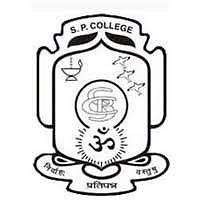
**SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE**



**SIR PARASHURAMBHAU COLLEGE**

**Tilak Road, Pune-30**



**DEPARTMENT OF STATISTICS**

# Project report on

# **Impact of Social Media in different virtues**

Submitted by

Ms.Prutha Nishikant Patil

Ms.Chinmai Santosh Agre

Mr.Mukund Ramesh Tambe

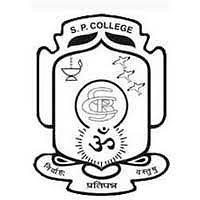
Ms.Kadambari Bhagvat Kharche

Ms. Gauri Satish Shivale

Ms. Aditi Nileshkumar Pawar

**( T. Y. B.SC – STATISTICS )**

**2023-24**



Shikshan Prasaraka Mandali’s

SIR PARASHURAMBHAU COLLEGE

TILAK ROAD, PUNE-30

# CERTIFICATE

This is to certify that student of T.Y.B.Sc. **Ms. Prutha Nishikant Patil (02210190)**, **Ms.Chinmai Santosh Agre (02210005) ,Mr. Mukund Ramesh Tambe (02210284) ,Ms. Gauri Satish Shivale (02210088),Ms. Kadambari Bhagvat Kharche(02210133),Ms. Aditi Nileshkumar Pawar (02210192)** have completed their project on **Impact of Social Media on different virtues** under the guidance of Mr. Santosh Kamble during Bsc in academic Year 2023-24.

Mr. Santosh Kamble Mrs. Pallavi Turkunde

(Asst. prof. Statistics Dept. Project Guide) ( HOD, Statistics Dept.)

**Internal Examiner External Examiner**

**Date: 29/04/2024**

**Place: Pune**

## 

## **Acknowledgement**

We would like thank the Principal of our college for permitting us to use all the facilities available in the institute for our project work.

We would like to thank the Head of department of Statistics for allowing us to use computer laboratory for our project. We would like to thank my project guide Mr. Santosh Kamble sir for his guidance throughout our project. We express thankfulness to other teaching faculty & non-teaching staff for their support & guidance.

We are thankful to all those who helped me directly or indirectly in completion of our Project.

Ms.Prutha N. Patil Ms.Chinmai S. Agre

(T.Y.B.SC Statistics 2023-24) (T.Y.B.SC Statistics 2023-24)

Mr.Mukund R. Tambe Ms. Kadambari B. Kharche

(T.Y.B.SC Statistics 2023-24) (T.Y.B.SC Statistics 2023-24)

Ms. Aditi N. Pawar Ms. Gauri S. Shivale (T.Y.B.SC Statistics 2023-24) (T.Y.B.SC Statistics 2023-24)

INDEX

1.INTRODUCTION

2.BACKGROUND

3.OBJECTIVE

4.DATA SUMMARY

5.LIMITATIONS

6.METHODOLOGY

7.THEORY

8.EXPLORATORY DATA ANALYSIS

9.FITTING OF LOGISTIC REGRESSION MODEL

10.NON-PARAMETRIC TESTS

11.APPLICATIONS OF M.L. ALGORITHMS

* 1. KNN
  2. NAÏVE BAYES ALGORITHM

11.3 K MEANS ALGORITHM

12.COMPARISIONS OF MODEL

13.CONCLUSION

14.REFERENCE

**INTRODUCTION**

Social media refers to a broad category of online platforms and technologies that enable users to create, share, and exchange content, as well as participate in social networking. These platforms facilitate interaction among users who may create profiles, post text updates, share photos, videos, and links, and engage in conversations through comments, likes, and messages. Social media has transformed the way people communicate, access information, and even influence public opinion.

Examples of social media platforms include Facebook, Twitter, Instagram, LinkedIn, YouTube and WhatsApp each serving different purposes and demographics but all emphasizing user interaction and content sharing.

Around 95% of young people between the ages of 13–17 use at least one social media platform. More than one-third of those report using social media nearly constantly.

The most common social media platforms these young people use are Facebook, Instagram, and Snapchat. Our project discusses the benefits and risks of social media use. It also gives tips for how to manage the risks and find balance.

**BACKGROUND**

* What is social Media?

Social media refers to websites and applications designed for users to create and share content or to participate in social networking. It includes a wide range of online, interactive platforms where individuals, groups, and organizations can share ideas, discuss interests, post photos, videos, and updates, and engage with others.

Social media is digital technology that allows the sharing of ideas and information, including text and visuals, through virtual networks and communities.

* What factors get affected by social media?

Social media exists mainly because of a few simple reasons:

1. Connection: People want to stay in touch with friends and family, meet new people, and share moments of their lives. Social media makes all of this easy and accessible from anywhere.

2. Sharing: It gives everyone a space to share their thoughts, photos, videos, and ideas with others. Whether it's a personal story or a creative piece of work, social media is a platform where anyone can broadcast their content.

3. Information: People use social media to stay informed. Whether it’s news about the world, updates from friends, or learning about events, it serves as a quick way to get information.

4. Entertainment: Social media is also a source of entertainment. From funny videos to interesting articles, it has a lot of content that keeps users engaged and entertained.

5. Business and Marketing: For businesses, social media is a tool to reach and engage with customers. It helps in marketing products, getting customer feedback, and building brand loyalty.

In simple terms, social media is popular because it fulfills basic human needs: the need to connect with others, the need to share and express ourselves, the need to be informed, and the need for entertainment.

* Positive and Negative Impact Of Social Media

The impact of social media on society, culture, and individual lives is profound and multifaceted. Here are some key aspects to consider:

Positive Impacts:

Connectivity: Social media platforms enable people to connect with friends, family, and communities worldwide. It has made communication more accessible and immediate.

Information Sharing: News and information spread rapidly through social media, keeping people informed about current events, trends, and developments globally.

Business and Marketing: Social media provides businesses with a powerful platform for marketing and advertising. It allows for targeted advertising, customer engagement, and brand building.

Education: Many educational institutions and teachers use social media to share resources, conduct online classes, and facilitate learning beyond traditional classrooms.

Social Movements: Social media has been instrumental in mobilizing people for social and political causes, providing a platform for activism and advocacy.

Negative Impacts:

Privacy Concerns: Users often share personal information unknowingly, leading to privacy issues and potential misuse of data by third parties.

Mental Health: Excessive use of social media can contribute to mental health issues like anxiety, depression, and low self-esteem, especially among young people.

Fake News and Misinformation: Social media platforms can be a breeding ground for fake news, misinformation, and propaganda, which can have serious consequences on public opinion and democratic processes.

Cyberbullying: Online harassment and bullying have become prevalent on social media platforms, affecting many users, particularly teenagers and young adults.

Addiction: The addictive nature of social media can lead to excessive screen time, sleep disturbances, and a decrease in real-world social interactions.

Societal and Cultural Impacts:

Cultural Exchange: Social media facilitates cultural exchange and understanding by connecting people from different backgrounds and regions.

Influence on Youth: Social media plays a significant role in shaping the opinions, values, and behaviours of young people, often influencing fashion trends, language, and social norms.

Political Influence: Social media has transformed political communication, campaigning, and mobilization, influencing elections and public opinion in unprecedented ways.

**OBJECTIVE**

The aim of our project to study the data of people with their opinion on Social media platform and understand the various underlying factor that may play role in their causation.

Through this project we hope to find:

1. Links between different factors like
   1. Physical health
   2. Mental health
   3. Education
   4. Career opportunities
   5. Commercial effect
   6. Data Privacy
   7. Communication Skills
   8. Political Decisions
2. Fitting a Logistic Regression Model, degerming whether a person needs a health check-up.
3. Checking a randomness of data.

**DATA SUMMARY**





**COLUMN DESCRIPTION**

• Gender: gender of the respondent

• Age

• Earning: whether a person is earning by himself/herself?

•Time spent: Average number of hours spent on social media platforms

•Sleep difficulty : rating sleep difficulty issues

•Physical discomfort: rating if they feel physical discomfort after using social media

•Body posture: effect of social media on body posture

•Exhausted feeling: rating if a person feels exhausted after use of social media

•Life comparison: rating if one compares their life with others

•Break: rating if a person prefers a life with or without social media

•Online motivation: rating if a person feels motivated after watching motivational content

•Promotion: rating if social media is used excessively for business promotions.

•Advertisement: rating if people purchase products based on social image

•Understanding concepts: rating the level of understanding through social content

•Misinformation: analysing the chances of misinformation through social media

•Hiring decision: rating if the people feel that companies depend on social media platforms for hiring decisions.

•Job search: rating if people use social media for job search

•Personal branding: rating if people use social media to promote themselves

•Personal info: rating if people feel that their personal information may be leaked or misused.

•Scams: rating the level of cybercrimes happening due to social media

•Communication: rating if social media use affects the

communication skills

•Political issue: rating if people feel social platforms are somewhat responsible for political issues

•Political decisions: rating if social media platforms affect the political decisions in a way.

**LIMITATIONS**

The data for this project was not collected manually. As a

result, the distribution of responses received is slightly

positively skewed in accordance to the ‘Age’ variable.

Out of 460 respondents, 342 of them lie between the

ages of 18-25. The remaining respondents are spread

over the ages of below 18 and above 25.

Hence the data may be biased. Also, as it was a survey with a Likert scale responses, perception of people to choose the neutral responses is very high compared to other responses.

**METHODOLOGY**

The primary data for this project was collected using a Google forms survey from the students of Sir Parashurambhau College, Pune.

A questionnaire was prepared about their age, gender, Mental health, Physical health, Education, Career opportunities, Cybercrime, Data privacy, commercial, communication and political participation.

Results related to social Media were recorded.

A total number of 460 students participated in the survey. The data collected was with voluntary participation and was used for research purposes only.

The link to the google form survey is given below:

<https://forms.gle/8eWFjrcnBQQXd2GX7>

The QR to the google form survey is given below:



**Data cleaning:**

After collecting the survey data, the null and inappropriate values were removed.

**Data manipulation:**

The survey data responses were in the textual and Likert scale form. They were converted into numerical Likert scale format, and binary responses as required.

The question wise scores and individual scores were also calculated for the analysis purpose.

**Scores calculation:**

Understanding Likert scales is best done through an example. Let’s assume the survey is administered to two people, and they answered the following way:

Person 1: Q1. Strongly agree, Q2. Disagree, and Q3. Strongly agree.

Person 2: Q1. Strongly Disagree, Q2. Neutral, Q3. Strongly Disagree

With these answers, we can then derive a number that describes what these two customers feel about this shop. We can call this a sentiment score:

Step 1: For each question on the questionnaire, calculate the total number of responses for each sentiment level (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).

Multiply the numerical value of each sentiment level by the amount of respondents. Let’s start with Question 1:

Step 2: Add the totals, and divide by the total number of respondents:

1 + 0 + 0 + 0 + 5 = 6 / 2 respondents = 3.

As the numerical value for the “Neutral” sentiment level is 3, this means that respondents generally feel neutral about item availability at the store.

We repeat steps 1 and 2 above for question 2:

If we calculate the average of the 3 questions, we get a sentiment score for the entire questionnaire:

Question 1 (3) + Question 2 (2.5) + Question 3 (3) = 8.5 / 3 questions = 2.83

**SAMPLING THEORY**

If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample, and in turn, the sample mean will serve as a good estimator of the population mean. Thus, if the population is homogeneous with respect to the characteristic under study, then the sample drawn through simple random sampling is expected to provide a representative sample. Moreover, the variance of the sample mean not only depends on the sample size and sampling fraction but also on the population variance. In order to increase the precision of an estimator, we need to use a sampling scheme which can reduce the heterogeneity in the population. If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is a stratified sampling.

**The basic idea behind the stratified sampling is to**

• divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and

• heterogeneous with respect to the characteristic under study between/among the subpopulations. Such subpopulations are termed as strata.

• Treat each subpopulation as a separate population and draw a sample by SRS from each stratum. [Note: ‘Stratum’ is singular and ‘strata’ is plural].

Notations: We use the following symbols and notations:

N: Population size

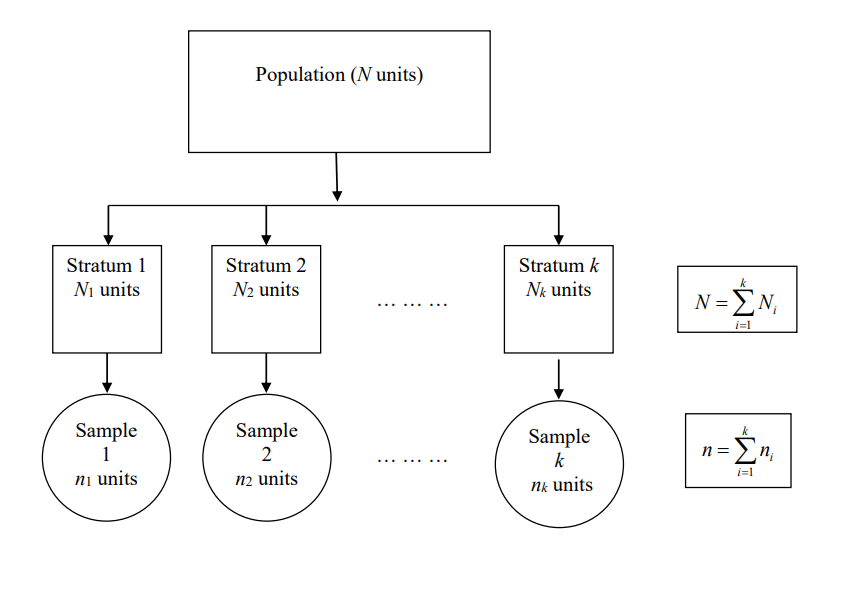
k: Number of strata

Ni: Number of sampling units in i th strata

N: ∑ Ni

ni : Number of sampling units to be drawn from i th stratum.

n= ∑ni : Total sample size.



**Procedure of stratified sampling**   
Divide the population of N units into k strata. Let the i th stratum has 1 N i k , 1,2,..., = number of units.

• Strata are constructed such that they are non-overlapping and homogeneous with respect to the characteristic under study such that Ni = ∑N

• Draw a sample of size i n from i th ( 1,2,..., ) i k = stratum using SRS (preferably WOR) independently from each stratum.

• All the sampling units drawn from each stratum will constitute a stratified sample of size n=∑ni

**Advantages of stratified sampling**

1.Data of known precision may be required for certain parts of the population. This can be accomplished with a more careful investigation to a few strata. Example: In order to know the direct impact of the hike in petrol prices, the population can be divided into strata like lower income group, middle-income group and higher income group. Obviously, the higher income group is more affected than the lower-income group. So more careful investigation can be made in the higher income group strata.

2. In the case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units, which in turn reduces the sampling variability.

3. When estimates are required not only for the population but also for the subpopulations, then the stratified sampling is helpful.

4. When the sampling frame for subpopulations is more easily available than the sampling frame for the whole population, then stratified sampling is helpful.

With reference to the project, the values are as follows;

N = total population= 12000 (aprox.)

N1 = no. of student with age below 18 in the population= 1512

N2 = no. of student age 18 – 25 in the population = 8922

N3= no. of student above 25  in the population = 1566

k = 3

n = sample size= 460

n1= no. of in the sample = 58

n2=no. of student age 18 – 25 in the sample = 342

n3= no. of student above 25  in the sample = 60

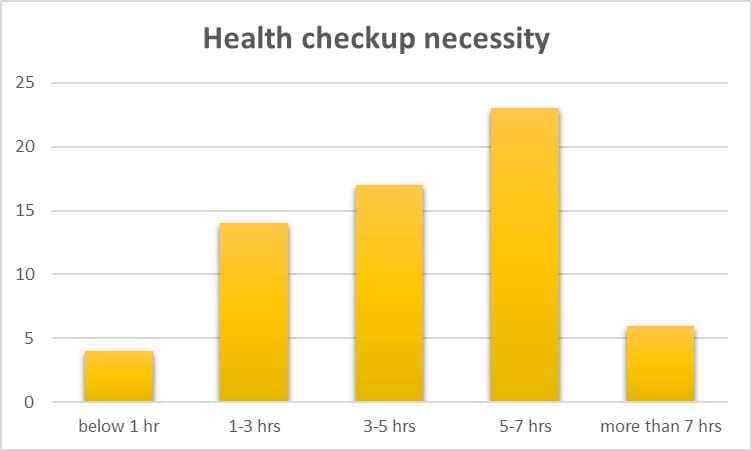
EXPLORATORY DATA ANALYSIS

1. EXPLORATORY DATA ANALYSIS

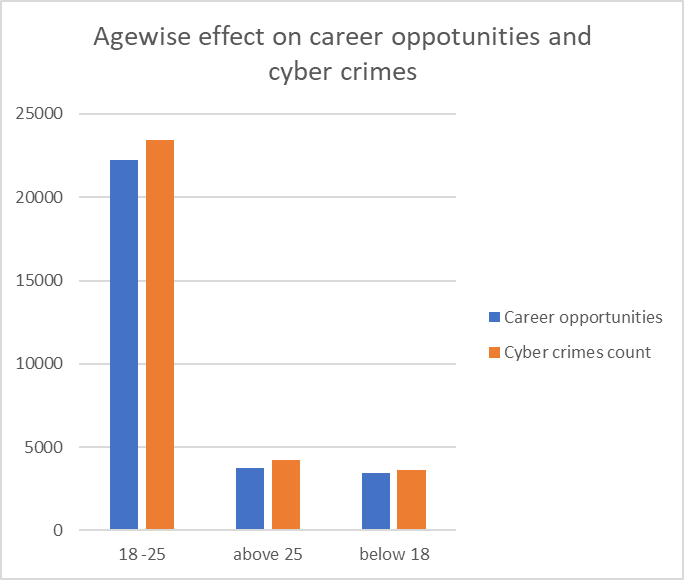
* What is the exploratory data analysis?
* The first step in any Data Science project is to understand the data that is given to us.
* Exploratory data analysis (EDA) is the process of examining and analysing data to understand its characteristics, uncover patterns and identify any anomalies or outliers.
* EDA is a crucial step in the data analysis process as it helps to inform the selection of appropriate statistical models, identify the data pre-processing requirements, and generate hypothesisfor further analysis.
* The main goal of EDA is to gain insight into the structure and features of the data, which includes identifying the distributionsof variables.
* There are few standard questions you can ask in order to get the understanding:
* What is the size of the dataset?
* What is the time range of the data?
* What is the meaning of each column and its values in the dataset? (can get this information from Data Dictionary)
* What is the distribution and unique values for each column?
  + - (univariate analysis)
* What is the distribution and unique values for somecombinations of columns? (bivariate & multivariate analysis)

If you try to answer these questions with the help of data, then you would get a good understanding of the data which can be leveraged in Model Building part of the project.

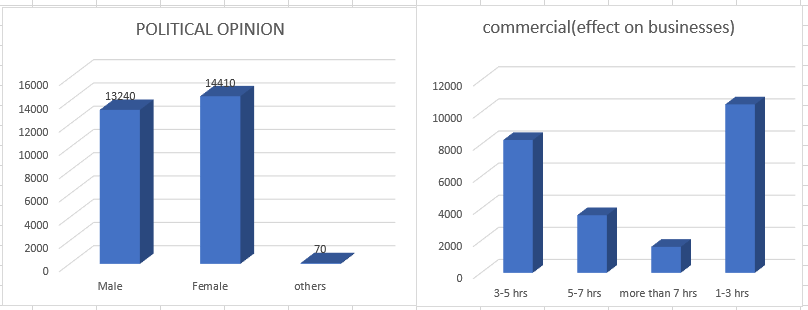
From the above graph, we can say that in most of the cases,i.e. for many factors like commercial purchases, data privacy,etc, most of the people agree with the fact that social media has a effect on them.



from the above graph we can observe that the people with screen time more than 3 hrs need to have a health check-up i.e. scree time less than 3hrs is safe for our body and mental health .

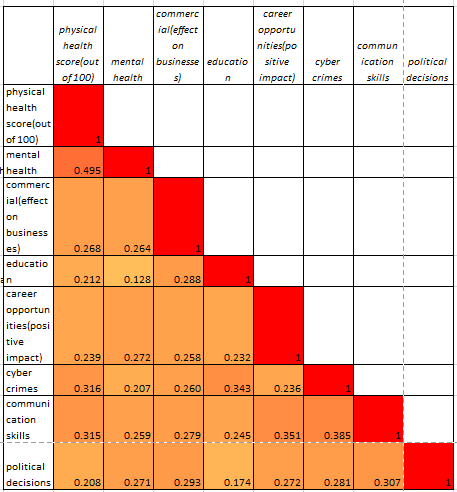


We can observe that the people of age between 18-25 years are teenagers and working youth. They seem to have more benefit with respect to career opportunities rather than other age groups. Also, there is increase in the number of crimes due to social media and youth is more prone to the cyber crimes.

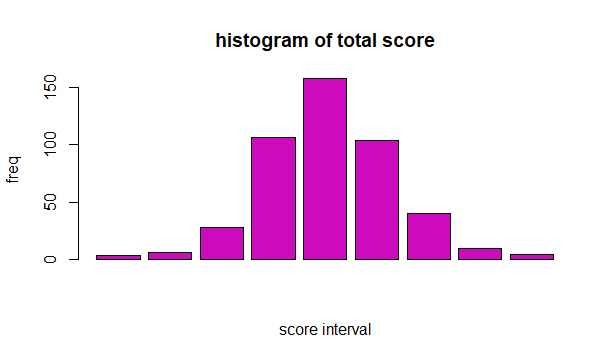


Regarding the political factor,we can say that more females agree to the fact that social media affects the political opinions.

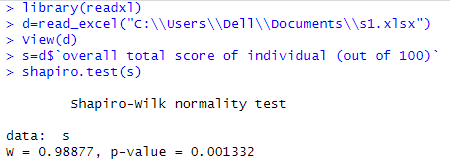
**CORRELATION PLOT BETWEEN DIFFERENT FACTORS AFFECTED BY SOCIAL MEDIA**



There is weak positive correlation among almost all the factors.



We can see that the data seems to follow Normal Distribution. But , when we perform the normality test (shapiro.test), we conclude that the data is not normally distributed.



**FITTING OF LOGISTIC REGRESSION MODEL.**

LOGISTIC REGRESSION MODEL:

In Statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

Formula:

П(X)= (exp (β 0 + β 1 x 1 + β 2 x 2 + …. .βnXn)) / (1+ exp(β 0 + β 1 x 1 + β 2 x 2 + …..βnXn))

П(X) = 1 Probability of success & П(X)= 0 Probability of failure

Where, when the value of П(X) is greater than 0.5 is considered as success and vice versa.An **odds ratio** (**OR**) is a [statistic](https://en.wikipedia.org/wiki/Statistic) that quantifies the strength of the [association](https://en.wikipedia.org/wiki/Association_(statistics)) between

two events, A and B. The odds ratio is defined as the ratio of the [odds](https://en.wikipedia.org/wiki/Odds) of A in the presence of

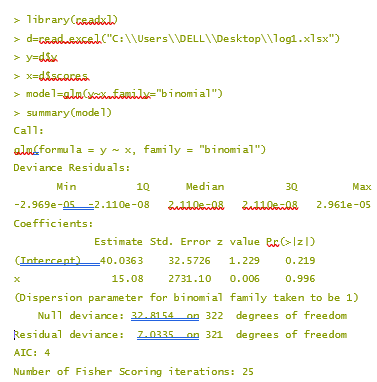
B and the odds of A in the absence of B, or equivalently (due to [symmetry](https://en.wikipedia.org/wiki/Odds_ratio#Symmetry)), the ratio of the

odds of B in the presence of A and the odds of B in the absence of A.

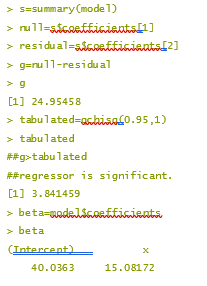
Here, X= predictor variable=physical and mental health score of an individual

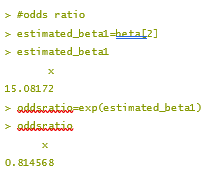
Binary Response variable Y: a person requires health checkup or not , (if score is >=70 , then we recommend health checkup ,otherwise not .)

Firstly, we divide the data into train -test (70%-30%) We fit logistic model on train data(70% data) and estimate the coefficients (beta0 ,beta1,etc.)



Then, test the model accuracy on the test data(remaining 30% data).





Here, the odds ratio value is 0.814568 and model accuracy is 79.56204.

Conclusion:We can conclude that the chances of requiring a health checkup increases by 81.45 % per unit increase in the health score of individual.

**NON- PARAMETRIC TESTS**

Non-parametric tests are experiments that do not require the underlying population for assumptions. It does not rely on any data referring to any particular parametric group of probability distributions. Non-parametric methods are also called distribution-free tests since they do not have any underlying population.

1. **Run test**

Run : Given an ordered sequence of one or more types of symbols a run is defined to be a succession of one or more type of symbols which are followed or preceded by different symbols or no symbols at all.

To test,

H0: The sequence of observation is random

against

**H1:** The sequence of observation is not random

R: Total number of runs

The Distribution of R is given by

For even

f(R=r0) = {2\* (n1-1)C(r/2-1) \* (n2-1)C(r/2-1) / (n1+n2)Cn1}

for odd

( (n1-1)C(r-1/2) \* (n2-1)C(r-3/2) )( (n1-1)C(r-3/2) \* (n2-1)C(r-5/2))/ (n1+n2)Cn1

n= n1+n2 total observation

n1= no. of observation of type 1

n2= no. of observation of type 2

r0= r1+r2

r1=no. of runs of observation of type 1

r2= no. of runs of observation of type 2

reject H0 , if R<= R1 or R >= R2

where R1 is lower critical value

R2 is upper critical value

R1 and R2 can be obtained from the statistical table .

HYPOTHESIS:

H0 : The arrangement is random.

V/s

H1 : The arrangement is not random.

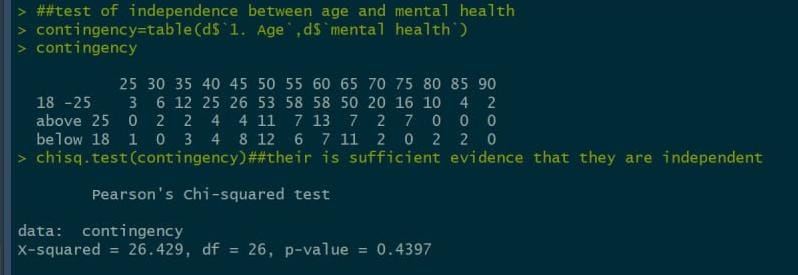


1. **Chi- square test of independence**

The Chi-square test of independence checks whether two variables are likely to be related or not. We have counts for two categorical or nominal variables. We also have an idea that the two variables are not related. The test gives us a way to decide if our idea is plausible or not.

H0: (null hypothesis) The two variables are independent.

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



Since the p-value (0.4397) of the test is not less than 0.05, we fail to reject the null hypothesis. Two variables are independent.

Similarly test was conducted for other variables.

**NAÏVE BAYES**

Naïve Bayes is a straightforward and also fast algorithm for classification. Its working process is based on Bayes theorem. It is represented below:

P(X|Y) =P (Y|X) \* P(X)P(Y)

Naive Bayes assumes that the presence of a particular feature in a class is independent of the presence of any other feature. This is a strong assumption and hence the term "naive".

Naive Bayes calculates the probability of each class given the input features using Bayes' theorem and the assumption of feature independence. It selects the class with the highest probability as the prediction.

Confusion Matrix

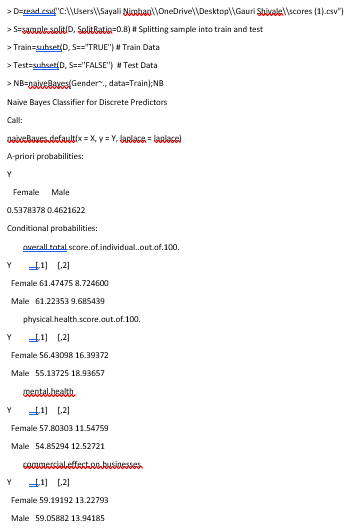
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows visualization of the performance of an algorithm, particularly in binary classification problems, where there are two possible outcomes. Here’s how a confusion matrix is structured:

True Positive (TP): These are the cases where the model correctly predicts the positive class.

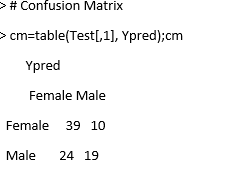
True Negative (TN): These are the cases where the model correctly predicts the negative class.

False Positive (FP): Also known as a Type I error, these are the cases where the model incorrectly predicts the positive class when it's actually negative.

False Negative (FN): Also known as a Type II error, these are the cases where the model incorrectly predicts the negative class when it's actually positive.



The Conditional probability for each feature or variable is created by model separately. The prior probabilities are also calculated which indicates the distribution of our data.



So, 39 Females are correctly classified as Female. Out of 43 Males, 19 Males are correctly classified as Male, and 24 are classified as Female.



The model achieved 66.04% accuracy with a p-value of less than 1.

**KNN**

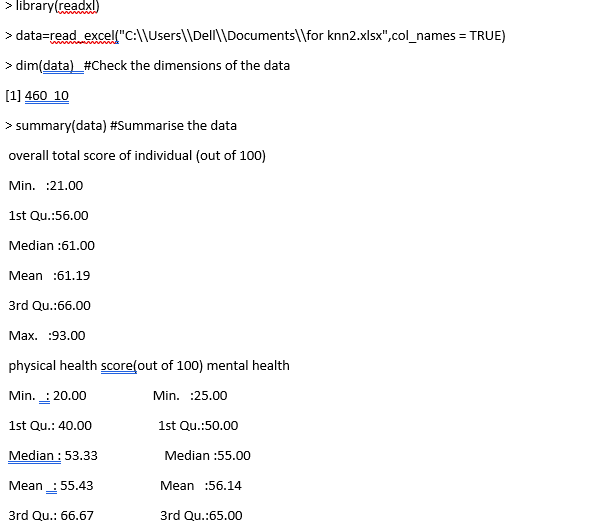
K-Nearest Neighbour (KNN)

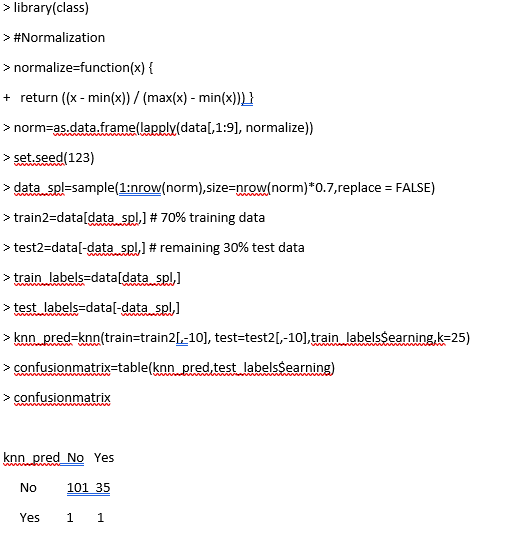
KNN is a supervised learning technique that means the label of the data is identified before making predictions. Clustering and regression are two purposes to use it.

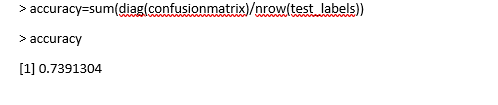
K represents a numerical value for the nearest neighbours. KNN algorithm does not have a training phase.

Predictions are made based on the Euclidean distance to k-nearest neighbours.

The label is classified according to the nearest neighbour to the class labels of its neighbours.



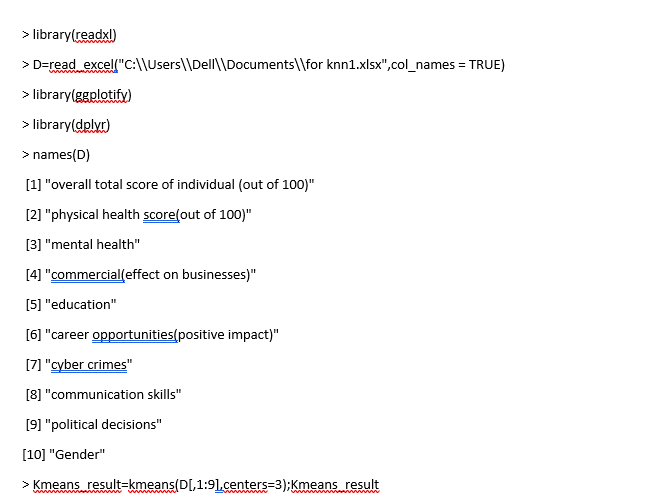


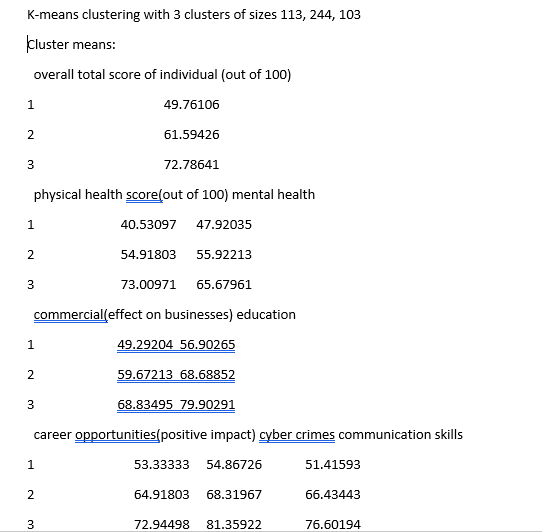


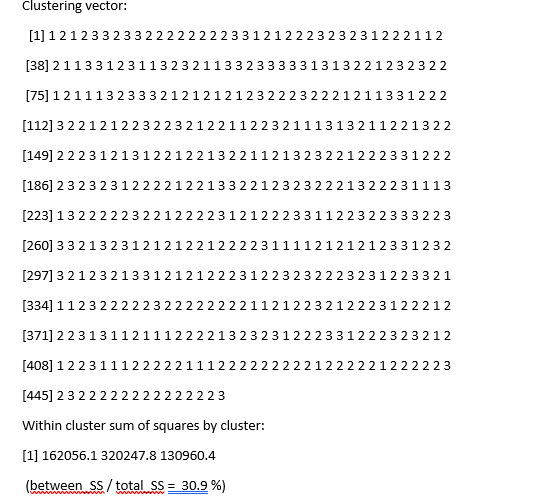
**K-MEANS**

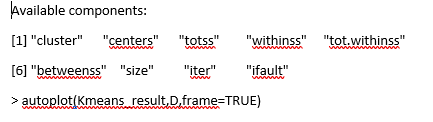
**K-means clustering** is one of the most popular and widely used clustering algorithms. It's a partition-based clustering algorithm that aims to divide a dataset into k clusters, where each data point belongs to the cluster with the nearest mean (centroid).

Selecting the appropriate number of clusters (k) is crucial in k-means clustering. Common methods for determining k include the elbow method, silhouette score, and gap statistics. For each data point, k-means assigns it to the cluster with the nearest centroid. Analyzing the cluster membership can help understand which data points are similar to each other and how they group together.





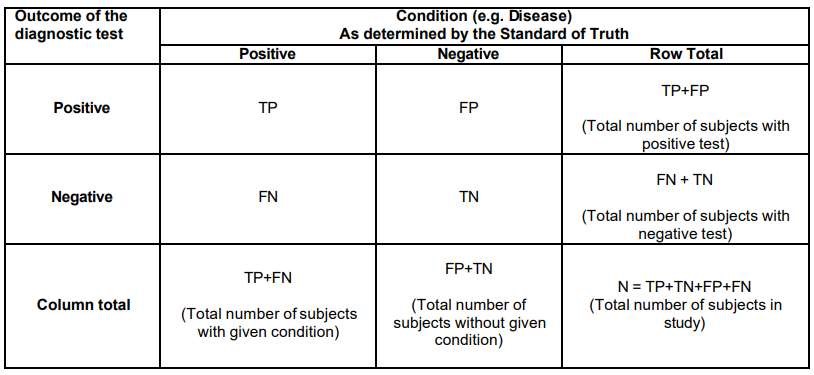






### **COMPARISION OF MODELS**

Here the algorithms used by us for classification of prediction are LR, KNN, NB and DT. The classification and comparison of algorithms is done on the basis of Accuracy, Recall, F1 and Precision.



There are several terms that are commonly used along with the description of sensitivity, specificity and accuracy. They are true positive (TP), true negative (TN), false negative (FN), and false positive (FP).

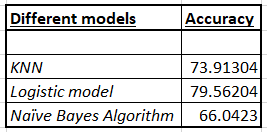
Sensitivity: In machine learning, sensitivity, also known as recall or true positive rate, measures the proportion of actual positive instances that are correctly identified by the model, indicating its ability to minimize false negatives.

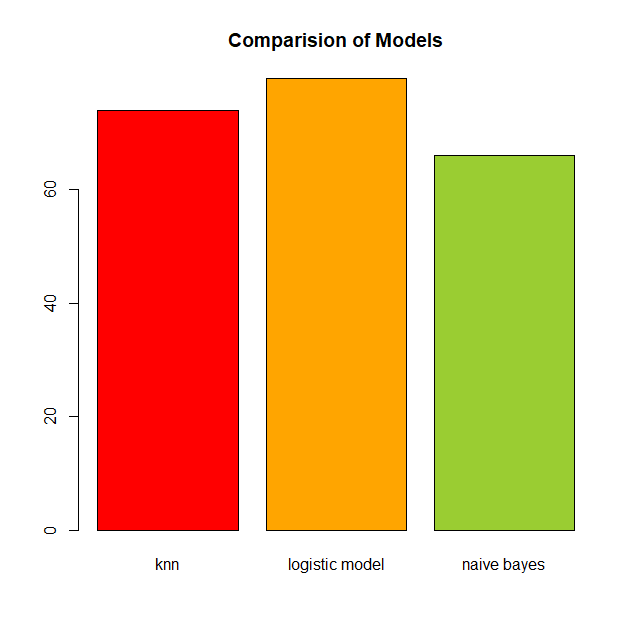
Precision: Precision in machine learning measures the proportion of correctly predicted positive instances out of all instances predicted as positive, highlighting the model's ability to minimize false positives.

Specificity: In machine learning, specificity measures the proportion of actual negative instances that are correctly identified by the model, indicating its ability to minimize false positives and correctly identify true negative instances.

Accuracy: In machine learning, accuracy is a metric that measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances to the total number of instances in the dataset.

F1 Score: In machine learning, the F1 score is a metric that combines precision and recall (sensitivity) into a single value. It provides a balanced measure of a model's performance by considering both the true positive rate and the positive prediction accuracy.





**CONCLUSION**

From the above table of model accuracy we can declare that our Logistic Model algorithm model is the best model for handling and prediction of the social media survey data.

The people of age between 18-25 years are teenagers and working youth. They seem to have more benefit with respect to career opportunities rather than other age groups. Also, there is increase in the number of crimes due to social media and youth is more prone to the cyber crimes.

Also from the logistic regression model we can conclude that chances of requiring health check up increases by 81.45 % per unit increase in the physical and mental health score of an individual.

The social media has both positive and negative impact on youth.

We can try finding balance between the both effects and use the social platforms wisely.

Some ways of managing risks :

Creating tech-free zones and encouraging in-person relationships: This can include restricting the use of screens like phones and tablets 1 hour before bed and keeping times like family meals device-free.

Protect others and themselves: If a young person witnesses scams or cyberbullying, they should tell a trusted adult about it.

**REFERENCES**

https://www.ombea.com/resources/articles/5-point-likert-scale-the-key-to-easily-understanding-your-audience

https://www.questionpro.com/blog/types-of-sampling-for-social-research/

https://surveysparrow.com/blog/stratified-sampling/#section4