



# CSE271 Project

## Phone Addiction

### (REPORT)

#### Abstract:

This study aims to explore how various behavioral, social, physical, and psychological features influence smartphone addiction, and in turn, how addiction to smartphones affects these dimensions of daily life. By analyzing relationships between key variables in our dataset, we strive to uncover patterns that can enhance our understanding of digital dependency. This insight will not only help in identifying individuals at risk but also assist in developing predictive models to guide interventions and healthier usage habits.

#### 1. Introduction:

Addiction has long been one of humanity's most persistent challenges. The human brain is wired to seek pleasure and repeat behaviors that provide instant gratification. Scientific advances have revealed that addiction hijacks the brain's reward system, releasing a flood of hormones that overpower logical thinking. Even when individuals recognize their addictive behaviors, breaking free often requires immense willpower and support. As W. Edwards Deming once said, "*The big problems are where people don't realize they have one in the first place.*"

While addiction is often associated with substances like drugs or alcohol, **the most pervasive form of addiction in the modern era is smartphone addiction**. People of all ages, genders, and cultural backgrounds are affected. The danger lies not just in the hours spent scrolling, but in the profound psychological and behavioral changes that excessive smartphone use can trigger. Research has shown that phone addiction impairs focus, reduces productivity, and disrupts mental and physical well-being.

Unlike other addictions, complete abstinence from smartphones is neither practical nor possible in today's digital world. Therefore, the solution lies in understanding and regulating our usage patterns. This project investigates a dataset containing features related to smartphone use and daily life activities. By analyzing this data, we aim to better understand the underlying factors of phone addiction and its consequences on individuals. Through this

understanding, we hope to contribute to more informed strategies for managing digital dependency in a healthy, balanced way.

## 2. Dataset

- Technology\_Affects\_Us uploaded by Izhah01

### 2.1. Data Munging

- Dataset size before data wrangling à [1086 rows x 15 columns]
- Data set size after filling in missing values and deleting duplicates à [1056 rows x 15 columns]

### 2.2. Columns in dataset (Features):

- **User\_ID**
- **Age** (mean= **38.67**)

*Individuals in this Dataset are prone to being adults or elderly rather than teens*

- **Gender** (mode= **Female**)
- **Total\_App\_Usage\_Hours** (mean= **7.31**)
- **Daily\_Screen\_Time\_Hours** (mean= **7.22**)
- **Number\_of\_Apps\_Used** (mean= **16.89**)
- **Social\_Media\_Usage\_Hours** (mean= **3.23**)
- **Productivity\_App\_Usage\_Hours** (mean= **2.32**)
- **Gaming\_App\_Usage\_Hours** (mean= **2.59**)
- **Location** (mean= **Los Angeles**)
- **Attention\_Span\_Min** (mean= **11.75**)
- **Relationship\_Status** (mean= **Single**)
- **Sleep\_Quality** (mean= **5.79**)
- **Physical\_Health\_Score** (mean= **7.08**)
- **Online\_Shopping\_Frequency** (mean= **5.88**)

## 2.3. Brief description of the dataset

Understood through various generated graphs (graphs available in the dashboard)

- Individuals in this Dataset are prone to being adults or elderly rather than teens.
- unbiased towards a specific gender, both genders have similar numbers of contributions.
- unbiased towards a specific location, all locations have a similar number of contributions.
- females avg. age in the dataset is significantly higher than the males avg. age.
- The number of married males is close to none(single and divorced men are the highest existing category in our dataset). Number of married females is the second largest category in our dataset( the number of married females is significantly higher than the number of divorced and single females combined).

## 3. Visual graphs and Relations:

Our dataset contains many features which describe the personal behavior of some individuals. These features can be classified into three different types of features.

- features which describe individuals' phone usage, [*Total\_App\_Usage\_Hours*, *Daily\_Screen\_Time\_Hours*, *Number\_of\_Apps\_Used*, *Social\_Media\_Usage\_Hours*, *Productivity\_App\_Usage\_Hours*, *Gaming\_App\_Usage\_Hours*]
- features which describe social, health, and consumption activity. [*Attention\_Span\_Min*, *Relationship\_Status*, *Sleep\_Quality*, *Physical\_Health\_Score*, *Online\_Shopping\_Frequency*]
- identification information. [*Age*, *Gender*, *Location*]

Our goal is to use data to understand how different phone usage rates may impact on one's personal traits and activity, to achieve this we need to learn which features are correlated and how they affect one another

### 3.1. Phone addiction and daily screen time

#### 3.1.1. How age, gender and location affect daily screen time

- **Age x Daily screen time**

We can conclude from the graph generated that younger individuals (20 to 40) have higher screen time than older ones (>40)

This can be explained by the different nature of each generation, younger generations are more attached to phones than older generation who are mostly digital immigrants

- **Gender x Daily screen time**

We can see from the generated bar graph that the average screen time is significantly higher in males than females.

This can be explained by the huge difference in men's preference for gaming compared to women's (will be displayed in (gender x gaming app usage))

- **Location x Daily screen time**

We classified the available locations ['Phoenix', 'Los Angeles', 'Houston', 'Chicago', 'New York'] into two teams, in two different graphs. We classified them according to the Region and according to their technological advancement. These classifications were made to understand how the difference in the geographical and environmental state of the city may affect the individual's daily screen time.

**Team classification Based on the Region:**

1. **West Team:** Phoenix and Los Angeles
  - Both cities are in the western United States, characterized by a warmer climate and significant population growth in recent decades.
2. **East Team:** Houston, Chicago, and New York
  - This team comprises cities situated in the central to the eastern United States, with Chicago and New York being prominent cities in their respective regions, contributing to their historical significance and economic prowess.

**Team classification based on technological advancements:**

1. **Team A (Advanced):** New York, Los Angeles, and Chicago

**New York** – Major tech and finance hub, strong presence of tech startups, smart city initiatives.

**Los Angeles** – Strong in entertainment tech, startups, AI, biotech, and green tech.

**Chicago** – Emerging tech scene with investment in fintech, healthtech, and smart infrastructure.

2. **Team B (Developing)** : Houston and Phoenix

**Houston** – Known more for energy, starting to grow in biotech and aerospace tech.

**Phoenix** – Growing tech scene with focus on semiconductors and autonomous vehicles but still maturing.

These classifications are based on overall ecosystem maturity and investment in tech sectors.

Both created teams in the two done experiments resulted in nearly similar average screen time usage. There are no signs of a relation between the location and the daily screen time (the Change between different locations available in this dataset won't affect the individuals screen time)

### **3.1.2. Daily screen time effects on consumption behavior, social, physical, and mental activity**

- **Daily screen time x Attention span**

We can conclude from the graph that the relation between them is an inverse relationship. As screen time increases, attention span decreases.

- **Relationship status x Daily screen time** (can be cause or effect)

We can deduce from the bar plot that those who have the least average screen time are married people, second least are those who are single, and the team with the largest screen time is divorced.

This shows that excessive screen time may cause loneliness and social introversion

- **Daily screen time x Sleep quality**

We can clearly see from the line plot that the relation between them is an inverse relationship. As screen time increases, Sleeping Quality decreases

- **Daily screen time x Physical health**

We can clearly see from the graph that the relation between them is an inverse relationship. As screen time increases, physical health decreases

- **Daily screen time x Online shopping frequency**

There are no signs of a relation between daily screen time and Online Shopping Frequency. Online Shopping Frequency may be correlated to

other features like gender or relationship status.

### **3.1.3. Conclusion (Daily screen time):**

From the visual graphs generated, an individual's age has proven to be a feature which can contribute to the determination of individuals' daily screen time (from 20 to 40 higher screen time than above 40). Graphs also showed that males have significantly higher average screen time compared to females. This can be explained by the (gender and relationship status) graph [if more of the men in the dataset are single or divorced than women then they may have extra free time, feel boredom or loneliness which may result in excessive screen time], it can also be explained by the (gender x age) graph [females in our dataset have higher average age than men, and age is directly correlated with screen time (younger individuals have higher screen time)]. Excessive daily screen time is proven by the generated graphs to decrease Attention span, sleep quality score, and physical health

### **Note: We chose to analyze data with 'Daily\_Screen\_Time\_Hours' instead of 'Total\_App\_Usage\_Hours'**

There are phone activities that add up to screen time but not app usage hours like:

1. Staying on the Home Screen or Lock Screen
  - You unlock your phone often and leave it on the home screen without opening apps.
2. Reading Notifications or Quick Replies
  - Interacting with notifications (replying to messages, dismissing alerts) doesn't always count toward specific app time.
3. Watching Content Without App Focus
  - Watching YouTube through a browser or previewing videos without the app being fully open.
4. Using Built-in Tools or Features
  - Flashlight, calculator, camera, or settings menu don't count toward app usage time but keep the screen on.

### **3.2. Phone addiction and social media usage**

### 3.2.1. How age, gender and location affect social media usage

- **Age x social media usage**

We can conclude from the graph displayed that younger individuals (20 to 50) have higher social media usage than older ones (>50)

This can be explained by the different nature of each generation, younger generations are more attached to social media than older generation who prefer real life interactions

- **Gender x social media usage**

We can see from the generated bar graph that the avg social media usage time is higher in males than females

this can be explained by (gender x relationship status) graph ) [if more of the men in the dataset are single or divorced than women then they may have extra free time, feel boredom or loneliness which may result in excessive screen time]

- **Location x social media usage**

We classified the available locations ['Phoenix', 'Los Angeles', 'Houston', 'Chicago', 'New York'] into two teams, in two different graphs. We classified them according to the Region and according to their technological advancement. These classifications were made to understand how the difference in the geographical and environmental state of the city may affect the individual's social media usage.

Both teams gave approximately equal avg. social media usage time in both graphs (Location doesn't affect social media usage)

### 3.2.2. Social media usage effect on consumption behavior, social, physical, and mental activity

- **Social media usage x Attention span**

We can clearly see from the graph that at low social media usage hours (<2), the attention span of individuals is higher than 15 minutes. At higher social media usage hours (>3.5), attention span is 15 minutes or lower

(increasing social media usage hours decreases an individual's attention span)(inverse relation).

- **Relationship status x Social media usage** (can be cause or effect)

We can deduce that those who have the least average social media usage are the married people, second least are those who are divorced, and the team with the largest social media usage are the single

This shows that excessive social media usage may cause loneliness and social introversion

- **Social media usage x Sleep quality**

We can clearly see from the graph that at low social media usage hours(<2),sleep quality score is 9 or higher. At higher social media usage hours(>3.5), sleep quality score is 8 or lower

(increasing social media usage hours decreases sleep quality score)(inverse relation)

- **Social media usage x Physical health**

We can clearly see from the graph that at low social media usage hours(<2),physical health score is 9 or higher. At higher social media usage hours(>3.5), physical health score is 9 or lower

(increasing social media usage hours decreases physical health score)(inverse relation)( Note: both features aren't highly correlated,there are many individuals who have high social media usage and high physical health score ) social media may act as a fitness motivator for some individuals

- **Social media usage x Online shopping frequency**

As the social media usage hours increase the online shopping frequency increases (direct relation).

### **3.2.3. Conclusion (Social media usage):**

From the visual graphs generated, an individual's age has proven to be a feature which can contribute to the determination of individuals' Social media usage (from 20 to 50 higher social media usage than above 50).Graphs also showed that males have higher average social media usage time compared to females. This can be explained by the (gender and relationship status) graph [if more of the men in the dataset are single or divorced than women then they may have extra free time,feel boredom or loneliness which may result in excessive social media usage], it can also be explained by the (gender x age) graph [females in our dataset have higher average age than men ,and age is directly correlated with social media usage time (younger individuals have higher social media usage)]. Excessive social media usage time is proven by the generated graphs to decrease Attention span, sleep quality score, physical health score(weak correlation), and increase online shopping frequency.



### 3.3. Phone addiction and Gaming app usage

#### 3.3.1. How age, gender, and location affect gaming app usage

- **Age x Gaming app usage**

We can conclude from the bar graph displayed that younger individuals (20 to 30) have the highest gaming app usage. As age increases average gaming app usage decreases. Individuals from (50 to 60) have the least average gaming app usage time.

This can be explained by the different nature of each generation, younger generations are more attracted to gaming than older generations who prefer real life interactions.

- **Gender x Gaming app usage**

We can see from the generated bar graph that the avg gaming usage time is higher in males than females.

this can be explained by (gender and relationship status) graph ) [if more of the men in the dataset are single or divorced than women then they may have extra free time, feel boredom or loneliness which may result in excessive gaming] (it can also be related to the males nature as they naturally prefer gaming more than females).

- **Location x Gaming app usage**

We classified the available locations ['Phoenix', 'Los Angeles', 'Houston', 'Chicago', 'New York'] into two teams, in two different graphs. We classified them according to the Region and according to their technological advancement. These classifications were made to understand how the difference in the geographical and environmental state of the city may affect the individual's social media usage.

- 1) **Classified based on region:** East region has a slightly higher average gaming app usage than west region
- 2) **Classified based on technological advancement:** Both teams approximately have similar average gaming app usage time

#### 3.3.2. Gaming app usage effect on consumption behavior, social, physical, and mental activity

- **Gaming app usage x Attention span**

We can clearly see from the line graph generated that there are signs of a relation between gaming app usage time and attention span. (inverse relation pattern)

Attention span score when gaming app usage is (less than 1.5 hours) is (16 min or higher), and Attention span score when gaming app usage is (3.5 hours or more) is (13 min or lower)

As the number of hours spent on gaming app usage increases, attention span decreases.

- **Relationship status x Gaming app usage** (can be cause or effect)

We can deduce from the generated bar graph that those who have the least average gaming app usage are the married people, second least are those who are divorced, and the team with the largest gaming app usage are the single.

- **Gaming app usage x Sleep quality**

We can clearly see from the line graph generated that there are signs of a relation between gaming app usage time and sleep quality. (inverse relation pattern).

Sleep quality score when gaming app usage is (less than 1.5 hours) is (8 or higher), and sleep quality score when gaming app usage is (3.5 hours or more) is (6 or lower)

As the number of hours spent on gaming app usage increases, sleep quality decreases.

- **Gaming app usage x Physical health**

We can clearly see from the line graph generated that there are signs of a relation between gaming app usage time and sleep quality. (inverse relation pattern)

Physical health score when gaming app usage is (less than 1.5 hours) is (9 or higher), and physical health score when gaming app usage is (3.5 hours or more) is (8 or lower)

As the number of hours spent on gaming app usage increases, physical health decreases.

- **Gaming app usage x Online shopping frequency**

We can clearly see from the graph that there are signs of a relation between gaming app usage time and Online Shopping Frequency. (direct relation pattern).

As the number of hours spent on gaming app usage increases, Online Shopping Frequency increases.

### **3.3.3. Conclusion (Gaming app usage):**

From the visual graphs generated, an individual's age has proven to be a feature which can contribute to the determination of individuals' Gaming app usage (from 20 to 30 is the highest Gaming app usage and from 50 to 60 is the lowest). Graphs also showed that males have higher average Gaming app usage time compared to females. This can be explained by the (gender x relationship status) graph [if more of the men in the dataset are single or divorced than women then they may have extra free time, feel boredom or loneliness which may result in excessive

Gaming app usage ], it can also be explained by the (gender x age) graph [females in our dataset have higher average age than men ,and age is directly correlated with Gaming app usage time (younger individuals have higher Gaming app usage)]. (it can also be related to the male's nature as they naturally like gaming more than females do). Excessive gaming app usage time is proven by the graph generated to decrease attention span, sleep quality, and physical health. The deterioration in these features becomes obvious when gaming app usage exceeds 3.5 hours, these features were seen to be at normal rates when gaming app usage is less than 1.5 hours. Excessive gaming app usage results in an increase in online shopping frequency

### **3.4. Phone addiction and Gaming app usage**

#### **3.4.1. How age, gender, and location affect Productivity app usage**

- **Age x Productivity app usage**

We can conclude from the graph shown that older individuals (50 to 60) have the highest productivity app usage. As age decreases average productivity app usage decreases. Individuals from (20 to 30) have the lowest average productivity app usage time.

This can be explained by the different nature of each generation, older generations are more focused on utilizing technology for their benefit rather than entertainment

- **Gender x Productivity app usage**

We can see from the generated bar graph that the avg productivity usage time is higher in females than males

This can be explained by (gender x relationship status) graph ). if more of the women in the dataset are married than men then they are more busy, disciplined and mature which may result in high productivity app usage and low gaming and social media usage (also, it can be due to that the female's average age in the dataset is higher than men (proved in the (gender x avg. age) graph ) ,and productivity app usage was proven to be directly correlated to individuals age(as the age increases productivity app usage increases ).

- **Location x Productivity app usage**

We classified the available locations ['Phoenix', 'Los Angeles', 'Houston', 'Chicago', 'New York'] into two teams, in two different graphs. We classified them according to the Region and according to their technological advancement. These classifications were made to understand how the difference in the geographical and environmental state of the city may affect the individual's social media usage.

- 1) **Classified based on region:** East team have higher average productivity app usage time than West team
- 2) **Classified based on technological advancement:** team A (Advanced) have higher average productivity app usage time than team B (developing)

Explanation: team A is the more advanced team, so they have more experience in utilizing technology to serve their personal benefits (more useful and productive usage of phones than team B).

### **3.4.2. Productivity app usage effect on consumption behavior, social, physical, and mental activity**

- **Productivity app usage x Attention span**

We can clearly see from the line graph that there are signs of a relation between productivity app usage time and attention span. as productivity app usage increases,

attention span seems to increase as well

This can be explained by studying the behavior of individuals who have higher productivity app usage correlated to screen time, because

attention span is mainly related to screen time (productivity app usage x screen time )

- **Relationship status x Productivity app usage** (can be cause or effect)

We can deduce that those who have the least average productivity app usage are the single people, second least are those who are divorced, and the team with the the largest productivity app usage are the married. (married people are mostly more mature,responsible,and disciplined)

- **Productivity app usage x Sleep quality**

We can clearly see from the graph that there are signs of a relation between productivity app usage time and attention span. as productivity app usage increases,

sleep quality seems to increase as well

This can be explained by studying the behavior of individuals who have higher productivity app usage correlated to screen time, because sleep quality is mainly related to screen time (productivity app usage x screen time ). People with high productivity app usage time have low daily screen time, so as a result they have better sleep quality.

- **Productivity app usage x Physical health**

We can observe that the relation is unclassifiable for low productivity app usage but for individuals with higher productivity app usage (>3.5 hours) mostly are considered fit with a physical score (=>7).

- **Productivity app usage x Online shopping frequency**

We can clearly see from the line graph that there are signs of a relation between productivity app usage time and Online Shopping Frequency. As productivity app usage increases, Online Shopping Frequency seems to decrease (inverse relation).

This can be explained by studying the behavior of individuals who have higher productivity app usage. individuals who have higher productivity app usage are mostly elders(>50) and married people.(individuals that have high Online Shopping Frequency are mostly young and single or divorced individuals) Proved by Online Consumption Behavior graphs

### **3.4.3. Conclusion (Productivity app usage)**

From the visual graphs generated, an individual's age has proven to be a feature which can contribute to the determination of individuals' Productivity app usage. Older individuals (50 to 60) have the highest productivity app usage. As age decreases, average productivity app usage decreases. Graphs also showed that females have higher average productivity app usage time compared to males. The East team has higher average productivity app usage time than the West team, and team A (Advanced) has a higher productivity app usage time than team B (developing). High productivity app usage is proven by the graphs generated to increase Attention span, sleep quality score, and physical health. High productivity app usage results in a decrease in online shopping frequency.

### **3.5. Productive app usage effect on Daily screen time**

Does having high productive app usage mean that the individual should have high daily screen time?

- **Productivity app usage x Daily screen time**

We can observe from the generated line plot that most individuals with low productivity app usage time have high daily screen time ( $\Rightarrow >10$  hours), and those with high productivity app usage time have low daily screen time ( $\Rightarrow <4$  hours)

Explanation: those with higher productivity app usage are individuals who utilize technology in a useful and productive manner, their behavior shows they don't waste unnecessary time on their phone (more disciplined), unlike people with lower productivity app usage who've shown to have higher screen time wasted on gaming and social media (entertainment)

### **3.6. The number of apps used effect on attention span**

- **Number of apps used x Attention span**

This line plot describes the relation between the number of apps used and attention span, we can observe that individuals who use (less than 10) apps have higher attention span than those who use (more than 10) apps

### **3.7. Age, gender, and location effect on online shopping frequency**

- **age x online shopping frequency**

younger individuals have high Online Shopping Frequency while older ones have low online shopping frequency

Explanation: younger generations are more active online shoppers unlike older generations who aren't used to online shopping( mostly are digital immigrants).

- **gender x online shopping frequency**

males have higher online shopping frequency than female

(explained by the individuals' relationship status in the dataset) [males are mostly single and divorced while females are mostly married].

- **relationship status x online shopping frequency**

We can deduce that those who have the least online shopping frequency are the married people, second least are those who are divorced, and the team with the largest online shopping frequency are the single.

## 4. Methods

I've picked Naïve Bayes and k-Nearest Neighbors, although I know that Naïve Bayes won't give good results but one of my objectives is to illustrate why k-Nearest Neighbors is better in our case.

### 4.1. Regression

To analyze how Daily Screen Time impacts Attention Span, we employed two regression techniques: Linear Regression and Polynomial Regression. Each was selected to evaluate different types of relationships (linear and nonlinear) in the dataset.

#### 4.1.1. Linear Regression

Linear Regression models a relationship between a dependent variable  $y$  and an independent variable  $x$  using a straight line. The model aims to fit the best linear equation that minimizes the prediction error using the Least Squares Method.

##### Mathematical Formula:

$$y = a + b * x$$

Where:

- $y$ : Predicted value of the dependent variable (Attention Span)
- $x$ : Independent variable (Daily Screen Time)

- a: Intercept (value of y when  $x = 0$ )
- b: Slope (rate of change of y with respect to x)

This model assumes:

- A linear relationship between x and y
- Homoscedasticity (constant variance of errors)
- Independence of observations
- Normally distributed residuals

Linear Regression serves as a baseline model to assess whether a more complex model is necessary.

#### 4.1.2. Polynomial Regression

Polynomial Regression is a form of linear regression where the relationship between the independent variable x and the dependent variable y is modeled as an nth-degree polynomial. It allows for more flexibility in capturing curved trends in the data.

Mathematical Formula (degree n):

$$y = a + b_1 * x + b_2 * x^2 + \dots + b_n * x^n$$

Where:

- $x^n$ : Higher-order terms (squared, cubed, etc.) that allow the model to bend and better fit non-linear data
- a,  $b_1$ , ...,  $b_n$ : Coefficients learned during training

In implementation, we used PolynomialFeatures to transform the original input variable into multiple polynomial terms, and then fit a standard linear regression model on these expanded features.

Polynomial regression maintains linearity in the coefficients, but introduces non-linearity in the feature space, making it suitable for datasets with curvature.

We trained polynomial models from degree 1 (equivalent to linear regression) up to degree 6, and compared performance using  $R^2$  score and Mean Squared Error to avoid overfitting while improving accuracy.



Both regression models were trained and evaluated using train-test splits to ensure generalizability.

## 4.2. Classification

### 4.2.1 Naïve Bayes Classifier:

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming that all features are conditionally independent given the class label.

#### Formula:

$P(C|X) = \frac{P(X|C) * P(C)}{P(X)P(C|X)}$ : Posterior probability of class given features

- $P(X|C)$ : Likelihood of features given class
- $P(C)$ : Prior probability of class
- $P(X)$ : Evidence (normalizing constant)

This Model uses Gaussian Naïve Bayes, where feature likelihoods are modeled using a Gaussian distribution:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2}$$

### 4.2.2 k-Nearest Neighbors (k-NN):

k-NN is a non-parametric algorithm that classifies a data point based on the majority label of its k closest neighbors in feature space.

#### Distance Metric (Euclidean):

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

In this model, k = 5 was used.

### 4.2.3. Classification Cases:

- Young (20-30)
- Middle (31-50)
- Older (51-60)

#### 4.2.4. Neural Network Classification

## 5. Evaluation

### 5.1 Regression

#### 5.1.1. Linear Regression Evaluation

The Linear Regression model was trained on 80% of the dataset and tested on the remaining 20%. The performance was evaluated using two key metrics: the  $R^2$  Score and Root Mean Squared Error (RMSE).

Metric	Definition	Formula
$R^2 = \text{Score}$	Indicates how well the model explains the variance in the target variable. A value closer to 1 means better fit.	$R^2 = 1 - (\text{SSres} / \text{SStot})$ , where $\text{SSres} = \sum (y - \hat{y})^2$ and $\text{SStot} = \sum (y - \bar{y})^2$
RMSE	Measures the average prediction error between the actual and predicted values. Lower values indicate more accurate predictions.	$\text{RMSE} = \sqrt{(\sum (y - \hat{y})^2 / n)}$ , where $n$ is the number of observations

- The  **$R^2$  score of 0.9521** means that 95.21% of the variability in attention span can be explained by daily screen time.
- The **RMSE of 0.8661** indicates that the typical prediction error is less than 1 minute, showing good accuracy.

#### 5.1.2. Polynomial Regression Evaluation

To assess the benefit of modeling a nonlinear relationship, we trained polynomial regression models with degrees ranging from 1 to 6. All models were evaluated using  **$R^2$  Score** and **Mean**

**Squared Error (MSE).** The **R<sup>2</sup> score increased** gradually with the degree of the polynomial, reaching its maximum at **degree 6 (0.962)**.

Although Polynomial Regression models non-linear relationships by introducing higher-degree terms (e.g.,  $x^2$ ,  $x^3$ ), it is still evaluated using the same metrics as Linear Regression. The **R<sup>2</sup> Score** and **RMSE** are calculated in the same way

- The **MSE** also decreased, showing improved prediction accuracy.
- However, the performance gains between degree 3 and degree 6 were minimal, while the complexity increased.
- Degree **3 to 5** provided a good trade-off between accuracy and generalization, avoiding overfitting while improving performance.

These evaluations show that while Linear Regression already fits the data well, a carefully chosen Polynomial Regression model (degree 3–5) can offer even better results without overfitting.

## 5.2. Classification

### 5.2.1. Numerical Evaluation:

Metric	Definition	Formula
Accuracy	Overall correctness of the model	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	How many predicted positives are truly positive	$\frac{TP}{TP+FP}$
Recall	How many actual positives are correctly identified	$\frac{TP}{TP+FN}$
F1-Score	Harmonic mean of precision and recall	$2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$
Support	Number of actual samples in each class	Count of true classes

## 5.2.2. Graphical Evaluation:

**5.2.2.1. Confusion Matrix:** visualizes the performance of a classification model by comparing actual vs predicted classes

- a. **True Positives (TP):** Correctly predicted class instances (diagonal values).
- b. **False Positives (FP):** Wrong predictions where the model over-predicted a class.
- c. **False Negatives (FN):** Missed predictions where the model failed to predict the correct class.
- d. **Error distribution:** Which classes are often confused with each other.

**5.2.2.2. Scatter Plot of True vs Predicted:** compares the predicted labels with the true labels across each data sample.

- a. **Circles (o)** → Actual labels
- b. **Crosses (x)** → Predicted labels

## 6. Results

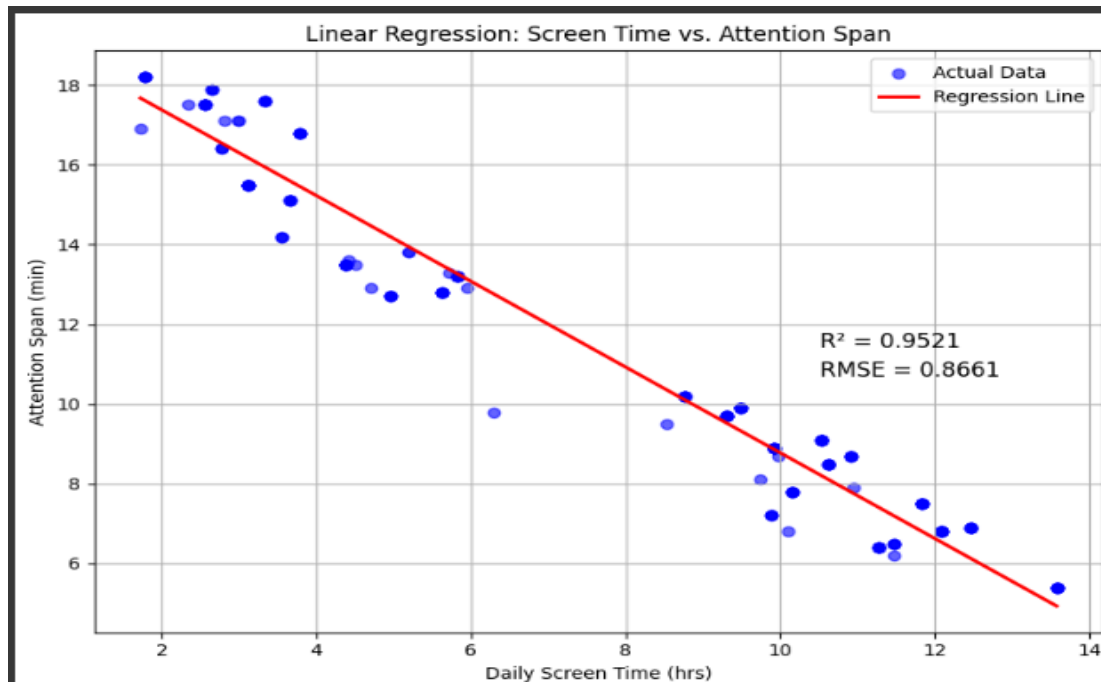
### 6.1 Regression

#### 6.1.1. Linear Regression Results

The Linear Regression model revealed a strong negative relationship between daily screen time and attention span. The regression equation derived from the model was:

$$\text{Attention Span} = 19.52 - 1.07 \times \text{Daily Screen Time}$$

This means that for every 1-hour increase in screen time, the attention span is predicted to decrease by approximately 1.07 minutes. The model achieved a  $R^2$  score of 0.9521, indicating that it explains over 95% of the variance in the target variable. The low RMSE of 0.8661 further confirms the model's high accuracy, with prediction errors being less than 1 minute on average.



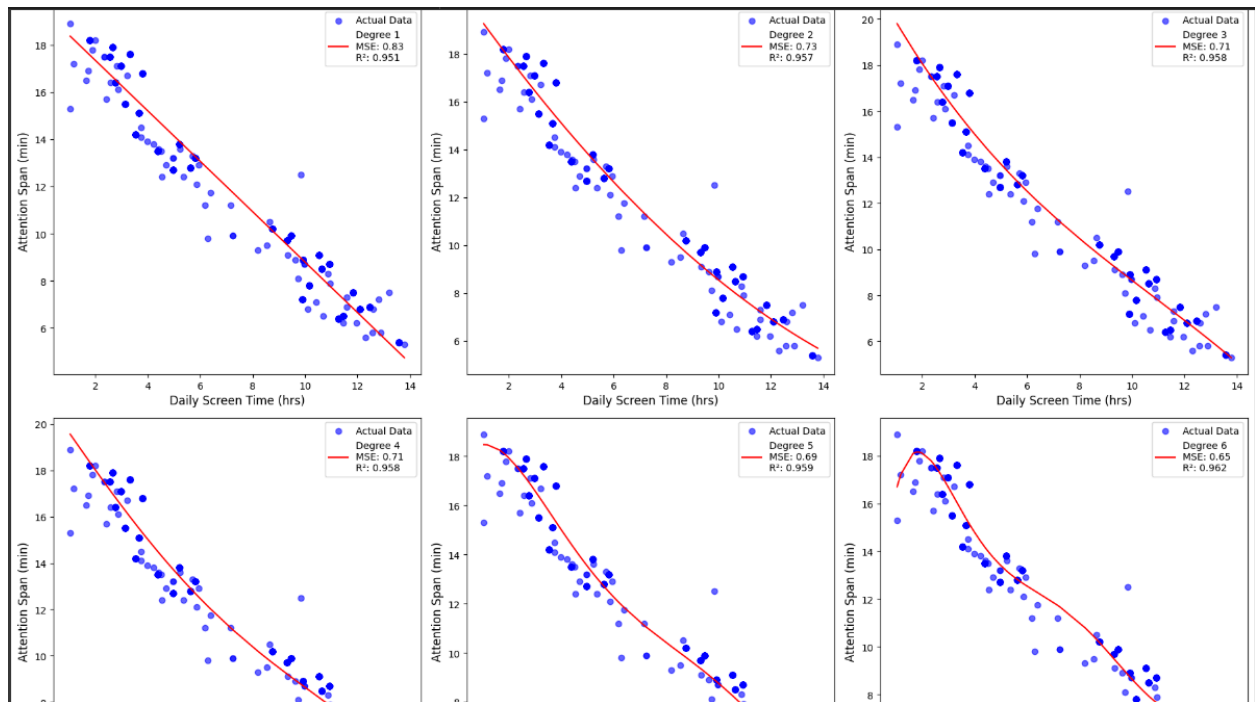
### 6.1.2. Polynomial Regression Results

Degree	R <sup>2</sup> Score	MSE	Remarks
1	0.951	0.83	Baseline (Linear Regression)
2	0.957	0.73	Slight improvement, captures curvature
3	0.958	0.71	Smooth and accurate fit
4	0.958	0.71	Same as degree 3
5	0.959	0.69	Slightly better performance
6	0.962	0.61	Best performance, risk of overfitting

Polynomial Regression models with degrees ranging from **2 to 6** were tested to determine whether a curved relationship would yield better results. The model with **degree 6** produced the highest **R<sup>2</sup> score of 0.962** and the **lowest MSE of 0.61**, outperforming the linear model slightly.

However, improvements beyond **degree 3** were marginal, and **increasing the degree** further introduced a risk of **overfitting**, where the model fits noise rather than the

underlying pattern. Models with **degrees 3 to 5** struck a balance between accuracy and simplicity, achieving high performance while maintaining generalizability.



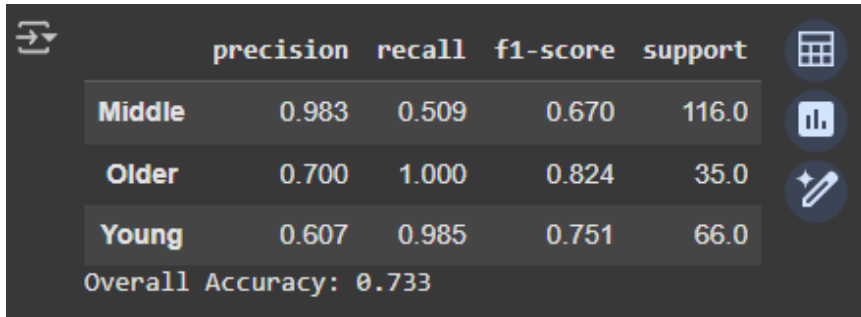
**In summary:**

- Linear Regression provides a strong baseline with interpretable results.
- Polynomial Regression (degree 3–5) slightly improves predictive accuracy and captures the non-linear pattern more effectively.

## 6.2. Classification

### 6.2.1. Numerical Evaluation:

#### 6.2.1.1. Naïve Bayes Classifier:



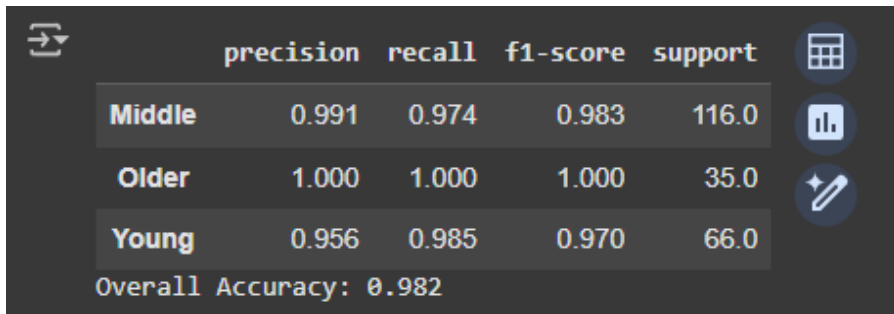
A screenshot of a model performance dashboard. It features a table with columns for precision, recall, f1-score, and support for three classes: Middle, Older, and Young. The overall accuracy is 0.733. To the right of the table are three icons: a grid, a bar chart, and a pencil.

	precision	recall	f1-score	support
Middle	0.983	0.509	0.670	116.0
Older	0.700	1.000	0.824	35.0
Young	0.607	0.985	0.751	66.0

Overall Accuracy: 0.733

- The model achieved perfect recall for the "Older" class, identifying all its samples correctly.
- The "Middle" class had high precision (0.98) but very low recall (0.51), indicating the model rarely predicted "Middle" and misclassified many as "Young".
- This is confirmed by the confusion matrix, where 42 samples of "Middle" were predicted as "Young".
- Overall, the scatter plot shows high variation in predicted values, reflecting inconsistent model performance across classes.

### 6.2.1.2. k-Nearest Neighbors (k-NN):



A screenshot of a model performance dashboard for a k-NN model. It features a table with columns for precision, recall, f1-score, and support for three classes: Middle, Older, and Young. The overall accuracy is 0.982. To the right of the table are three icons: a grid, a bar chart, and a pencil.

	precision	recall	f1-score	support
Middle	0.991	0.974	0.983	116.0
Older	1.000	1.000	1.000	35.0
Young	0.956	0.985	0.970	66.0

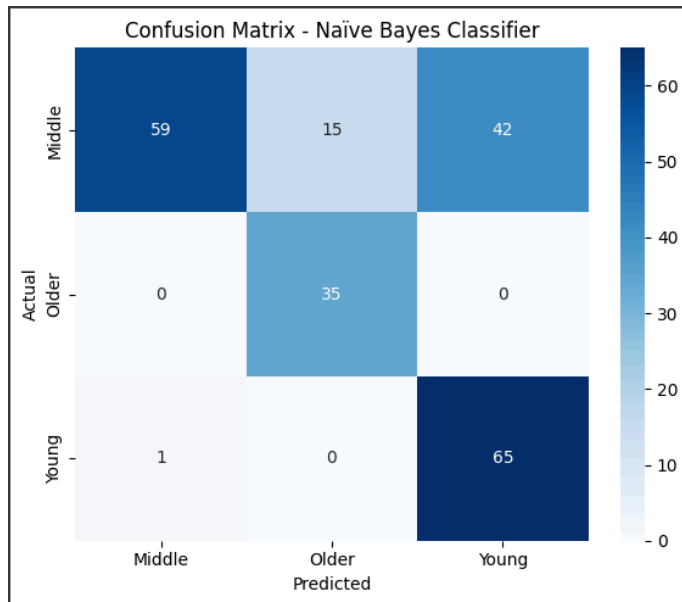
Overall Accuracy: 0.982

- All classes were classified with near-perfect precision, recall, and F1-score.
- Confusion matrix shows only 4 total misclassifications out of 217 samples.
- The scatter plot confirms this: predicted points nearly overlap true labels.
- k-NN showed excellent generalization and no bias toward any particular class.

## 6.2.2 Graphical Evaluation:

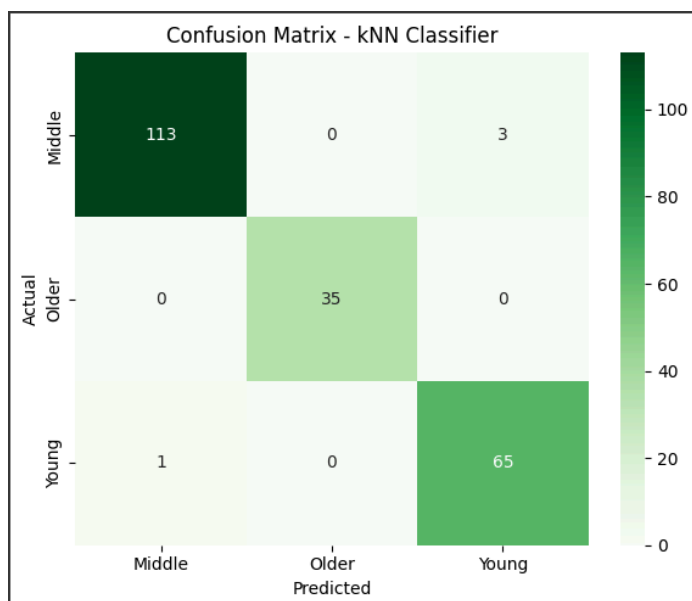
### 6.2.2.1. Confusion Matrix:

#### 6.2.2.1.1. Naïve Bayes Classifier



Many "Middle" samples were confused with "Young".

#### 6.2.2.1.2. k-Nearest Neighbors (k-NN)

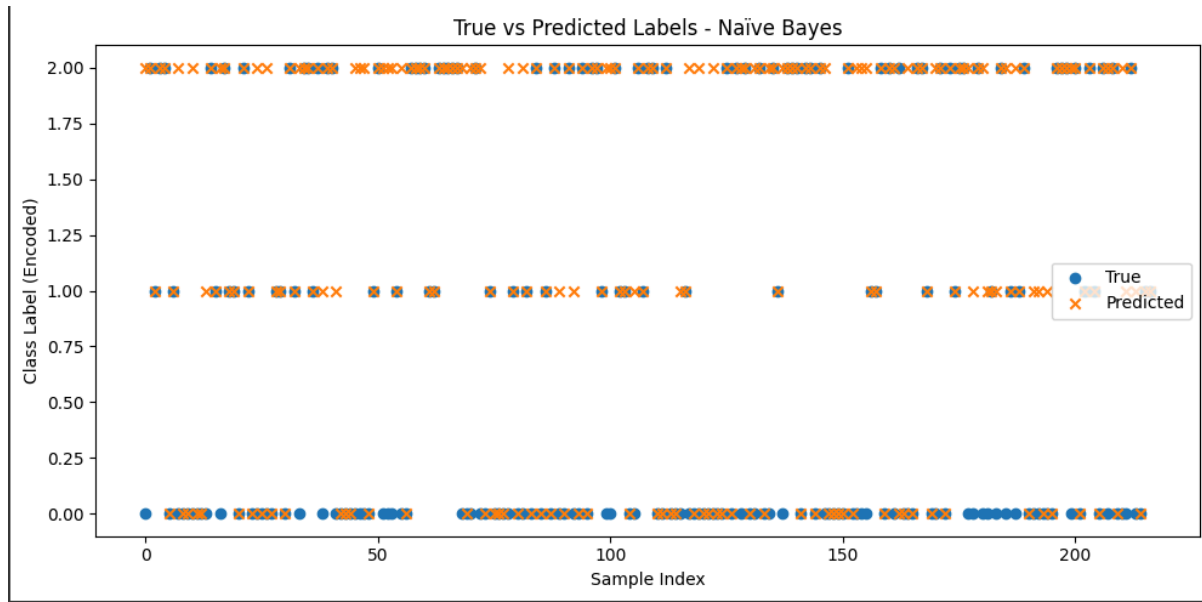


k-NN showed almost perfect class separations

#### 6.2.2.2 Scatter Plot of True vs Predicted

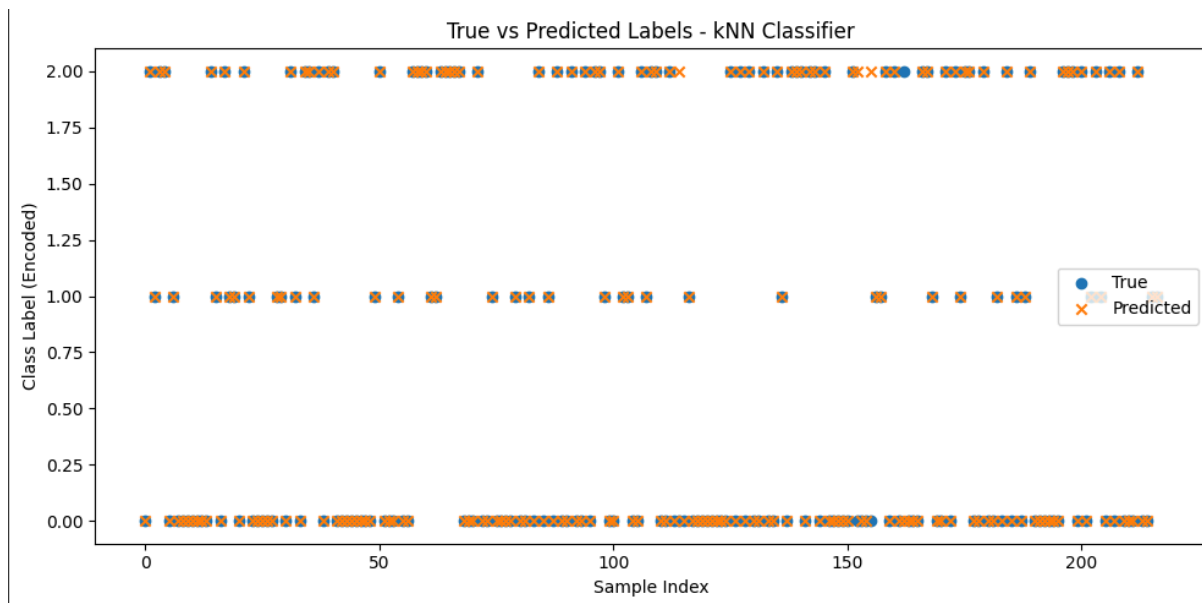
##### 6.2.2.2.1. Naïve Bayes Classifier





Many misalignments are visible, especially for the "Middle" class.

#### 6.2.2.2.2. k-Nearest Neighbors (k-NN)



k-NN, predicted labels closely match true labels, indicating strong performance.

## 7. Conclusion

## 7.1 Regression

## 7.2. Classification

### 1. Naïve Bayes (GaussianNB)

- Assumes features are independent and normally distributed.
- Simple, fast, and interpretable.
- Performs well on smaller or clean datasets.
- Decision boundaries are linear in the log-probability space (not necessarily in the feature space).
- This means Naïve Bayes behaves like a linear classifier (though it's a probabilistic one).
- It works well when classes are linearly separable or nearly so.

### 2. k-Nearest Neighbors (kNN with k=5)

- Non-parametric and non-linear.
- Instance-based learning: stores training data and classifies based on closest neighbors.
- Makes no assumptions about data distribution or linearity.
- Sensitive to irrelevant features and scaling.
- k-NN's decision boundary is highly flexible and non-linear, adapting to the shape of the data.

### ● k-NN significantly outperformed Naïve Bayes:

- Accuracy: 98.16% vs 73.27%
- F1-Score improvements across all classes
- Fewer misclassifications and better alignment in scatter plots

### ● Reason for k-NN's superiority:

- Naïve Bayes assumes independent features and Gaussian distribution, which may not hold in real-life user behavior data.
- k-NN leverages spatial proximity and is more adaptive to complex, non-linear decision boundaries.

- **Final Verdict:** For this classification problem, **k-NN is clearly the superior model**, offering excellent accuracy, robustness, and interpretability across all three age groups.

## **Reference:**

- **Dataset Link:** [https://github.com/Izaiah01/Technology\\_Affects-Us](https://github.com/Izaiah01/Technology_Affects-Us)

## **Group members :**

- Badr Eltaher
- Hassan Elfaransawy
- Adham Hisham
- Mohammed Alkardisy
- Habiba Ahmed
- Sara Aldahshan

