital information, opinions, precious time and their support.

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| Serial No. | Topics | | Page No. |
| 1. | Introduction | | 5 |
| 2. | Methodology | | 7 |
| 3. | Objectives | | 8 |
| 4. | Statistical techniques and Software | | 9 |
| 5. | Graphical Representation | | 15 |
| 6. | Analysis | |  |
|  | 6.1 | Chi - Square Test of Independence & Fisher’s Exact Test | 20 |
| 6.2 | Based on machine learning we used various classifiers | 32 |
| 6.3 | Kruskal Wallis Test | 40 |

**INTRODUCTION**

The emergence of Corona Virus disease (COVID-19) has led the world to an unprecedented public health crisis. Emergency protocols were implemented in India to control the spread of the virus which resulted in restrictions on all non-essential public movements. With the closure of educational institutions, the need for a rapid transition from physical learning to the digital sphere of learning emerged. Online learning has been observed as a possible alternative to conventional learning. However, according to a meta-analysis on e-learning, it is reported that online learning is better than nothing and similar to conventional learning. To improve the e-learning experience, the education institutions are required to comply with the guidelines and recommendations by government agencies, while keeping students encouraged to continue learning remotely in this tough environment addresses five high-impact guidelines for the efficient conduct of online education.

We investigated and analysed the potential consequences of the COVID-19 pandemic on the life of students.

Our research shows that there is a wide gap between the government's policy aspirations and the implementation of these online education policies at the grassroots level. Moreover, our study attempts to

assess the mental situation of students of different age groups using different parameters including sleeping habits, daily fitness routine, and social support. Further, we analyse different coping mechanisms used by students to deal with the current situation.

**Why this topic?**

This rapid evolution at such a large scale has influenced the students of all age groups. It is expected that the continued spread of the disease, travel restrictions and the closure of educational institutions across the country would have a significant effect on the education, social life, and mental health of the students. The students from the less privileged backgrounds have experienced larger negative impacts due to the Covid-19 outbreak. Reduction in family income, limited access to digital resources, and the high cost of internet connectivity have disrupted the academic life of the students.

Moreover, 1.5 billion students across the world are now deprived of basic education leading to a serious psychological impact on their health. Moreover, changes in daily routine including lack of outdoor activity, disturbed sleeping patterns, social distancing have affected the mental well-being of the students. uses 7-item Generalized Anxiety Disorder Scale as a diagnostic tool for the assessment of anxiety disorders, panic disorders, and social phobia. Further, analyses mediating roles of resilience, coping, and social support to deal with psychological symptoms.

**METHODOLOGY**

**1)Data collection**:

• We collected the Secondary data from Kaggle website. The data included 1182 records.

**2)Data Cleaning**:

• There data consisted of 75 missing values in the data.

• Using the Measure of Central Tendency (Mode) we filled up the missing values.

**3)Defining Objective**

**4)Exploratory Data Analysis**

**5)Analysis of data using various statistical technique**

**6)Data Coding and Data Entry:**

* Using software like R, Python, SPSS, MS. Excel, Minitab.

**7)Interpretation of results**

**8)Conclusion**

**9)Preparation of Project Report**

**Objectives**

1. **To study whether there is relationship between:**

* **Health issue v/s Change in weight**
* **No. of meals per day v/s Health issue**
* **Time spent on fitness v/s Change in weight**
* **Time spent on online class v/s Medium used for online class**
* **Time spent on social media v/s Health issue**

1. **To analyse whether there was effect on student’s health due to online mode of study during Covid -19 lockdown.**
2. **To analyse the time spent by students on their online class and the other daily activities performed by them including their age factor.**

**STATISTICAL TECHNIQUES AND SOFTWARE**

* **Techniques:**

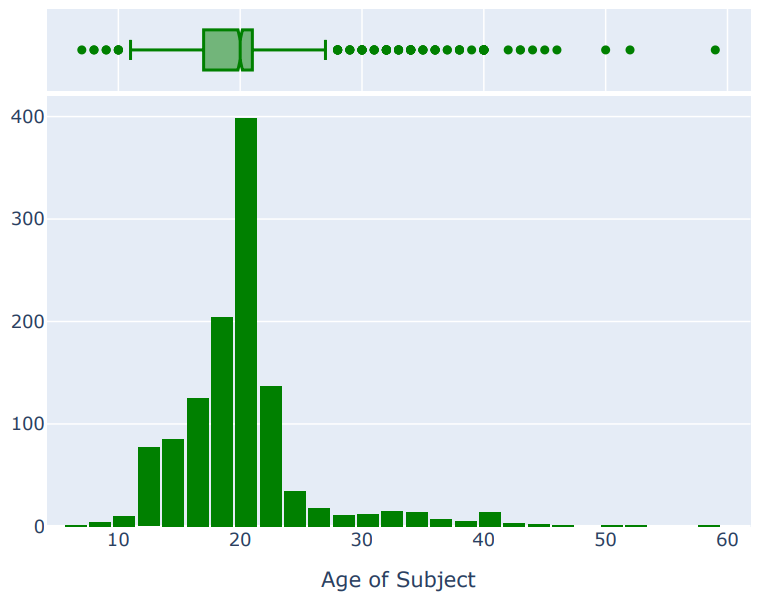
1. Chi-Square test of Independence
2. Based on machine learning we used various classifiers.
3. Kruskal Wallis Test

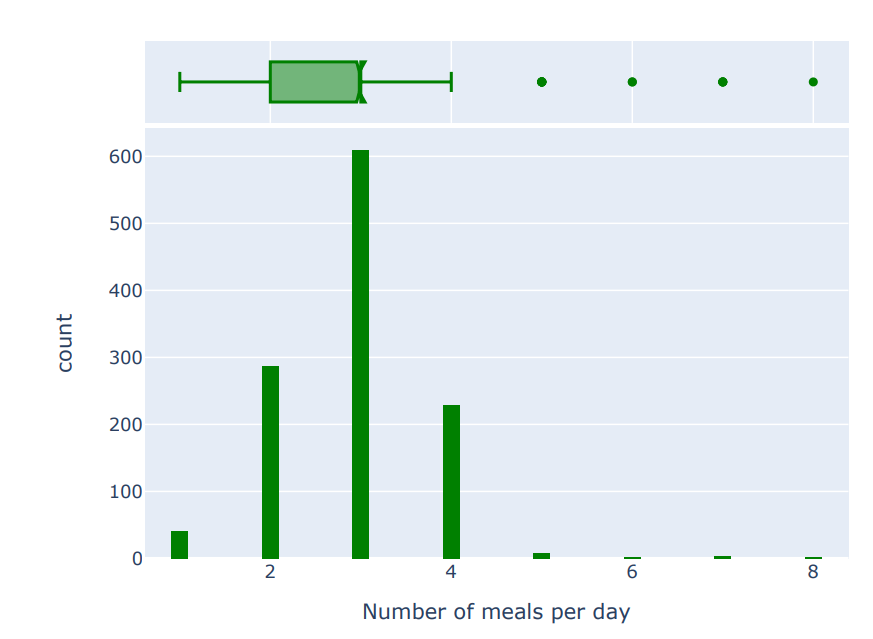
* **Software:**

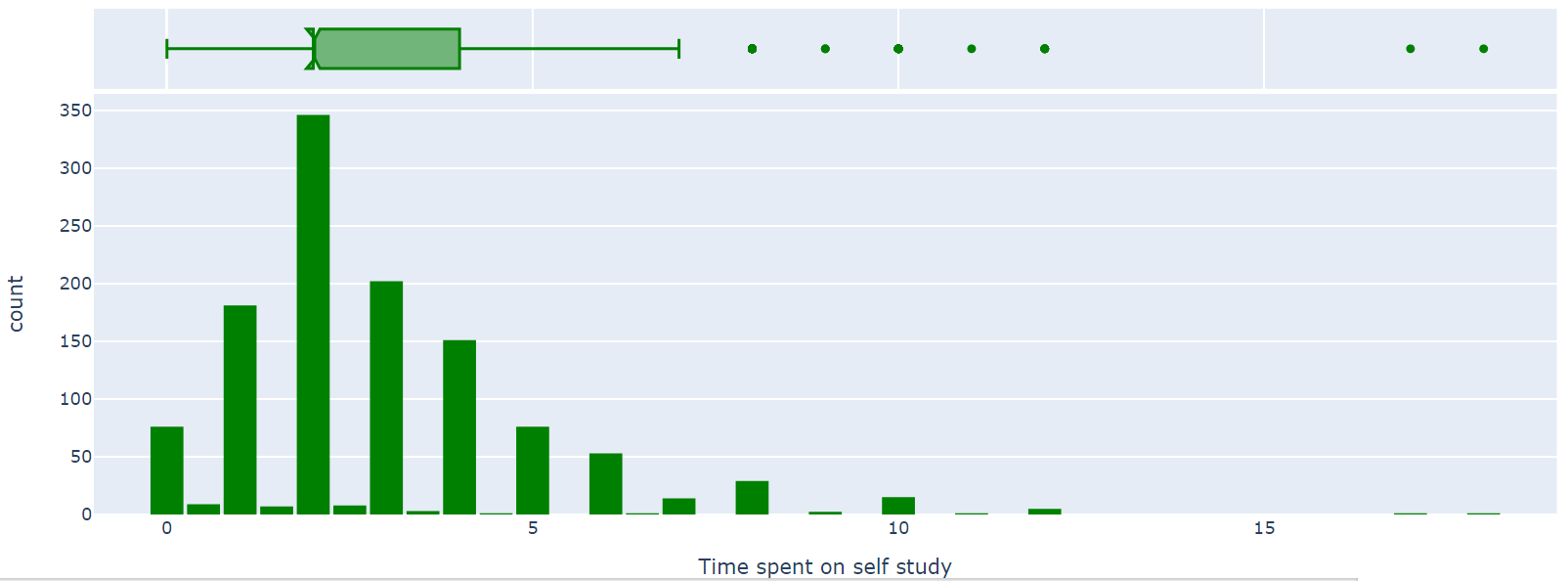
1. R Software
2. SPSS
3. Minitab
4. MS-Office

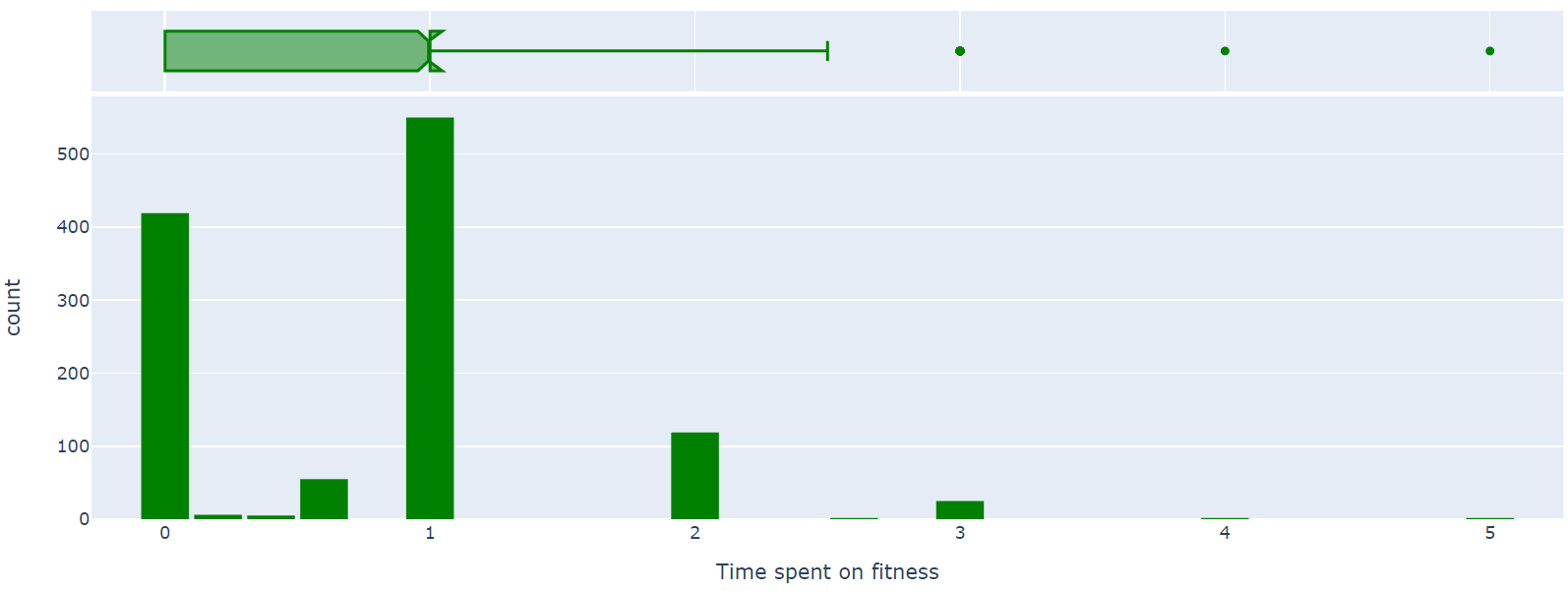
**Data Cleaning:**

* There were total 75 missing values in the data
* The variable which consists missing values were: Rating of Online Class experience, Medium for online
* Using the Measure of Central tendency i.e. Mode we filled up the missing values as the variable were categorical.
* To detect the presence are any outliers we plot boxplot**.**
* Also, we remove the outliers using Inter Quartile Range (IQR) technique.

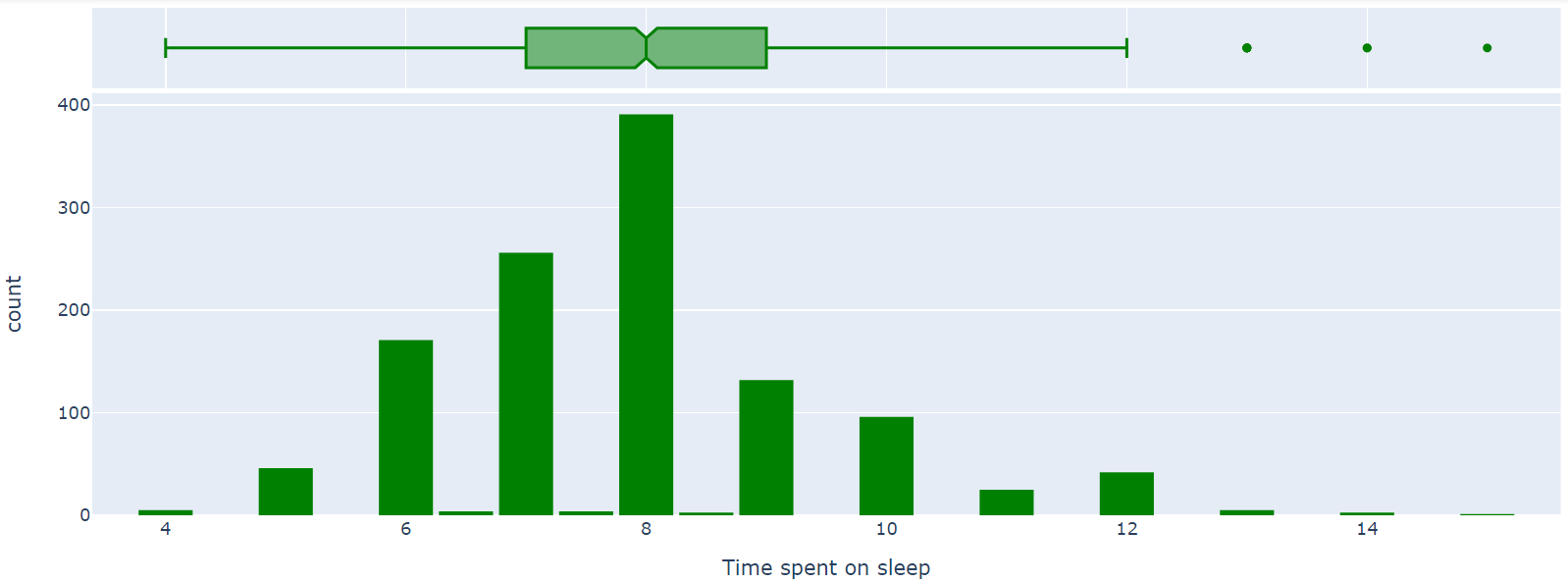
**Detecting of Outliers in the data:**



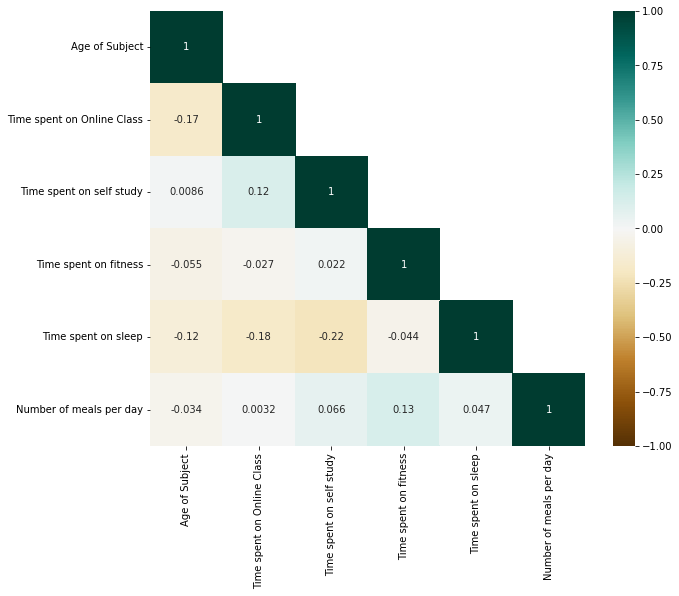




****

****

**To check multicollinearity using Heat map:**



**Conclusion:**

Here, we observe that VIF >10 for Time spent on sleep, Time spent on online class which implies multicollinearity exists for them.

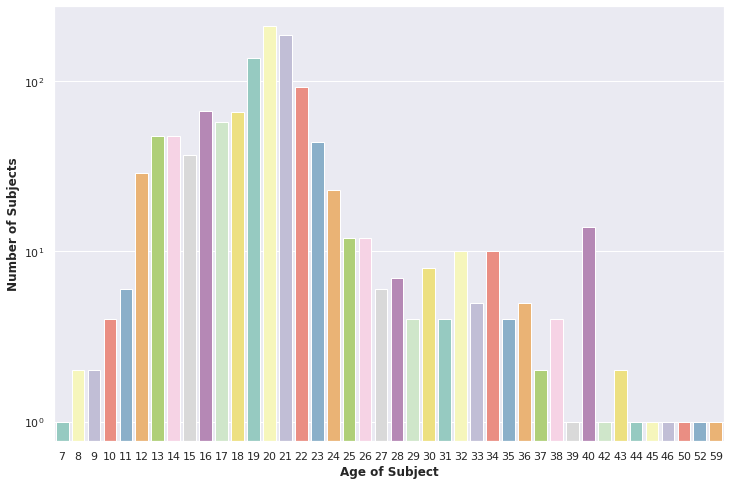
**Graphical Representation:**

Region of Residence

**Conclusion:**

**From the sample, we conclude that the data consist of about 61% people from Delhi-NCR and 39% people from outside Delhi-NCR.**

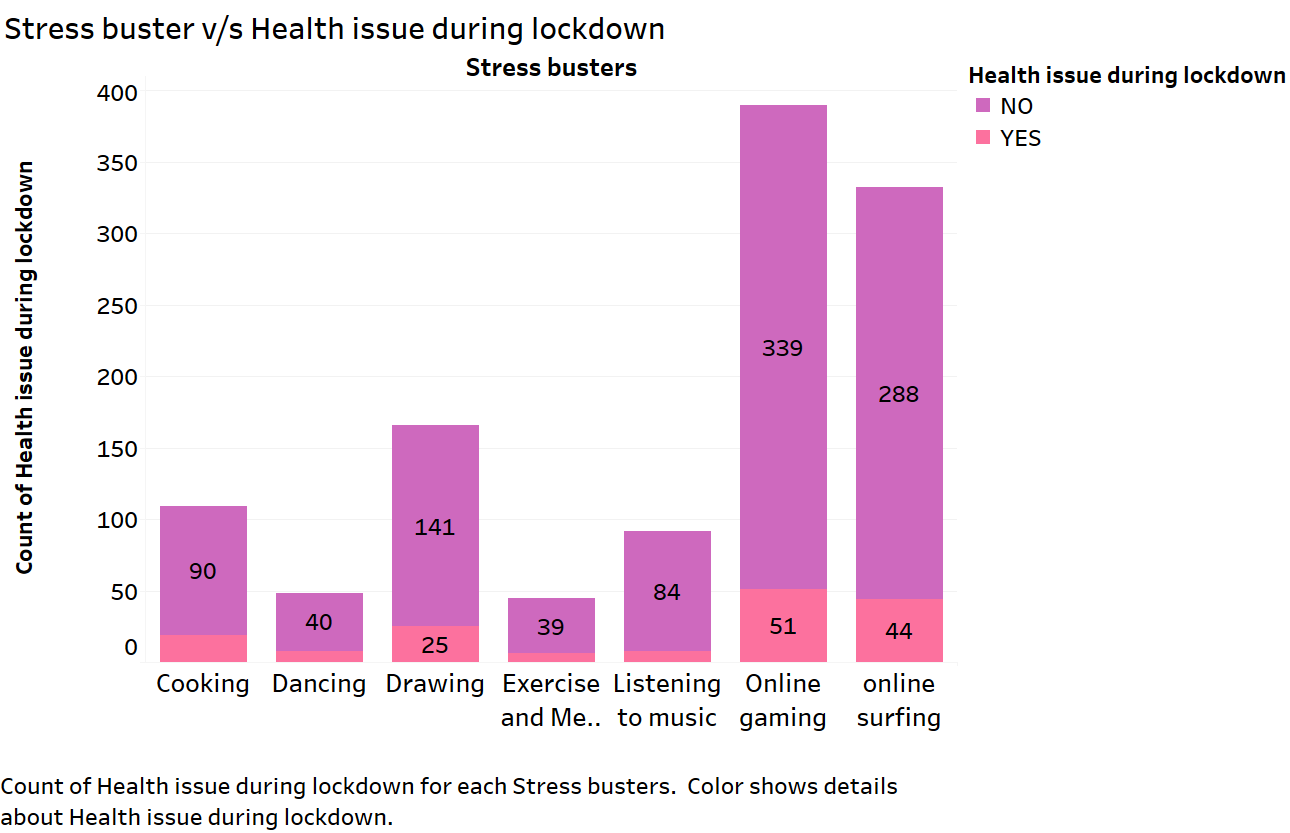
**Age wise distribution:**

**Conclusion:**

**Age of the subjects starts from 7 years end up to 59 years in which most of the subjects are from age group 17 years to 23 years.**

**Conclusion:**

**From our sample, we conclude that 3% of subject rating Poor for the online class, 8% rating Excellent, 19% rating Good, 33% rating Average and 37% rating very poor for the online class which is a higher % in all ratings.**

****

**Stress buster v/s Health issue during lockdown.**

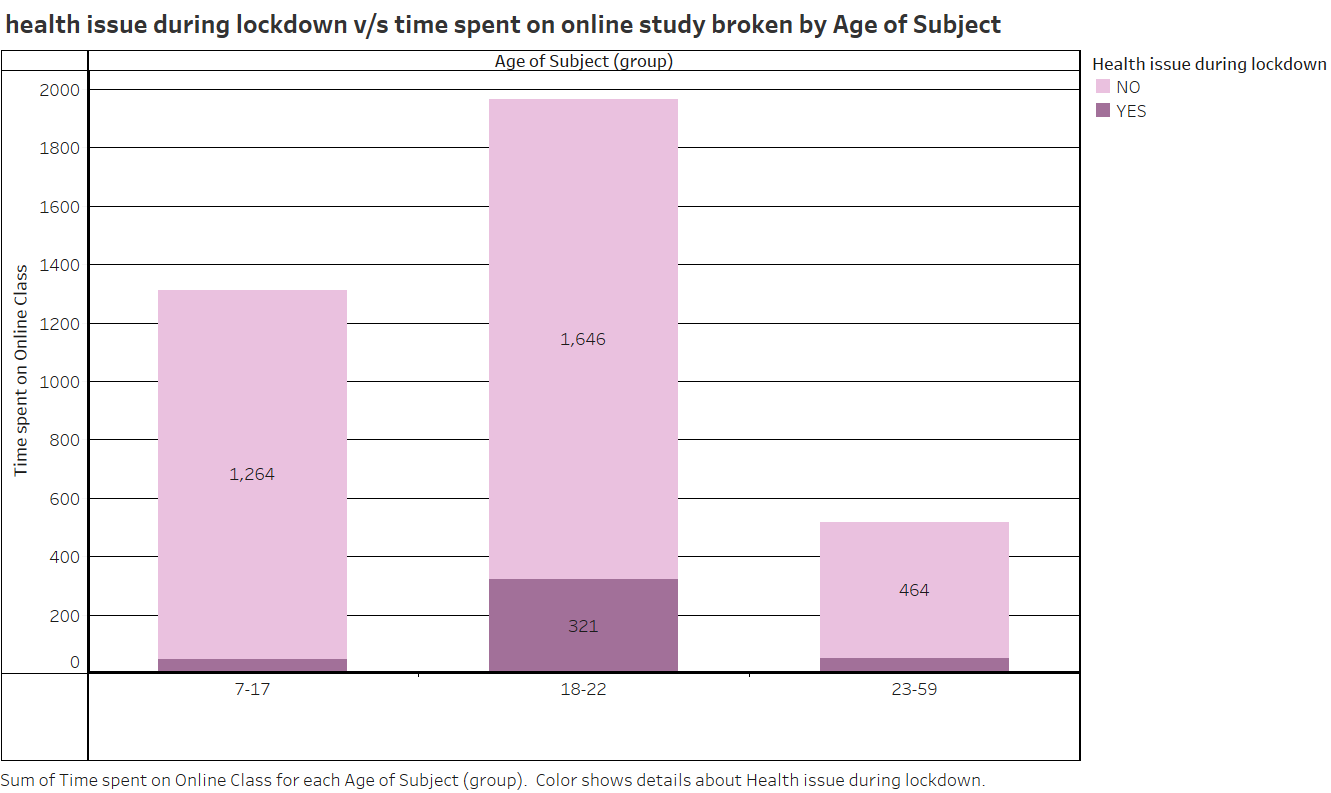
**Conclusion:**

**From the above graph, we conclude that most of the subjects doing Online gaming for a stress buster during lockdown followed by Online surfing and health issue is very less among the subjects.**

**Conclusion:**

* **Most popular social media platform**

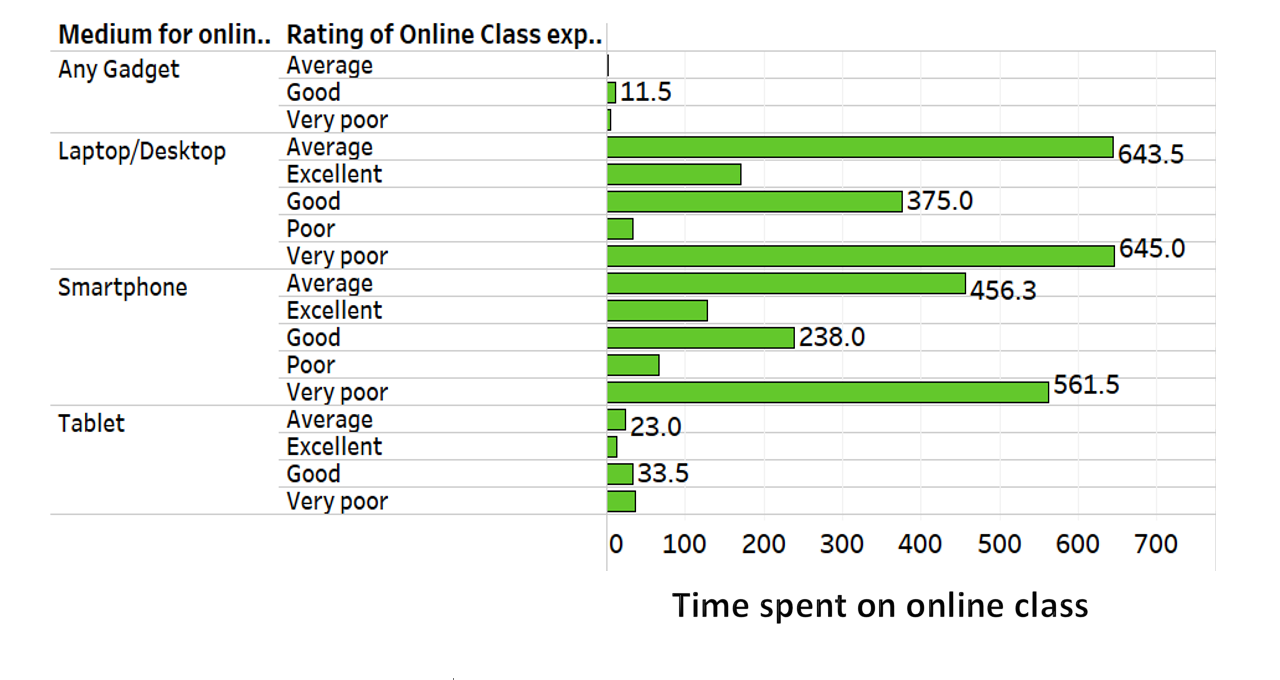
**From sample, we conclude that most popular social media used by students is Instagram followed by, WhatsApp, YouTube and other social media platform.**

****

**Health issue v/s Time spent on online study.**

**Conclusion:**

**The subject of age group (18-22) years collectively spent more time on online class as compared to age group (7-17) years and (23-59) years also we can conclude that subjects doesn’t face much health issues during lockdown which is differentiated by colour.**

****

**Conclusion:**

**We can conclude that Laptop and Smartphone are more convenient than other medium for online class.Objective 1:**

To study whether there is relationship between:

* **Health issue v/s Change in weight**
* **No. of meals per day v/s Health issue**
* **Time spent on fitness v/s Change in weight**
* **Time spent on online class v/s Medium used for online class**
* **Time spent on social media v/s Health issue**

Technique used: Chi-squared test and Fisher’s exact test

Tool used: Excel and R software

**Analysis:**

We find out observed frequency table by using MS. Excel and then create matrix, chi-squared test, Fisher’s exact test by using R-software.

**To check the Association between different factors:**

Independence tests are used to determine if there is a significant relationship between two categorical variables. There exist two different types of independence test:

**the Chi-square test (the most common)**

**the Fisher’s exact test**

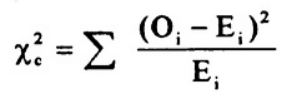
On the one hand, the Chi-square test is used when the sample is large enough (in this case the pp-value is an approximation that becomes exact when the sample becomes infinite, which is the case for many statistical tests). On the other hand, the Fisher’s exact test is used when the sample is small (and in this case the pp-value is exact and is not an approximation).

**CHI Square test to check the association:**

The Chi-Square Test for Association is **used to determine if there is any association between two variables**. It is really a hypothesis test of independence. The null hypothesis is that the two variables are not associated, i.e., independent. The alternate hypothesis is that the two variables are associated.

The procedure involves comparing the observed cells frequencies with expected cell frequencies. Expected frequencies are number of cases that should fall in each cell if there is no relationship between the two categorical variables.

Test statistic is given as:



Where,

𝑂𝑖: Observed frequency

𝐸𝑖: (Row total \* Column total)/N

N: Total number of observations

**Fisher’s Exact test to check the association:**

* **Fisher's exact test** is a [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance) test used in the analysis of [contingency tables](https://en.wikipedia.org/wiki/Contingency_table).[Although in practice it is employed when [sample](https://en.wikipedia.org/wiki/Sample_(statistics)) sizes are small, **it is valid for all sample sizes.**

It is used to determine whether or not there is a significant association between two categorical variables. It is typically used as an alternative to the [Chi-Square Test of Independence](https://www.statology.org/chi-square-test-of-independence/).

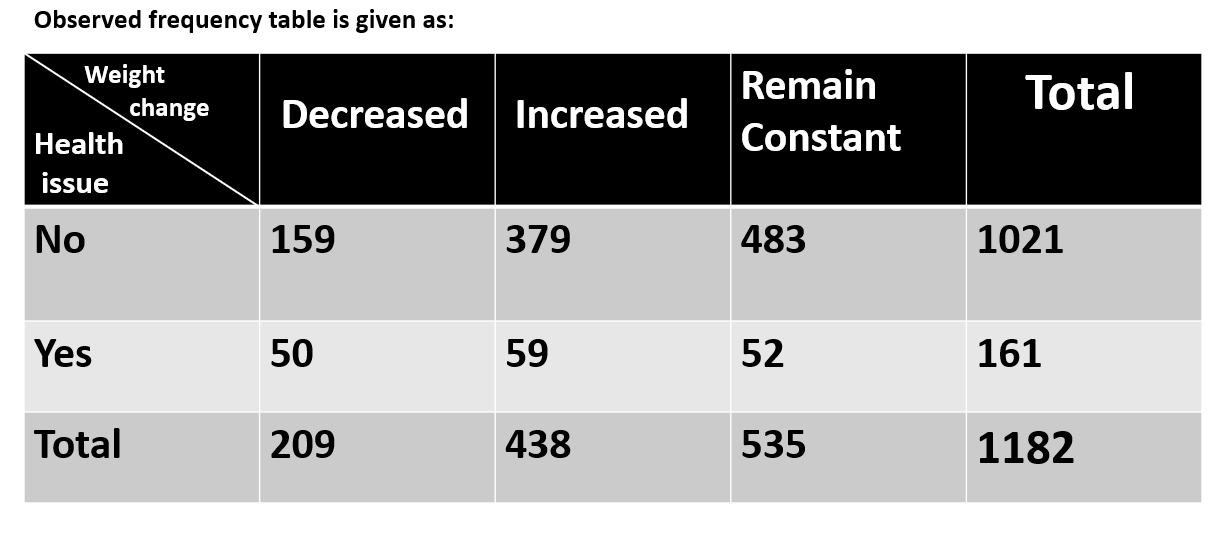
Fisher’s Exact Test uses the following null and alternative hypotheses:

* **H0: (null hypothesis)**The two variables are independent.
* **H1: (alternative hypothesis)**The two variables are *not* independent.

The one-tailed p value for Fisher’s Exact Test is calculated as:

**p = (a+b)! (c+d)! (a+c)! (b+d)! / (a!b!c!d!n!)**

**1)We want to know there is an association between health issue and change in weight during lockdown due to Covid-19.**



**Ho: Health issue and change in weight during lockdown are independent.**

**v/s**

**H1: Health issue and change in weight during lockdown are not independent.**

**Output Table:**

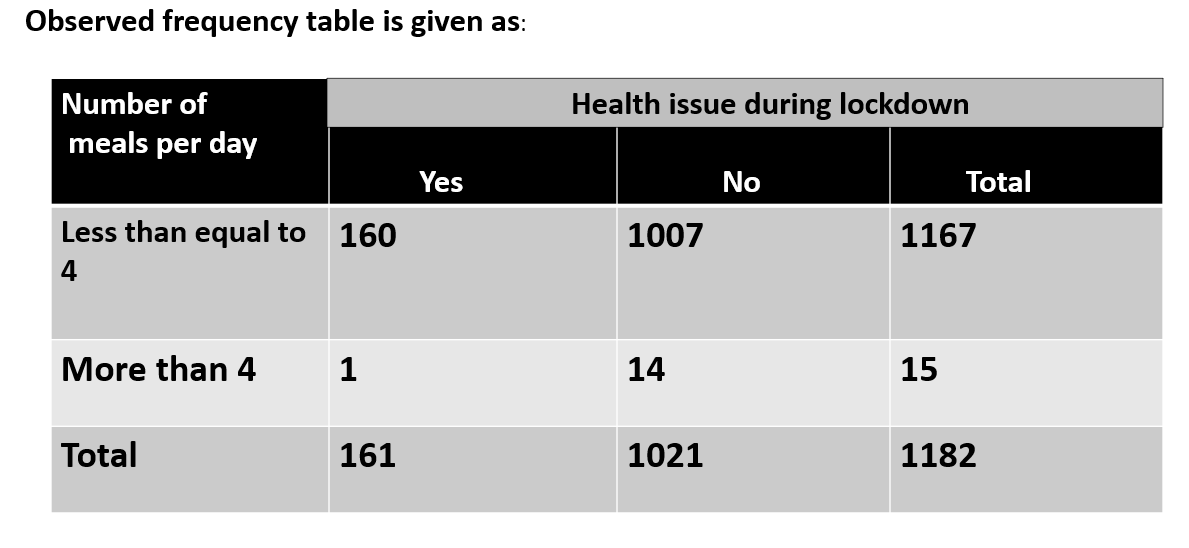
|  |  |
| --- | --- |
| Calculated Chi-squared | 25.784 |
| Df | 2 |
| p-value | 2.518e-06 |
| Fisher’s p-value | 6.95e-06 |

**Conclusion:**

p-value is less than 0.05, we reject Ho.

**Hence, there is an association between health issue during lockdown and change in weight.**

**2) We want to know there is an association between Number of meals per day and health issue during lockdown due to Covid-19.**



**Ho: Number of meals per day and Health issue during lockdown are**

**independent.**

**v/s**

**H1: Number of meals per day and Health issue during lockdown are**

**not independent.**

**R code :**

# Number of meals v/s health issue

n=matrix(data=c(160,1007,1,14), nrow=2,ncol=2,byrow=T)

chisq.test(n)

fisher.test(n)

|  |  |
| --- | --- |
| Calculated Chi-squared | 0.16931 |
| Df | 1 |
| p-value | 0.6807 |
| Fisher’s p-value | 0.7078 |

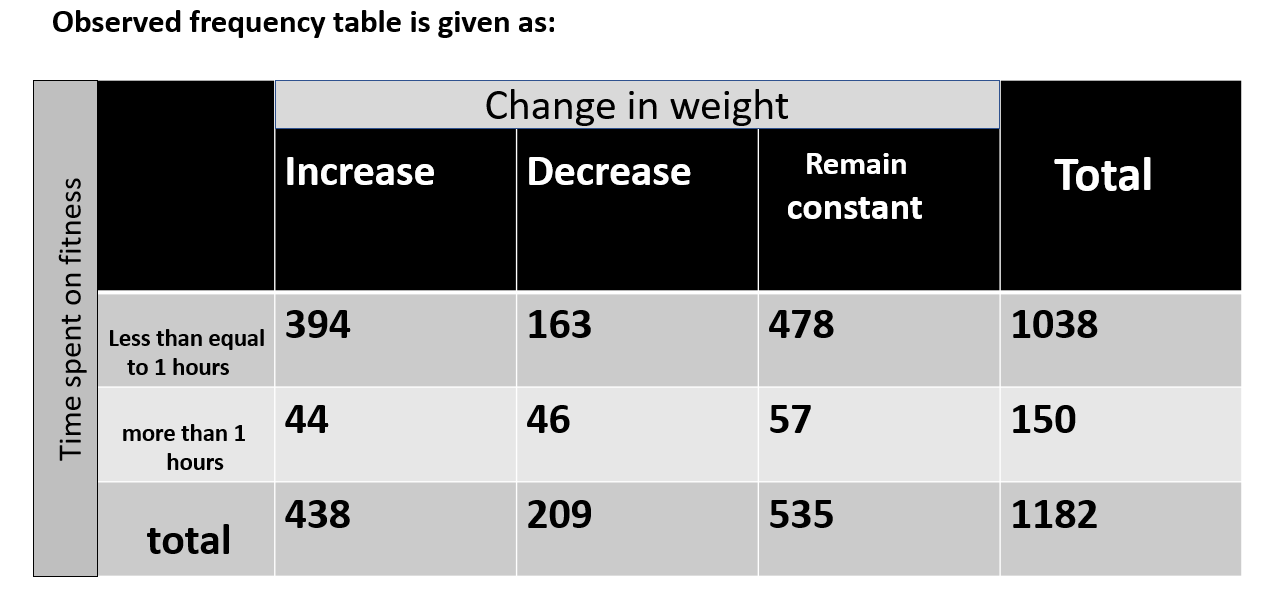
**Output Table:**

**Conclusion:**

Two tailed P-Value is greater than 0.05, Since we do not reject the null hypothesis**.**

**We do not have sufficient evidence to say that there is any statistically significant association between number of meals per day and health issue during lockdown.**

**3) We want to know there is an association between time spent on fitness and change in weight during lockdown due to Covid-19.**

****

**Ho: Time spent on fitness and change in weight during lockdown are independent.**

**v/s**

**H1: Time spent on fitness and change in weight during lockdown are not independent.**

**R code:**

**# Time spent on fitness v/s change in weight**

**p=matrix(data=c(394,163,478,44,46,57),nrow=2,ncol = 3,byrow=T)**

**chisq.test(p)**

**fisher.test(p)**

**Output Table:**

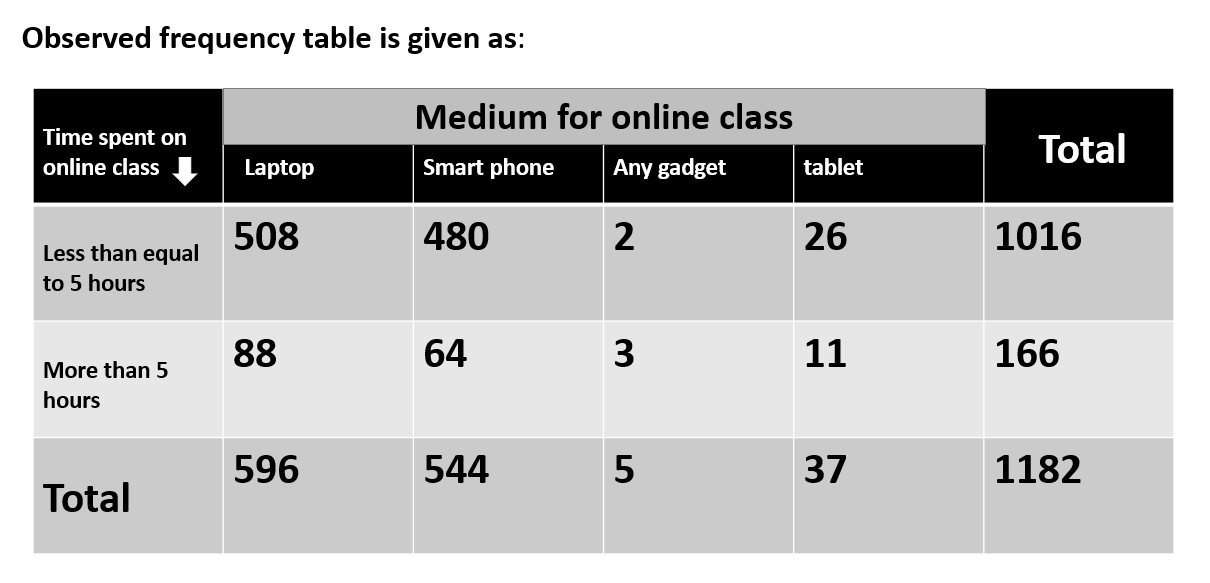
|  |  |
| --- | --- |
| Calculated Chi-squared | 21.448 |
| Df | 2 |
| p-value | 2.201e-05 |
| Fisher’s p-value | 7.174e-05 |

**Conclusion:**

Two tailed P-Value is less than 0.05, Since we reject the null hypothesis.

**Hence, there is an association between Time spent on fitness and change in weight during lockdown**

1. **We want to know there is an association between Time spent on online class and Medium used for online class.**

**Ho: Time spent on online class and Medium using for online class are independent.**

**v/s**

**H1: Time spent on online class and Medium using for online class are not independent.**

**R code:**

**#time spent on online class v/s medium use for online class**

**t=matrix(data=c(508,480,2,26,88,64,3,11),nrow=2,ncol=4,byrow=T)**

**chisq.test(t)**

**fisher.test(t)**

**Output Table:**

|  |  |
| --- | --- |
| Calculated Chi-squared | 18.887 |
| Df | 3 |
| p-value | 0.0002885 |
| Fisher’s p-value | 0.0009477 |

**Decision Criteria:**

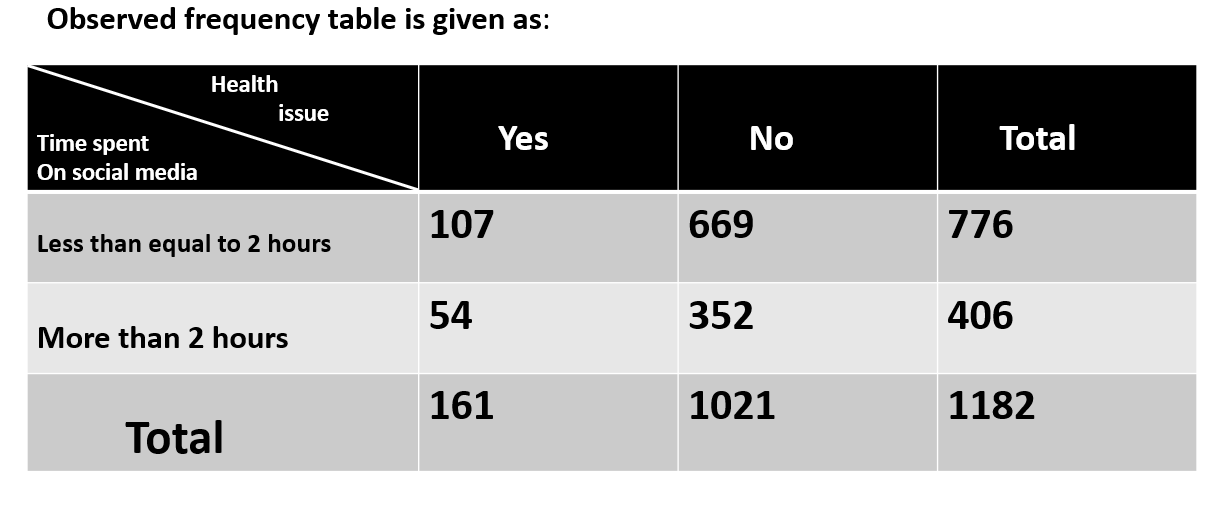
**Reject 𝐻0, if p-value < 0.05**

**Conclusion:**

p-value is less than 0.05, Since we reject the null hypothesis.

**Hence, Time spent on online class and Medium used for online class are not independent.**

**5) We want to know there is an association between time spent on social media and Health issue during lockdown.**



**Ho: Time spent on social media and Health issue during lockdown are independent.**

**v/s**

**H1: Time spent on social media and Health issue during lockdown are not independent.**

**R code:**

**#Time spent on social media v/s health issue**

**b=matrix(data=c(107,669,54,352),nrow=2,ncol=2,byrow=T)**

**chisq.test(b)**

**fisher.test(b)**

**Output Table:**

|  |  |
| --- | --- |
| Calculated Chi-squared | 0.020468 |
| Df | 1 |
| p-value | 0.8862 |
| Fisher’s p-value | 0.8585 |

**Decision Criteria:**

**Reject 𝐻0, if p-value < 0.05**

**Conclusion:**

Two tailed P-Value is greater than 0.05, Since we don’t reject the null hypothesis**.**

**We conclude that, there is no association between Time spent on social media and health issue during lockdown.**

**Objective 2:**

**To analyse whether there was effect on student’s health due to online mode of study during Covid -19 lockdown.**

# Technique used: Binary Logistic Regression, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) Using machine learning.

Tools used: Python

**Dependent Variable**:

Health Issues during lockdown: 1= Yes

0= No

**Independent Variables:**

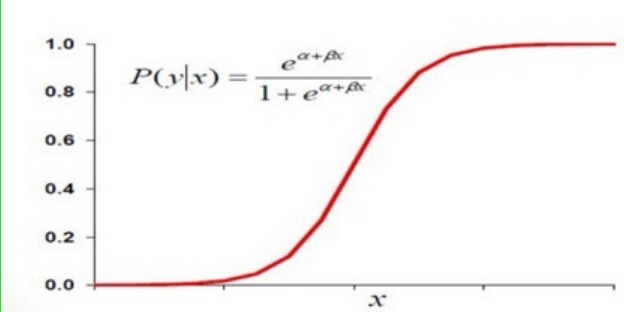
X1: Age of subject (quantitative)  
X2: Time spent on Online Class (quantitative)  
X3: Medium for online class (qualitative)  
X4: Time spent on fitness (quantitative)   
X5: Time spent on sleep (quantitative)  
X6: Time spent on social media (quantitative)  
X7: Time spent on TV (quantitative)   
X8: Number of meals per day (quantitative)  
X9: Change in your weight (qualitative)

**Analysis:**

* Steps in Building a model like Data cleaning, EDA, feature scaling, handling class imbalance problems, training, prediction, and evaluation of model on the test dataset.
* In Model Fitting the we are splitting our data set into train and test split. The whole data set generally split into 80% train and 20% test data set (general rule of thumb).

1)**Model 1: Binary Logistic Regression**

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* Logistic Regression is a classification algorithm which is used when we want to predict a categorical variable (Yes/No, Pass/Fail) based on a set of independent variable(s). Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



**Logistic Function (Sigmoid Function):**

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.

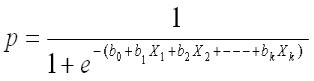
The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

We used binary logistic regression since our dependent variable is categorical.

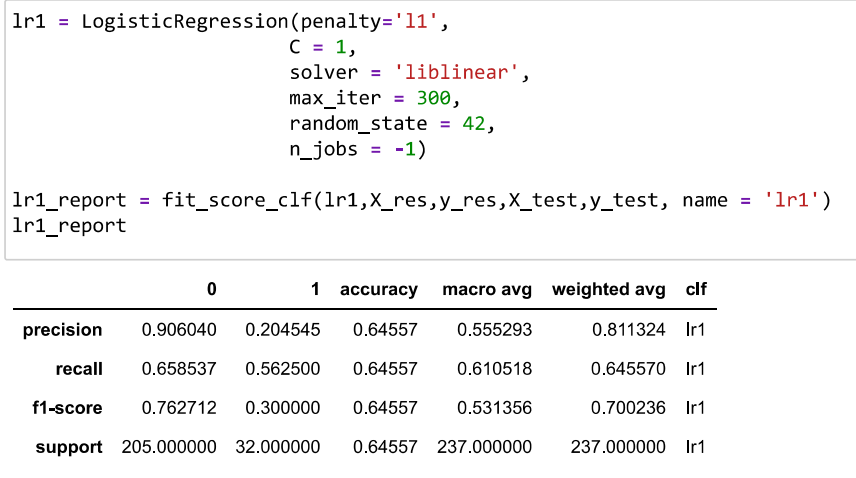
Logistic Regression Equation

***p:***

*the probability of the dependent variable equaling a "success" or "event".*

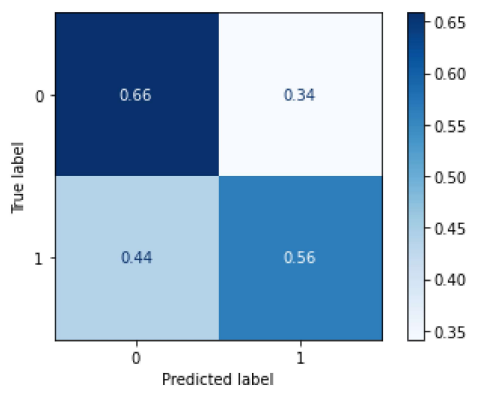


**Output:**



**Confusion Matrix :**

It is a tabular representation of Observed vs Predicted values. It helps to quantify the efficiency (or accuracy) of the model.

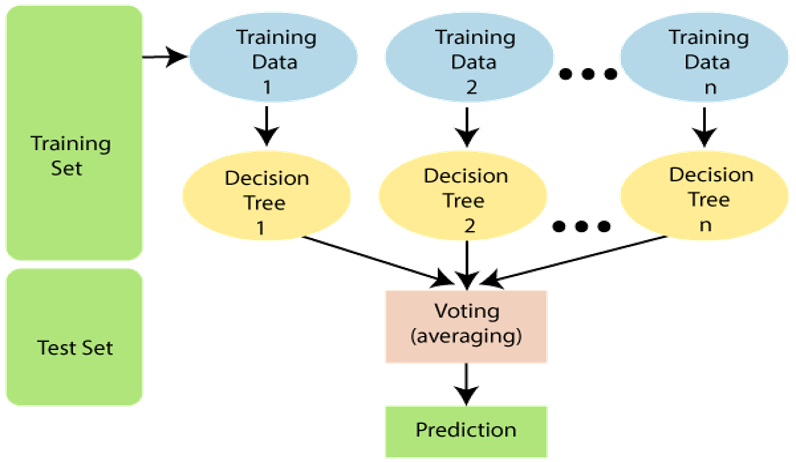
****

**Here it is observed that, 66% of the students who don't have issues are correctly predicted as well 56% students who have health issues are correctly predicted.**

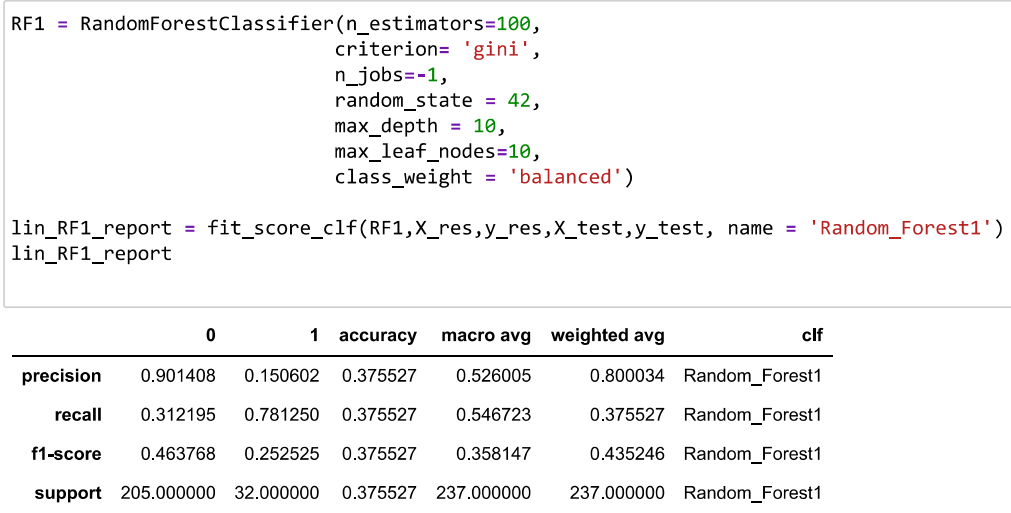
**2) Model 2: Random Forest**

* Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.
* It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

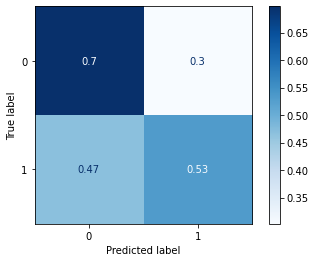
The greater number of trees in the forest leads to higher accuracy.

****

**Output:**

****

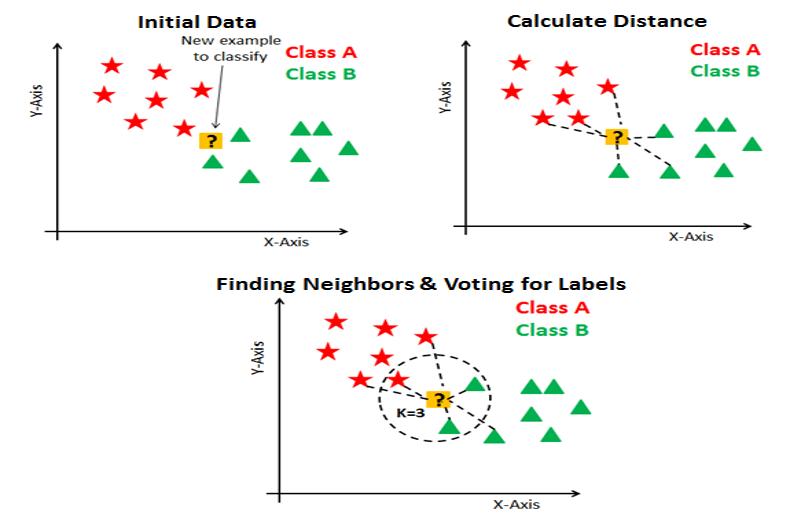
**Confusion Matrix:**

****

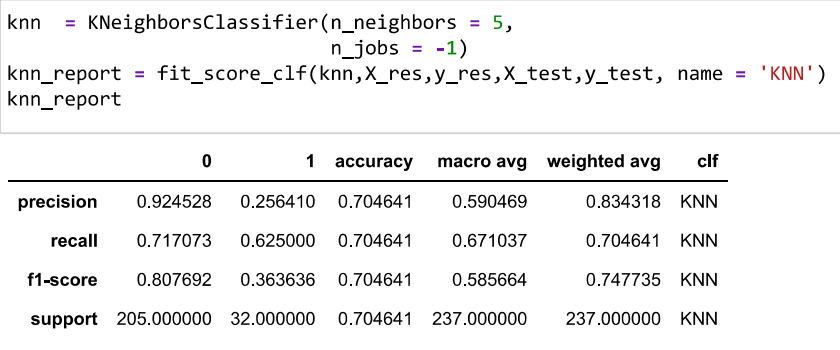
**Here it is observed that, 70% of the students who don't have issues are correctly predicted as well 53% students who have health issues are correctly predicted.**

**3)Model 3: K-Nearest Neighbor(KNN)**

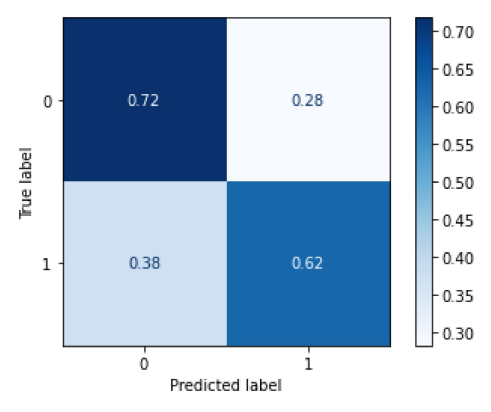
* It is one of the simplest and widely used classification algorithms in which a new data point is classified based on similarity in the specific group of neighboring data points.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.



**Output:**

****

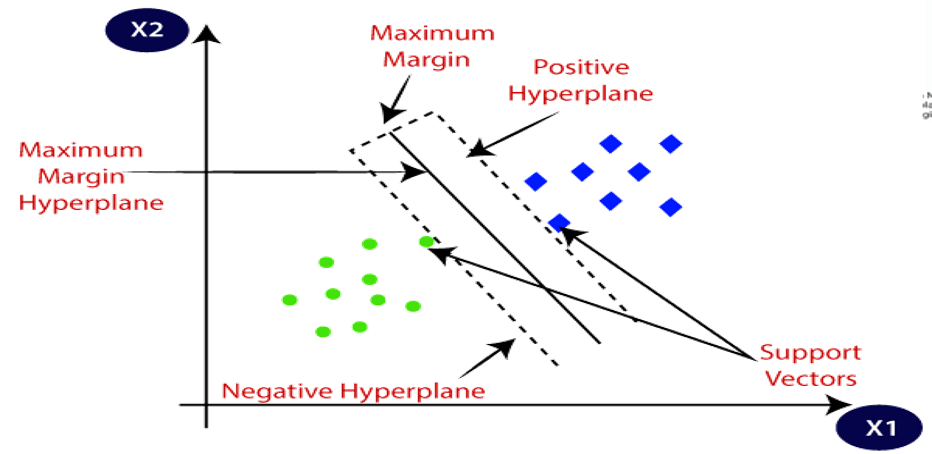
**Confusion Matrix:**

****

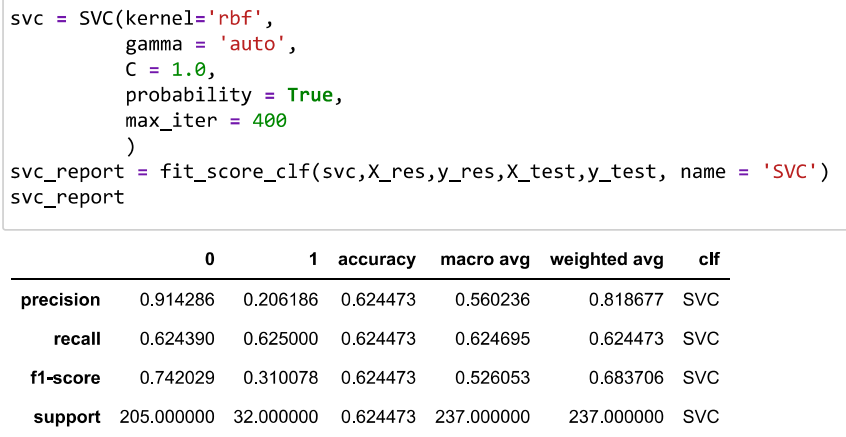
**Here it is observed that, 72% of the students who don't have issues are correctly predicted as well 62% students who have health issues are correctly predicted.**

**4)Model 4: Support Vector Machine (SVM)**

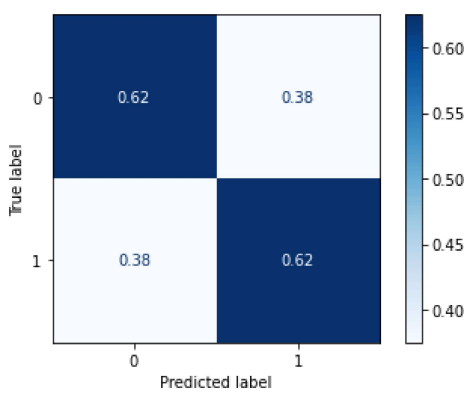
* The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
* SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.
* Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



**Output:**

****

**Confusion Matrix:**

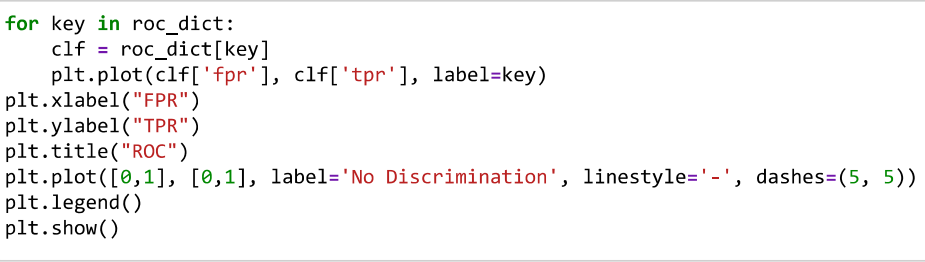
****

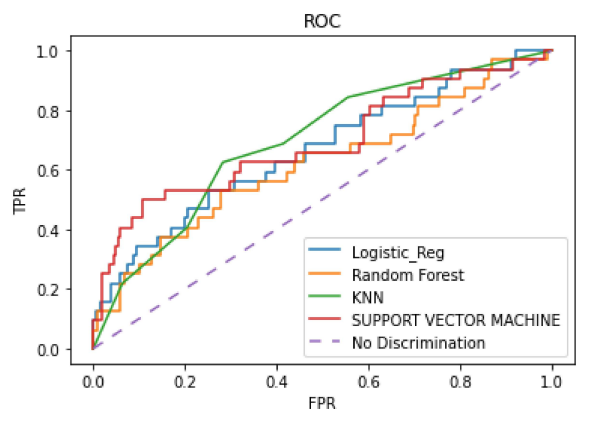
**Here it is observed that, 62% of the students who don't have issues are correctly predicted as well 62% students who have health issues are correctly predicted.**

**ROC Curve**

ROC (Receiver Operating Characteristic) Curve which displays the percentage of true positives predicted by the model as the prediction probability cutoff is lowered from 1 to 0.

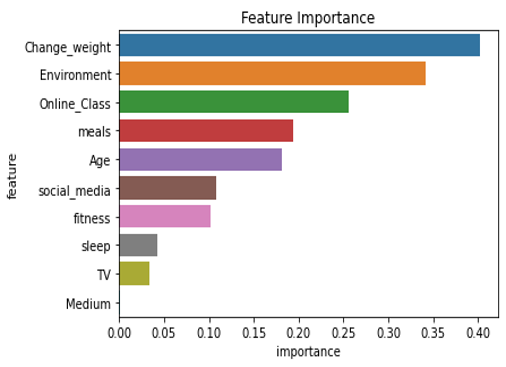
It explains the model’s performance by evaluating Sensitivity vs Specificity.

****

****

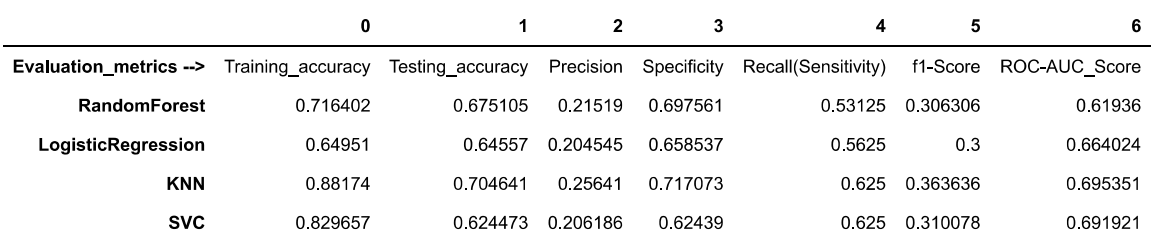
* **The more that a ROC curve hugs the top left corner of the plot, the better the model does at classifying the data into categories.**
* **To quantify this, we can calculate the AUC – area under the curve – which tells us how much of the plot is located under the curve.**
* **The closer AUC is to 1, the better the model**.

**Feature Importance:**

****

**Summary:**

summary = pd.DataFrame(df)



From our plot we can see the following AUC metrics for each model:

ROC\_AUC \_Score of Logistic Regression model: 0.6640.

ROC\_AUC \_Score of Random Forest model:0.61936.

ROC\_AUC \_Score of KNN model: 0.70.

ROC\_AUC \_Score of SVM model: 0.6919.

**Clearly the K-Nearest Neighbor(KNN)does a better job of classifying the data into categories compared to the Logistic regression model and Random Forest, Support Vector Machine (SVM).**

**Objective 3:**

To analyse the time spent by students on their online class and the other daily activities performed by them including their age factor.

* **Assumption Checking**:

Normality of Data using Shapiro Wilk Test (using SPSS software)

H0: The data is normally distributed.

H1: The data is not normally distributed.

* **Output**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | |
|  | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| Online class | 0.244 | 1182 | <.001 | 0.835 | 1182 | <.001 |

Since the p-value is 0.0010 which is < 0.05, we reject null hypothesis.

Thus the data is not normally distributed. Hence, instead of ANOVA we perform nonparametric **Kruskal-Wallis Test**.

**Dependent variable:**  Time spent on Online Class (quantitative)

**Independent Variable:**

X1: Age of subject (quantitative)

X2: Time spent on Self study(quantitative)

X3: Time spent on fitness (quantitative)

X4: Time spent on Social media(quantitative)

X5: Time spent on sleep (quantitative)

X6: Time spent on TV (quantitative)

X7: Number of meals per day (quantitative)

**Hypothesis Defined:**

H0: There is no significant difference between the time spent on online class and the time spent on other daily activities and the age factor.

H1: There is significant difference between the time spent on online class and the time spent on other daily activities and the age factor.

**Output:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Statisticsa,b** | | | | | | | |
|  | Age | self\_study | fitness | sleep | social\_media | TV | meals |
| Kruskal-Wallis H | 84.921 | 41.576 | 10.605 | 45.423 | 35.174 | 7.083 | 6.073 |
| df | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| Asymp. Sig. | <.001 | <.001 | 0.389 | <.001 | <.001 | 0.718 | 0.809 |
| a. Kruskal Wallis Test | | | | | | | |

We Reject H0, if p-value < 0.05 else accept H0.

So It is observed that p-value for factors Age and the time spent on Self study, Sleep and Social media is < 0.05 and for remaing factors the time spent on fitness, TV and the no. of meals consumed per day is > 0.05

**Conclusion:**

* There is an adverse affect of time spent on online class and the time spent on Self study, Sleep, Social media and age factor as well.
* There is no affect of time spent on online class and the time spent on their fitness, TV and the no. of meals consumed per day.

**Suggestion:**

**For study:**

To overcome such health issues and education issues government should find some scheme so that they can arrange or provide mini laptops/mini tabs to those students who are from less privileged background and have experienced larger negative impacts due to limited access to digital resources and high cost of internet connectivity so in future if such condition of online education arises then there academic life should not be disrupted.

**For health:**

One of the most important health issues was Stress due to sudden change in daily routine and adaptation of fully online learning so to overcome such issues universities or colleges or schools should organize some supportive and motivational sessions to motivate the students and they should organize some activities every week so that students will be active.

**APPENDIX**

Python codes**:**

**Python libraries:**

import numpy as np # linear algebra

import pandas as pd #data processing(e.g.pd.read\_csv)

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

from sklearn.preprocessing import LabelEncoder

from sklearn.model\_selection import train\_test\_split

**Data cleaning:**

data = pd.read\_csv('/Data/covid 19 data')

print(data.isnull().sum())#To find out the missing values

#Finding missing values use MODE (measure of central tendency)

#To plot box plot

fig = px.histogram(Numeric\_col,

x = column,

marginal = 'box',

color\_discrete\_sequence = ['green'],

nbins = 47,

title = 'Histogram and Boxplot for '+ column)

fig.update\_layout(bargap = 0.1)

fig.show()

#To remove outliers using IQR technique

**#Correlation matrix:**

plt.figure(figsize=(10, 8))

sns.heatmap(Numeric\_col.corr(), annot=True, vmax=1, vmin=-1, square=True, cmap='BrBG', mask=mask);

plot.show()

R Codes

**# To check the association between different factor (Using R codes):**

**# Health issue V/S change in weight**

e=matrix(data=c(159,379,483,50,59,52),nrow=2,ncol = 3,byrow = T)

chisq.test(e)

fisher.test(e)

**#number of meals v/s health issue**

n=matrix(data=c(160,1007,1,14),nrow=2,ncol=2,byrow=T)

chisq.test(n)

fisher.test(n)

**#time spent on fitness v/s change in weight**

p=matrix(data=c(394,163,478,44,46,57),nrow=2,ncol = 3,byrow=T)

chisq.test(p)

fisher.test(p)

**#time spent on online class v/s medium use for online class**

t=matrix(data=c(508,480,2,26,88,64,3,11),nrow=2,ncol=4,byrow=T)

chisq.test(t)

fisher.test(t)

**#Time spent on social media v/s health issue**

b=matrix(data=c(107,669,54,352),nrow=2,ncol=2,byrow=T)

chisq.test(b)

fisher.test(b)

**# Based on machine learning we used various classifiers:**

#Data Encoding

enc = LabelEncoder()

Cat\_col = cat\_col.apply(LabelEncoder().fit\_transform)

new\_data = pd.concat([num\_col,Cat\_col],axis =1)

#Traning and Testing data

X\_train, X\_test = feature\_col.iloc[train\_index], feature\_col.iloc[test\_index]

y\_train, y\_test = target\_col.iloc[train\_index], target\_col.iloc[test\_index]

#Classification Report:

def fit\_score\_clf(clf,X\_train,y\_train,X\_test,y\_test, name = 'clf'):

clf.fit(X\_train,y\_train)

y\_pred = clf.predict(X\_test)

clf\_report = pd.DataFrame(classification\_report(y\_test, y\_pred, output\_dict=True))

clf\_report['clf'] = name

return clf\_report

#Confusion\_matrix(clf):

y\_pred = clf.predict(X\_test)

plot\_confusion\_matrix(clf, X\_test , y\_test , normalize='true', cmap = 'Blues')

##To plot ROC curve:

ROC dict for FPR and TPR (FPR = False Positive Rate & TPR = True Positive Rate)

clf = roc\_dict[key]

plt.plot(clf['fpr'], clf['tpr'], label=key)

plt.xlabel("FPR")

plt.ylabel("TPR")

plt.title("ROC")

plt.plot([0,1], [0,1], label='No Discrimination', linestyle='-', dashes=(5, 5))

plt.legend()

plt.show()

#Feature Importance:

plt.title('Feature Importance')

sns.barplot(data=importance\_df.head(25), x='importance', y='feature');

#Summary of Data:

summary = pd.DataFrame(data1)

**# For Kruskal Wallis Test:**