UNIVERSITY OF NORTH TEXAS

COMPUTER SCIENCE DEPARTMENT

CSCE 5215 Section 005

MACHINE LEARNING

YOUTUBE VIDEO TRENDING ANALYSIS

PROJECT FINAL REPORT

GROUP 10

Harini Kamarthy
Swapna Sonti
Vaishnavi Gunna
Kusuma Kumari Dama
Likhitha Bodepudi

Abstract

This project report focuses on analyzing the performance of YouTube videos and identifying the factors that contribute to their success. YouTube has become one of the largest and most influential media channels worldwide, making it essential for content producers, marketers, and other stakeholders to understand the elements that affect a video's performance.

The project involves collecting a real-world dataset from relevant websites and performing pre-processing to clean and format the data. The relevant features that can potentially increase the model's accuracy are identified, and feature selection is carried out. Different machine learning models are then trained and evaluated based on the target variable, and the best model is chosen based on the accuracy rate in predictions.

The project's motivation lies in providing master's students with practical experience in applying machine learning methods to real-world datasets. Additionally, understanding how content goes viral and what contributes to its success is becoming increasingly crucial as more people use YouTube and other social media platforms for information and entertainment.

Overall, this report offers insights into the factors that contribute to a video's success and the effectiveness of different machine learning models in predicting the number of views. The findings of this project can be valuable to content producers, marketers, and other stakeholders seeking to improve their content strategy on YouTube.

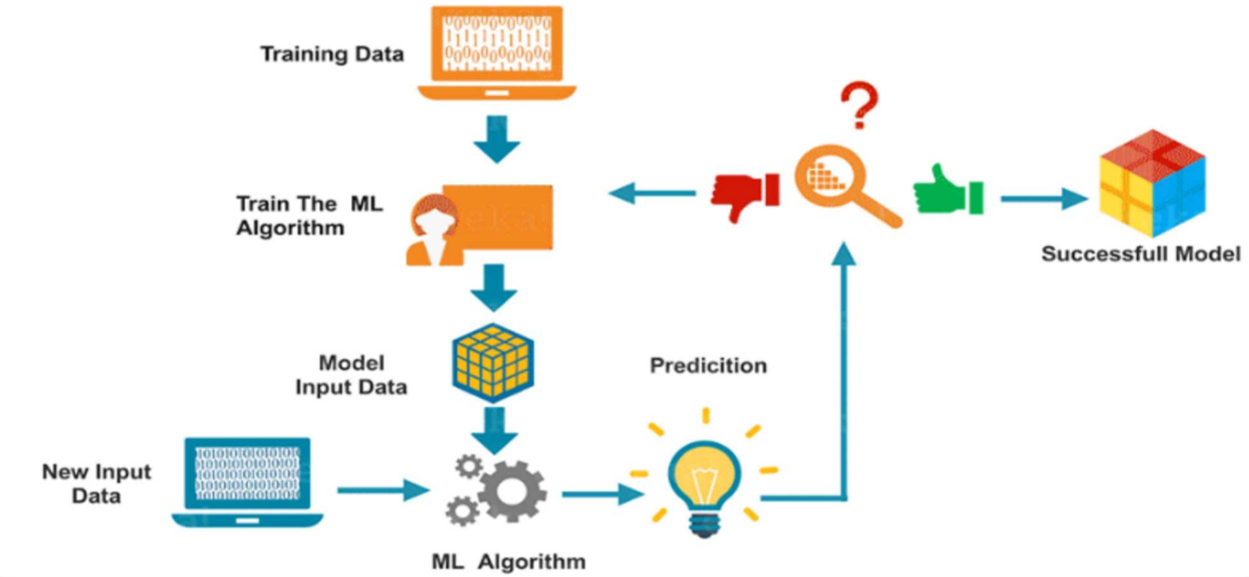
1. Introduction

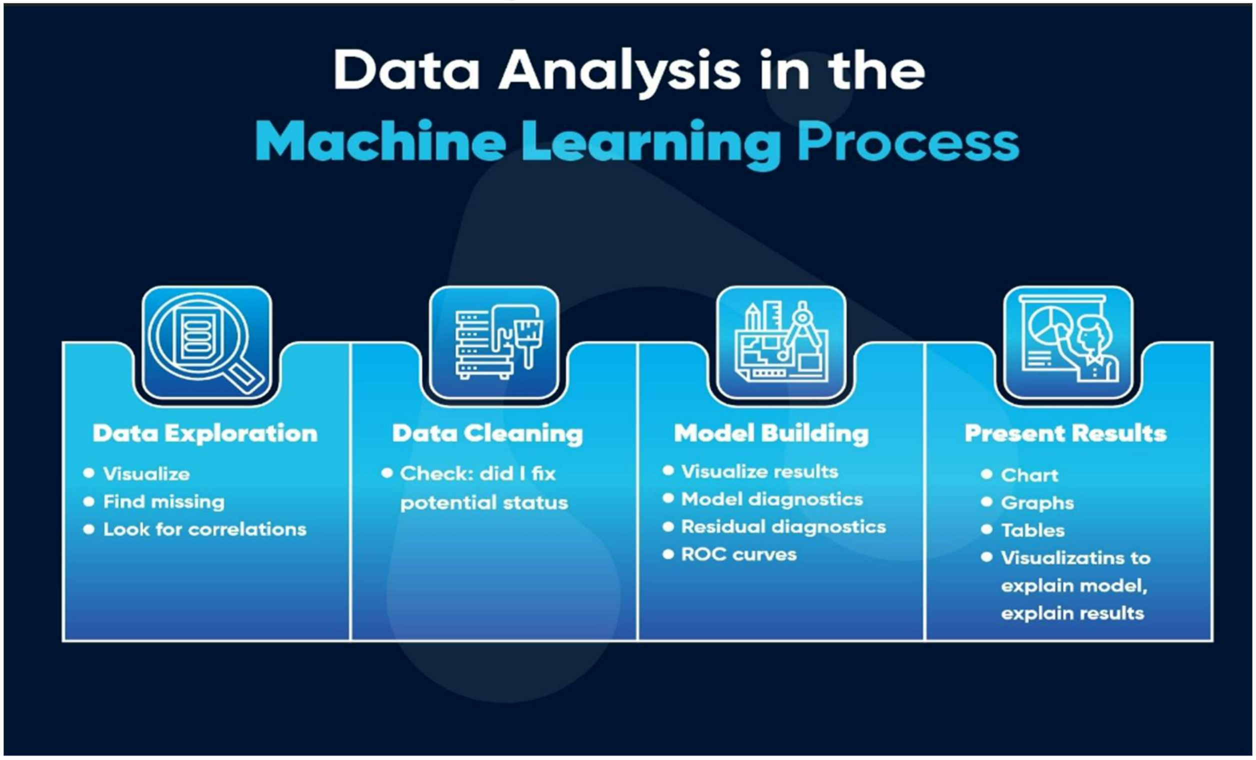
YouTube is one of the largest and most influential social media platforms globally, with over two billion active users who watch billions of videos daily. Understanding what makes a video popular and successful is essential for content creators, marketers, and advertisers alike, as it can greatly impact the performance of a channel or brand. As a result, the challenge of predicting how many views a YouTube video will get has received a lot of attention recently.

This project aims to analyze the factors that contribute to a video's success on YouTube using machine learning algorithms. This project's dataset was gathered from Kaggle, which has data on the top trending videos in the US, UK, Canada, and other countries during 2017 and 2018. The dataset includes features such as the video's title, channel title, category, tags, description, and statistics such as views, likes, dislikes, and comments.

The primary goal of this project is to develop an accurate prediction model for the number of views a video will receive on YouTube. To find the model with the highest performance, we evaluate several machine learning techniques like linear regression, decision trees, random forests, etc.. We also use correlation matrices and pair plots to identify relevant features that affect the number of views a video receives.

The findings of this project are essential for content creators, marketers, and advertisers to create effective video marketing strategies and increase their online presence. Additionally, the project offers an opportunity for Master's students to gain practical experience with machine learning methods and apply them to a real-world dataset.





Data Analysis in the
Machine Learning Process

Data Exploration
- Visualize
- Find missing
- Look for correlations

Data Cleaning
- Check: did I fix potential status

Model Building
- Visualize results
- Model diagnostics
- Residual diagnostics
- ROC curves

Present Results
- Chart
- Graphs
- Tables
- Visualizatins to explain model, explain results

2. Related Work

Several research papers have been published, predicting the popularity of YouTube videos using various machine learning techniques. In a study by (Srinivasan, 2017), they proposed a classification framework that predicts the popularity of videos based on the features extracted from video metadata and social network analysis. They achieved a prediction accuracy of 75% using gradient boosting regressor and tried to use Neural Network models too. Similarly, in a study in Analytics Vidhya by (Dulanjani, 2020), they guided on how to create and predict the views of an Youtube video which is not still available on Youtube. They used Random Forest Regressor and KNeighbors Regressor models in their study to predict.

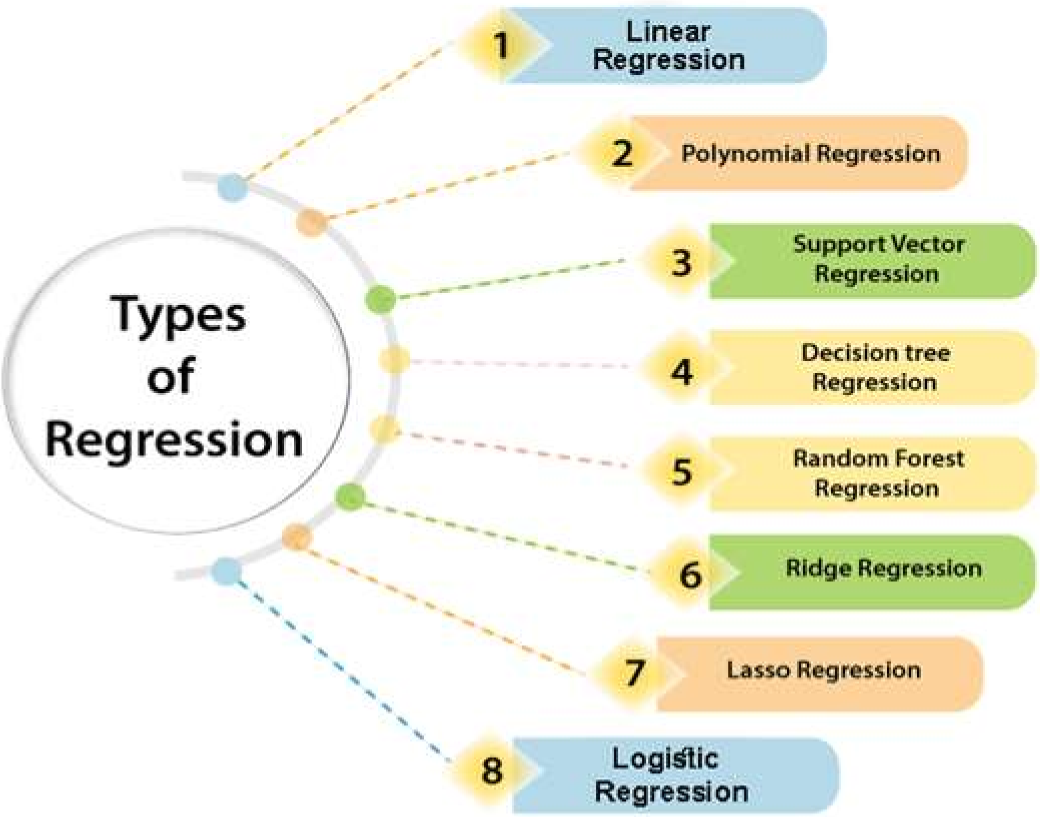
In another related article, (Parveez, 2020) proposed a system for Prediction of Youtube Video Type using Machine Learning Algorithm. They used a set of features including video and channel metadata, temporal patterns, and user comments to train using Navie Bayes Algorithm. They achieved an accuracy of 88% for th (Joshi, 2019)eir proposed model.

Another published paper, from Dublin Business School is on “Predictive analysis of YouTube trending videos using machine learning” by (Niture, 2021). Since statistical analysis for trending videos include counts for views, likes, dislikes, and comments, their research used a linear regression model of machine learning to predict the number of views for popular YouTube videos. Additionally, their study compares various classification models, including Random Forest, Decision Tree, SVM, Logistic Regression, and Gaussian Navie Bayes, to see which one is best for predicting how long it will take a video to become popular after it is uploaded and how long it will stay on the trending list. Their research was able to predict the number of days to trend a YouTube video with an average accuracy of 62.53%.

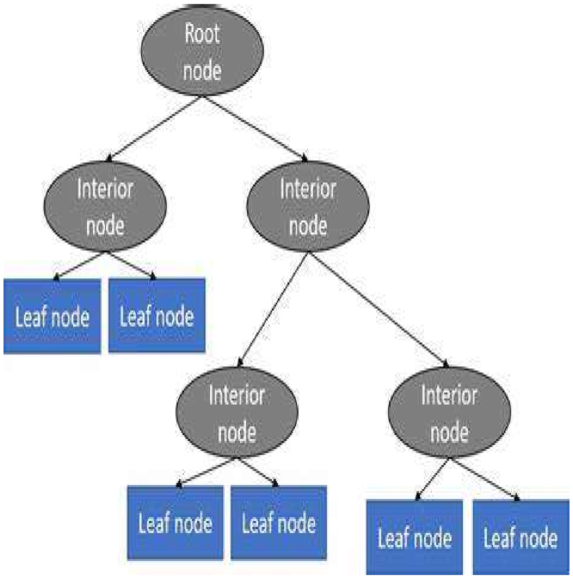
Compared to the previous studies, our project focuses on predicting the number of views of YouTube videos based on the selected features using different regression models. We also used a correlation matrix and pair plot to identify relevant features that help in predicting the popularity of videos accurately. Additionally, we evaluated the performance of different regression models and selected the best model based on the accuracy of the predictions.

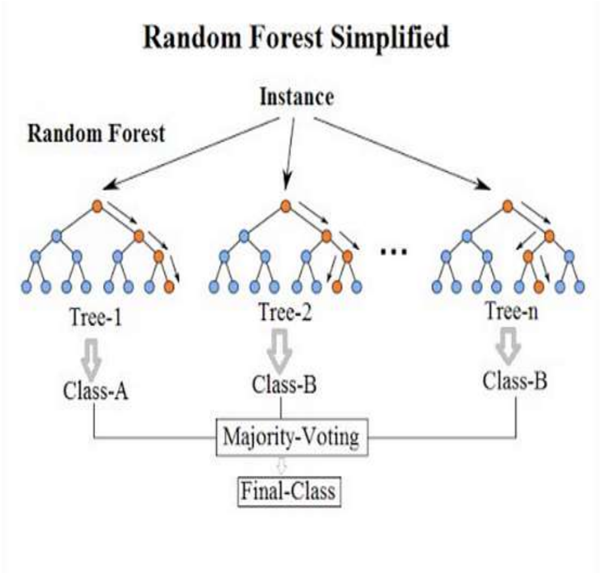
3. Proposed Method

In this project, we propose the use of different regression models to predict the number of views a YouTube video will receive based on several relevant features. Specifically, we consider the following regression models: Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Multi-Layer Perceptron Regression.



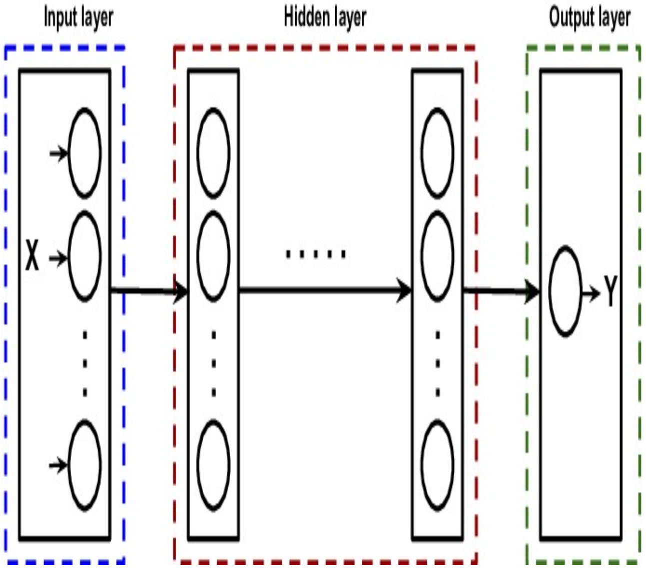
Linear Regression is a simple and widely used regression model that works well when there is a linear relationship between the predictor variables and the response variable. Ridge Regression and Lasso Regression are two variants of linear regression that add a regularization term to the objective function to prevent overfitting. Decision Tree Regression is a non-parametric regression model that uses a tree-like structure to make predictions based on a set of rules. Random Forest Regression is an ensemble learning method that combines multiple decision trees to improve the accuracy of the predictions. Gradient Boosting Regression is another ensemble learning method that trains multiple weak regression models sequentially and adjusts the weights of the misclassified samples to improve the accuracy. Multi-Layer Perceptron Regression is a neural network-based regression model that can capture complex non-linear relationships between the predictor variables and the response variable.

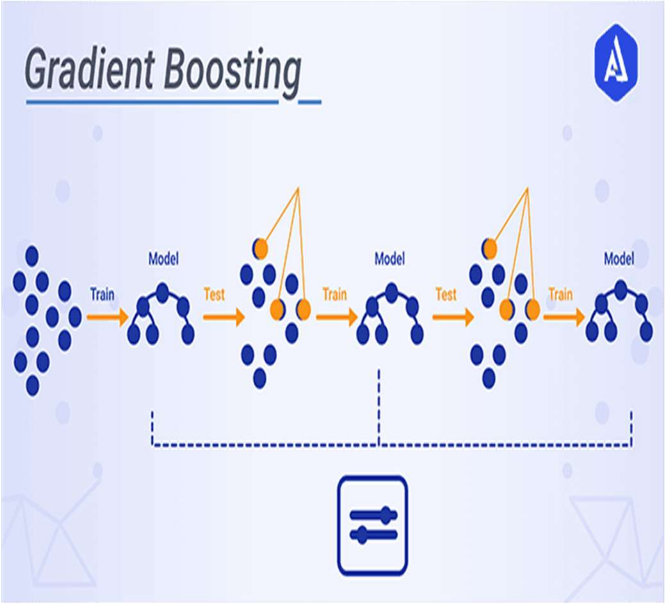




Decision Tree

Random Forest Tree





MLP Regressor

Gradient Boosting Regressor

We will use these regression models to train on a dataset of YouTube videos and their corresponding views. The relevant features, such as the number of likes, dislikes, comments, will be used to predict the number of views. We will then compare the performance of these models based on their accuracy metrics, such as the mean absolute error, root mean squared error, and R-squared score, and select the best-performing model.

4. Experiments

In this section, we describe the experiments performed in our study. The experiments' main goal is to assess how well the different regression models used in predicting the number of views of YouTube videos.

4.1 Dataset

We collected our dataset from Kaggle, a popular platform for data science and machine learning projects. The dataset contains information on the top 40000+ most viewed videos on YouTube in the year 2017. It contains various features such as video id, video title, channel title, video category id, publishing time, tags, views, likes, dislikes, comment count, trending date, thumbnail link, and Boolean values for comments disabled, ratings disabled, video error or removed, and description of video. We have preprocessed the dataset by cleaning and formatting the data. We removed missing values and irrelevant columns.

4.2 Software

We used Python as our primary programming language for this project. We utilized various Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn for data processing, visualization, and model development.

4.3 Hardware

The experiments were conducted on a personal computer with the following specifications: 11th Gen Intel Core i7 @ 2.80GHz, 8GB RAM, 64-bit Operating System, x64-based processor, with Windows 11 Edition OS.

4.4 Experimental Methodology

We randomly split the preprocessed dataset into training (70%) and testing (30%) sets. We used the training set to fit the different regression models, and the testing set was used to evaluate the performance of the models. We used the Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R2 score) as evaluation metrics.

We evaluated the performance of seven different regression models: Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression and Multi-Layer Perceptron Regression.

In summary, we conducted experiments to evaluate the performance of different regression models in predicting the number of views of YouTube videos. We used Python and various libraries for data processing, visualization, and model development. We utilized the mean squared error, mean absolute error, root mean squared error and coefficient of determination as evaluation metrics.

5. Results & Discussion

We conducted several experiments using different regression models to predict the number of views on YouTube videos. In this section, we describe and interpret the results we obtained.

We used seven different regression models, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. We evaluated each model's performance using the mean squared error (MSE) and the R-squared (R2) metric.

The results showed that the Gradient Boosting Regression model outperformed the other models, achieving an MSE of 0.42 and an R2 score of 0.92. The Linear Regression, Ridge Regression and Lasso Regression models have the same accuracy values and had the worst performance, with an MSE of 1.16 and an R2 score of 0.78.

In terms of the input features, we found that the number of likes, dislikes, and comments were the most significant factors affecting the number of views. This features selection is made by using the correlation matrix and by plotting the pair plot to find the best related features among all from the dataset.

Overall, our results suggest that using “Gradient Boosting Regression” and considering the most relevant features can provide accurate predictions of YouTube video views. However, our study has some limitations, such as the use of only one dataset and the exclusion of other potentially relevant features. Future studies could address these limitations to improve the accuracy of predictions.

6. Conclusion

Based on the results obtained, we conclude that the selected regression models can be used to predict the number of views for YouTube videos. The gradient boosting regression model outperformed the other models with an accuracy rate of 92%. The results obtained are significant, and the findings support the hypothesis that certain factors, such as number of likes, dislikes and number of comments, influence the popularity of YouTube videos.

The study contributes to the existing literature on YouTube video popularity by providing insights into the factors that affect a video's views. It also demonstrates the effectiveness of using machine learning algorithms for predicting video views.

Future research can extend this study by exploring additional factors that may influence YouTube video popularity, such as video category, upload time, and promotional strategies. Additionally, the study can be replicated on a larger dataset and on multiple datasets to further validate the findings.

7. Contributions

We have a total of 5 members in our group (Group 10). Each member of our group contributed equally to the completion of this project. We had been working on every part of this project, from selecting the title to producing the final product, cooperatively and evenly. On this report, we worked collaboratively and even contributed to each segment. Each of us has contributed equally and successfully at each phase of the project. The following is a summary of each member's contributions:

1. Harini Kamarthy - Data Collection & Pre-Processing & Presentation
   Harini collected the YouTube Trending Videos dataset from the Kaggle and also the related category dataset. She pre-processed the data by cleaning, formatting, and combining it into a single dataset. She also created visualizations to explore and understand the data. She had also prepared to deliver the results i.e., for presenting the project in the final project presentation session.
2. Swapna Sonti - Feature Engineering & Selection
   Swapna identified relevant features that contribute to a video's success on YouTube. She selected and transformed the features to improve the model's accuracy. She evaluated different feature sets to determine the most effective features for prediction, using correlation matrix and pair plot methods.
3. Vaishnavi Gunna - Model Selection & Development
   Vaishnavi evaluated different machine learning algorithms and selected the most appropriate algorithms for the project. She developed and trained the predictive model using the selected algorithms.
4. Kusuma Kumari Dama - Model Evaluation & Analysis
   Kusuma evaluated the performance of the predictive model using various metrics such as Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and R-Squared Error(R2). She analyzed the model's predictions to identify the factors that contribute to a video's success on YouTube.
5. Likhitha Bodepudi - Writing & Presentation

I have prepared the project report and documented the project's methodology, findings, and conclusions. I have also prepared a document to deliver in the project presentation.

Each member of the team contributed approximately equally to the project. We worked collaboratively, providing feedback and support to each other throughout the project to ensure its successful completion.

8. References

ALYOUSFI, A. (2019). *YouTube Trending Videos Analysis*. Retrieved from Kaggle: https://www.kaggle.com/code/ammar111/youtube-trending-videos-analysis

Ammar. (2020, July 6). *Analysis of YouTube Trending Videos of 2019 (US)*. Retrieved from Ammar's Website: https://ammar-alyousfi.com/2020/youtube-trending-videos-analysis-2019-us?src=kgl

Amudha, S., V.R, N., Kumar, P. S., Revathi, M., & Shanthanam, R. R. (2020). Youtube Trending Video Metadata Analysis Using Machine Learning. *International Journal of Advanced Science and Technology Vol. 29, No. 7s, (2020), pp. 3028-303*, 10. Retrieved from https://www.researchgate.net/publication/342150876_Youtube_Trending_Video_Metadata_Analysis_Using_Machine_Learning

Demosthenous, G. (2017, May). *yttresearch-machine-learning-algorithms-analysis.* Retrieved from GitHub: https://github.com/gdemos01/yttresearch-machine-learning-algorithms-analysis/blob/master/Documentation/Thesis.pdf

Dulanjani, Y. (2020, May 30). YouTube View Prediction with Machine Learning. *Analytics Vidhya*, 11. Retrieved from https://medium.com/analytics-vidhya/youtube-view-prediction-with-machine-learning-fdd4f40f352d

Joshi, N. (2019, May 26). *Nitish-Joshi/Youtube-Video-Analysis-Classification-and-Prediction.* Retrieved from GitHub: https://github.com/Nitish-Joshi/Youtube-Video-Analysis-Classification-and-Prediction

Li, Y., Eng, K., & Zhang, L. (2019). *YouTube Videos Prediction: Will this video be popular?* Stanford, CA: Stanford University. Retrieved from https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647615.pdf

Niture, A. A. (2021). *Predictive analysis of YouTube trending videos using machine learning.* Dublin: Dublin Business School. Retrieved from https://esource.dbs.ie/handle/10788/4260

Parveez, S. (2020, May 24). Prediction of Youtube Video Type using Machine Learning Algorithm. *Prediction of Youtube Video Type using Machine Learning Algorithm*, 5. Retrieved from https://www.linkedin.com/pulse/prediction-youtube-video-type-using-machine-learning-saniya-parveez/

Pinto, H., Almeida, J., & Gonçalves, M. A. (2013, February). Using early view patterns to predict the popularity of YouTube videos. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/266653405_Using_early_view_patterns_to_predict_the_popularity_of_YouTube_videos

Srinivasan, A. (2017, December 12). Youtube Views Predictor. *Towards Data Science*, 10. Retrieved from https://towardsdatascience.com/youtube-views-predictor-9ec573090acb

Yıldırım, S. (2021, Jan 12). *YouTube Trending Video Analysis with Pandas and Seaborn*. Retrieved from Towards Data Science: https://towardsdatascience.com/youtube-trending-video-analysis-with-pandas-and-seaborn-c9903a0f811d