**PROJECT REPORT**

**ON**

**YOUTUBE ADVIEW PREDICTION**

Submitted in partial fulfillment for the requirement of the award of Internship

in

**Machine Learning**

Submitted By

**Soumya K C**

**KLS GOGTE INSTITUTE IF TECHNOLOGY BELAGAVI**

# ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my project mentor, Sanjoy Bhargab, Machine Learning Trainer, whose guidance and unwavering support made this report possible. His valuable insights and constructive feedback were instrumental in helping me navigate through challenges and complete this project successfully. I am sincerely grateful for his assistance and corrections whenever I needed them. I also wish to thank my wonderful parents, whose encouragement and moral support have been a constant source of motivation throughout this journey. Finally, I hereby declare that, to the best of my knowledge.

## INTRODUCTION

YouTube advertisers pay content creators based on ad-views and clicks for the goods and services being marketed. They want to estimate the adview based on other metrics like comments, likes etc. The problem statement is therefore to train various regression models and choose the best one to predict the number of ad-views. The data needs to be refined and cleaned before feeding in the algorithms for better results.

## Objective

The objective of this project is to develop a machine learning model to predict YouTube adview counts based on various YouTube metrics such as views, likes, dislikes, comments, published date, duration, and category of the video.

**Data Description**

* **train.csv**: Contains metrics and details for approximately 15,000 YouTube videos, including:
* Views: The number of unique views for each video.
* Likes: The number of likes for each video.
* Dislikes: The number of dislikes for each video.
* Comments: The number of unique comments for each video.
* Published Date: The upload date of the video.
* Duration: The length of the video (in minutes and seconds).
* Category: The category or niche of each video.
* Ad-view: The target variable representing the number of adviews for each video.
* **test.csv**: Used to evaluate the model's performance.

## Technology and Concepts

**Machine Learning:** A subset of artificial intelligence, machine learning involves training algorithms to learn from data and make predictions or decisions without explicit programming.

**Linear Regression:** A supervised learning technique used to predict a continuous target variable by fitting a linear relationship between the dependent and independent variables. Types include:

**Simple Regression:** Involves a single independent variable.

**Multiple Regression:** Involves multiple independent variables.

**Support Vector Machine (SVM):** A supervised learning algorithm used for both classification and regression. SVM works by finding a hyperplane that best separates the data points in an n-dimensional space. It is primarily used for classification but can be adapted for regression tasks.

**Decision Tree:** A decision-support tool that uses a tree-like graph of decisions and their possible consequences. It is used for both classification and regression tasks and helps break down complex decision-making processes into simpler parts.

**Artificial Neural Network (ANN):** A computational model inspired by the human brain, consisting of interconnected layers of nodes. ANNs learn from data to solve complex problems, making them foundational to artificial intelligence.

**Steps For adview prediction**

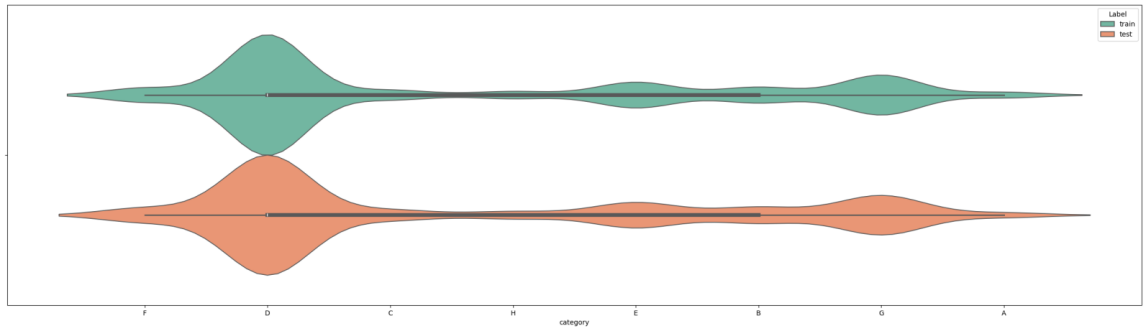
1. Import the datasets and libraries, check shape and datatype.
2. Visualize the dataset using plotting using heat maps and plots.
3. Clean the dataset by removing missing values and other things.
4. Transform attributes into numerical values and other necessary transformations
5. Normalize your data and split the data into training, validation and test set in the appropriate ratio.
6. Use linear regression, Support Vector Regressor for training and get errors.
7. Use Decision Tree Regressor and Random Forest Regressors.
8. Build an artificial neural network and train it with different layers and hyper-parameters.

Experiment a little. Use keras.

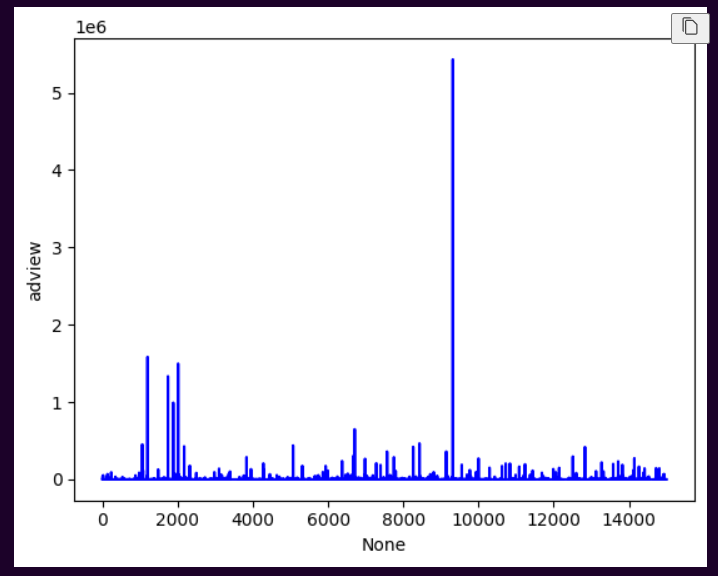
1. Pick the best model based on error as well as generalization.
2. Take the test dataset test.csv
3. Clean the test dataset by removing missing values
4. Remove unnecessary columns which has no impact to target variable
5. Transform the categorical attribute to numerical attribute.
6. Find prediction using the best algorithm
7. Save it into a new csv file by naming as Predictions\_Submission.csv

## Visualization

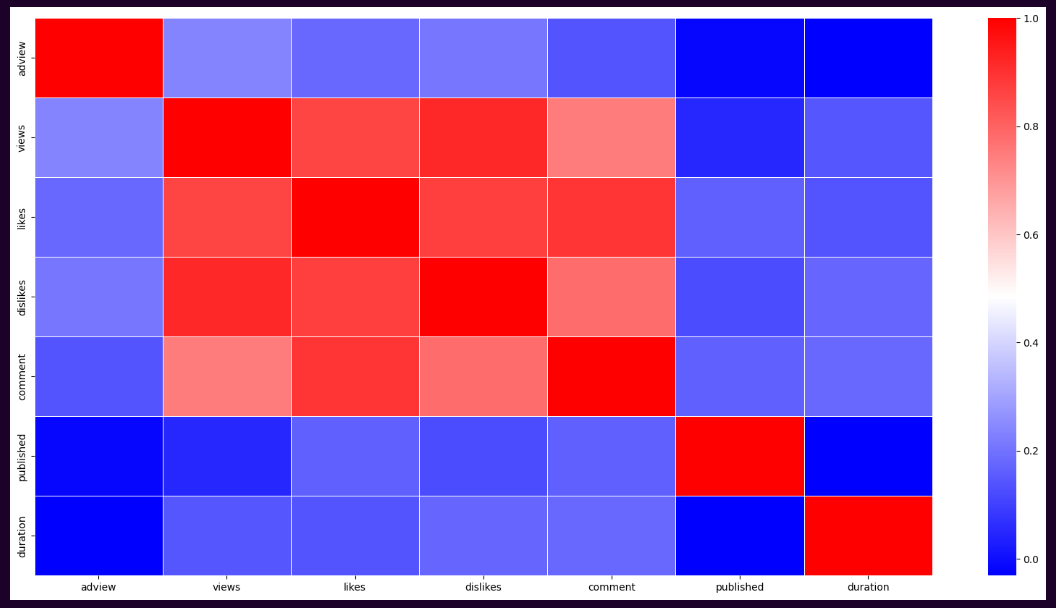
This is the violin plot-of “Category” column



This is the Line Chart of “adview” column



This is the heat-map which shows the co-relation of all columns with each other.



## Best Model

After training the dataset with multiple algorithms, we determined that the "Random Forest Regressor" outperformed others by achieving the lowest Root Mean Squared Error (RMSE). A lower RMSE indicates a more accurate model, which is why we chose to use the "Random Forest" algorithm for predicting the test dataset.

## Conclusions

We began this project with many ideas, some of which were initially too ambitious for our objectives. Our primary aim was to predict the view count of advertisements, allowing us to estimate the ad-views effectively. Although we made significant progress, there are still several aspects we could explore further given more time.