ITL PUBLIC SCHOOL
SUMMER ENGAGEMENT PROGRAM
ARTIFICIAL INTELLIGENCE (843)-XII

Project Title

Predicting YouTube Video Views Using Computational Thinking and Linear Regression

---

Problem Statement

In this project, students are required to build a simple machine learning model to predict the number of views a video may receive on YouTube using historical data. The dataset will include features such as likes, comments, subscriber count, and engagement metrics, along with the target variable "Views". Some entries in the dataset may have missing values, which students must identify using isnull() and handle appropriately using either deletion or imputation techniques like fillna().

After preparing the data, students are expected to apply Computational Thinking to break down the problem, identify patterns between features and views, and design a logical approach for prediction. They must then split the dataset into training (80%) and testing (20%) sets, train a Linear Regression model using the training data, and use it to predict video views on the test data. Finally, students must compare the predicted values with actual values to evaluate how well their model performs.

Objective

The objective of this project is to help students understand how missing data is identified and handled in real-life datasets using Python file handling concepts and basic data analysis. Students will explore how functions like isnull() can be used to detect absent or incomplete values, interpret why data may be missing (such as absence due to illness or participation in events), and apply computational thinking to clean, organize, and make meaningful conclusions from the given dataset.

Computational Thinking Approach

1. Decomposition

Break the problem into smaller tasks:

- Data collection
- Data cleaning
- Feature selection
- Model training
- Prediction and evaluation

2. Pattern Recognition

Identify relationships:

- More likes → more views
- Higher subscribers → higher baseline views
- More comments → higher engagement

3. Abstraction

Ignore non-numerical irrelevant details (like video titles) and focus on measurable features.

4. Algorithm Design

Develop a step-by-step process to clean data, train model, and evaluate results.

Model Implementation: Linear Regression

Step 1: Data Preparation

- **Features (X):** likes, comments, subscribers, etc.
- **Target (y):** views

Step 2: Train-Test Split (80/20 Rule)

- 80% → Training data
- 20% → Testing data

Step 3: Model Training

from sklearn.linear_model import LinearRegression

```
m = LinearRegression()
m.fit(x_train, y_train)
```

Step 4: Prediction

```
y_pred = m.predict(x_test)
```

Evaluation of the Model

1. Actual vs Predicted Comparison

Actual Views	Predicted Views
15000	14500
8000	8200
30000	29000

Project Deliverables

Students must submit the following:

1. Project Report File

Should include:

- Title page
- Objective
- Problem statement
- Computational Thinking explanation
- Dataset description
- Steps of implementation
- Model explanation
- Evaluation results
- Conclusion

---

2. Python Code File (.py or Notebook)

Must include:

- Data loading
- Handling missing values (isnull, fillna, dropna)
- Train-test split (80/20 rule)
- Linear Regression model training
- Prediction using m.predict()
- Output comparison

---

3. Dataset File (CSV)

- Original dataset used OR
- Self-created sample dataset

---

4. Output Screenshots

Include screenshots of:

- Code execution
- Model training output
- Actual vs Predicted table
- Final results

---

5. Reflection Sheet

Students should answer:

- Why is data cleaning important?
- What does Linear Regression do?
- How does YouTube-like platforms use prediction models?
