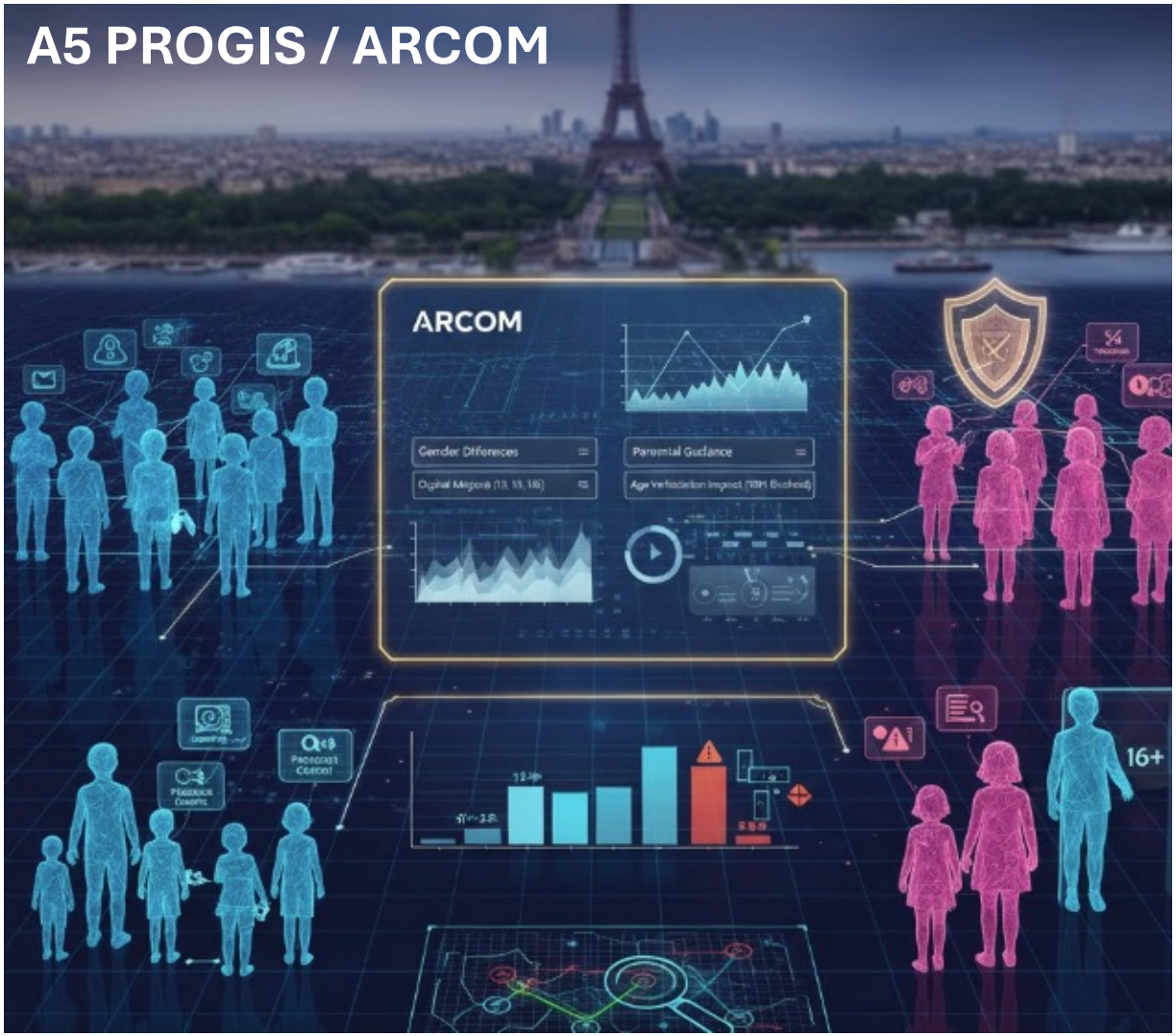
A5 PROGIS / ARCOM



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

A Quantitative Deep-Dive into 2,000 Teens (Ages 11-17)

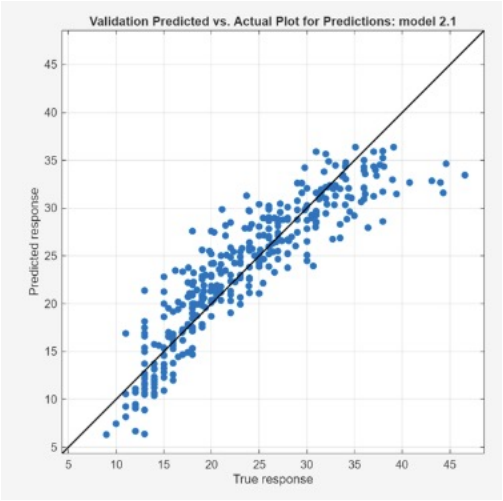
Project Goals

- **Comparing Girls and Boys:**
  - We are looking at how genders differ in what they do online, how their parents supervise them, and how often they run into trouble.
- **The "Digital Age" Question:**
  - Is it better to set the legal age for social media at 13, 15, or even 16? We are comparing younger kids (11-12) with older ones (13-14) to see who handles risks better.
- **Does Age Checking Work?:**
  - Since nearly 1 in 5 teens (18%) have been blocked or asked for ID, we want to know if these checks actually keep them safer or change how they act.
- **The Power of Parents:**
  - We are diving deeper into how much of a difference a parent’s involvement really makes in a child's digital life.
- **Being "Street Smart" Online:**
  - We want to understand what makes a teen "cautious" and if being careful actually prevents them from facing online risks

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Regression (Predicting a Value)

- Use this if you want to understand **how much** one factor influences another.
- **The Goal:** Predict a **continuous "score."**
- **Example:** Can we predict the "Prudence Score" of a teenager based on their age, the level of parental supervision, and the number of hours spent online?
- **Key Columns needed:** A numerical target (like a scale of 1-10 on safety awareness) and numerical/categorical predictors.

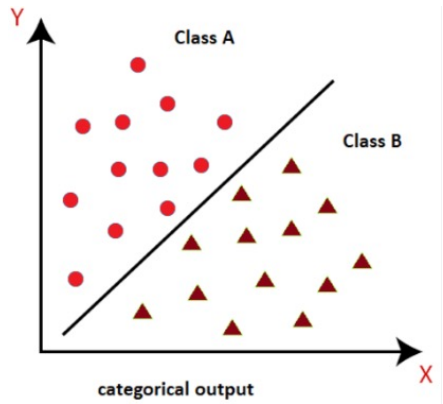


https://fr.mathworks.com/help/stats/regressionlearner-app.html

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Classification (Predicting a Category)

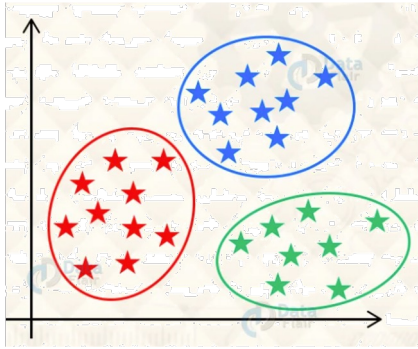
- **The Goal:** Predict a "Yes/No" or "Group A/Group B."
- **Example:**
  - Based on their behaviors, can we classify whether a teenager has been exposed to online risks? Or, can we predict if a teen is likely to support the "Digital Majority" at 13 vs. 15?
- **Key Columns needed:**
  - **A "Label"** (e.g., Exposed_to_Risk: 0 or 1)



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Clustering (Finding Hidden Patterns)

- Use this if you want to see if there are **natural "tribes"** of teenagers that the researchers didn't pre-define.
- **The Goal:** Discover **groups with similar profiles**.



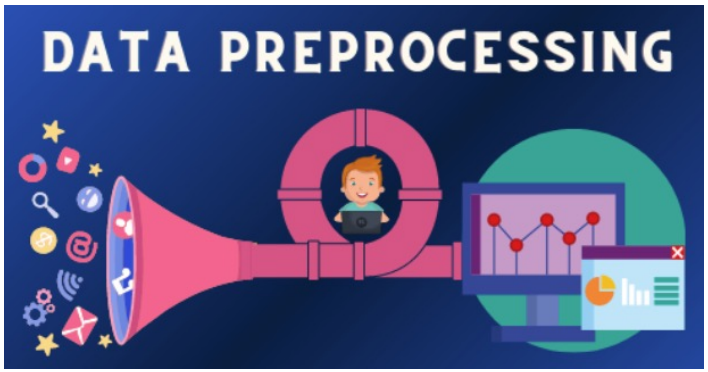
https://data-flair.training/blogs/scipy-clustering/

- **Example:**
  - Are there distinct "profiles" of online users? (e.g., "The Supervised Socialite," "The Unregulated Gamer," "The Cautious Researcher").
- **Key Columns needed:** All behavioral and parental control columns.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Data Pre-processing: Preparing for Analysis

- **Removing Redundancy:** If two columns ask the same thing in different ways, we keep one.
- **Handling Missing Values:** Survey data often has "I don't know" or skipped answers.
- **Encoding:** Turning "Boy/Girl" or "Yes/No" into numbers (0 and 1) so the computer can read them.



https://www.linkedin.com/pulse/beginners-guide-data-preprocessing-machine-learning-cleaning-subash-a/

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Classification (The "Predictor")

- **Use case:** To answer the "Digital Majority" (13 vs 15 years old) or "Age Verification" questions.
- **How:** You can set a target variable like Age_Verification_Status (Did they get blocked/checked?).
- **Goal:** Use columns like S2_R (Household size) or S4_R (Child age) to see which factors best predict if a child is subject to stricter controls.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Regression (The "Influence Tracker")

- **Use case:** To study the "Role of Parents" or "Online Prudence."
- **How:** If you create a "Prudence Score" (by summing up various safety behaviors in the dataset), you can run a regression.
- **Goal:** Quantify the link between Parental_Control (Independent variable) and Risk_Exposure (Dependent variable).

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Clustering (The "Persona Builder")

- **Use case:** To find "differences between girls and boys" or general behavioral "types."
- **How:** Use K-Means clustering on the behavioral columns (social media use, gaming, etc.).
- **Goal:** Group the 2,000 teens into 3 or 4 "Profiles."
- **Insight:** You might discover a group like "High-Risk Explorers" (mostly boys, 14-15, low parental control) vs. "Safe Socializers."

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Example of k-means clustering

```
df = pd.DataFrame(data)

# 2. Scaling (Important for K-Means!)
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)

# 3. Running K-Means
# Let's try to find 3 groups (e.g., Low, Medium, High Risk)
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(df_scaled)
```

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Scaling:

- scaling is a crucial step because machine learning models are "sensitive to distance.
- Since "Age" (11-17) and "Number of people in house" (1-6) are on different scales, you "scale" them so the model doesn't think Age is more important just because the number is bigger.
- The two main method:
  - 1. Standardization (StandardScaler): This centers the data around 0 with a standard deviation of 1. It is best for **Regression** and **Clustering**.
  - 2. Normalization (MinMaxScaler): This shrinks the data strictly between 0 and 1.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

```
df = pd.DataFrame(data)

# 2. Initialize the Scaler (StandardScaler is usually the safest bet)
scaler = StandardScaler()

# 3. Apply the scaling
# We "fit" (learn the average) and "transform" (apply the math) at once
df[['S4_R_scaled', 'S2_R_scaled']] = scaler.fit_transform(df[['S4_R', 'S2_R']])

print(df)
```

Age of child Household size

Since your file uses **Codes** (e.g., 1 = Male, 2 =Female), **do not scale categorical codes.**

- You only scale **Continuous** numbers (Age, Number of hours, Score out of 10).
- For categories (Gender, Region), you use **One-Hot Encoding** instead.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Vectorization vs. Encoding: What's the difference?

- **Embeddings/Vectorization:** These are used for *unstructured* data, like turning a whole paragraph of text or an image into a list of numbers so an AI can "understand" the meaning.

- **Encoding (What you need):** Your data is *structured*. You already have categories. You just need to turn "Woman" into 1 and "Man" into 0. This is called **One-Hot Encoding** or **Label Encoding**.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Convert Excel spreadsheet into a DataFrame

```
import pandas as pd

# Load the file
df = pd.read_excel('your_file_name.xlsx')

# Look at the first 5 rows
print(df.head())
```

Simplest way to load the data from xlsx file

```
# To read the sheet with the technical names
df_variables = pd.read_excel('bva_study.xlsx', sheet_name='VARIABLES')

# To read the sheet with the labels/descriptions
df_texts = pd.read_excel('bva_study.xlsx', sheet_name='TEXTS')
```

Handling Multiple Sheets

End