ONLINE EDUCATION IN TIMES OF COVID-19

“Education is the most powerful weapon which you can use to change the world.” ~ Nelson Mandela

ACKNOWLEDGEMENT.

1. We would like to express our heartfelt thanks and gratitude to all, who out of concern had come forward in completing our project.

2. We would like to express our special thanks to the management and college for providing us such a good platform like Microsoft teams during these difficult times which truly turned out to be a quite helpful source for us group members to contact amongst ourselves.

2. We would like to express our special word of thanks to our coordinator and mentor Mr. Chaitanya Alshi who has helped us, guided us, and being there to give us the right direction to conduct and complete our project with the right insights.

3.We thank Mr. Amrit Rajwadkar (Head of Statistics Department) for his timely guidance and support till the completion of our project work.

4. We would also like to thank our teachers Mrs. Manjiri Vartak, Mr. Pratik Patil for always guiding and helping us in our analysis.

5. And a huge thank you to everyone who has helped us in every way possible right from the start of our project by means of pitching in their ideas, collection of data to the completion of our project.

TABLE OF CONTENT:

INTRODUCTION :-

"Change is the end result of all true learning." But what happens when learning in itself goes through a sudden change?

Today, our education system has gone through a paradigm shift not just in terms of resources and content, but also in terms of mode. E-books replaced the hardcovers, Google Classrooms replaced project files, tablet screens replaced whiteboards, and classroom interaction replaced by virtual meetings.

Today, the COVID-19 crisis has disrupted many sectors, including the educational area. Among many casualties of COVID-19, there is also demise of the traditional classrooms. While online courses have existed even before the pandemic, they served a different purpose to learners, like granting them accessibility to modules in case they do not have means or supplementing what individuals were taught in physical classrooms. Online mode was never meant to be the fundamental mode of learning back then.

Social distancing guidelines and lockdowns, however, have made online classrooms the primary source of educational instruction for students of all ages. Sugarcoating the ongoing situation doesn't help or make it any easier for us: online learning is just hard; Period. Education is meant to be passed through human connection. The very first form of education was storytelling: people all around the world have told stories as a way of passing down knowledge, history, beliefs, and traditions. Stories are at the heart of all that makes us human. While the platform of online learning is incredibly helpful, it takes away what we as humans are made for social thinking, positive relationships, and authentic connection.

Just as human beings have a basic need for food and shelter, we also have a fundamental need to belong to a group and form relationships. The desire to be in a loving relationship, to fit in at school, to avoid rejection, to be well-liked, to have fun with friends, to get along with family, and to check in on social media — these things motivate most of our thoughts, actions, and feelings. Life is complicated, humans are complex, and it is no small feat to go from traditional education to online education.

We are not against Online Classes the problem is the way it is going on. Let's rise up and think about this issue because it just becoming a waste of time and reason for more & more stress. If a person gets Graduated through Online Classes and gets a job without some important knowledge then we are just wasting our human capital.

OBJECTIVE'S :

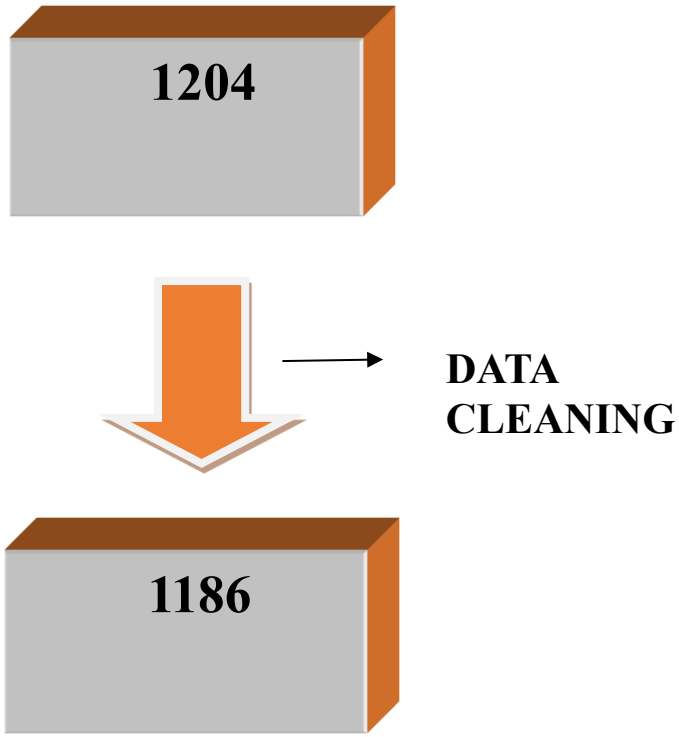
Sr no	Objective	Technique
1	To determine prime underlying factors affecting the overall outcome of online education.	Factor Analysis
2	To determine major variable which need to be given proper attention to boost online education.	Pareto Analysis
3	To analyze people's opinion on online education.	Sentiment and Text Analysis
4	To understand characteristics of responders who will continue with online education if given a chance in future based on various demographics.	Decision Tree
5	To understand which of the demographics affects the effectiveness of online education.	Naive Bayes algorithm.
6	To understand which of the demographics impacts responder’s decision to choose - online, blended or offline education.	Naive Bayes algorithm.
7	To understand Strengths, Weakness, Opportunities and Challenges for online education.	SWOC analysis
8	To check whether online education has affected one's intellectual growth in terms of self-study time.	Paired T test
9	To check which method of education is better in terms of doubt/query resolution, syllabus coverage and concept clearance.	Paired T test

STEPS INVOLVED IN CONDUCTING THE SURVEY

- Step 1: Defining our objectives.
- Step 2: Specifying information needs.
- Step 3: Designing questionnaires.
- Step 4: Pilot Survey.
- Step 5: Modifying Questionnaires.
- Step 6: Actual Survey.
- Step 7: Data Cleaning.
- Step 8: Coding of Data.
- Step 9: Analysis of Data.
- Step 10: Interpretation of Data.

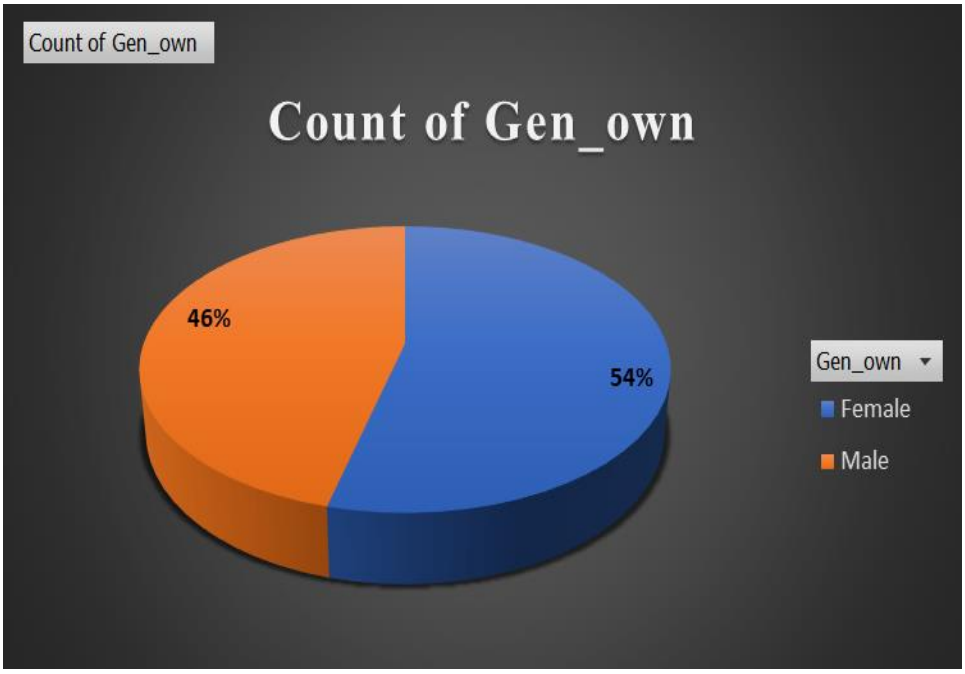
Data Collection:

- *The entire data has been collected through online mode.*

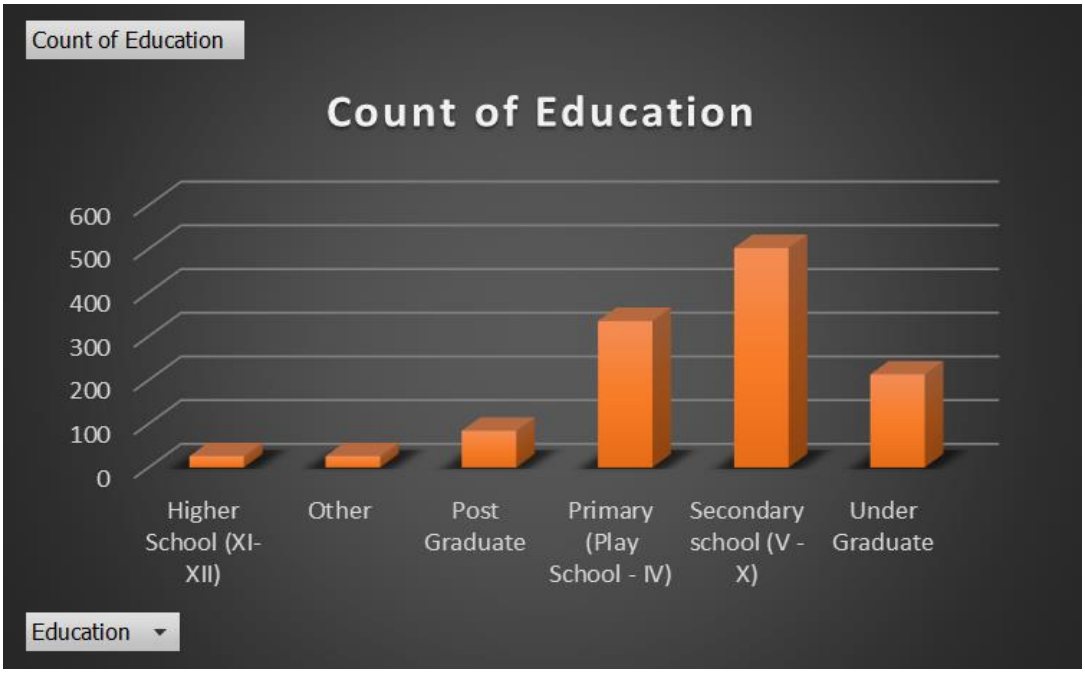


Graphical Representation:

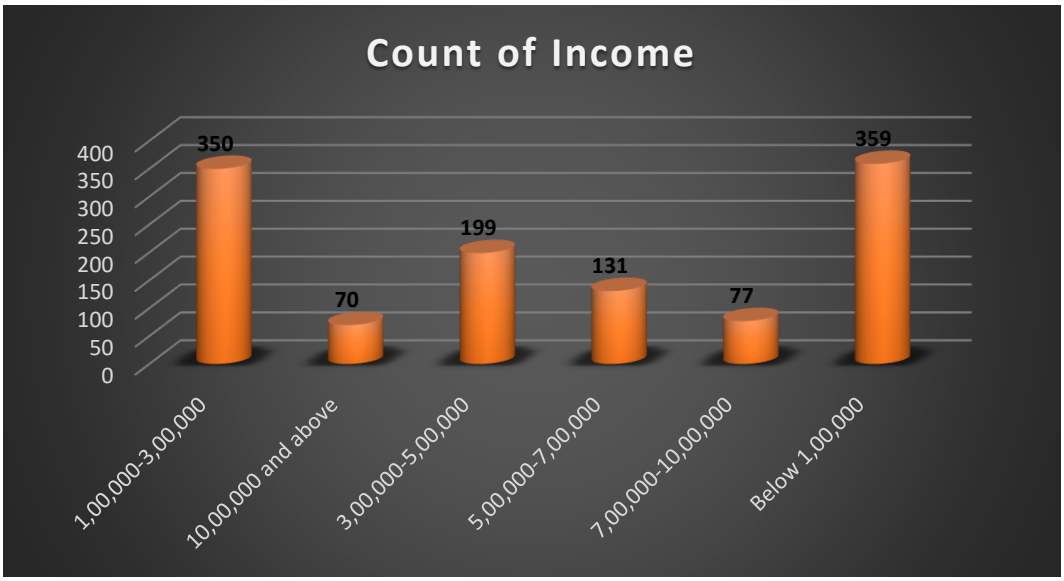
1.Gender



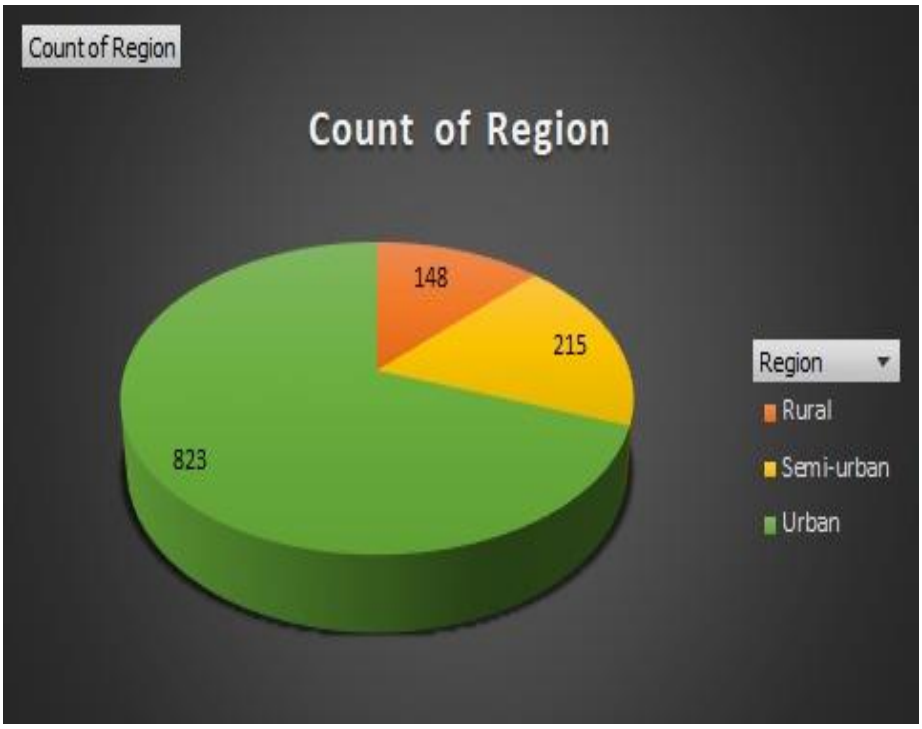
2.Education



3.Income



4.Region



FACTOR ANALYSIS

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. Factor analysis is part of general linear model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no perfect multicollinearity, it includes relevant variables into analysis and there is true correlation between variables and factors. Several methods are available, but principle component analysis is used most commonly.

Key concepts and terms:

Principal component analysis: This is the most common method used by researchers. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

Correlation matrix: A correlation matrix is a lower triangle showing the sample correlations, 'r ' between all possible pairs of variables included in the analysis.

Communality: Communality is the amount of variance a variable shares with all the other variables being considered. This is also the proportion of variance explained by common factors.

Factor loading: Factor loading is basically the correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that particular factor. In the SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

Eigen-values: Eigen values are also called characteristic roots. Eigen values shows variance explained by that particular factor out of the total variance. From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For example, if our first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

Factor score: The factor score is also called the component score. This score is of all row and columns, which can be used as an index of all variables and can be used for further analysis. We can standardize this score by multiplying a common term. With this factor score, whatever analysis we will do, we will assume that all variables will behave as factor scores and will move.

Criteria for determining the number of factors: According to the Kaiser Criterion, Eigen values are good criteria for determining a factor. If Eigen values is greater than one, we should consider that a factor and if Eigen values is less than one, then we should not consider that a factor. According to the variance extraction rule, it should be more than 0.6. If variance is less than 0.6, then we should not consider that a factor.

Rotation method: Rotation method makes it more reliable to understand the output. Eigen values do not affect the rotation method, but the rotation method affects the Eigen values or percentage of variance extracted.

There are a number of rotation methods available:
(1) Varimax rotation method
(2) Quartimax rotation method
(3) Equamax rotation method
(4) Direct oblimin rotation method
(5) Promax rotation method.

Each of these can be easily selected in SPSS, and we can compare our variance explained by those particular methods. Direct Oblimin Method: A method for oblique (non-orthogonal) rotation. When delta equals 0 (the default), solutions are most oblique. As delta becomes more negative, the factors become less oblique. To override the default delta of 0, enter a number less than or equal to 0.8.

Assumptions:
1. No outlier: Assume that there are no outliers in data.
2. Adequate sample size: The case must be greater than the factor.
3. Homoscedasticity: Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.
4.Linearity: Factor analysis is also based on linearity assumption. Nonlinear variables can also be used. After transfer, however, it changes into linear variable.
5. Interval Data: Interval data are assumed.

Output:
Variables used in Factor Analysis:
L1: Online lecture materials provided to you are helpful.
L2: Social media is a distraction while attending online lectures.
L3: Is language a barrier while attending online lectures
L4: Feeling hesitate to ask questions/doubts in online lectures.
L5 : Lack of direct contact with other students/colleague/friends/teachers.
L6: Insufficient practical knowledge through online education.
L7: Indiscipline and disturbance from outsiders in online lecture.
L8 : Technical glitches during online lectures affects your attendance and concentration.
L9 : Home environment is suitable for participating online lectures.
L10: Feeling frustrated when unable to submit your assignments or paper on time.
L11: Excessive use of mobiles or laptops led to eye pain, ear pain, headache.
L12: Sitting continuously for online lectures led to back pain and neck pain.
L13: Piled up assignments and too many upcoming tests depresses you.
L14: Network issues faced while writing online exams stresses you which directly affects your score.
L15: Does power cut affect online education.
L16: Does limited availability of data restrict online education.
L17: Does ones financial background restricts online education.
L18: Online education has made todays generation tech-savvy.
L19: Online lectures are flexible and comfortable to attend.
L20: Reduced travelling time and money spent on travelling expenses.
L21: Online education is giving you more time for other activities.
L22: In online education notes are available at just one click.
L23: Online lectures are accessible at any time because of recording feature.
L24: Getting proper meals at proper time has improved ones physical health.
L25: Time management and increased self motivation.
L26: Online education was perfect alternative in covid time instead of waste of academic year.
Before proceeding with factor analysis on the variables we need to check whether factor analysis is appropriate for our data and are the variables correlated with each other which are the basic assumption for factor analysis.

KMO and Bartlett's test of sphericity:
The Kaiser-Meyer-Olkin measure of Sampling Adequacy is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors. High values (close to 1 .0) generally indicate that a factor analysis may be useful with your data. If the value is less than 0.50, the results of the factor analysis probably won't be very useful. Bartlett's test of sphericity tests the hypothesis that your correlation matrix is an identity matrix, which would indicate that your variables are: unrelated and therefore unsuitable for structure detection. Small values (less than 0.05) of the significance level indicate that a factor analysis may be useful with your data.

Kaiser-Meyer-Olkin factor adequacy
Overall MSA = 0.91
MSA for each item =

L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
0.91	0.94	0.80	0.90	0.91	0.92	0.93	0.92	0.90	0.94	0.89	0.88	0.95

L14	L15	L16	L17	L18	L19	L20	L21	L22	L23	L24	L25	L26
0.94	0.91	0.93	0.93	0.93	0.91	0.92	0.91	0.91	0.89	0.89	0.87	0.92

The overall KMO for data is 0.91 , which is acceptable and this suggest that data is appropriate for factor analysis. i.e. variables and sample size are enough to proceed for factor analysis.

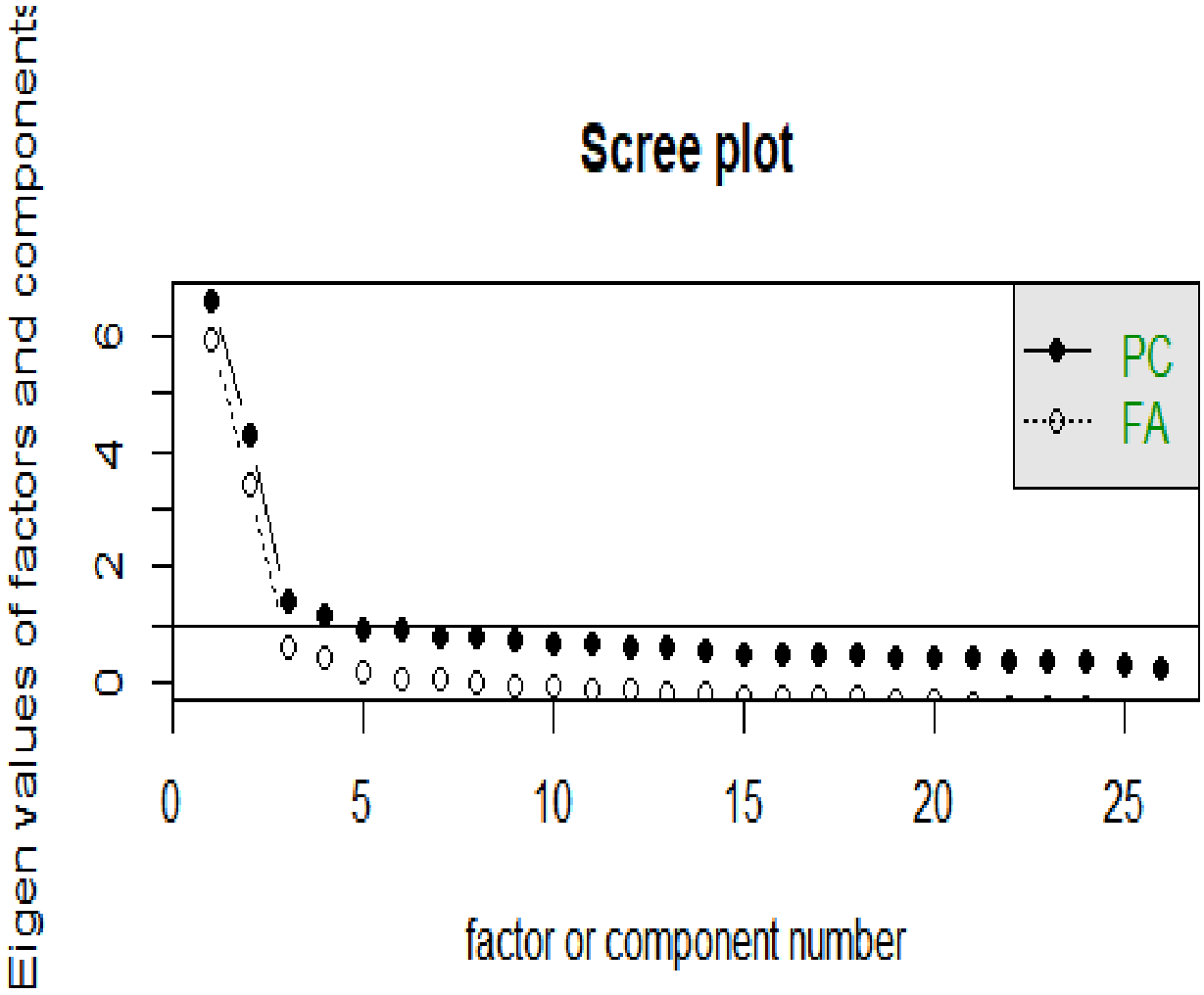
Bartlett's test for sphericity test the null hypothesis that the correlation matrix is an identity matrix.
H_0: correlation matrix is an identity matrix.
H_1: correlation matrix is not an identity matrix.
$chisq
[1] 11460.5
$p.value
[1] 0
$df
[1] 325
Bartlett test is statistically significant, suggesting that correlation matrix is different from identity matrix. There is enough correlation between variable to proceed for factor analysis.

Correlation Matrix

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22	L23	L24	L25	L26
L1	1.00	0.10	0.06	0.01	-0.07	-0.10	-0.07	0.02	0.36	0.02	-0.03	-0.06	-0.06	-0.02	0.05	0.04	0.02	0.28	0.39	0.25	0.34	0.38	0.28	0.37	0.39	0.32
L2	0.10	1.00	0.18	0.25	0.27	0.30	0.28	0.32	0.00	0.29	0.32	0.32	0.25	0.25	0.20	0.26	0.24	0.23	0.11	0.15	0.10	0.10	0.15	0.10	0.05	0.13
L3	0.06	0.18	1.00	0.37	0.15	0.15	0.26	0.12	0.09	0.18	0.13	0.11	0.26	0.17	0.28	0.22	0.27	0.12	0.07	0.04	0.05	0.05	0.16	0.06	0.15	-0.02
L4	0.01	0.25	0.37	1.00	0.36	0.33	0.34	0.28	0.04	0.29	0.25	0.28	0.32	0.27	0.27	0.28	0.27	0.10	0.03	0.10	0.01	0.05	0.13	0.00	0.02	0.05
L5	-0.07	0.27	0.15	0.36	1.00	0.53	0.35	0.39	-0.02	0.37	0.41	0.37	0.33	0.36	0.24	0.26	0.24	0.22	0.01	0.20	0.03	0.08	0.04	0.07	-0.01	0.16
L6	-0.10	0.30	0.15	0.33	0.53	1.00	0.43	0.46	-0.02	0.41	0.41	0.40	0.37	0.39	0.28	0.34	0.30	0.19	-0.04	0.14	0.01	0.00	0.06	0.04	-0.07	0.13
L7	-0.07	0.28	0.26	0.34	0.35	0.43	1.00	0.48	-0.04	0.39	0.39	0.39	0.35	0.34	0.29	0.32	0.32	0.13	-0.03	0.09	0.04	0.06	0.08	0.02	-0.03	0.09
L8	0.02	0.32	0.12	0.28	0.39	0.46	0.48	1.00	0.05	0.50	0.45	0.44	0.37	0.50	0.30	0.41	0.32	0.23	0.02	0.21	0.12	0.10	0.10	0.12	0.09	0.19
L9	0.36	0.00	0.09	0.04	-0.02	-0.02	-0.04	0.05	1.00	0.05	0.02	0.01	-0.03	-0.02	0.02	0.04	0.02	0.26	0.36	0.22	0.28	0.27	0.21	0.35	0.38	0.23
L10	0.02	0.29	0.18	0.29	0.37	0.41	0.39	0.50	0.05	1.00	0.45	0.45	0.41	0.41	0.24	0.32	0.29	0.27	0.01	0.19	0.04	0.10	0.08	0.14	0.04	0.14
L11	-0.03	0.32	0.13	0.25	0.41	0.41	0.39	0.45	0.02	0.45	1.00	0.72	0.48	0.48	0.31	0.40	0.35	0.26	0.00	0.22	0.08	0.12	0.14	0.11	0.05	0.21
L12	-0.06	0.32	0.11	0.28	0.37	0.40	0.39	0.44	0.01	0.45	0.72	1.00	0.50	0.52	0.34	0.37	0.36	0.21	-0.04	0.20	0.03	0.10	0.09	0.06	0.02	0.18
L13	-0.06	0.25	0.26	0.32	0.33	0.37	0.35	0.37	-0.03	0.41	0.48	0.50	1.00	0.49	0.40	0.41	0.41	0.20	-0.02	0.16	0.04	0.08	0.18	0.05	0.00	0.08
L14	-0.02	0.25	0.17	0.27	0.36	0.39	0.34	0.50	-0.02	0.41	0.48	0.52	0.49	1.00	0.42	0.47	0.41	0.18	-0.01	0.20	0.06	0.10	0.14	0.07	0.04	0.16
L15	0.05	0.20	0.28	0.27	0.24	0.28	0.29	0.30	0.02	0.24	0.31	0.34	0.40	0.42	1.00	0.51	0.47	0.19	0.09	0.20	0.13	0.08	0.18	0.04	0.10	0.11
L16	0.04	0.26	0.22	0.28	0.26	0.34	0.32	0.41	0.04	0.32	0.40	0.37	0.41	0.47	0.51	1.00	0.50	0.26	0.11	0.23	0.13	0.09	0.22	0.10	0.09	0.13
L17	0.02	0.24	0.27	0.27	0.24	0.30	0.32	0.32	0.02	0.29	0.35	0.36	0.41	0.41	0.47	0.50	1.00	0.25	0.06	0.19	0.06	0.11	0.25	0.07	0.07	0.13
L18	0.28	0.23	0.12	0.10	0.22	0.19	0.13	0.23	0.26	0.27	0.26	0.21	0.20	0.18	0.19	0.26	0.25	1.00	0.39	0.41	0.35	0.29	0.25	0.31	0.34	0.33
L19	0.39	0.11	0.07	0.03	0.01	-0.04	-0.03	0.02	0.36	0.01	0.00	-0.04	-0.02	-0.01	0.09	0.11	0.06	0.39	1.00	0.47	0.50	0.46	0.40	0.41	0.45	0.33
L20	0.25	0.15	0.04	0.10	0.20	0.14	0.09	0.21	0.22	0.19	0.22	0.20	0.16	0.20	0.20	0.23	0.19	0.41	0.47	1.00	0.51	0.41	0.33	0.38	0.36	0.39
L21	0.34	0.10	0.05	0.01	0.03	0.01	0.04	0.12	0.28	0.04	0.08	0.03	0.04	0.06	0.13	0.13	0.06	0.35	0.50	0.51	1.00	0.46	0.41	0.48	0.54	0.35
L22	0.38	0.10	0.05	0.05	0.08	0.00	0.06	0.10	0.27	0.10	0.12	0.10	0.08	0.10	0.08	0.09	0.11	0.29	0.46	0.41	0.46	1.00	0.49	0.48	0.43	0.39
L23	0.28	0.15	0.16	0.13	0.04	0.06	0.08	0.10	0.21	0.08	0.14	0.09	0.18	0.14	0.18	0.22	0.25	0.25	0.40	0.33	0.41	0.49	1.00	0.41	0.45	0.24
L24	0.37	0.10	0.06	0.00	0.07	0.04	0.02	0.12	0.35	0.14	0.11	0.06	0.05	0.07	0.04	0.10	0.07	0.31	0.41	0.38	0.48	0.48	0.41	1.00	0.62	0.45
L25	0.39	0.05	0.15	0.02	-0.01	-0.07	-0.03	0.09	0.38	0.04	0.05	0.02	0.00	0.04	0.10	0.09	0.07	0.34	0.45	0.36	0.54	0.43	0.45	0.62	1.00	0.40
L26	0.32	0.13	-0.02	0.05	0.16	0.13	0.09	0.19	0.23	0.14	0.21	0.18	0.08	0.16	0.11	0.13	0.13	0.33	0.33	0.39	0.35	0.39	0.24	0.45	0.40	1.00

Scree Plot

To determining the number of factor to extract ,we use eigenvalue criterion and scree plot to extract important factors.



Eigen values are a measure of the amount of variance accounted for by a factor.in the above graph , eigenvalue criterion states that only factors with eigenvalue grater than 1 should be retained. Therefore 4 factors are retained for factor analysis.

Standardized loadings (pattern matrix) based upon correlation matrix: The higher the absolute value of the loading, the more the factor contributes to the variable. We suppressed all loadings less than 0.40

Component Matrix

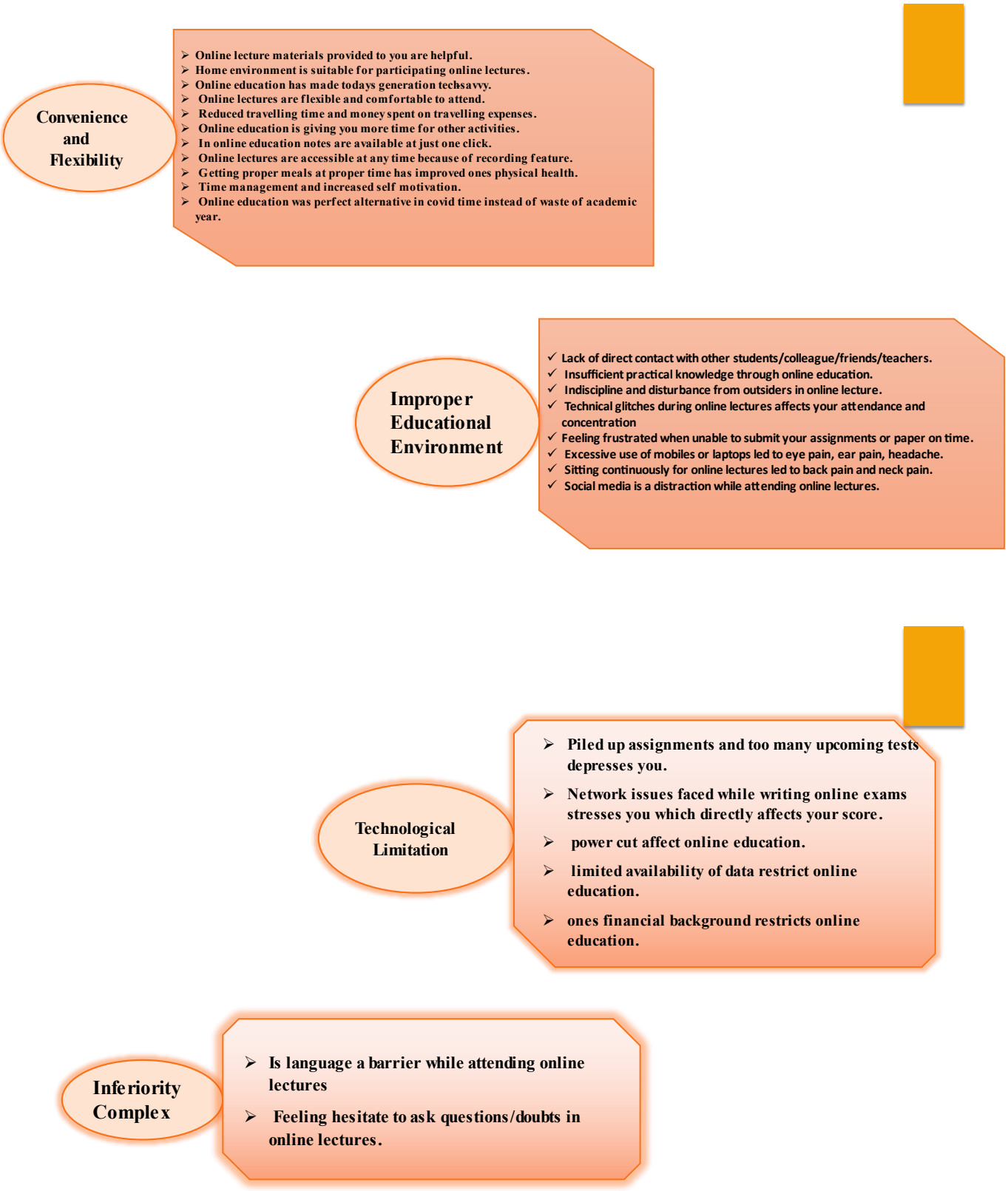
	1	2	3	4	h2	u2
L1	0.61	-0.08	-0.06	0.13	0.4	0.6
L2	0.13	0.49	0.09	0.2	0.3	0.7
L3	0.08	0.49	0.09	0.77	0.67	0.33
L4	0.01	0.43	0.14	0.61	0.58	0.42
L5	0.03	0.71	0.03	0.12	0.52	0.48
L6	-0.04	0.72	0.12	0.12	0.54	0.46
L7	0.03	0.6	0.2	0.26	0.46	0.54
L8	0.1	0.68	0.23	0	0.53	0.47
L9	0.54	0.01	-0.13	0.18	0.34	0.66
L10	0.08	0.67	0.17	0.06	0.49	0.51
L11	0.07	0.66	0.38	-0.18	0.62	0.38
L12	0.01	0.65	0.4	-0.18	0.62	0.38
L13	-0.01	0.45	0.56	0.08	0.52	0.48
L14	0.03	0.5	0.56	-0.09	0.57	0.43
L15	0.08	0.17	0.72	0.18	0.59	0.41
L16	0.11	0.29	0.69	0.09	0.59	0.41
L17	0.08	0.22	0.7	0.17	0.58	0.42
L18	0.53	0.29	0.13	0	0.38	0.62
L19	0.73	-0.07	0.03	0.07	0.54	0.46
L20	0.62	0.21	0.18	0.15	0.49	0.51
L21	0.73	0	0.09	-0.04	0.55	0.45
L22	0.7	0.05	0.09	-0.04	0.5	0.5
L23	0.59	-0.05	0.34	0.14	0.48	0.52
L24	0.74	0.09	-0.03	-0.04	0.56	0.44
L25	0.76	-0.06	0.06	-0.08	0.59	0.41
L26	0.59	0.26	0	-0.22	0.47	0.53

	1	2	3	4
SS loadings	0.47	4.43	2.86	1.4
Proportion Var	0.18	0.17	0.11	0.05
Cumulative Var	0.18	0.35	0.46	0.52

Factor Loading

	1	2	3	4
L1	0.61			
L9	0.54			
L18	0.53			
L19	0.73			
L20	0.62			
L21	0.73			
L22	0.7			
L23	0.59			
L24	0.74			
L25	0.76			
L26	0.5			
L5		0.71		
L6		0.72		
L7		0.6		
L8		0.68		
L10		0.67		
L11		0.66		
L12		0.45		
L13			0.56	
L14			0.56	
L15			0.72	
L16			0.69	
L17			0.7	
L3				0.77
L4				0.61
L2		0.49		

On the basis of factor loadings obtained from the matrix we can classify the variables into 4 factors as follows:



Convenience and Flexibilit

Improper Educational Environment

FACTORS EXTRACTED

Technological Limitation

Inferiority Complex

Paired Sample T-Test

The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations. Common applications of the paired sample t-test include case-control studies or repeated-measures designs. Suppose you are interested in evaluating the effectiveness of a company training program. One approach you might consider would be to measure the performance of a sample of employees before and after completing the program, and analyze the differences using a paired sample t-test.

Hypotheses:

Like many statistical procedures, the paired sample t-test has two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis assumes that the true mean difference between the paired samples is zero. Under this model, all observable differences are explained by random variation. Conversely, the alternative hypothesis assumes that the true mean difference between the paired samples is not equal to zero. The alternative hypothesis can take one of several forms depending on the expected outcome. If the direction of the difference does not matter, a two-tailed hypothesis is used. Otherwise, an upper-tailed or lower-tailed hypothesis can be used to increase the power of the test. The null hypothesis remains the same for each type of alternative hypothesis. The paired sample t-test hypotheses are formally defined below:

- The null hypothesis (H0) assumes that the true mean difference (μd) is equal to zero.
- The two-tailed alternative hypothesis (H1) assumes that μd is not equal to zero.
- The upper-tailed alternative hypothesis (H1) assumes that μd is greater than zero.
- The lower-tailed alternative hypothesis (H1) assumes that μd is less than zero.

The mathematical representations of the null and alternative hypotheses are defined below:

$H0: \mu d = 0$
$H1: \mu d \neq 0$ (two-tailed)
$H1: \mu d > 0$ (upper-tailed)
$H1: \mu d < 0$ (lower-tailed)

T-Test for Self-Study

Objective: To check whether online education has affected one's intellectual growth in terms of self-study time.

Hypothesis:

H0: Average self-study time for online and offline is same.

H1: Average self-study time for online and offline is not same.

Output and Conclusion:

t-Test: Paired Two Sample for Means

Self Study time	*Offline*	*Online*
Mean	2.84738617	2.59780776
Variance	2.32521257	2.69633274
Observations	1186	1186
Pearson Correlation	0.50358553	
Hypothesized Mean Difference	0	
df	1185	
t Stat	5.4363432	
P(T<=t) one-tail	3.2995E-08	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	6.599E-08	
t Critical two-tail	1.96196791	

- **Since; P-value < 0.05**
- **We Reject the null hypothesis and conclude that average self study time for online and offline is not same.**
- **Hence we now go for one tail test to check which average time is more.**

One Tail T-Test

Hypothesis:
H0: Average self-study time for online and offline is same.
H1: Average score of self-studies during online is less than self-study during offline.

Output and Conclusion for One Tail

t-Test: Paired Two Sample for Means

Self Study time	*Offline*	*Online*
Mean	2.847386172	2.597807757
Variance	2.325212571	2.696332743
Observations	1186	1186
Pearson Correlation	0.503585526	
Hypothesized Mean Difference	0	
df	1185	
t Stat	5.436343202	
P(T<=t) one-tail	3.2995E-08	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	6.599E-08	
t Critical two-tail	1.961967913	

- **Since; P-value < 0.05**
- **We Reject the null hypothesis and conclude that average score of self study during online is less than self study during offline.**

T-Test:To check which method of education is better in terms of doubt/query resolution, syllabus coverage and concept clearance.

t-Test: Paired Two Sample for Means		
Doubt Solving	*Online*	*Offine*
Mean	3.013490725	3.6037099
Variance	1.336526708	1.4732014
Observations	1186	1186
Pearson Correlation	0.376679823	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-15.3536864	
P(T<=t) one-tail	5.83689E-49	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	1.16738E-48	
t Critical two-tail	1.961967913	

- **Two Tail:**
- **H0: Average rating of doubt solving for offline and online is same.**
- **H1: Average rating of doubt solving for offline and online is not same.**
- **Conclusion:**
- **Since, P value < 0.05**
- **We reject null hypothesis and conclude that average rating of doubt solving for offline and online is not same.**

t-Test: Paired Two Sample for Means		
Doubt Solving	*Online*	*Offine*
Mean	3.013490725	3.603709949
Variance	1.336526708	1.473201415
Observations	1186	1186
Pearson Correlation	0.376679823	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-15.35368643	
P(T<=t) one-tail	5.83689E-49	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	1.16738E-48	
t Critical two-tail	1.961967913	

- **One Tail:**
- **H0: Average rating of doubt solving for offline and online is same.**
- **H1: Average rating of doubt solving for offline is greater than online.**
- **Conclusion:**
- **Since, P value < 0.05**
- **We reject null hypothesis and conclude that average rating of doubt solving for offline is greater than online.**

Syllabus Coverage

t-Test: Paired Two Sample for Means

Syllabus Coverage	*Online*	*Offline*
Mean	3.285834739	3.6888702
Variance	1.281098754	1.3706271
Observations	1186	1186
Pearson Correlation	0.423800187	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-11.2264811	
P(T<=t) one-tail	3.63937E-28	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	7.27874E-28	
t Critical two-tail	1.961967913	

- **Two Tail:**
- H0: Average rating of syllabus coverage for offline and online is same.
- H1: Average rating of syllabus coverage for offline and online is not same.
- Conclusion:
- Since, P value < 0.05
- We reject null hypothesis and conclude that average rating of syllabus coverage for offline and online is not same.

Syllabus Coverage

t-Test: Paired Two Sample for Means

Syllabus Coverage	*Online*	*Offline*
Mean	3.285834739	3.688870152
Variance	1.281098754	1.370627077
Observations	1186	1186
Pearson Correlation	0.423800187	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-11.22648113	
P(T<=t) one-tail	3.63937E-28	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	7.27874E-28	
t Critical two-tail	1.961967913	

- **One Tail:**
- H0: Average rating of syllabus coverage for offline and online is same.
- H1: Average rating of syllabus coverage for offline is greater than online.
- Conclusion:
- Since, P value < 0.05
- We reject null hypothesis and conclude that average rating of syllabus coverage for offline is greater than online.

Concept Clearance

t-Test: Paired Two Sample for Means		
Concept Clearance	*Online*	*Offline*
Mean	2.986509275	3.7057336
Variance	1.410788311	1.5091148
Observations	1186	1186
Pearson Correlation	0.289342958	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-17.1926293	
P(T<=t) one-tail	1.28376E-59	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	2.56752E-59	
t Critical two-tail	1.961967913	

- **Two Tail:**
- H_0: Average rating of concept clearance for offline and online is same.
- H_1: Average rating of concept clearance for offline and online is not same.
- Conclusion:
- Since, P value < 0.05
- We reject null hypothesis and conclude that average rating of concept clearance for offline and online is not same.

Concept Clearance

t-Test: Paired Two Sample for Means		
Concept Clearance	*Online*	*Offline*
Mean	2.986509275	3.705733558
Variance	1.410788311	1.509114778
Observations	1186	1186
Pearson Correlation	0.289342958	
Hypothesized Mean Difference	0	
df	1185	
t Stat	-17.19262931	
P(T<=t) one-tail	1.28376E-59	
t Critical one-tail	1.64614052	
P(T<=t) two-tail	2.56752E-59	
t Critical two-tail	1.961967913	

- **One Tail:**
- H0: Average rating of concept clearance for offline and online is same.
- H1: Average rating of concept clearance for offline is greater than online.
- Conclusion:
- Since, P value < 0.05
- We reject null hypothesis and conclude that average rating of concept clearance for offline is greater than online.

PARETO ANALYSIS

Pareto Analysis is a Statistical Technique in decision making that is used for the selection of a limited number of tasks that produce a significant overall effect. It uses the Pareto Principle. It is also known as 80/20 rule. The idea is that by doing 20% of the work you can generate 80% of the benefit of doing the whole job. This is also known as 'vital few' and the 'trivial many' effect.

The Pareto Principle has many applications in quality control. It is the basis for the Pareto Diagram, one of the key tools used in total quality control and Six Sigma. A Pareto chart is used to graphically summarize and display the relative importance of the differences between groups of data. Pareto chart organizes and displays information to show the relative importance of the differences between groups of data. Pareto chart organizes and displays information to show the relative importance of various problems or causes of problems.

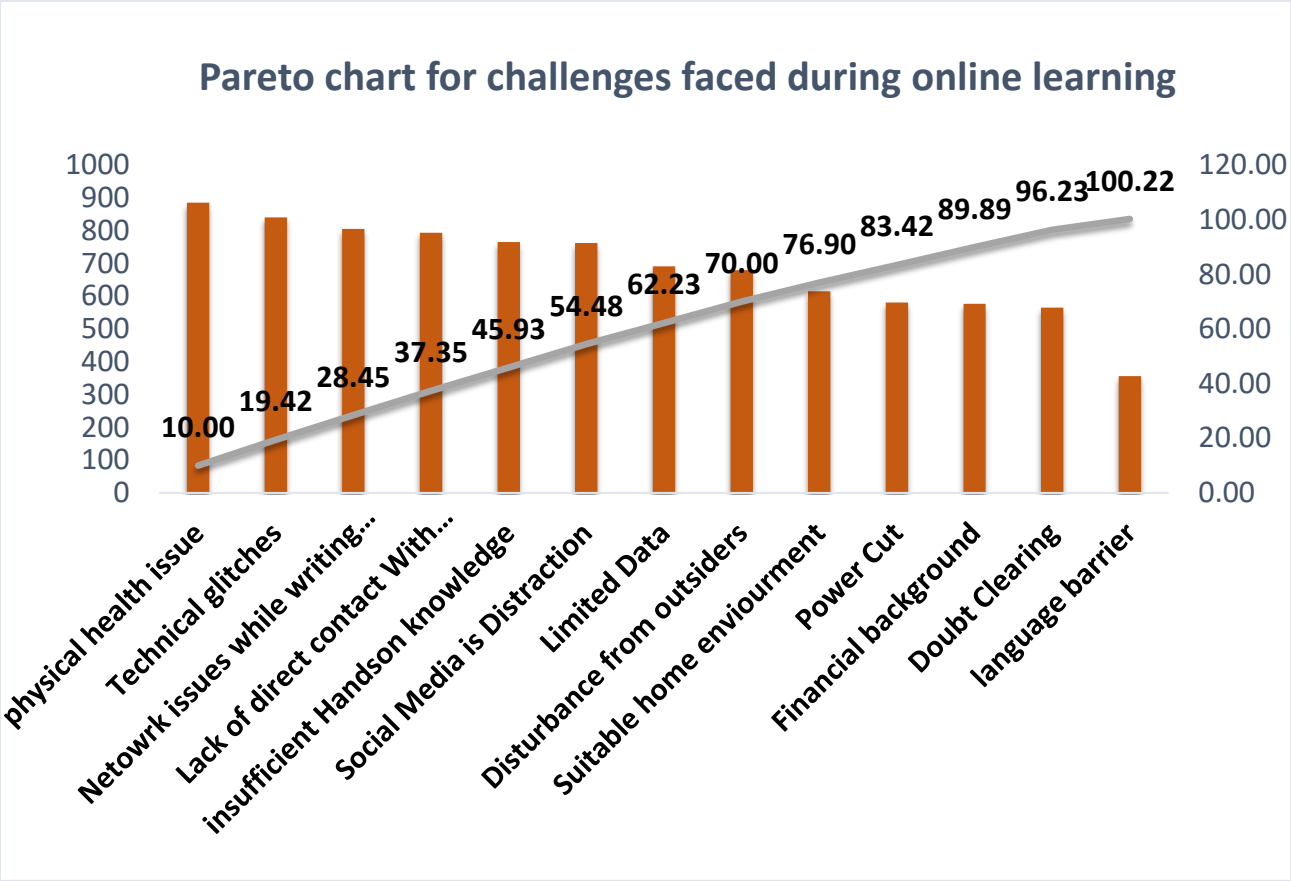
It is essentially a special form of vertical bar chart that puts items in order (highest to lowest) relative to some measurable effect of interest such as frequency, cost or time. The chart is based on the Pareto Principle, which states that when several factors affect a situation, a few factors will account for most of the impact. The Pareto describes a phenomenon in which 80 percent of variation observed can be explained by a mere 20 percent of the causes of that variation. The Pareto curve makes it clear as to where effort must be concentrated to give maximum effect. The Pareto Chart is a very simple but effective tool for prioritizing problem causes, which is why it is widely used for problem-solving in the manufacturing industry.

The Pareto Chart is a descending bar graph that shows the frequencies of occurrences or relative sizes of the various problems or causes of a particular problem. The problem categories or causes are shown on the x-axis of the bar graph. Aside from its main bar graph, the Pareto chart may also include a line graph that indicates the cumulative percentage of occurrences at each bar of the bar graph. This line graph, referred to as the 'cumulative percentage line', is used to determine which of the bars belong to the 'vital few' and which ones are relegated to the 'trivial many'.

Objective:

To find out the major challenges faced by people during online education in the times of Covid-19.

Graph:



From Graph we can see that first 8 challenes containing 70% impact.

Output:

From pareto analysis we can conclude that the major challenges faced by people during online education are:

- **Physical health issues.**
- **Technical glitches.**
- **Network issues while writing exams affecting scores.**
- **Lack of direct contact With colleagues/teachers.**
- **Insufficient Hands on knowledge.**
- **Social Media is Distraction.**
- **Limited Data.**
- **Disturbance from outsiders.**

If we can tackle the above challenges then we can improve online education.

SENTIMENT/ TEXT ANALYSIS.

What is a Sentiment?

- It is a view or opinion that is held or expressed by an individual.

What is Sentiment Analysis?

- Sentiment Analysis or Opinion Mining is a technique used to analyze the emotion in a text. We can extract the attitude or the opinion of a piece of text and get insights on it.
- In the context of machine learning, you can think of Sentiment Analysis as a Classification problem where the text can either have a positive sentiment, a negative sentiment or a neutral one.

What are the applications of Sentiment Analysis in the industry?

- In the age of social media, it is extremely common to comment about
    1. a movie you liked or
    2. a book you didn't like or
    3. a product you bought was not up to the mark.
- Therefore, a lot of companies use sentiment analysis for their products since it provides direct feedback of the customer's opinion.
- It is also important to detect and remove hateful content from social media and companies like Twitter, Facebook, etc. extensively use sentiment analysis on a daily basis.

On what kind of projects would I implement sentiment analysis?

There are a wide variety of projects where you can use Sentiment Analysis. Here are a couple of popular use cases:

- Sentiment Analysis can not only be used for customer reviews or product feedback, but in other domains as well.
- Analyzing the sentiments on social media on the US Elections, for example, gives useful insights on which candidates are favored by the public and which are not.

STEPS IN PERFORMING SENTIMENT ANALYSIS: -

1. Data Collection.
2. Forming a Corpus.
3. Pre-processing text.
4. Tokenization.
5. Forming a Term Document Matrix.
6. Visualization of the output.

EXPLANATION OF EACH OF THE TERMS:

1. Data Collection.
The data collection is the primary step before carrying out any analysis. There is no such compulsion regarding the collection of data for the analysis.
In our project the data was collected by means of a questionnaire (Online), where the respondents were told to express their opinion and views on "Online Learning" in the space provided in the questionnaire.

2. Forming a Corpus.
A Corpus is a collection of documents gathered at a single instance. Since we had only one column of entries, it was treated as a corpus i.e. a main body consisting of records from different users.

3. Pre-processing text.
Now the data required for analysis need to be made clean as it contains a lot of unnecessary information which maybe not required for the analysis.
It is done in the following manner: -
a) Stemming: it is required to identify the root of the word from which it is formed. For instance, sailed, sailing, sailor is stemmed to sail.
b) Removing punctuation: It is required to remove the punctuations such as "! @&<>, /" etc.
c) Removing Stop word: Stop words are words which are used in sentences as connectors or conjunctions like the, is, that, for, which, etc. that don't have any specific meaning.
d) Converting to lowercase: It is very necessary to convert the entire data into lowercase, because the computer can take up to similar words ship and SHIP as two different words.
e) Replacing special words: like don't to do not and wouldn't too would not etc. in order to have a clearer context of the words used.

4. Tokenization.
It is the most important step which involves splitting of sentence into separate words. This is done in order to perform the text analysis that is the main step.

5. Forming a Term Document Matrix.
We then formulate the Term Document Matrix which is basically a matrix of words and their respective frequencies i.e. it displays in a tabular format the no. of times each word has occurred in a sentence, which can be further used in computing the Word Cloud and many much more analysis.

6. Visualization of the output.
Here we obtain the final output of our analysis and then we interpret the results representing it in the form of Graphs and Cloud.

WORDCLOUD

- **What is a Word Cloud?**

A Word Cloud is a popular visualization of words typically associated with text data. They are most commonly used to highlight popular or trending terms based on frequency of use of prominent and important. A Word Cloud is a beautiful, Informative image that communicates much in a single glance.

- **What is the purpose of the Word Cloud?**

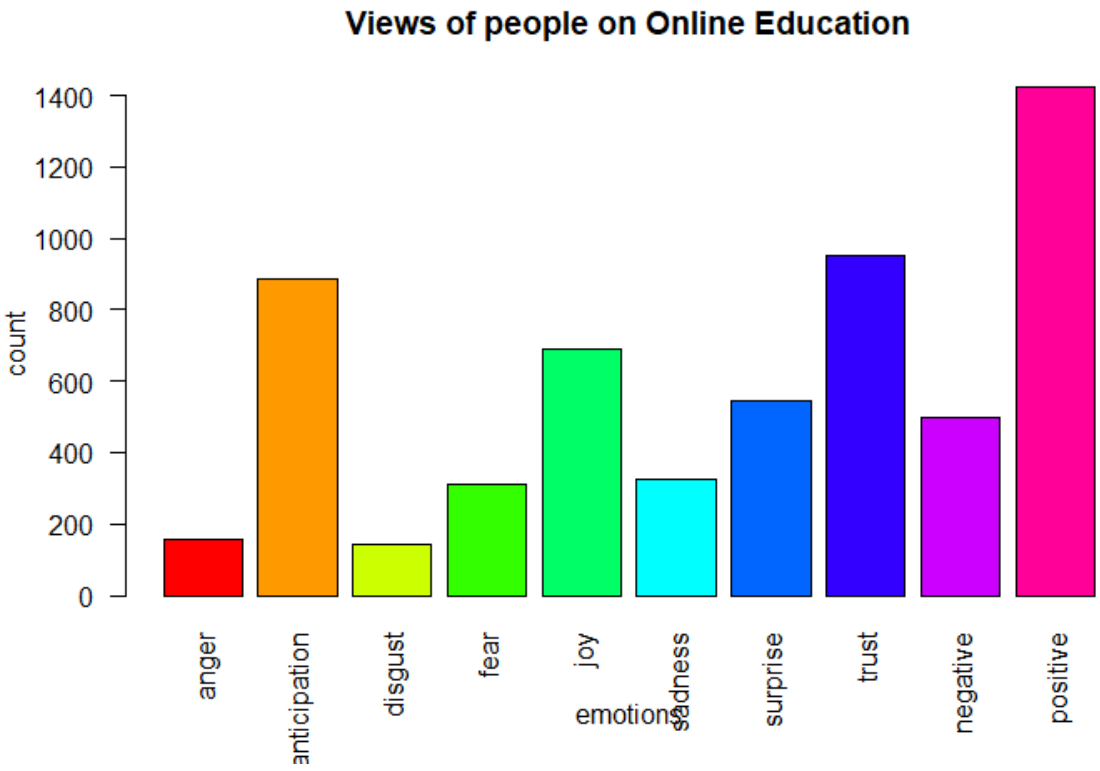
The Word Cloud displays the most common words found in that text and shows them in a way that lets the viewer know what words are used in a text and with what kind of frequency.

Objective: To study the people's opinion and views on "Online Education"
The sentiment analysis was conducted with the help of two lexicons: -
• BING lexicon: This lexicon categorizes words into a binary fashion. (E.g. Positive and Negative).
• NRC lexicon: This lexicon categorizes words into a multiple category fashion (E.g. Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust.)
(* A lexicon is a vocabulary of words)

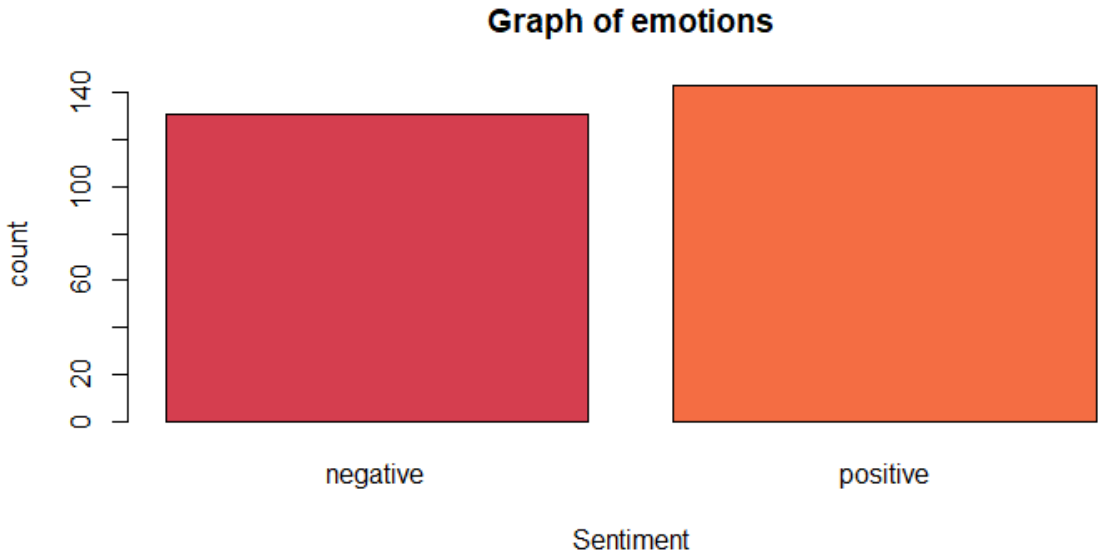
• Trust: Firm Belief.
• Anger: Displeasure.
• Fear: Feeling of threat.
• Anticipation: Predict.
• Disgust: Strong Disapproval.
• Sadness: Unhappy.
• Joy: Feeling of happiness.
• Surprise: Unexpectedness.

To accomplish this task, first the overall responses were collected, cleaned, pre-processed and visualized so as to understand what views were being expressed by the people on ONLINE LEARNING.

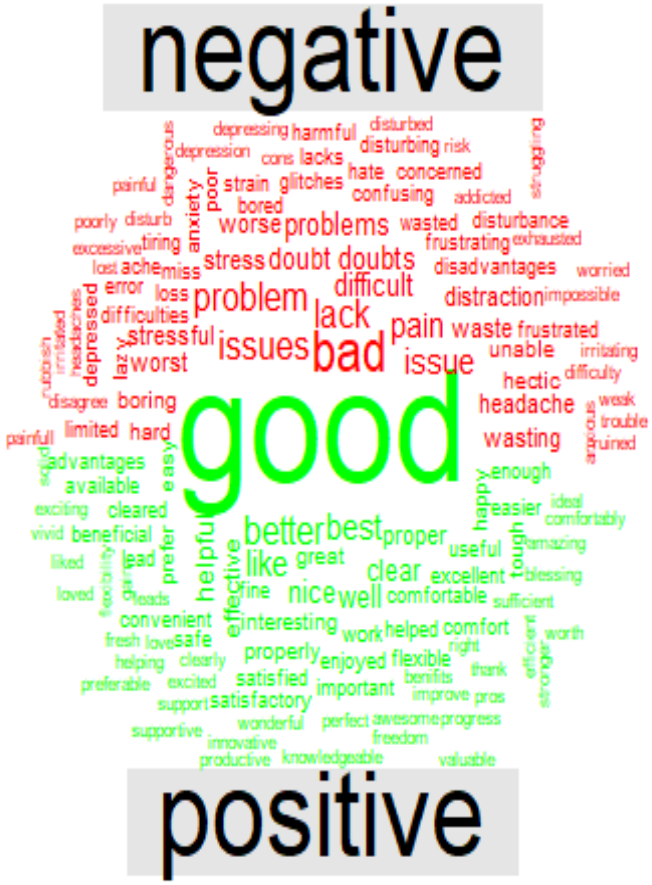
• The above graph is obtained by NRC lexicon which showcases the emotions expressed in terms of counts classified in multiple categories.
• The responses obtained by the people expressed the emotion of "trust" to "disgust", "sadness", "joy"," surprise".
• The people have strongly put forth their consent regarding the Online learning experience which can be seen in the graph corresponding to "trust".
• The people have strongly expressed the emotions of "Joy", "Surprise" and "Positivity" which are positive emotions, to a greater extent as compared to the emotion of "Anger", "fear", "Negativity" showing their consent that they approve that Online mode of learning was best alternative in these unprecedented times.
• The emotion of "Disgust" is expressed to a vey lesser extent indicating that very few people had mixed opinions on online learning system that express their view which contradicts with the emotion of "Trust".



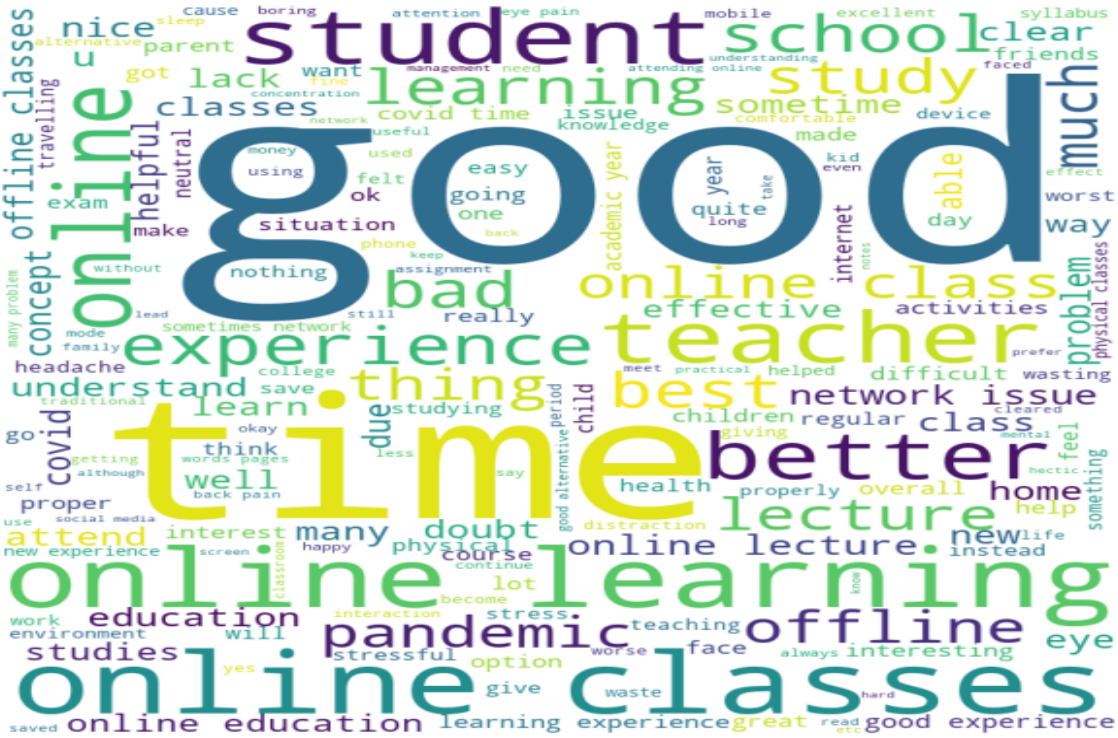
- The below graph is obtained called the BING lexicon, shows the emotions expressed in count classified in a Binary Fashion.



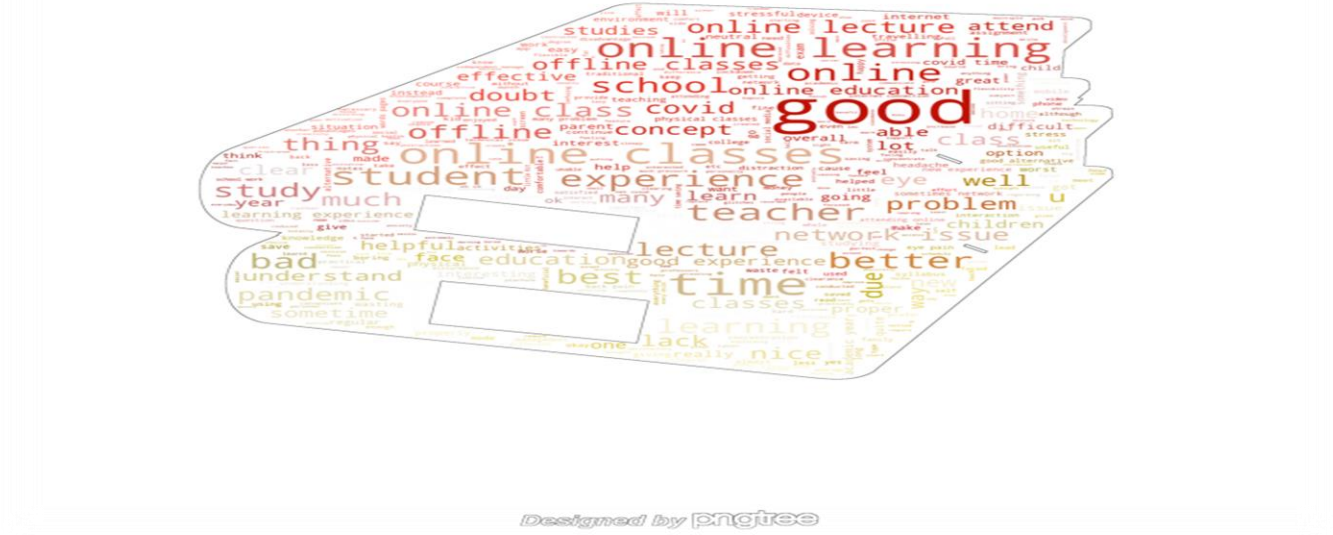
- **The positive sentiment is expressed to a greater extend as compared to negative sentiment.**
- **This gives us clear evidence that most of the people were happy with online education or at least it served as best alternative at the times of Covid-19.**
- **Negative sentiments are also expressed at quiet a high level, giving us an idea of people struggling with technical and health issues due to online learning schedule, which gives frustration of people for online learning.**



Python Wordcloud:



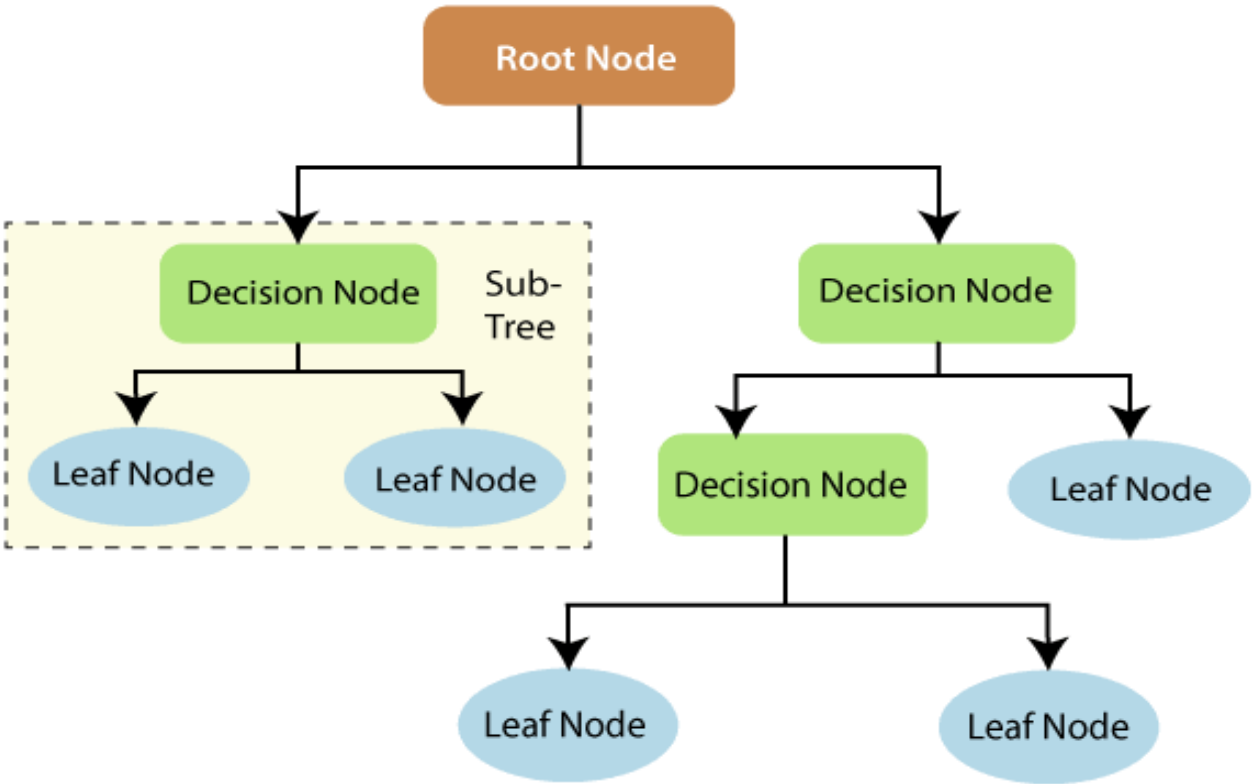
Masked Word cloud:



DECISION TREES

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables. Decision tree are used when:
1. We have high dimensional data.
2.Intuitive representation that is easily understood by humans.
3.Learning and classification are simple and fast.
4.We have a good accuracy data.
There are two types of decision tree: -
Classification Tree: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.
Regression Tree: where the target variable is continuous and tree is use to predict its value.

Terminology related to Decision Trees



Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.

Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.
Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.

Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree

Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

There are 4 different types of tree growing methods:
1. Gini
2. Entropy
3. Chi square
4. Reduction in variance (used for continuous target data)

Gini: - Gini works with categorical target variable. It performs only binary splits. Higher the value of Gini higher the homogeneity. CART (Classification and Regression Trees) uses Gini method to create binary splits.

Entropy: - A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. Entropy uses to calculate the homogeneity of a sample. If a sample is completely homogeneous then entropy is zero and if the sample is equally divided then it has entropy of one.
The lower the entropy the better split, the pure segmentation. The higher the entropy more impurity, more diversity in segment, the more uncertainty. Therefore, entropy should be less.

Chi-square: - It works with categorical target variable. It can perform two or more splits. Higher the value of chi-square higher the statistical significance of differences between sub-node and the parent node. It generates tree called CHAID (Chi-square Automatic Interaction Detector)

Reduction in Variance: - Reduction in variance is an algorithm used for continuous target variables.

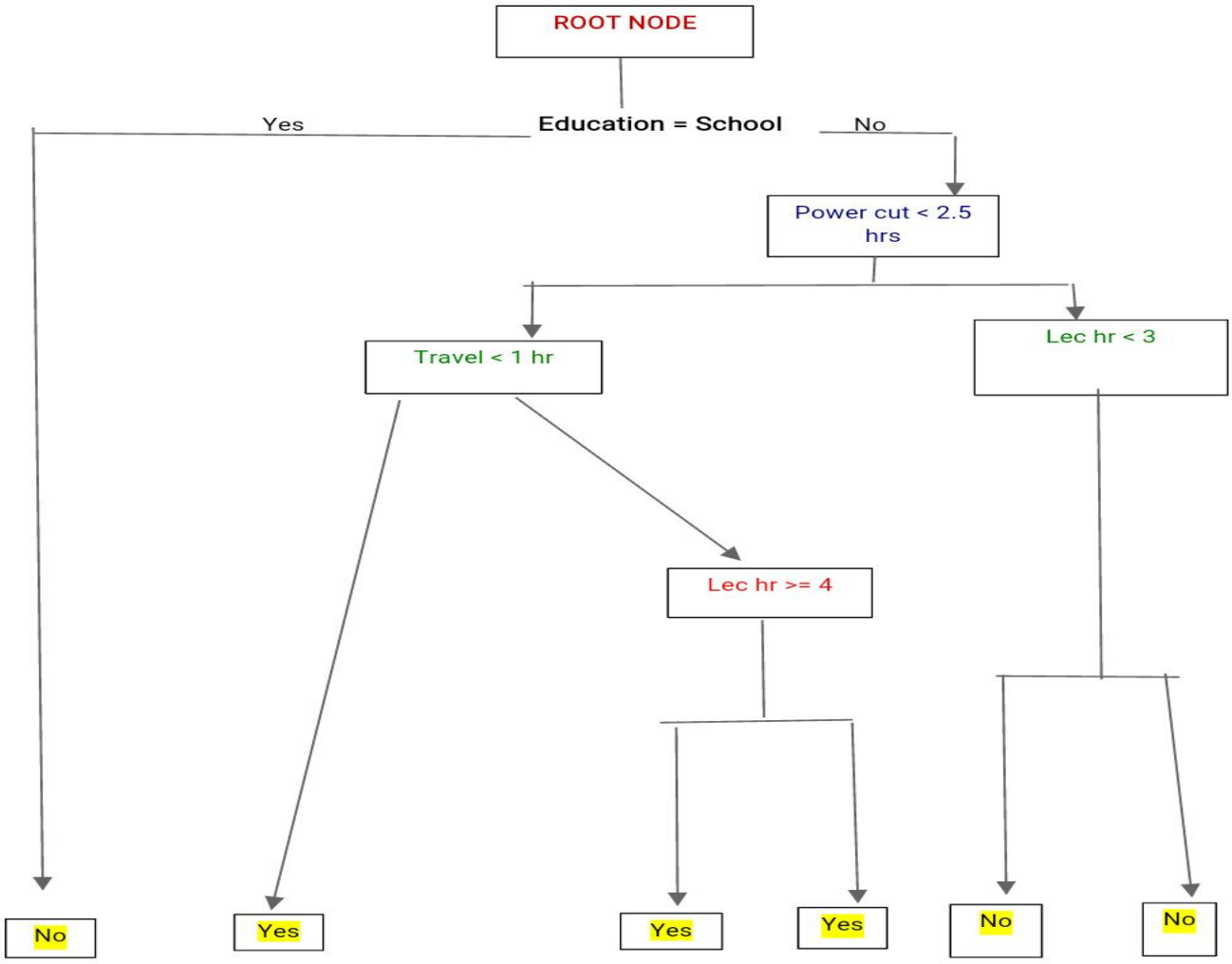
Pruning: - It should reduce the size of decision tree by removing nodes (opposite of splitting) without reducing predictive accuracy. There are many techniques for the tree pruning.
Reduced error pruning: - One of the simplest forms of pruning is reduced error pruning. Pruning and subtree selection based on minimizing the error rate in the validation partition at each pruning step and then in the overall subtree sequence. This is usually based on the misclassification rate for a categorical response variable, but ASE can also be used.

Types of Data partition:
Training Dataset: - The sample data use to fit model.
Testing Dataset: - The sample of data use to provide unbiased evaluation of final fit model on Training dataset.

<u>OBJECTIVE</u>:

TO UNDERSTAND CHARACTERISTICS OF RESPONDERS WHO WILL CONTINUE WITH ONLINE EDUCATION IF GIVEN A CHANCE IN FUTURE BASED ON VARIOUS DEMOGRAPHICS.

WE HAVE BUILD DECISION TREE TO PREDICT WHETHER A PERSON WILL CONTINUE WITH ONLINE EDUCATION IN FUTURE BASED ON VARAIABLES: EDUACTION (SCHOOL AND COLLEGE), LECTURE HOURS, TRAVEL TIME TO REACH SCHOOL OR COLLEGE IN NON COVID TIMES.



<u>INTERPRETATION DRAWN</u>:

1) School going students do not prefer to continue with online education in future.

2)For college going students:

{a} If Power cut is less than 2.5 hours (2 hrs. 30 minutes) and travel time was less than 1 hour then the person would continue with online education.

{b} If Power cut is less than 2.5 hours and travel time is greater than 1 hour and lecture timings are greater than 4 hours daily then the person would continue with online education.

{c} If Power cut is less than 2.5 hours and travel time is greater than 1 hour and lecture timings are less than 4 hours daily then the person would continue with online education.

{d} If Power cut is more than 2 hours 30 minutes and Lecture hours are less than 3 hours the person wouldn't continue with online education.

{e} If Power cut is more than 2 hours 30 minutes and Lecture hours are greater than 3 hours the person wouldn't continue with online education.

NAIVE BAYES ALGORITHM:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
It is mainly used in text classification that includes a high-dimensional training dataset.
Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as:
Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes Theorem:

Bayes' theorem is also known as Bayes Rule or Bayes law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

where,
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:
Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular

document belongs to which category such as Sports, Politics, education, etc.The classifier uses the frequency of words for the predictors.
Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

<u>Assumptions made by Naïve Bayes:</u>
The fundamental Naïve Bayes assumption is that each feature makes an:
1)independent
2)equal contribution to the outcome.

<u>Working of Naïve Bayes' Classifier:</u>
Working of Naïve Bayes' Classifier can be understood with the help of the below example:
Suppose we have a dataset of weather conditions and corresponding target variable "Play". So, using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So, to solve this problem, we need to follow the below steps:
Convert the given dataset into frequency tables.
Generate Likelihood table by finding the probabilities of given features.
Now, use Bayes theorem to calculate the posterior probability.
Problem: If the weather is sunny, then the Player should play or not?

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likelihood table weather condition:

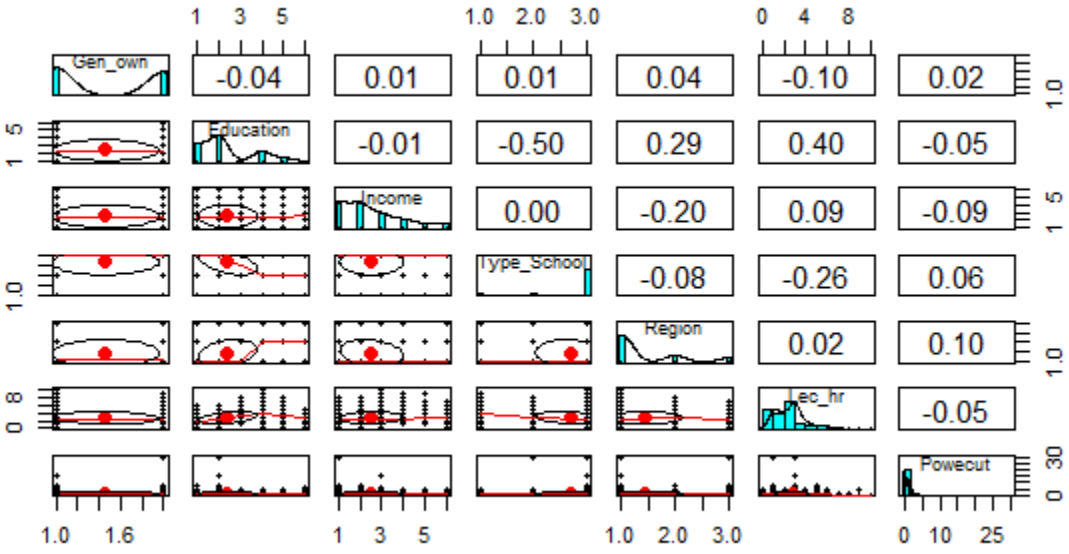
Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

Applying Baye's theorem:

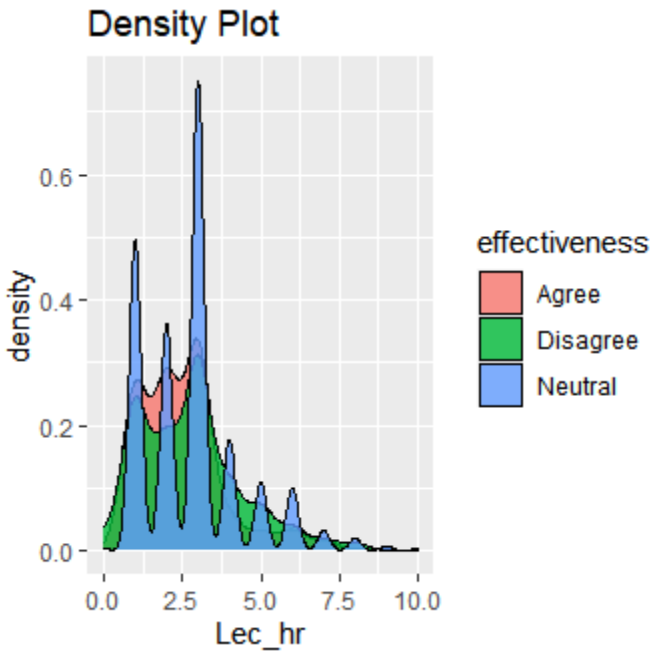
P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)
P(Sunny|Yes)= 3/10= 0.3
P(Sunny)= 0.35
P(Yes)=0.71
So , P(Yes|Sunny) = 0.3*0.71/0.35= 0.60
P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)
P(Sunny|NO)= 2/4=0.5
P(No)= 0.29
P(Sunny)= 0.35
So, P(No|Sunny)= 0.5*0.29/0.35 = 0.41
So, as we can see from the above calculation that P(Yes|Sunny)>P(No|Sunny)
Hence on a Sunny day, Player can play the game.

<u>Application of Naive Bayes model in our project :</u>

We have used Naive Bayes to predict whether a person will find online education effective or not based on various demographics - Gender, Education, Family Income, Region, Lecture hours, Power cut.
We check whether the predictor variables are independent i.e. not that strong correlation between them.



All the predictor variables are independent.



Significant amount of overlapping is there. Similarly, we can do for remaining variables. So, there is some potential to develop classification model , but due to overlapping the model won't be 100% accurate.

Output of trained data :

```
 A priori probabilities:

    Agree  Disagree   Neutral
0.2576558 0.3210137 0.4213305

-------------------------------------------------------------------
---------------------------------------

 Likelihood probabilities

-------------------------------------------------------------------
---------------------------------------
 ::: Gen_own (Bernoulli)
-------------------------------------------------------------------
---------------------------------------

Gen_own      Agree  Disagree   Neutral
  Female 0.5368852 0.4934211 0.5714286
  Male   0.4631148 0.5065789 0.4285714

-------------------------------------------------------------------
---------------------------------------
 ::: Education (Categorical)
```

```
------------------------------------------------------------------
----------------------------------------

Education                          Agree   Disagree    Neutral
  Higher School (XI-XII)     0.01229508 0.03618421 0.01253133
  Other                      0.02868852 0.01315789 0.02255639
  Post Graduate              0.05737705 0.08223684 0.08020050
  Primary (Play School - IV) 0.29508197 0.29605263 0.26566416
  Secondary school (V - X)   0.43442623 0.43421053 0.39849624
  Under Graduate             0.17213115 0.13815789 0.22055138

------------------------------------------------------------------
----------------------------------------
 ::: Income (Categorical)
------------------------------------------------------------------
----------------------------------------

Income                        Agree   Disagree    Neutral
  1,00,000-3,00,000   0.31147541 0.28618421 0.27568922
  10,00,000 and above 0.04508197 0.08552632 0.05012531
  3,00,000-5,00,000   0.19262295 0.14473684 0.19047619
  5,00,000-7,00,000   0.10245902 0.14144737 0.10025063
  7,00,000-10,00,000  0.04508197 0.05921053 0.07268170
  Below 1,00,000      0.30327869 0.28289474 0.31077694

------------------------------------------------------------------
----------------------------------------
 ::: Type_School (Categorical)
------------------------------------------------------------------
----------------------------------------

Type_School             Agree   Disagree    Neutral
  Government      0.10245902 0.10855263 0.11278195
  Private         0.77459016 0.79605263 0.78446115
  Semi-Government 0.12295082 0.09539474 0.10275689

------------------------------------------------------------------
----------------------------------------
 ::: Region (Categorical)
------------------------------------------------------------------
----------------------------------------

Region             Agree   Disagree    Neutral
  Rural       0.13524590 0.08881579 0.14536341
  Semi-urban  0.17622951 0.19736842 0.18045113
  Urban       0.68852459 0.71381579 0.67418546
```

Interpretation:

A priori probabilities:

```
    Agree  Disagree   Neutral
0.2576558 0.3210137 0.4213305
```

Implies that from our trained data 25.76% found online education effective 32.10% didn't found online education effective and 42.13% sticked to neutral.

```
Gen_own      Agree  Disagree   Neutral
  Female 0.5368852 0.4934211 0.5714286
  Male   0.4631148 0.5065789 0.4285714
```

P(Gen_own = Female | effectiveness= Agree)=0.5368

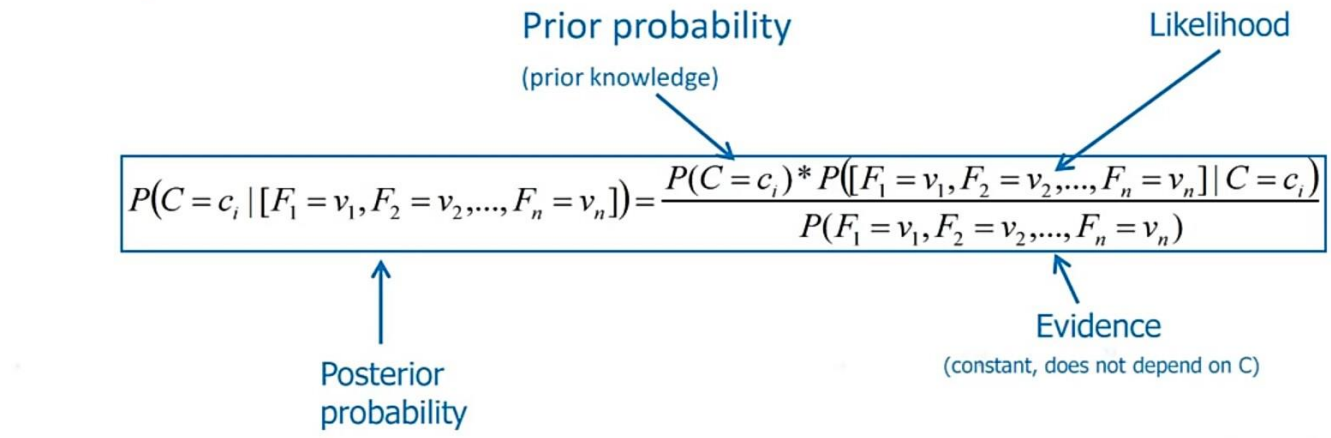
LIKELIHOOD PROBABILTIES:

	Agree	Disagree	Neutral
Gender			
Female	0.5368852	0.4934211	0.5714286
Male	0.4631148	0.5065789	0.4285714
Education			
Higher School (XI-XII)	0.01229508	0.03618421	0.01253133
Other	0.02868852	0.01315789	0.02255639
Post Graduate	0.05737705	0.08223684	0.0802005
Primary (Play School - IV)	0.29508197	0.29605263	0.26566416
Secondary school (V - X)	0.43442623	0.43421053	0.39849624
Under Graduate	0.17213115	0.13815789	0.22055138
Income			
1,00,000-3,00,000	0.31147541	0.28618421	0.27568922
10,00,000 and above	0.04508197	0.08552632	0.05012531
3,00,000-5,00,000	0.19262295	0.14473684	0.19047619
5,00,000-7,00,000	0.10245902	0.14144737	0.10025063
7,00,000-10,00,000	0.04508197	0.05921053	0.0726817
Below 1,00,000	0.30327869	0.28289474	0.31077694
Type_School			
Government	0.10245902	0.10855263	0.11278195
Private	0.77459016	0.79605263	0.78446115
Semi-Government	0.12295082	0.09539474	0.10275689
Region			
Rural	0.1352459	0.08881579	0.14536341
Semi-urban	0.17622951	0.19736842	0.18045113
Urban	0.68852459	0.71381579	0.67418546

FOR NUMERICAL DATA:

	mean	Standard deviation
Lec_hr		
Agree	2.528689	1.458337
Disagree	2.769737	1.662915
Neutral	2.877193	1.64804
Power cut		
Agree	1.034016	2.181618
Disagree	1.149671	2.238366
Neutral	0.8915038	1.106329

From Bayes' rule:



PREDICTIVE MODEL (It is not 100% accurate)

Predicted	Gen_own	Education	Income	Type_School	Region	Lec_hr	Powercut	effectiveness
Neutral	Female	Under Graduate	3,00,000-5,00,000	Government	Semi-urban	5	0	Neutral
Neutral	Female	Primary (Play School - IV)	Below 1,00,000	Private	Urban	2	2	Neutral
Disagree	Male	Under Graduate	10,00,000 and above	Semi-Government	Semi-urban	6	4	Disagree
Agree	Female	Secondary school (V - X)	1,00,000-3,00,000	Semi-Government	Urban	3	1	Disagree

Agree	Disagree	Neutral	Predicted	Gen_own	Education
0.123	0.08635	0.791	Neutral	Female	Under Graduate

Income	Type_School	Region	Lec_hr	Powercut	effectiveness
3,00,000-5,00,000	Government	Semi-urban	5	0	Neutral

Misclassification :
Training Data set : 0.40248
Test Data set : 0.38025

NAIVE BAYES ALGORITHM:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
It is mainly used in text classification that includes a high-dimensional training dataset.
Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as:
Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes Theorem:

Bayes' theorem is also known as Bayes Rule or Bayes law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
The formula for Bayes' theorem is given as:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

where,
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

Types of Naïve Bayes Model:

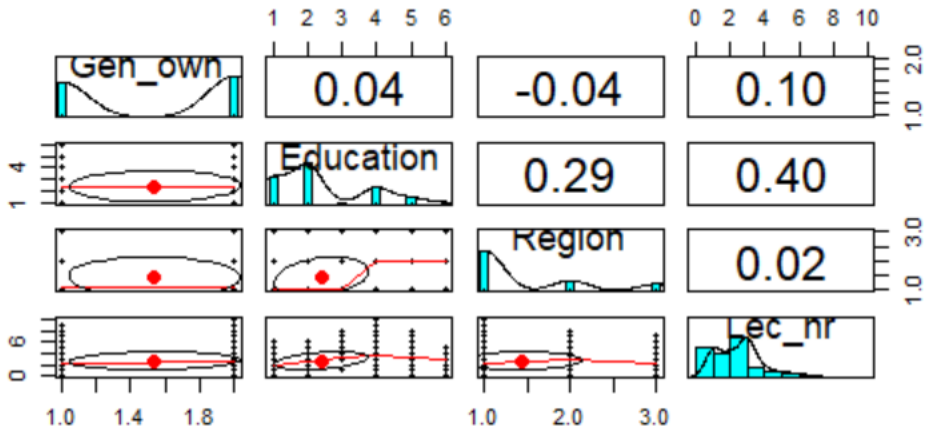
There are three types of Naive Bayes Model, which are given below:
Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.The classifier uses the frequency of words for the predictors.
Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

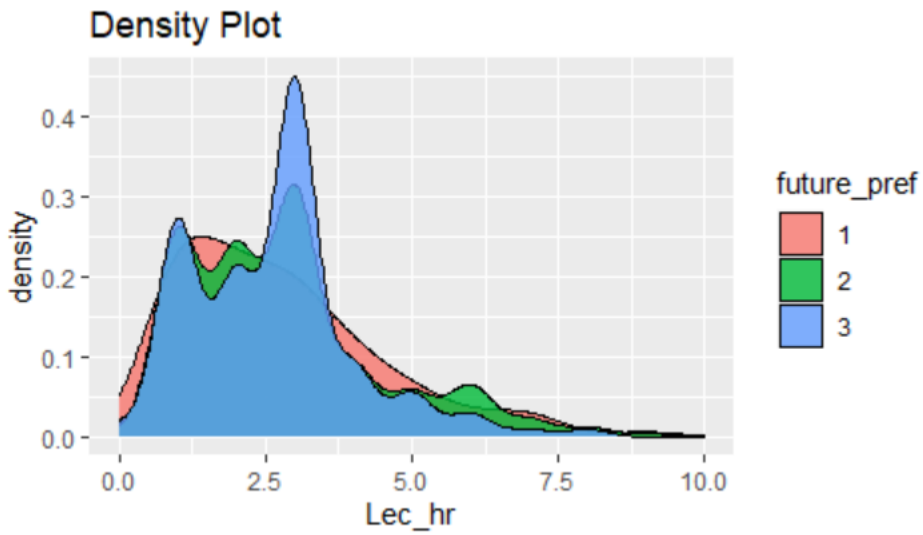
<u>Assumptions made by Naïve Bayes:</u>
The fundamental Naïve Bayes assumption is that each feature makes an:
1)Independent
2) Equal contribution to the outcome.

<u>Working of Naïve Bayes' Classifier:</u>
Working of Naïve Bayes' Classifier can be understood with the help of the below example:
Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:
Convert the given dataset into frequency tables.
Generate Likelihood table by finding the probabilities of given features.
Now, use Bayes theorem to calculate the posterior probability.

<u>Application of Naive Bayes model in our project :</u>
We have used Naive Bayes to predict whether a person will refer Online, Blended (Mixture of online and offline) or Offline education system once this pandemic is over based on various demographics - Gender, Education, Region, Lecture hours.
We check whether the predictor variables are independent i.e. not that strong correlation between them.



All the predictor variables are independent.



Significant amount of overlapping is there. Similarly, we can do for remaining variables. So, there is some potential to develop classification model , but due to overlapping the model won't be 100% accurate.

Output of trained data:
A priori probabilities:

Blended	Offline	Online
0.42977825	0.48468849	0.08553326

Tables:

Gen own (Bernoulli)

Gen_own	Blended	Offline	Online
Female	0.5281174	0.5531453	0.4939759
Male	0.4718826	0.4468547	0.5060241

Education (Categorical)

Education	Blended	Offline	Online
Higher School (XI-XII)	0.01452785	0.02580645	0.04597701
Other	0.03389831	0.01075269	0.04597701
Post Graduate	0.09200969	0.06666667	0.05747126
Primary (Play School - IV)	0.32445521	0.25806452	0.19540230
Secondary school (V - X)	0.34866828	0.49462366	0.29885057
Under Graduate	0.18644068	0.14408602	0.35632184

Region (Categorical)

Region	Blended	Offline	Online
Rural	0.1243902	0.1017316	0.2738095
Semi-urban	0.2146341	0.1580087	0.2023810
Urban	0.6609756	0.7402597	0.5238095

Interpretation:

A priori probabilities:

Blended	Offline	Online
0.42977825	0.48468849	0.08553326

Implies that from our trained data 8.55% wants to continue with Online education 42.97% wants to continue with Blended (Mixture of Online and Offline) education and 48.46% wants to continue with Offline education.

Gen_own	Blended	Offline	Online
Female	0.5281174	0.5531453	0.4939759
Male	0.4718826	0.4468547	0.5060241

P(Gen_own = Female | future_pref= Online)=0.5060

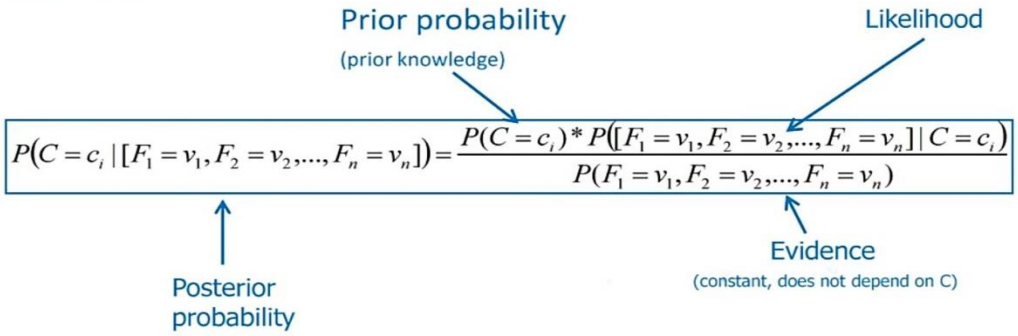
LIKELIHOOD PROBABILTIES:

	Online	Blended	Offline
Gen_own			
Female	0.4939759	0.5281174	0.5531453
Male	0.5060241	0.4718826	0.4468547
Education			
Primary (Play School - IV)	0.1954023	0.32445521	0.25806452
Secondary school (V - X)	0.29885057	0.34866828	0.49462366
Higher School (XI-XII)	0.04597701	0.01452785	0.02580645
Under Graduate	0.35632184	0.18644068	0.14408602
Post Graduate	0.05747126	0.09200969	0.06666667
Other	0.04597701	0.03389831	0.01075269
Region			
Urban	0.5238095	0.6609756	0.7402597
Semi-urban	0.202381	0.2146341	0.1580087
Rural	0.2738095	0.1243902	0.1017316

FOR NUMERICAL DATA:

	Mean	Standard Deviation
Lec_hr		
Online	2.728395	1.837201
Blended	2.85258	1.732872
Offline	2.668845	1.446019

From Bayes' rule:



PREDICTIVE MODEL (It is not 100% accurate):

Predicted	Gen_own	Education	Region	Lec_hr	future_pref
Blended	Female	Under Graduate	Semi-urban	6	Blended
Blended	Female	Under Graduate	Semi-urban	1	Offline
Blended	Male	Post Graduate	Semi-urban	6	Blended
Blended	Male	Post Graduate	Semi-urban	6	Blended
Offline	Female	Secondary school (V - X)	Urban	1	Online
Offline	Female	Secondary school (V - X)	Urban	3	Offline

	Blended	Offline	Predicted	Gen_own
0.05050806	0.7158631	0.2336288	Blended	Male

Education	Region	Lec_hr	future_pref
Post Graduate	Semi-urban	6	Blended

SWOC ANALYSIS

SWOC analysis is a strategic planning method used to research external and internal factors which affect company success and growth. Firms use SWOC analysis to determine the strengths, weaknesses, opportunities, and challenges of their firm, products, and competition.

SWOC analysis is relevant to SWOT analysis. SWOT examines strengths, weaknesses, and opportunities. But it focuses on threats rather than challenges. The two are similar but they do have their differences, which is why firms may choose to use SWOC or SWOT.

How to use SWOC analysis

When beginning a SWOC analysis of a product or firm, you must go through each section individually. Starting with…

Strengths:

Strengths are features which benefit the company, such as product sales. For example, sales of Product X is growing 3% each month. But Product Z is seeing a 3% monthly decline. In this case, Product X, which brings in more revenue, is where the firm should focus their efforts to continue profit growth.

Strengths can also be more abstract. If you've decided to build a product because you know you can offer it cheaper than your competitor, this is an overall strength of the company. Or if you have records of better customer service via positive reviews online, this is a strength you can use to your advantage. Strengths can be documented through statistics, customer service reviews, and surveys.

Weaknesses:

The next step is noticing weaknesses. Weaknesses cause a company to struggle. For example, if you've decided to target a younger audience but your packaging is still dedicated to senior citizens, the new consumer base will struggle to connect to the product. This will show in reports, and cause an internal struggle within the company.

Weaknesses need to be documented and acknowledged to handle them promptly before it spreads and leads to overall destruction.

Opportunities:

Opportunities are often external. They provide ways for firms to grow successfully. For example, a digital marketing agency helps a client develop an effective email marketing strategy. The agency has been thinking of doing graphic design so they offer a reduced fee to re-do the existing client's logo. This is an opportunity for the agency to develop a new section of their business without having to devise a marketing plan because they can reach out to existing clients.

Being open to opportunities, knowing when to look for them, and how to act on them can boost a firm's success. Documenting past opportunities can help create a plan on how to capitalize future opportunities.

Challenges:

The final step in SWOC analysis is acknowledging challenges. This is how SWOC and SWOT analysis differ because SWOT analysis focuses on threats.

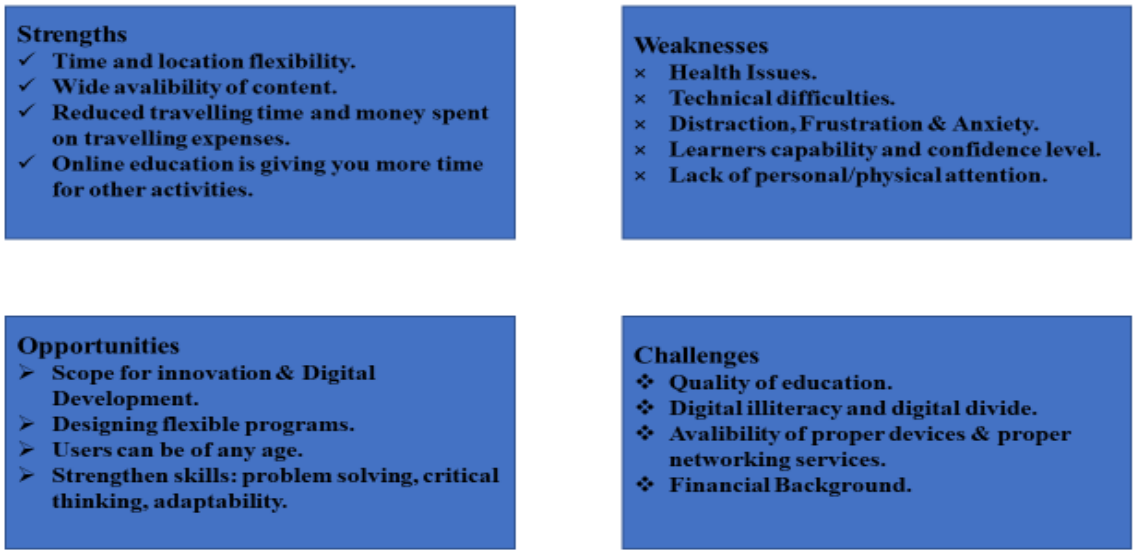
Challenges are similar to threats but have the chance of being overcome. Threats have the potential to damage a firm, but challenges often already exist and need to be handled appropriately.

This step is crucial. If you've already examined the strengths, weaknesses, and opportunities but skip assessing challenges, you may be on the path to failure. Challenges can greatly undermine any progress you've made, so by ignoring this step, you've opened yourself up to potential failure.

When to use SWOC analysis:

Use SWOC analysis whenever you have a business idea. Whether it's starting a brand new business, a product, or a product upgrade. You can do SWOC analysis annually, quarterly, or monthly; it depends on what product or idea you're using SWOC analysis for.

But if you choose to do SWOC analysis, remember it's a great cost-effective way to reduce challenges and deter failure of a business venture or product.



Conclusion

From our analysis we have observed that people have mixed opinions regarding online education i.e. it is neither good nor bad, so in future if government wants to adopt online education they should rather opt for blended education.
Online Education was convenient & flexible for people to attend but at times the technical glitches were causing them problem. Also, improper educational environment does curb the one's experience. Online Education also gives rise to inferiority complex into an individual due to lack of confidence indulged into them due to virtual lectures.
Online education did affect one's intellectual growth since self-study time reduced during this online academic year. "Brick & Mortar (traditional) education was effective in terms of doubt solving, concept clearance, syllabus coverage.
If we tackle the challenges below, we can boost online education in India.

Physical health issues, Technical glitches, Network issues while writing exams affecting scores, Lack of direct contact with colleagues/teachers, insufficient Hands-on knowledge, Social Media is Distraction, Limited Data, Disturbance from outsiders.
Most of our focus was on school going kids and we observed that online education is not quite helpful for these kids so government or the educational institutes need to focus on some better option for the school going kids.

SCOPE

- **Online learning is booming in current times. Aided by the widespread availability of high-speed internet, making use of new technologies such as 4G and the soon-to-be-released 5G, online learning is expected to grow by leaps and bounds in the foreseeable future. The worldwide market size of online learning was approximately $187.87 billion in 2019, a 400% increase over what it was just six years ago. This phenomenal growth has been made possible not just by the rapidly evolving scenario in the world of technology, but also by the spread of education in the developing world. Experts predict that the next wave of online education will occur not in North America and Europe, but newly emerging markets like Africa, India, and China.**
- **Online learning is no longer just limited to colleges and universities. Right since primary school, online learning is gradually being incorporated into the curriculum. The recent COVID-19 pandemic further illustrates the importance of online learning in today's school system, as it has proven to be a boon to both students and teachers alike who are unable to attend school due to the risk of disease spread. Beyond high school, online learning is steadily increasing its market share at the pre-university level. Furthermore, e-learning is expanding in presence beyond the traditional fields as well.**

SUGGESTIONS

1)A further studies can be conducted to study online education from the point of view of lecturer / teacher.
2) Online education can be studied from economic perspective by considering variables : Income, School fees, Devices brought, etc.
3) Effectiveness of online education can be studied separately for different educational qualifications , different regions.
4)Study the overall Market of Educational Industry.
5)To Identify the problems and get proper solutions based on Different States/Different Cities/Different Boards/Different Universities.
6)Study Educational industry in two different ways, based on foundational courses and other courses.
7)Identify problems and thereby suggest solutions for practical based courses.

CODES

R code for Factor Analysis:

```
install.packages("corpcor")
install.packages("GPArotation")
install.packages("psych")
library(corpcor)
library(GPArotation)
library(psych)
setwd()

online=read.csv('factor analysis 2.csv',sep=',',header = T)
View(online)
Correlation = cor(online)
head(round(online,2))
KMO(online)

cortest.bartlett(online)
  a = scree(online)
a

pc1 =    principal(online, nfactors=26, rotate="none")
plot(pc1$values, type="b")

pc2 = principal(online, nfactors=4, rotate="none")
fact=principal(online,nfactors = 4,rotate = "varimax",sort = TRUE)
fact
View(fact)

loading = print(loadings(fact),digits = 2,cutoff = .4,sort = TRUE)
fa.diagram(fact);
```

R CODE FOR WORDCLOUD:

```
library(lubridate)
library(reshape2)
library(syuzhet)
library(NLP)
library(ggplot2)
library(tidyverse)
library(scales)
library(stringr)
library(dplyr)
library(tidyverse)
library(tidytext)
library(dplyr)
library(wordcloud)
library(sqldf)

docs=read.csv("Project.csv" ,header = F)
text=iconv(PS$experience)
sentiment=get_nrc_sentiment(text)
s=get_nrc_sentiment(text)
set.seed(1234)
barplot(colSums(s),las=2,col=rainbow(10),ylab = "count",xlab =
            "emotions",main="views and opinions of people")

get_sentiments('bing') %>%
  filter(sentiment %in% c("negative ","positive")) %>%
```

```
   count(sentiment)

tokens=data_frame(text=text)%>%
   unnest_tokens(word,text)



bing_word_count= tokens %>%
   inner_join(get_sentiments("bing")) %>%
   count(word,sentiment,sort=TRUE) %>%
   ungroup()
sqldf("SELECT sentiment,COUNT(sentiment) as count FROM bing_word_count
GROUP BY sentiment")
y=c("negative","positive")
x=c(131,143)
z=data.frame(
   y=c("negative","positive"),
   x=c(131,143)
barplot(x,names.arg=y,main="Graph of emotions",xlab="sentiment",ylab =
              "count",col=brewer.pal(8,"Spectral"))
set.seed(1)
tokens %>%
   inner_join(get_sentiments("bing")) %>%
   count(word,sentiment,sort=TRUE) %>%
   acast(word ~ sentiment,value.var = "n",fill = 0) %>%
   comparison.cloud(colors = c("red","green"),max.words = 100)
```

PYTHON CODES:

```
# importing all necessary modules
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv("project.csv")
comment_words = ''
stopwords = set(STOPWORDS)

# iterate through the csv file
for val in df.experience:

	# typecaste each val to string
	val = str(val)

	# split the value
	tokens = val.split()

	# Converts each token into lowercase
	for i in range(len(tokens)):
		tokens[i] = tokens[i].lower()

	comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 800, height = 800,
			background_color ='white',
			stopwords = stopwords,
			min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Python Code for Masking Image:

```
import pandas as pd
import matplotlib.pyplot as plt
import wordcloud
from PIL import Image
from wordcloud import WordCloud, STOPWORDS , ImageColorGenerator
import numpy as np
df = pd.read_csv("project.csv" )
comment_words = ' '
stopwords = set(STOPWORDS)
# iterate through the csv file
for val in df.experience:
    # typecaste each val to string
      val = str(val)
  # split the value
      tokens = val.split()
      # Converts each token into lowercase
      for i in range(len(tokens)):
            tokens[i] = tokens[i].lower()
      for words in tokens:
            comment_words = comment_words + words + ' '
image_file = Image.open("BO.png")
# create mask

df_mask = np.array(image_file, dtype='uint8' )
# generate word cloud
wc = WordCloud(width=400,height=200,background_color="white",
max_words=500,min_font_size=4, mask=df_mask,
stopwords=stopwords,contour_width=0.5, contour_color='black')
wc.generate(comment_words)
# plot the word cloud
image_colors = ImageColorGenerator(df_mask)
plt.figure(figsize=(25,14),facecolor=None, dpi=None)
title=plt.title("Online Learning Experience")
plt.imshow(wc.recolor(color_func=image_colors), interpolation="bilinear")
plt.imshow(wc)
plt.axis("off")
plt.show()
```

DECISION TREE R CODE:

```
#Based on various demographies to predict whether given person will continue online
#eduaction or not.
setwd("F:/PROJECT/Practice codes")
dir()
(decision_tree_data=read.csv("Decision_final.csv"))
library(rpart)
library(rpart.plot)
str(decision_tree_data)
(decision_tree_data$Gen_own=as.factor(decision_tree_data$Gen_own))
(decision_tree_data$Education=as.factor(decision_tree_data$Education))
(decision_tree_data$Income=as.factor(decision_tree_data$Income))
(decision_tree_data$indi_dev =as.factor(decision_tree_data$indi_dev ))
(decision_tree_data$con_ol =as.factor(decision_tree_data$con_ol ))

set.seed(1234)
ind=sample(2,nrow(decision_tree_data),replace=T,prob=c(0.8,0.2))
train=decision_tree_data[ind==1,]
test=decision_tree_data[ind==2,]
#Building the tree
(tree=rpart(con_ol~Education+Lec_hr+Powecut+Travel+indi_dev,data=decision_tree_dat
a,control=rpart.control(minsplit=5,minbucket=5,maxdepth=5)))
rpart.plot(tree,extra=4)
```

Naïve Bayes Algorithm (for effectiveness):

```
#The libraries needed
install.packages("naivebayes")
library(naivebayes)
install.packages("ggplot2")
library(ggplot2)
install.packages("pipeR")
library(dplyr)
install.packages("psych")
library(psych)

#Data
setwd("F:/PROJECT/Practice codes")
dir()
(data_naive=read.csv("Naive.csv"))
str(head(data_naive))
(data_naive$Gen_own=factor(data_naive$Gen_own,levels=c("Female","Male"),labels=c("0","1")))
(data_naive$Education=factor(data_naive$Education,levels=c("Primary (Play School - IV)","Secondary school (V - X)","Higher School (XI-XII)","Under Graduate","Post Graduate","Other"),labels=c("0","1","2","3","4","5")))
(data_naive$Income=factor(data_naive$Income,levels=c("Below 1,00,000","1,00,000-3,00,000","3,00,000-5,00,000","5,00,000-7,00,000","7,00,000-10,00,000","10,00,000 and above"),labels=c("0","1","2","3","4","5")))
(data_naive$Type_School=factor(data_naive$Type_School,levels=c("Government","Semi-Government","Private"),labels=c("0","1","2")))
(data_naive$Region=factor(data_naive$Region,levels=c("Urban","Semi-urban","Rural"),labels=c("0","1","2")))
(data_naive$effectiveness=as.factor(data_naive$effectiveness))

#The xtabs() function creates contingency tables in frequency-weighted format.
#Cross tabulation should have more than 5 values in each row and column.
xtabs(~effectiveness+Gen_own,data=data_naive)
xtabs(~effectiveness+Education,data=data_naive)
xtabs(~effectiveness+Income,data=data_naive)
xtabs(~effectiveness+Type_School,data=data_naive)
xtabs(~effectiveness+Region,data=data_naive)

#Visualisation
pairs.panels(data_naive[-8])
#All the predictors are uncorrelated that is they are independent of each other
```

```
#Box plots
data_naive %>%
   ggplot(aes(x=effectiveness,y=Lec_hr,fill=effectiveness)) +
   geom_boxplot() +
   ggtitle("Box plot for effectiveness and Lec_hr")
data_naive %>% ggplot(aes(x=Lec_hr,fill=effectiveness)) +
   geom_density(alpha=0.8,color='black') +
   ggtitle("Density Plot")
#There is significant amount of overlapping.

#Data Partition
set.seed(1234)
ind=sample(2,nrow(data_naive),replace=T,prob=c(0.8,0.2))
train=data_naive[ind==1,]
test=data_naive[ind==2,]

#Naive Bayes model
model=naive_bayes(effectiveness~.,data=train,laplace=1 , usekernel=T)    #used laplace
smoothing
model
train %>%
   filter(effectiveness=="Agree") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))
train %>%
   filter(effectiveness=="Disagree") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))
train %>%
   filter(effectiveness=="Neutral") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))

train %>%
   filter(effectiveness=="Agree") %>%
   summarise(mean(Powecut),sd(Powecut))
train %>%
   filter(effectiveness=="Disagree") %>%
   summarise(mean(Powecut),sd(Powecut))
train %>%
   filter(effectiveness=="Neutral") %>%
   summarise(mean(Powecut),sd(Powecut))
plot(model)
```

```
#Predictive model
p=predict(model)
head(cbind(p,train))
tail(cbind(p,train))

p=predict(model,train,type='prob')
head(cbind(p,train))

#confusion matrix for training
p1=predict(model,train)
tab1=table(p1,train$effectiveness)
1-sum(diag(tab1))/sum(tab1)    #misclassifiction

#confusion matrix for testing
p2=predict(model,test)
tab2=table(p2,test$effectiveness)
1-sum(diag(tab2))/sum(tab2)    #misclassifiction
```

Naïve Bayes Algorithm (for future preference):

```
#The libraries needed
install.packages("naivebayes")
library(naivebayes)
install.packages("ggplot2")
library(ggplot2)
install.packages("pipeR")
library(dplyr)
install.packages("psych")
library(psych)

#Importing Dataset
naive=read.csv("C:/Users/Omkar/OneDrive/Desktop/Data/naive.csv")
head(naive)

# Encoding categorical data
(naive$Gen_own=factor(naive$Gen_own,levels=c("Female","Male"),labels=c("0","1")))
(naive$Education=factor(naive$Education,levels=c("Primary (Play School - IV)","Secondary school (V - X)","Higher School (XI-XII)","Under Graduate","Post Graduate","Other"),labels=c("0","1","2","3","4","5")))
(naive$Region=factor(naive$Region,levels=c("Urban","Semi-urban","Rural"),labels=c("0","1","2")))
(naive$future_pref=as.factor(naive$future_pref))


#The xtabs() function creates contingency tables in frequency-weighted format.
#Cross tabulation should have more than 5 values in each row and column.
xtabs(~future_pref+Gen_own,data=naive)
xtabs(~future_pref+Education,data=naive)
xtabs(~future_pref+Lec_hr,data=naive)
xtabs(~future_pref+Region,data=naive)

#Visualisation
pairs.panels(naive[-5])
#All the predictors are uncorrelated that is they are independent of each other

#Box plots
naive %>%
   ggplot(aes(x=future_pref,y=Lec_hr,fill=future_pref)) +
   geom_boxplot() +
   ggtitle("Box plot for future_pref and Lec_hr")
```

```
naive %>% ggplot(aes(x=Lec_hr,fill=future_pref)) +
   geom_density(alpha=0.8,color='black') +
   ggtitle("Density Plot")
#There is significant amount of overlapping.


#Data Partition
set.seed(1234)
ind=sample(2,nrow(naive),replace=T,prob=c(0.8,0.2))
train=naive[ind==1,]
test=naive[ind==2,]

#Naive Bayes model
model=naive_bayes(future_pref~.,data=train,laplace=1  ,  usekernel=T)    #used  laplace
smoothing
model
train %>%
   filter(future_pref=="1") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))
train %>%
   filter(future_pref=="2") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))
train %>%
   filter(future_pref=="3") %>%
   summarise(mean(Lec_hr),sd(Lec_hr))
plot(model)

#Predictive model
p=predict(model)
head(cbind(p,train))
tail(cbind(p,train))
p=predict(model,train,type='prob')
tail(cbind(p,train))

#confusion matrix for training
p1=predict(model,train)
tab1=table(p1,train$future_pref)
1-sum(diag(tab1))/sum(tab1)     #misclassifiction

#confusion matrix for testing
p2=predict(model,test)
tab2=table(p2,test$future_pref)
1-sum(diag(tab2))/sum(tab2)     #misclassifiction
```

Bibliography

1. Blogs for introductory part

https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

https://www.google.com/amp/s/elearningindustry.com/advantages-and-disadvantages-online-learning/amp

https://www.google.com/amp/s/www.downtoearth.org.in/blog/economy/amp/covid-19-how-viable-is-online-education--73487

2. Blogs for Techniques

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

https://www.wallstreetmojo.com/pareto-analysis/

https://www.google.com/search?q=sentiment+analysis+analytics+vidhya&oq=sentiment+analysis+ana&aqs=chrome.1.69i57j0l3j0i22i30.32502j1j4&client=ms-android-huawei-rev1&sourceid=chrome-mobile&ie=UTF-8

https://www.journaldev.com/49875/factor-analysis-in-r

https://statistics.berkeley.edu/computing/r-t-tests#:~:text=One%20of%20the%20most%20common,normal%20distributions%20with%20equal%20variances.&text=The%20function%20t.,R%20for%20performing%20t-tests

https://www.google.com/amp/s/pestleanalysis.com/what-is-swoc-analysis/

3. YouTube videos

https://youtu.be/gSy7KPfaA8g

https://youtu.be/RLjSQdcg8AM

https://youtu.be/HmEPCEXn-ZM

https://youtu.be/-SeyrC4yZF4

https://youtu.be/otoXeVPhT7Q

https://youtu.be/Ilf1XR-K3ps

https://youtu.be/q0ckcKsSPXU

Questionnaire:

ONLINE EDUCATION IN TIMES OF COVID-19

The spread of COVID-19 has led to the closure of educational institutions all over the world. Such closure accelerated the development of the online learning environments within those institutions so that learning would not be disrupted. The coronavirus pandemic has tested the readiness of centers to deal with a crisis that requires online and remote measures. Many were not prepared, but it is important to review the reasons for offering students online classes, which go beyond periods of confinement.

Therefore, I thought of conducting research to identify influencing factors for the Effectiveness of Online Learning during COVID-19 Pandemic.

To further understanding the scope for Blended Education. * Required

1.Are you a parent or student? *

If you are a parent fill the information of your child.

Mark only one oval.

- ◯ Parent *Skip to question 3*
- ◯ Student*Skip to question 2*

Your gender.

Gender *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Other

Skip to question 4

Gender of your child.

Gender *

If you are a parent fill the information of your child.

Mark only one oval.

- Male
- Female Other
-

Skip to question 4 Section 1

2.Education (Currently Pursuing). * *Mark only one oval.*

- Primary (Play School - IV)
- Secondary school (V - X)
- Higher School (XI-XII)
- Under Graduate
- Post Graduate
- Other

3.Type of School or University or Institute *

Mark only one oval.

- Government
- Semi-Government
- Private

4.Annual family income *

Mark only one oval.

- Below 1,00,000
- 1,00,000-3,00,000
- 3,00,000-5,00,000
- 5,00,000-7,00,000
- 7,00,000-10,00,000
- 10,00,000 and above

5.Number of students attending online lectures in your family ? *

6.Number of available devices in your family ? *

7.Region *

Mark only one oval.

- Urban
- Semi-urban
- Rural

8.Does your locality face power-cut, if yes how many hours daily ? *

If no put 0 & If in minutes for instance 30min=0.30 & 1 hour= 1.

9.Source of internet. *

Check all that apply.

- WiFi
- Mobile data Hotspot
- Dongle
-

10.Mode of accessing online lectures. *

You can select more than one option.

Check all that apply. laptop

- Mobile phone
- Desktop
- Tablet
-

11.Which software/platform you use for attending online lectures ? *

You can select more than one option *Check all that apply.*

- [ ] Zoom
- [ ] Google meet
- [ ] MS Teams
- [ ] Other

12.How many hours you used to spent daily on self study while attending regular offline lectures? *

13.How many hours you spent daily on self study while attending online lectures? *

14.How many hours you spent daily in online learning ?(apart from your social media uses) *

In Exact Hours. Eg= 1 ,2 ...

15.Number of hours you used to spend travelling to college/school in non covid times ? *

In Exact Hours. Eg= 1 ,2 ... , If in minutes for instance 30min=0.30 & 1 hour= 1

16.Have you faced any these problems while using online lectures? *

1=Minimum, 5=Maximum

Mark only one oval per row.

	1	2	3	4	5
Stress	◯	◯	◯	◯	◯
Lack of sleep	◯	◯	◯	◯	◯
Reduced thinking power	◯	◯	◯	◯	◯
Loss of interest	◯	◯	◯	◯	◯
Obesity	◯	◯	◯	◯	◯
Depression	◯	◯	◯	◯	◯
Anxiety	◯	◯	◯	◯	◯

17.Did you buy any new device for online learning after lockdown was imposed in the India ? *

Mark only one oval.

◯ Yes

◯ No

18.Do you have your individual device to attend your lectures? *

Mark only one oval.

◯ Yes

◯ No

19.Did your School/College ask for immediate payment of fees? *

Mark only one oval.

◯ Yes

◯ No

20.Do you think online learning provided good value of money? *

Mark only one oval.

Yes No

Effectiveness.

21.According to you is online education effective to you ? *

Mark only one oval.

Agree

Neutral

Disagree

- 5= Strongly Agree
- 4= Agree
- 3= Neutral
- 2= Disagree
- 1= Strongly Disagree

22.Likert Scale

● 5= Strongly Agree ● 4= Agree ● 3= Neutral ● 2= Disagree ● 1= Strongly Disagree *

Mark only one oval per row.

	5	4	3	2	1
Online lecture materials provided to you are helpful.	◯	◯	◯	◯	◯
Social media is a distraction while attending online lectures.	◯	◯	◯	◯	◯
Is language a barrier while attending online lectures	◯	◯	◯	◯	◯
Feeling hesitate to ask questions/doubts in online lectures.	◯	◯	◯	◯	◯
Lack of direct contact with other students/colleague/friends/teachers.	◯	◯	◯	◯	◯
Insufficient practical knowledge through online education.	◯	◯	◯	◯	◯
Indiscipline and disturbance from outsiders in online lecture.	◯	◯	◯	◯	◯
Technical glitches during online lectures affects your attendance and concentration.	◯	◯	◯	◯	◯
Home environment is suitable for participating online lectures.	◯	◯	◯	◯	◯
Feeling frustrated when unable to submit your assignments or paper on time.	◯	◯	◯	◯	◯
Excessive use of mobiles or laptops led to eye pain, ear pain, headache.	◯	◯	◯	◯	◯
Sitting continuously for online lectures led to back pain and neck pain.	◯	◯	◯	◯	◯
Piled up assignments and too many upcoming tests depresses you.	◯	◯	◯	◯	◯
Network issues faced while writing	◯	◯	◯	◯	◯

online exams stresses you which directly affects your score.					
Does power cut affect online education.	⬭	⬭	⬭	⬭	⬭
Does limited availability of data restrict online education.	⬭	⬭	⬭	⬭	⬭
Does ones financial background restricts online education.	⬭	⬭	⬭	⬭	⬭
Online education has made todays generation tech-savvy.	⬭	⬭	⬭	⬭	⬭
Online lectures are flexible and comfortable to attend.	⬭	⬭	⬭	⬭	⬭
Reduced travelling time and money spent on travelling expenses.	⬭	⬭	⬭	⬭	⬭
Online education is giving you more time for other activities.	⬭	⬭	⬭	⬭	⬭
In online education notes are available at just one click.	⬭	⬭	⬭	⬭	⬭
Online lectures are accessible at any time because of recording feature.	⬭	⬭	⬭	⬭	⬭
Getting proper meals at proper time has improved ones physical health.	⬭	⬭	⬭	⬭	⬭
Time management and increased self motivation.	⬭	⬭	⬭	⬭	⬭
Online education was perfect alternative in covid time instead of waste of academic year.	⬭	⬭	⬭	⬭	⬭

- 1= Worse
- 2= Bad
- 3= Neutral
- 4= Good
- 5= Best

23.Rating Scale Questions.

● 1= Worse ● 2= Bad ● 3= Neutral ● 4= Good ● 5= Best *

Mark only one oval per row.

	1	2	3	4	5
Doubt solving/query resolution in online lectures.	◯	◯	◯	◯	◯
Syllabus coverage in online lectures.	◯	◯	◯	◯	◯
Concept clearance in online lectures.	◯	◯	◯	◯	◯

● 1= Worse ● 2= Bad ● 3= Neutral ● 4= Good ● 5= Best *

Mark only one oval per row.

	1	2	3	4	5
Doubt solving/query resolution in regular physical lectures.	◯	◯	◯	◯	◯
Syllabus coverage in regular physical lectures.	◯	◯	◯	◯	◯
Concept clearance in regular physical lectures.	◯	◯	◯	◯	◯

Describe in Brief.

Open Ended Questions.

24.Describe your online learning experience in few words. *

25.Will you prefer to continue with this online lectures once this pandemic is over and why? *

Mark only one oval.

- Yes *Skip to question 29*
- No*Skip to question 29*

Why ? *

26.What improvements should be made for the betterment of online education? *

27.If given a choice in future which will you prefer? *

Mark only one oval.

- Online
- Offline
- Blended (Mixture of online and offline)