YouTube Video Popularity Prediction and Engagement Analysis

**Author:** Ergi Demiraj

**Professor**: Janghoon Yang

**Class:** CMPSC 445

**Date:** 23/ October/ 2025

**1. Project Description**

This project aims to predict YouTube video engagement and identify the factors that influence popularity. Two datasets were collected — one through web scraping and another using the YouTube Data API. The goal is to compare how each data source performs when training machine learning models. Two algorithms were implemented: Linear Regression and k-Nearest Neighbors (k-NN). The project includes data preprocessing, model training, evaluation, and visualization of key insights.

**2. Data Collection**

Data was collected from two sources:  
1. Web Scraping: Attempted to scrape trending videos using Beautiful Soup. However, YouTube’s dynamic content limited direct scraping, so data extraction relied on metadata URLs.  
2. YouTube Data API: Used the official API to collect structured video metadata including title, views, likes, comments, duration, and publish date. Around 100 videos were used for model training.

Example attributes collected:  
- Title  
- Channel  
- Publish Date  
- Duration  
- Views  
- Likes  
- Comments  
- Category ID

**Sample Data**

Below is an example of two rows from the preprocessed dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Title** | **Channel** | **Duration** (min) | **Views** | **Likes** | **Comments** | **Engagement** **Rate** |
| Exploring Mars Rover Technology | NASA Official | 6.5 | 245,000 | 4,800 | 310 | 0.0209 |
| The Future of Electric Cars | EV News Daily | 8.3 | 158,000 | 2,700 | 185 | 0.0182 |

**3. Data Preprocessing and Feature Engineering**

The collected data was cleaned and normalized. Missing values were handled, and numeric fields such as views, likes, and comments were converted to integers. Durations were standardized in minutes.

During preprocessing, the engagement\_rate column — calculated as (likes + comments) / views — was cleaned and standardized along with other numeric features. All relevant columns were then prepared for the model, including duration, views, likes, comments, and engagement\_rate.

After preprocessing, all columns were prepared for model training. The following transformations were made:

- Converted views, likes, and comments from strings to integers.  
- Converted duration from ISO 8601 format into total minutes.  
- Dropped rows with missing essential fields (title, channel, duration).  
- Created the engagement\_rate column using (likes + comments) / views.  
- Removed unneeded columns like tags and descriptions.

The resulting dataset was saved as 'youtube\_data\_preprocessed.csv'. These features were used in the models:

|  |  |
| --- | --- |
| **Feature** | Description |
| **Duration** | Length of the video (in minutes) |
| **Views** | Total view count |
| **Likes** | Total number of likes |
| **Comments** | Total number of comments |
| **Engagement\_rate** | Target variable representing viewer interaction |

**4. Model Development**

Two regression models were trained using scikit-learn:  
- Linear Regression: A simple baseline model that learns relationships between features and engagement rate.  
- k-Nearest Neighbors (k-NN): A non-parametric model that predicts engagement based on the average of nearby samples.

The dataset was split into 80% training and 20% testing. Models were evaluated using Mean Squared Error (MSE) and R² (coefficient of determination).

**5. Model Evaluation**

After training, both models were evaluated on test data. The results were exported to evaluation\_results.csv. Linear Regression achieved higher performance, showing a stronger fit to the engagement rate target.

**6. Visualization of Results**

Several visualizations were created to summarize findings, including model comparison, feature correlation, and engagement trends over time.

A graph with a bar and text

AI-generated content may be incorrect.

Figure 1: Model Performance Comparison

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 2: Feature Correlation Heatmap

A graph with blue lines

AI-generated content may be incorrect.

Figure 3: Engagement Rate Over Time

**7. Discussion and Conclusions**

This project demonstrates how machine learning can be used to analyze and predict YouTube engagement. The Linear Regression model provided more stable and interpretable results than k-NN. Feature correlation analysis revealed that likes and comments had the strongest relationship with engagement rate.  
  
Challenges included limited access to YouTube’s dynamic data and API quota restrictions. Future improvements could involve collecting larger datasets, including more video features such as tags, description keywords, and upload frequency, and experimenting with additional models like Decision Trees or Support Vector Regression.  
  
Overall, this project provided hands-on experience in data collection, preprocessing, machine learning model development, and visualization using Python.