Preprocessing:

➢ **Drop rows contain null values.**

Features	Preprocessing technique
Video id, Channel_title	Category encoding technique
trending_date,publish_time	we will make a new column (days_to _be_trend)resulting from subtracting the trending_date from publish_time and insert this column(days_to _be_trend)to dataset and drop these colmns(trending_date,publish_time) from dataset
Title, Tags , Video_description	➢ convert to lowercase ➢ remove Special Characters ➢ remove Single Characters ➢ remove Single Characters from the start ➢ Replace multiple spaces with single space Removing prefixed 'b' ➢ Removing links ➢ Applying natural language processing(TfidfVectorizer)

Comment_disapled, Rating_disabled, video_error_or_removed	Category encoding technique
Category_id, views,comment_count, likes,video_id,channel_title,days_to_ be_trend	Normalization technique

Analysis:

Apply correlation to dataset

- Likes depend on (The first is the most depend)
    1. views, comment_count
    2. Category_id ,days_to_be_trend
    3. Tags , Video_description

The sizes of your training, testing:

Split dataset to 30% -> test and 70%-> train.

Features We use:

- video_id
- channel_title
- views
- comments_count
- comments_disabled
- rating_disabled
- video_error_or_removed
- days_to_be_trend

Regression techniques:

- Polynomial Regression(degree = 2):
  train_mean_square_error : 0.00013246
  test_mean_square_error : 0.00011235
- Polynomial Regression(degree = 3):
  train_mean_square_error : 9.5193979
  test_mean_square_error : 0.0002954
- Polynomial Regression(degree = 4):
  train_mean_square_error : 6.58554
  test_mean_square_error : 0.01929116
- Polynomial Regression(degree = 5):(Overfitting)
  train_mean_square_error : 3.894829
  test_mean_square_error : 668837.39
- Multiple Regression:
  train_mean_square_error : 0.0002726
  test_mean_square_error : 0.0002139

We Use for Mode1 ->Polynomial Regression(deg = 2)
We Use for Mode2 ->Multiple Regression
Polynomial Regression is the best model .

Further techniques that were used to improve the results:

- Using Ridge Regularization To Avoide Overfitting.
- Using Text in Features To predict likes.
- Using Cross-Validation To Avoide Overfitting and split train to train and validate.

Conclusion:

After Showing correlation figure we Found that Likes most dependent on views and comments_count.