YouTube Trending Video Metadata Analysis Using Machine Learning

A main Project Report submitted in the partial fulfillment of the

requirements for the award of the degree

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

G.V. Karthik (19471A05E6)

Under the esteemed guidance of

M.VenkataRao **Asst.Prof**

DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING

NARASARAOPETA ENGINEERING COLLEGE:

NARASARAOPET (AUTONOMOUS)

Accredited by NAAC with A+ Grade and NBA under Cycle-1
NIRF rank in the brand of 251-320 and an ISO 9001 : 2015 Certified
Approved by AICTE , New Delhi,Permanently Affliated to JNTUK, Kakinada
KOTAPPAKONDA ROAD,YALAMANDA VILLAGE, NARASARAOPET-522601
2022-2023

NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPETA (AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the main project entitled "YouTube Trending Video Metadata Analysis Using Machine Learning" is a bonafide work done by "T. Gopi (19471A05I8) ,G.V. Karthik (19471A0E6)" in partial fulfilment of the requirements for the award of the degree of BACHELOR OF TECHNOLOGY in the department of COMPUTER SCIENCE AND ENGINEERING during 2022-2023.

PROJECT GUIDE
M.Sireesha ,MTech.,Ph.D.

PROJECT CO-ORDINATOR
M.Sireesha ,MTech.,Ph.D.

HEAD OF THE DEPARTMENT
Dr.S.N.TirumalaRao,M.Tech.,Ph.D

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairman sir **M.V. Koteswara Rao**, **B.sc** who took keen interest in us in every effort throughout this course. We owe out gratitude to our principal **Dr.M.Sreenivasa Kumar, M.Tech,Ph.D(UK),MISTE,FIE(1)** for his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr.S.N.Tirumala Rao,M.Tech, Ph.D** H.O.D. CSE department and our guide **M.Venkata Rao** of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **Dr.M.Sireesha,M.Tech,PhD**. Assistant Professor of the project for extending her encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff to department for their cooperation and encouragement during our B.Tech degree. We have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents. We affectionately acknowledge the encouragement received from our friends those who involved in giving valuable suggestions had clarifying out all doubts which had really helped us in successfully completing our project.

By

G.V.Karthik (19471A05E6)

ABSTRACT

This study analysed the metadata of over 40,000 videos in YouTube's Trending section using the machine learning techniques to identify factors contributing to a video's Virality. Machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were trained on the dataset to predict the popularity of trending videos based on their metadata. Factors such as the video category, title, description, view count, and the comment count were evaluated to identify their impact on a video's virality. The models achieved high accuracy scores with a score of 100% in the Linear Regression model score and R-squared score, and 99.98% in Random Forest Regressor. Based on the results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. This study demonstrates the value of using machine learning techniques to analyse large datasets of metadata and provides insights into the factors that contribute to video virality on YouTube.

NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

INSTITUTE VISION AND MISSION

INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community,

INSTITUTION MISSION

M1: Provide the best class infra-structure to explore the field of engineering and research

M2: Build a passionate and a determined team of faculty with student centric teaching,imbibing experiential, innovative skills

M3: Imbibe lifelong learning skills, entrepreneurial skills and ethical values in students for addressing societal problems

NEC NARASARAOPETA ENGINEERING COLLEGE

(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research and ethical values to cater to the needs of industry and society.

MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

M1: Mould the students to become Software Professionals, Researchers and Entrepreneurs by providing advanced laboratories.

M2: Impart high quality professional training to get expertize in modern software tools and technologies to cater to the real time requirements of the Industry.

M3: Inculcate team work and lifelong learning among students with a sense of societal and ethical responsibilities.

Program Specific Outcomes (PSO's)

PSO1: Apply mathematical and scientific skills in numerous areas of Computer Science and Engineering to design and develop software-based systems.

PSO2: Acquaint module knowledge on emerging trends of the modern era in Computer Science and Engineering

PSO3: Promote novel applications that meet the needs of entrepreneur, environmental and social issues.

Program Educational Objectives (PEO’s)

The graduates of the programme are able to:

PEO1: Apply the knowledge of Mathematics, Science and Engineering fundamentals to identify and solve Computer Science and Engineering problems.

PEO2: Use various software tools and technologies to solve problems related to academia, industry and society.

PEO3: Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

PEO4: Pursue higher studies and develop their career in software industry.

Program Outcomes

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, research literature, and analyzecomplex engineering problems reaching substantiated conclusions using firstprinciples of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions:** Design solutions for complexengineering problems and design system components or processes that meetthe specified needs with appropriate consideration for the public health andsafety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems:** Use research-basedknowledge and research methods including design of experiments, analysis andinterpretation of data, and synthesis of the information to provide validconclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques,resources, and modern engineering and IT tools including prediction andmodeling to complex engineering activities with an understanding of thelimitations.

6. **The engineer and society:** Apply reasoning informed by the contextualknowledge to assess societal, health, safety, legal and cultural issues and theconsequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability:** Understand the impact of theprofessional engineering solutions in societal and environmental contexts,and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Project Course Outcomes (CO’S):

CO425.1: Analyse the System of Examinations and identify the problem.

CO425.2: Identify and classify the requirements.

CO425.3: Review the Related Literature

CO425.4: Design and Modularize the project

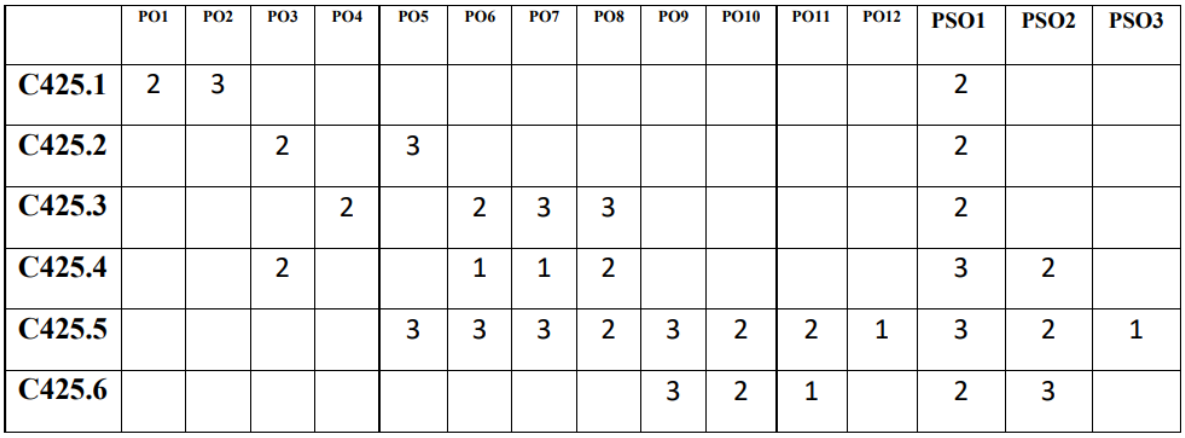
CO425.5: Construct, Integrate, Test and Implement the Project.

CO425.6: Prepare the project Documentation and present the Report using appropriate method.

Course Outcomes – Program Outcomes mapping

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1		√											√		
C425.2	√		√		√								√		
C425.3				√		√	√	√					√		
C425.4			√			√	√	√					√	√	
C425.5					√	√	√	√	√	√	√	√	√	√	√
C425.6									√	√	√		√	√	

Course Outcomes – Program Outcome correlation



	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1	2	3											2		
C425.2			2		3								2		
C425.3				2		2	3	3					2		
C425.4			2			1	1	2					3	2	
C425.5					3	3	3	2	3	2	2	1	3	2	1
C425.6									3	2	1		2	3	

Note: The values in the above table represent the level of correlation between CO's and PO's:

1. **Low level**

2. **Medium level**

3. **High level**

Project mapping with various courses of Curriculum with Attained PO's:

Name of the course from which principles are applied in this project	**Description of the device**	**Attained PO**
C3.2.4, C3.2.5	Gathering the requirements and defining the problem, plan to develop a **Youtube Trending video Metadata Analysis Using Machine Learning**	PO1, PO3
CC4.2.5	Each and every requirement is critically analyzed, the process model is identified and divided into **five modules**	PO2, PO3
CC4.2.5	Logical design is done by using the unified modelling language which involves individual team work	PO3, PO5, PO9
CC4.2.5	Each and every module is tested, integrated, and evaluated in our project	PO1, PO5
CC4.2.5	Documentation is done by all our four members in the form of a group	PO10
CC4.2.5	Each and every phase of the work in group is presented periodically	PO10, PO11
CC4.2.5	Implementation is done and the project will be handled by the **Predicting Youtube Tredning Videos using Machine Learning**	PO4, PO7
CC4.2.8 CC4.2.	**The physical design includes hardware components like Jupyter,Anaconda,Flask**	PO5, PO6

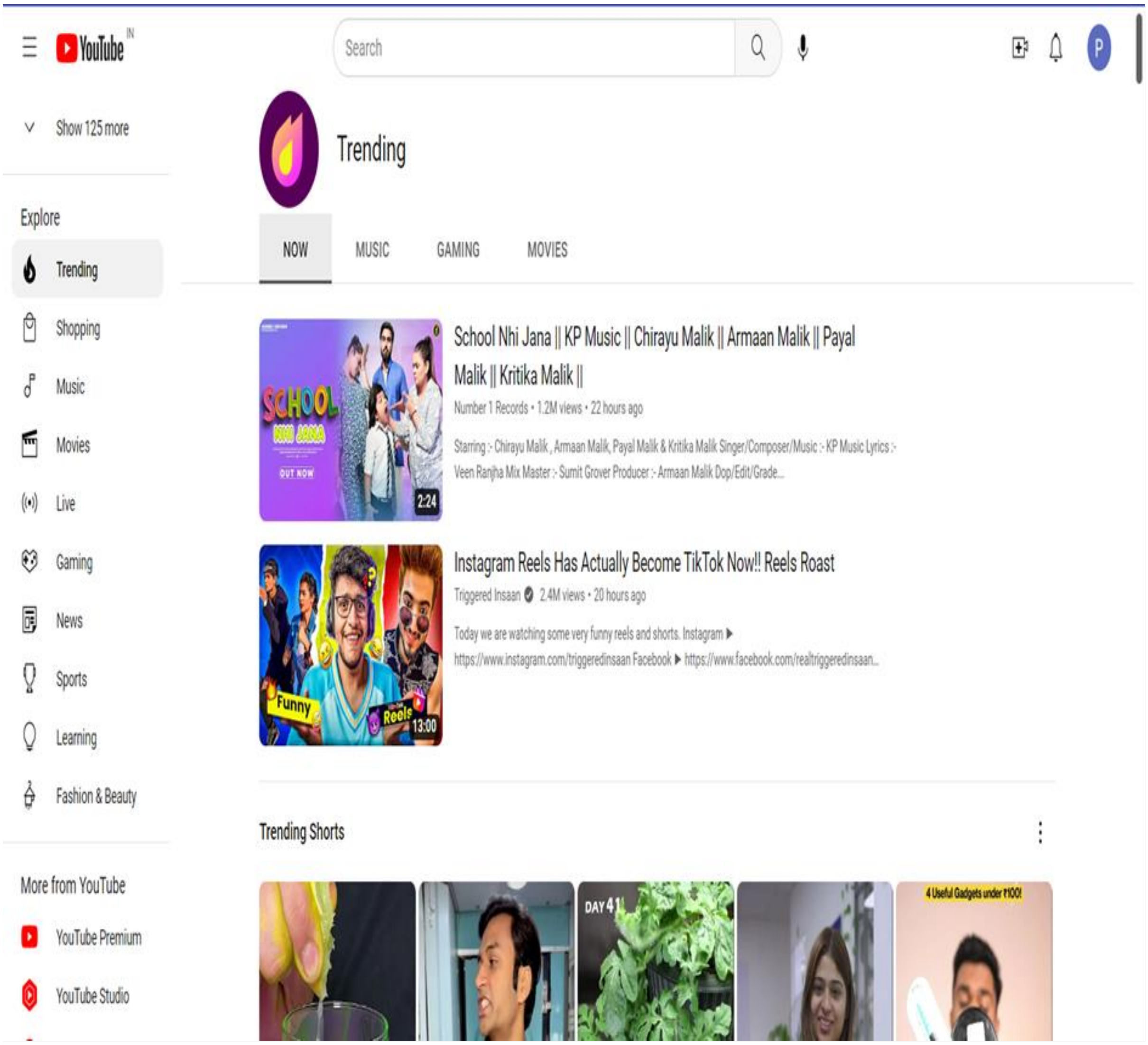
INDEX

List of Figures

1.INTRODUCTION

1.1 Introduction

YouTube is one of the most popular platforms for sharing videos and has become an important tool for content creators and marketers. The Trending section on YouTube highlights the most popular videos on the platform, making it a highly sought-after spot for creators to showcase their content. Understanding the factors that contribute to a video's virality can provide insights into YouTube's audience and content trends, and help creators and marketers optimize their content for maximum visibility and reach.



School Nhi Jana || KP Music || Chirayu Malik || Armaan Malik || Payal Malik || Kritika Malik ||

Number 1 Records • 1.2M views • 22 hours ago

Starring :- Chirayu Malik , Armaan Malik, Payal Malik & Kritika Malik Singer/Composer/Music :- KP Music Lyrics :- Veen Ranjha Mix Master :- Sumit Grover Producer :- Armaan Malik Dop/Edit/Grade...

Instagram Reels Has Actually Become TikTok Now!! Reels Roast

Triggered Insaan 2.4M views • 20 hours ago

Today we are watching some very funny reels and shorts. Instagram ▶ https://www.instagram.com/triggeredinsaan Facebook ▶ https://www.facebook.com/realtriggeredinsaan...

Figure 1.1 YouTube Trending Page

In recent years, analysts have increasingly turned to machine learning techniques to analyse large datasets of metadata, enabling insights into user behaviour and content trends on online platforms. Our study leverages these techniques, applying the Linear Regression, Random Forest Regressor, and Decision Tree Regressor to analyse metadata of over 40,000 videos in YouTube's Trending section. Our objective is to identify the factors that influence a video's Virality on YouTube and provide actionable recommendations to creators and marketers for optimizing their content on the platform. By doing so, we hope to provide insights into the value of using machine learning techniques to analyze large datasets of metadata and their potential to inform content strategies and marketing efforts on online platforms like YouTube.

Figure 1.2 YouTube Trending

1.2 Existing System

There are many existing systems for analysing YouTube video meta data .Some of the populare ones are:

1.Google Trends – a free tool that provides insights into search trends and popularity of keywords related to YouTube videos.

2.VidIQ - a paid tool that provides analytics and optimization tools for YouTube videos to help creators increase their views and subscribers.

3.TubeBuddy - a paid tool that offers similar features to VidIQ.

Disadvantages:

1.Google Trends:

- It doesn’t provide specific data about individual YouTube videos, only general keyword trends.
- It doesn’t provide any information about the performance of specific channels or creators.
- It may not be as accurate in reflecting the latest trends as real-time tools that focus specifically on YouTube data.

2.VidIQ:

- It is a paid tool, so not accessible to everyone.
- The basic plan may not provide all the features that some creators need, and the higher-tier plans can be expensive.
- It can be overwhelming for beginners and requires some technical knowledge to fully utilize.

3.TubeBuddy:

- It is also a paid tool, which may not be affordable for all creators.

- Some of the features may be redundant with what YouTube already offers for free, such as keyword research and video optimization.

1.3Proposed System

Building a tool that uses machine learning algorithms to predict the success of YouTube videos based on their metadata. Overall, proposed system should aim to help YouTube creators optimize their videos for success and increase their views and subscribers.

1.4. System Requirements

1.4.1 Hardware Requirements:

System type	:	intel®core™i7-7500UCPU@2.70gh
Cache memory	:	4 MB
RAM	:	12 GB
Hard Disc	:	8 GB

1.4.2 Software Requirements:

Operating system	:	windows 10, 64 bit OS
Coding language	:	Python
Python distribution	:	Anaconda, Spyder, Flask

2. LITERATURE SURVEY

2.1 Machine Learning

A literature survey on YouTube trending video metadata analysis using the machine learning reveals that previous studies have focused on different aspects of YouTube's platform and content. Researchers have used machine learning algorithms to analyze YouTube videos and predict their popularity, engagement, and sentiment analysis.

For instance, a study by Anusha et al. (2020) used machine learning techniques to predict the popularity of YouTube videos based on their thumbnail images. The authors found that image aesthetics, contrast, and emotion were important factors in predicting video popularity.

In another study by Abdelhaq etal. (2021), the machine learning algorithms were used to analyse the sentiment of YouTube comments on a particular video. The authors found that the use of deep learning algorithms significantly improved the accuracy of the sentiment analysis.

Furthermore, a study by Kim et al. (2019) used machine learning techniques to predict the success of YouTube channels based on their content and engagement metrics. The authors found that the length and frequency of videos, as well as likes and dislikes, were significant predictors of channel success.

Tejal Rathod and Mehul Barot explored trend analysis on twitter for predicting public opinion on ongoing events. Study implemented various classification algorithms to predict the positive and negative classes. SVM (Support Vector Machine) and Naïve Bayes 15 algorithms have the better performances. Study is limited to textual data due to social platforms limitation (Tejal Rathod and Mehul Barot, 2018).

Trzcinski and Rokita, developed a regression method for predicting the popularity of an online video based on its number of views. Study implements Support Vector Regression induced with Gaussian Radial Basis Functions. Robustness and the non-linear aspect of the developed method improves accuracy results. Research explored the impact of video’s visual features, such as outputs of deep neural networks, on popularity prediction. Study also denotes that popularity prediction accuracy can be improved by combining early distribution patterns with social and visual features (Trzcinski and Rokita, 2017). Study implements UL, ML, MRBF and SVR algorithms, Support Vector Regression algorithm gives the better results.

Flavio et al. studied the importance of UGC (User generated content) as features for predicting YouTube's trending videos. Study implemented a new time series clustering algorithm, called K-Spectral Clustering (KSC). And research implemented 5-fold cross validation method for results evaluation. Study finds that clusters having lower F1 score are harder to predict (Figueiredo et al., 2016).

Hoiles, Aprem and Krishnamurthy explored the mata-data features such as title, tag, thumbnail and description's impact on popularity and trendiness of a YouTube videos (Hoiles, Aprem and Krishnamurthy, 2017). Study implemented various Machine Learning algorithms to predict the popularity of a YouTube video based on video's mata features as well as other aspects such as number of subscribers etc., study shows that CI Random Forest algorithm has the highest r2 value for prediction much like the current thesis.

Ouyang, Li and Li, explore the prediction of popularity of online videos. Study parts the popularity forecasting problem into two tasks: video's popularity prediction and video's view count prediction. Research first predict the future popularity levels of videos, with key set of features and various classification algorithms. Then, according to the popularity levels, study implemented a specialized regression models to predict the view count. Study was implemented on Youku, a Chinese social media platform. SVM and KNN algorithms shows better performance (Ouyang, Li and Li, 2016).

A lot of research has been done on YouTube platform as it is one of the biggest user generated content platform. Text mining, Natural Language Processing, sentiment analysis are few research areas which are popular amongst peers to perform on YouTube. Despite its importance YouTube trending videos analysis have not been a well-researched area yet. YouTube recommendation system has been analysed by many, but trending video analysis still holds a lot of scope.

Zhou et al. studied the impact of YouTube recommendation system on video views and concluded that there is a strong correlation between view count of a video and average view count of its top referred video by recommendation system (Acm.org, 2010).

Also, Davidson et al. studied recommendation system through CTR (click through rate) of videos on home page. They conclude that recommendation by YouTube account for 60% of all video clicks on YouTube homepage (Davidson et al., 2010).

Now, a few approaches have been made towards YouTube trending videos research such as Prabha et. al. discussed predicting the popularity of trending videos in YouTube using sentiment analysis. Their study chose the NLP path to predict the popularity of trending videos. After discussing various classifiers such as naïve Bayes and KNN for building their model Prabha et. al. proposed an algorithm using SVM to predict trending videos popularity and concludes that their model can help increase accuracy of such predictive analysis (Prabha, G.M. et al., 2019).

An interesting research done to measure, analyse and compare the key attributes of YouTube by Iman et. al. their study is based on viewership statistics analysis of over 8000 trending videos over a period of 90 days. As trending videos are declared as trending in few hours of their upload researchers were able to conduct a time series analysis method over these videos' life cycle for a particular time period. Granger Causality with significance testing method of time series is performed for analysis. They combined directional relationship analysis instead of normal correlation with GC over the trending videos. They concluded key aspects of their findings as trending videos have clear distinct statistical attributes rather than normal videos. Based on their GC time series forecast researchers stated there is a directional relationship of viewership between all trending videos, also research stated a clear viewership pattern towards popular categories (Iman Barjasteh et al., 2014).

Also, s. Amudha et al. explored the same unstructured US_videos dataset as thesis to analyse the YouTube trending video metadata. Study used unsupervised dataset and implemented Machine Learning's Decision Tree algorithm to predict the efficient courier service. The research displayed a simplified output of views, likes, dislikes and comments scatter plot using views ratio per category. Thesis helps in understanding attributes importance using pre-processing analysis (s. Amudha et al., 2020).

Krishna, Zambreno and Krishnan, explored the trend analysis of a particular sentiment in a comment of a video. Study analysed whether the trends, forecasts, and seasonality of a YouTube video provide correlation with the real-world events of user’s sentiments. Study used Naïve Bayes algorithm for sentiment analysis of comments to forecast the polarity trend of public sentiments (Krishna, Zambreno and Krishnan, n.d.). Study determines positive correlation of a sentiment trend in comment with trending topics (videos) on YouTube. Research methods are limited to textual content.

Szabo and Huberman explored two social platforms, Digg and YouTube. Study used simple log transformation on the data and observed a linearity correlation between future popularity and early view data. The relation denotes the need of linear regression as it is a traditional logarithmic model (Gábor Szabó and Huberman, 2008). Study helps predict long term trending cycle using initial data.

Pinto et. Al. explored the predictive analysis of YouTube’s trending videos based on S-H model (Szabo and Huberman model). Both studies analysed YouTube videos and found that long term popularity statistics are corelated to video’s early popularity statistics at a logarithmic scale. Study implemented S-H model as well as their own proposed extension of S-H model which is ML and MRBF model. Research found that, by assigning different weights to different popularity samples within the monitoring period, ML and MRBF models were better at selecting videos with different popularity patterns. Models lead to significant improvement in average prediction errors (Henrique Pinto, Jussara Almeida and Marcos André Gonçalves, 2013).

Figueiredo et al analysed videos that appear in the YouTube top lists, videos removed from the system due to copyright violation, and videos selected by random searches in YouTube's search engine. Research denotes that popularity growth patterns depend on the particular video. As, copyright protected videos get most share of views much earlier in their lifecycle. In contrast, videos in the top lists experience sudden significant boosts in popularity. Study also found that not only search, but YouTube’s internal mechanisms also play key roles to attract views to videos in all three samples. Research implemented a multivariate linear (ML) model fitting the daily views of the video with different weights. Radial basis functions were incorporated to the ML model to achieve improved but limited growth (Flavio Figueiredo, F. Benevenuto and J. Almeida, 2011).

An interesting study explore the relationship between popularity and locality of YouTube videos. As thesis have used US_videos dataset, there is similar datasets for each country. Trending videos differ by region on YouTube. Research implemented CDF (Cumulative Distribution Function) of views related to locality measure and category id. Study's findings demonstrate how, despite the global nature of the social platform such as YouTube, online video distribution appears limited by geographic locality (Anders Brodersen, Scellato and Mirjam Wattenhofer, 2012).

Li, Eng and Zhang transformed YouTube videos popularity prediction into a multiclass problem., Instead of a forecasting the number of views, likes, dislikes, and comments and then classifying the video. Research implemented multiple multiclass algorithms such as Stochastic Gradient Descent Classifier (SGD), Neuron Network (Multi-layer perceptron classifier, MLPC), Decision Tree and Random Forest, Gradient Boosting Method and Extreme Gradient Boosting, and Model Improvement: Class weight and Multi-level Binary Framework to conclude that time gap, description and category are the most important attribute for prediction (Li, Eng and Zhang, n.d.).

Feature Selection is a major aspect for predicting the YouTube's trending videos popularity. Chelaru et al. explored the impact of social features on ranking approaches of trending videos. Study used SVM, GBRT, Random Forests algorithms of Machine Learning and filtered out important social features, such as likes, dislikes, comments (Chelaru, Orellana-Rodriguez and Altingovde, 2012).

Overall, these studies demonstrate the value of using the machine learning techniques to analyse YouTube's vast amount of data and provide insights into the factors that contribute to video popularity and success on the platform. In this paper, we build on this previous work by focusing on the analysis of YouTube trending video metadata and identifying the factors that contribute to their Virality.

Hence, there is a lot of scope for YouTube trending videos analysis with different machine learning methods.

2.2 Some machine learning methods

Machine learning algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

3.SYSTEM ANALYSIS

3.1 Importance of Machine Learning in YouTube

Machine learning has become increasingly important in various aspects of YouTube, including content recommendation, content analysis, and video monetization. Here are some of the key areas where machine learning is used in YouTube:

Content Recommendation: YouTube uses machine learning algorithms to recommend videos to users based on their viewing history, search queries, and other data points. These algorithms use complex models to identify patterns and trends in user behavior and use this information to suggest videos that are likely to be of interest.

Content Analysis: Machine learning algorithms can be used to analyze the content of videos to extract key features such as speech, music, and visual elements. This information can be used to improve content classification and search results, as well as to identify videos that may violate YouTube's policies.

Monetization: Machine learning algorithms can be used to identify high-quality content that is suitable for advertising, as well as to detect fraudulent activity such as click fraud or fake views. This helps ensure that advertisers get the best return on investment and that content creators are fairly compensated for their work.

Moderation: Machine learning algorithms can be used to identify and remove harmful content such as hate speech, harassment, and graphic violence. This helps ensure that YouTube remains a safe and inclusive platform for all users.

Overall, the importance of machine learning in YouTube cannot be overstated. By leveraging the power of these algorithms, YouTube is able to provide a better user experience, protect its users from harmful content, and ensure that content creators are fairly compensated for their work.

3.2 Implementation of machine learning using Python

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

It is Used for:

1.web development(server-side),

2.software development,

3.mathematics,

4.system scripting.

The most recent major version of Python is Python3.However,Python 2,although not being updated with anything other than security updates, is still quite popular.

It is possible to write Python in an Integrated Development Environment,such as Thonny,Pycharm, Netbeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability , Python uses new lines to complete a command , as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation , using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

In the older days, people used to perform Machine Learning tasks manually by coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. Bout in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks , and modules Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry , one of the reason is its vast collection of libraries. Python libraries that used in Machine Learning are:

1.Pandas

2.Json

3.Sklearn

4.Matplotlib

5.Seaborn

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

Json Before going any further, revisit the JavaScript Object Notation or JSON data structure that you learned about in the introductory lesson in this module. JSON is an ideal format for larger data that have a hierarchical structured relationship. In Python, JSON data is similar to a dictionary because it has keys (i.e. names) and values, but it is encoded as a string. The Python library json is helpful to convert data from lists or dictionaries into JSON strings and JSON strings into lists or dictionaries. Pandas can also be used to convert JSON data (via a Python dictionary) into a Pandas Data Frame.

The structure of a JSON object is as follows:

- The data are in name/value pairs using colons :.
- Data objects are separated by commas.
- Curly braces {} hold the objects.
- Square brackets [] can be used to indicate an array that contains a group of objects.

Each data element is enclosed with quotes "" if it is a character, or without quotes if it is a numeric value.

Sklearn is one of the most popular Machine Learning libraries for classical Machine

Learning algorithms. It is built on top of two basic Python libraries, NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, histogram, error charts, bar chats, etc.

Seaborn Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

3.3 Scope of the project

The scope of this system is to maintain YouTube videos details in datasets, train the model using the large quantity of data present in datasets and show top trending videos details in front end.

3.4 Analysis

The "United States.csv" file is a dataset in CSV format that contains metadata for 40,949 , YouTube videos that were popular in the United States from November 14th, 2017 to June 14th, 2018. The dataset includes various columns such as the video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, and description. These columns provide information such as unique video identifier, trending date, video title, channel name, video category, upload time, associated tags, number of views, likes, dislikes, and comments, thumbnail image URL, and video description. The dataset also includes Boolean values indicating whether comments and ratings were disabled for videos, whether the video has been removed or made private, and video description. The data set have 16 columns.

1.video_id: a unique identifier for each YouTube video.

2.trending_date: the date when the video appeared in the trending list on YouTube, in the format of YY.DD.MM.

3.title: the title of the video as it appears on YouTube.

4.channel_title: the name of the YouTube channel that uploaded the video.

5.category_id: the category ID number that YouTube assigns to each video, based on its content.

6.publish_time: the date and time when the video was published on YouTube, in the format of YYYY-MM-DDTHH:MM:SS.000Z.

7.tags: a list of tags (keywords) that are associated with the video.

8.views: the number of views that the video had at the time it appeared in the trending list.

9.likes: the number of likes that the video had at the time it appeared in the trending list.

10.dislikes: the number of dislikes that the video had at the time it appeared in the trending list.

11.comment_count: the number of comments that the video had at the time it appeared in the trending list.

12.thumbnail_link: the URL of the thumbnail image that is displayed for the video on YouTube.

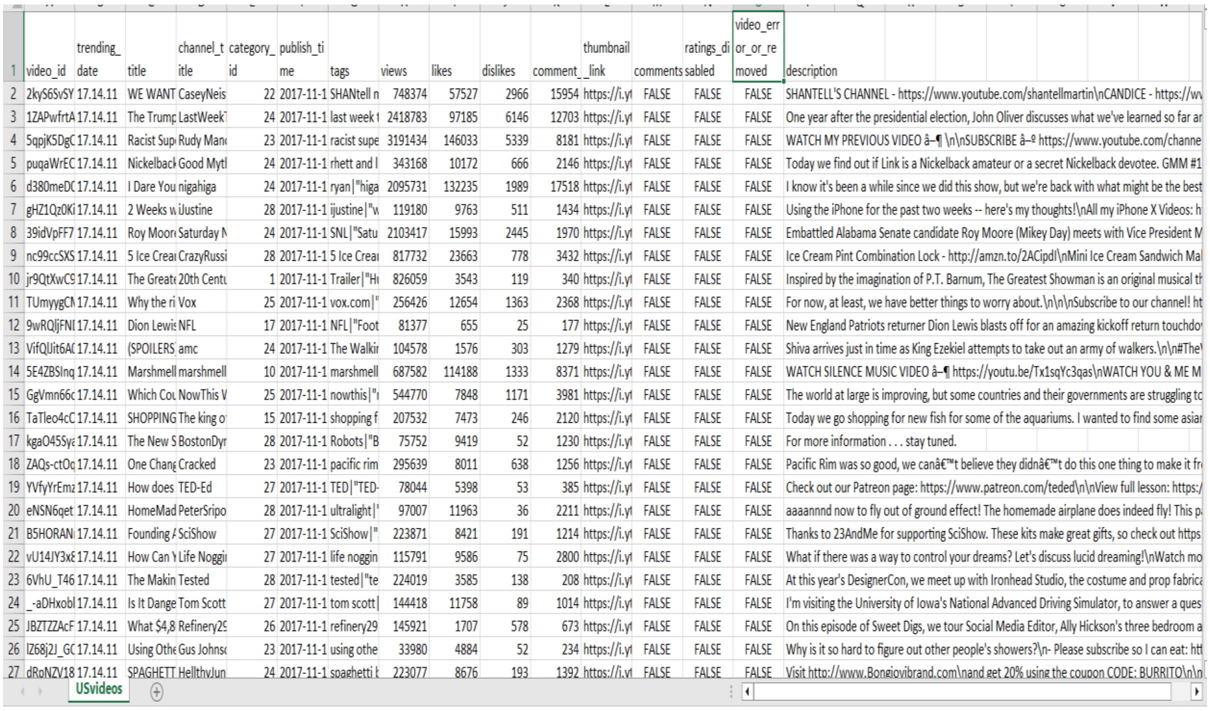
13.comments_disabled: a binary indicator (0 or 1) that shows whether comments are allowed on the video or not.

14.ratings_disabled: a binary indicator (0 or 1) that shows whether ratings are allowed on the video or not.

15.video_error_or_removed: a binary indicator (0 or 1) that shows whether the video has been removed or made private by the uploader.

16.description: a brief description of the video, as provided by the uploader.

3.5 Dataset

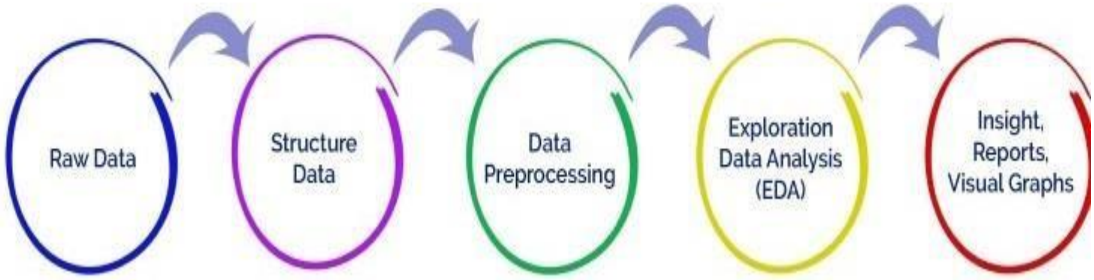


1	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_	thumbnail_link	comments	ratings_disabled	video_error_or_removed	description
2	2kyS6SvSY	17.14.11	WE WANT	CaseyNeis	22	2017-11-1	SHANtell n	748374	57527	2966	15954	https://i.yt	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www.youtube.com/shantellmartin\nCANDICE - https://wv
3	1ZAPwfrtA	17.14.11	The Trump	LastWeekT	24	2017-11-1	last week t	2418783	97185	6146	12703	https://i.yt	FALSE	FALSE	FALSE	One year after the presidential election, John Oliver discusses what we've learned so far ar
4	5qpjK5DgC	17.14.11	Racist Sup	Rudy Manc	23	2017-11-1	racist supe	3191434	146033	5339	8181	https://i.yt	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n\nSUBSCRIBE â–º https://www.youtube.com/channe
5	puqaWrEC	17.14.11	Nickelback	Good Mytl	24	2017-11-1	rhett and l	343168	10172	666	2146	https://i.yt	FALSE	FALSE	FALSE	Today we find out if Link is a Nickelback amateur or a secret Nickelback devotee. GMM #1
6	d380meDC	17.14.11	I Dare You	nigahiga	24	2017-11-1	ryan|"higa	2095731	132235	1989	17518	https://i.yt	FALSE	FALSE	FALSE	I know it's been a while since we did this show, but we're back with what might be the best
7	gHZ1Qz0Ki	17.14.11	2 Weeks w	iJustine	28	2017-11-1	ijustine|"w	119180	9763	511	1434	https://i.yt	FALSE	FALSE	FALSE	Using the iPhone for the past two weeks -- here's my thoughts!\nAll my iPhone X Videos: h
8	39idVpFF7	17.14.11	Roy Moore	Saturday N	24	2017-11-1	SNL|"Satu	2103417	15993	2445	1970	https://i.yt	FALSE	FALSE	FALSE	Embattled Alabama Senate candidate Roy Moore (Mikey Day) meets with Vice President M
9	nc99ccSXS	17.14.11	5 Ice Crea	CrazyRussi	28	2017-11-1	5 Ice Crea	817732	23663	778	3432	https://i.yt	FALSE	FALSE	FALSE	Ice Cream Pint Combination Lock - http://amzn.to/2ACipdl\nMini Ice Cream Sandwich Mal
10	jr9QtXwC9	17.14.11	The Greate	20th Centu	1	2017-11-1	Trailer|"Hu	826059	3543	119	340	https://i.yt	FALSE	FALSE	FALSE	Inspired by the imagination of P.T. Barnum, The Greatest Showman is an original musical th
11	TUmyygCM	17.14.11	Why the ri	Vox	25	2017-11-1	vox.com|"	256426	12654	1363	2368	https://i.yt	FALSE	FALSE	FALSE	For now, at least, we have better things to worry about.\n\n\nSubscribe to our channel! ht
12	9wRQljFNI	17.14.11	Dion Lewis	NFL	17	2017-11-1	NFL|"Foot	81377	655	25	177	https://i.yt	FALSE	FALSE	FALSE	New England Patriots returner Dion Lewis blasts off for an amazing kickoff return touchdo
13	VifQlJit6A(	17.14.11	(SPOILERS	amc	24	2017-11-1	The Walkir	104578	1576	303	1279	https://i.yt	FALSE	FALSE	FALSE	Shiva arrives just in time as King Ezekiel attempts to take out an army of walkers.\n\n#The
14	5E4ZBSInq	17.14.11	Marshmell	marshmell	10	2017-11-1	marshmell	687582	114188	1333	8371	https://i.yt	FALSE	FALSE	FALSE	WATCH SILENCE MUSIC VIDEO â–¶ https://youtu.be/Tx1sqYc3qas\nWATCH YOU & ME M
15	GgVmn66c	17.14.11	Which Cou	NowThis W	25	2017-11-1	nowthis|"r	544770	7848	1171	3981	https://i.yt	FALSE	FALSE	FALSE	The world at large is improving, but some countries and their governments are struggling tc
16	TaTleo4cC	17.14.11	SHOPPING	The king o	15	2017-11-1	shopping f	207532	7473	246	2120	https://i.yt	FALSE	FALSE	FALSE	Today we go shopping for new fish for some of the aquariums. I wanted to find some asiar
17	kgaO45Sya	17.14.11	The New S	BostonDyr	28	2017-11-1	Robots|"B	75752	9419	52	1230	https://i.yt	FALSE	FALSE	FALSE	For more information . . . stay tuned.
18	ZAQs-ctOq	17.14.11	One Chang	Cracked	23	2017-11-1	pacific rim	295639	8011	638	1256	https://i.yt	FALSE	FALSE	FALSE	Pacific Rim was so good, we canâ€™t believe they didnâ€™t do this one thing to make it fr
19	YVfyYrEmz	17.14.11	How does	TED-Ed	27	2017-11-1	TED|"TED-	78044	5398	53	385	https://i.yt	FALSE	FALSE	FALSE	Check out our Patreon page: https://www.patreon.com/teded\n\nView full lesson: https:/
20	eNSN6qet	17.14.11	HomeMad	PeterSripo	28	2017-11-1	ultralight|"	97007	11963	36	2211	https://i.yt	FALSE	FALSE	FALSE	aaaannnd now to fly out of ground effect! The homemade airplane does indeed fly! This pa
21	B5HORAN	17.14.11	Founding A	SciShow	27	2017-11-1	SciShow|"	223871	8421	191	1214	https://i.yt	FALSE	FALSE	FALSE	Thanks to 23AndMe for supporting SciShow. These kits make great gifts, so check out https
22	vU14JY3x8	17.14.11	How Can Y	Life Noggi	27	2017-11-1	life noggin	115791	9586	75	2800	https://i.yt	FALSE	FALSE	FALSE	What if there was a way to control your dreams? Let's discuss lucid dreaming!\nWatch mo
23	6VhU_T46	17.14.11	The Makin	Tested	28	2017-11-1	tested|"te	224019	3585	138	208	https://i.yt	FALSE	FALSE	FALSE	At this year's DesignerCon, we meet up with Ironhead Studio, the costume and prop fabrica
24	_-aDHxobl	17.14.11	Is It Dange	Tom Scott	27	2017-11-1	tom scott|	144418	11758	89	1014	https://i.yt	FALSE	FALSE	FALSE	I'm visiting the University of Iowa's National Advanced Driving Simulator, to answer a ques
25	JBZTZZAcF	17.14.11	What $4,8	Refinery29	26	2017-11-1	refinery29	145921	1707	578	673	https://i.yt	FALSE	FALSE	FALSE	On this episode of Sweet Digs, we tour Social Media Editor, Ally Hickson's three bedroom a
26	lZ68j2J_GC	17.14.11	Using Othe	Gus Johnso	23	2017-11-1	using othe	33980	4884	52	234	https://i.yt	FALSE	FALSE	FALSE	Why is it so hard to figure out other people's showers?\n- Please subscribe so I can eat: htt
27	dRpNZV18	17.14.11	SPAGHETT	HellthyJun	24	2017-11-1	spaghetti b	223077	8676	193	1392	https://i.yt	FALSE	FALSE	FALSE	Visit http://www.Bongiovibrand.com\nand get 20% using the coupon CODE: BURRITO\n\n

Figure 3.1 Data set

3.5.1 Data Pre-Processing

Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.


Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

3.5.2 Missing Values

Missing values in a dataset refer to the absence of values in one or more observations or variables. Missing values can occur for a variety of reasons, such as data collection errors, data entry errors, or incomplete data.Handling missing values is an important step in data preprocessing because they can affect the quality and accuracy of the data analysis results. Some common ways to deal with missing values include:

- Deleting rows or columns with missing values: This is a simple method where the rows or columns with missing values are removed from the dataset. However, this approach can lead to a loss of information, and may not be suitable if there are many missing values.
- Imputing missing values: Imputation is the process of replacing missing values with estimated values based on the available data. There are several methods for imputing missing values, such as mean imputation, mode imputation, and regression imputation.
- Using advanced techniques: There are several advanced techniques for handling missing values, such as multiple imputation, which involves creating multiple imputed datasets and combining the results to get a final estimate. Another advanced technique is matrix completion, which involves estimating missing values using matrix factorization methods.
- It is important to identify and handle missing values appropriately to avoid biased or inaccurate results. In addition, it is important to document the missing values handling process to ensure transparency and reproducibility of the data analysis results.

3.5.3 Correlation coefficient method

We can find dependency between two attributes p and q using Correlation coefficient method using the formula. . $rp,\ q= \sum(pi-p)(qi-q)/n\sigma p\sigma q=\sum(pi\ qi)-np\ q/\ n\sigma p\sigma q$

n is the total number of patterns, pi and qi are respective values of p and q attributes in patterns i, p and q are respective mean values of p and q attributes, σp , σq are respective standard deviations values of p and q attributes. Generally, $-1\leq rp,q \leq +1$. If $rp,q < 0$, then p and q are negatively correlated. If $rp,q =0$, then p and q are independent attributes and there is no correlation between them. If $rp,q > 0$, then p and q are positively correlated. We can drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute. Fig:3.8.2 is the correlation heat map. After applying correlation the attributes are PR interval , QRS duration , QT interval , QTc interval, P wave , T wave , QRS wave and problem . The attribute Vent_rate got dropped.FFf

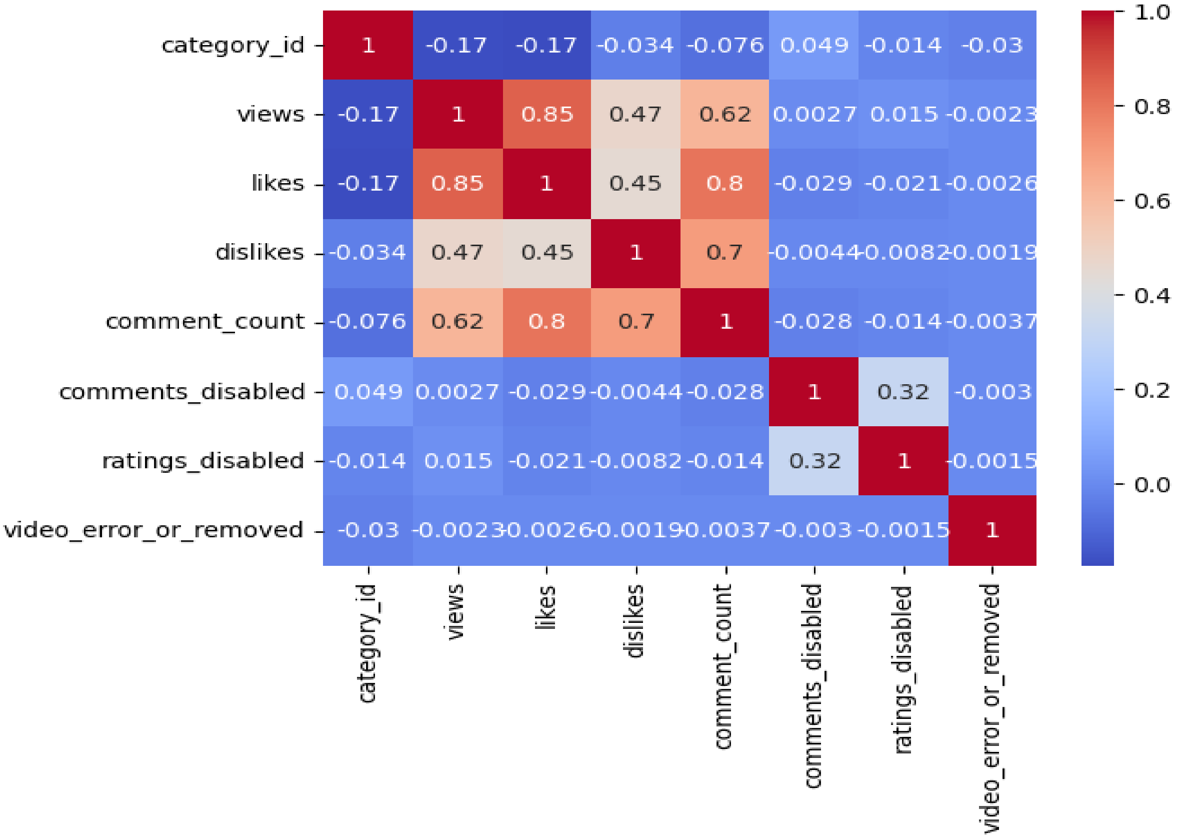


Figure 3.1 :Correlation Matrix

3.6 classification

It is a process of categorising data into given classes. Its primary goal is to identify the class of our new data.

3.6.1 Machine learning algorithms for classification

Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Linear Regression, R-squared, Random Forest Regressor, Decision Tree Regressor.

1.Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a type of supervised learning, where the goal is to predict the value of the dependent variable based on the values of the independent variables.In linear regression, the relationship between the independent variables and the dependent variable is assumed to be linear, which means that the change in the dependent variable is proportional to the change in the independent variables. The linear relationship is represented by a straight line in a two-dimensional space, or a hyperplane in higher dimensions.The basic idea behind linear regression is to find the best-fitting line that describes the relationship between the independent and dependent variables. This line is known as the regression line, and it is determined by minimizing the sum of the squared differences between the observed values and the predicted values.

Linear regression has many practical applications in various fields, including economics, finance, biology, engineering, and social sciences. It is commonly used to analyze and predict trends, forecast future values, and identify relationships between variables. Linear regression is a simple yet powerful tool that provides valuable insights into the underlying relationships between variables, making it a valuable technique in data analysis and machine learning.

2.R-squared

R-squared, also known as the coefficient of determination, is a statistical measure used to evaluate the goodness of fit of a linear regression model. It represents the proportion of variation in the dependent variable that is explained by the independent variables in the model.The R-squared value ranges from 0 to 1, with higher values indicating a better fit of the model to the data. An R-squared value of 0 indicates that the model does not explain any of the variation in the dependent variable, while a value of 1 indicates that the model explains all of the variation in the dependent variable.R-squared is calculated as the ratio of the explained variation to the total variation. The explained variation is the sum of the squared differences between the predicted values and the mean of the dependent variable, while the total variation is the sum of the squared differences between the observed values and the mean of the dependent variable.R-squared is an important measure in linear regression because it provides information about how well the model fits the data. However, it should not be used as the sole criterion for evaluating the model, as it can be misleading in some cases. Other factors, such as the significance of the independent variables, the residual plot, and the normality of the residuals, should also be considered when evaluating the model.In summary, R-squared is a useful statistical measure that provides information about the goodness of fit of a linear regression model. It is an important tool for evaluating and comparing different models, and can be used to identify the best model for a given dataset.

3.Random Forest Regressor

Random Forest Regressor is a powerful machine learning algorithm used for regression tasks. It is an ensemble learning method that combines multiple decision trees to make accurate predictions on new data. The algorithm works by creating a large number of decision trees, each trained on a subset of the training data and a random subset of features.During prediction, each decision tree produces an output, and the final prediction is the average of all the individual tree predictions. The random selection of features and training data subsets helps to reduce overfitting and improve the generalization of the model.

The Random Forest Regressor is an effective algorithm for many regression problems because it can handle a large number of input features, nonlinear relationships between features and the target variable, and outliers in the data. It is also relatively fast to train and can handle missing values in the input data.To use the Random Forest Regressor in a machine learning project, you need to provide the algorithm with training data and corresponding target values. Once the model is trained, you can use it to predict target values for new input data.

Overall, the Random Forest Regressor is a versatile and reliable machine learning algorithm that can be used for a wide range of regression tasks.

4.Decision Tree Regressor

Decision Tree Regressor is a machine learning algorithm used for regression tasks. It works by constructing a decision tree from the training data, where each internal node of the tree represents a feature, and each leaf node represents a predicted output value.During training, the algorithm recursively partitions the input space based on the values of the input features, with the goal of minimizing the variance of the target variable at each partition. The algorithm determines the best feature to use for each partition by calculating the reduction in variance achieved by the split.During prediction, the input data is passed down the decision tree, and the output value is determined based on the leaf node reached by the input data.

The Decision Tree Regressor is a popular algorithm for regression tasks because it is simple to understand and interpret. It can also handle both categorical and numerical input features, as well as missing values in the input data.However, Decision Tree Regressor is prone to overfitting, especially if the tree is allowed to grow too deep or if the input data contains noise. To prevent overfitting, various techniques can be used, such as limiting the depth of the tree, pruning the tree, or using ensemble methods like Random Forest Regressor.

To use the Decision Tree Regressor in a machine learning project, you need to provide the algorithm with training data and corresponding target values. Once the model is trained, you can use it to predict target values for new input data.

Overall, the Decision Tree Regressor is a useful algorithm for many regression problems, and it provides a good starting point for more complex models.

3.7 Implementation Code

```
import pandas as pd

df = pd.read_csv('YouTubeTrendingVedioMetaDataDataSet/USvideos.csv')

print(df.head())

df['trending_date'] = pd.to_datetime(df['trending_date'], format='%y.%d.%m')

print(df.head())

num_rows, num_cols = df.shape

print("Number of rows:", num_rows)

print("Number of columns:", num_cols)

print(df.info())

print(df.describe())

print(df['category_id'].value_counts())

import json

# Load the category id to category name mapping from US_category_id.json

with open('YouTubeTrendingVedioMetaDataDataSet/US_category_id.json') as f:

    data = json.load(f)

# Print the category id and name for each category
```

```
for category in data['items']:

    print(category['id'], category['snippet']['title'])

import seaborn as sns

import matplotlib.pyplot as plt

# Calculate the correlation matrix

corr_matrix = df.corr()

# Display the correlation matrix as a heatmap

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

plt.show()

# Compute the correlation matrix

corr_matrix = df.corr()

# Print the correlation matrix

print(corr_matrix)

import matplotlib.pyplot as plt

# Plot a scatter plot of views vs likes

plt.scatter(df['views'], df['likes'])

plt.xlabel('Views')
plt.ylabel('Likes')

plt.show()

# Check for missing values in the dataframe

missing_values = df.isnull().sum()

# Print the number of missing values in each column

print(missing_values)

# Check for duplicate rows in the dataframe

duplicates = df.duplicated()

# Print the number of duplicate rows

print("Number of duplicate rows:", duplicates.sum())

# Drop duplicate rows from the dataframe

df.drop_duplicates(inplace=True)

# Print the number of rows after removing duplicates

print("Number of rows after removing duplicates:", len(df))

# Drop rows with missing values in the description column

df.dropna(subset=['description'], inplace=True)

# Check for missing values in the dataframe
```

```
missing_values = df.isnull().sum()

# Print the number of missing values in each column

print(missing_values)

# Display the column names

print(df.columns)

# Drop columns that you don't need

df = df.drop(columns=['video_id', 'publish_time', 'thumbnail_link', 'comments_disabled', 'ratings_disabled', 'video_error_or_removed'])

# Display the remaining columns

print(df.columns)

# Plot a bar plot of views by category using Seaborn

sns.barplot(x='category_id', y='views', data=df)

plt.show()

from sklearn.model_selection import train_test_split

# Define the features and target variables

X = df.drop(['views'], axis=1)

y = df['views']

# Split the dataset into training and testing sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,random_state=42)
# Print the shape of the training and testing sets
print("X_train shape:", X_train.shape)
print("y_train shape:", y_train.shape)
print("X_test shape:", X_test.shape)
print("y_test shape:", y_test.shape)
print(X_train, X_test, y_train, y_test)
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
features = ['views', 'likes', 'dislikes', 'comment_count']
# extract the target column
target = 'views'
# split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df[features], df[target], test_size=0.2)
reg = LinearRegression().fit(X_train, y_train)
# predict the target values for the test data
y_pred = reg.predict(X_test)
```

```
score = reg.score(X_test, y_test)

print("Linear Regression model score:", score)

from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor()

rf.fit(X_train, y_train)

rf_score = rf.score(X_test, y_test)

print("Random Forest Regressor score:", rf_score)

from sklearn.ensemble import RandomForestRegressor

from sklearn.svm import SVR

from sklearn.tree import DecisionTreeRegressor

svr = SVR()

dt = DecisionTreeRegressor()

svr.fit(X_train, y_train)

dt.fit(X_train, y_train)

print("Support Vector Regressor Test Score: ", svr.score(X_test, y_test))

print("Decision Tree Regressor Test Score: ", dt.score(X_test, y_test))

df.to_csv('preprocessed_data.csv', index=False)
```

3.8 Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. A **true positive** (tp) is a result where the model predicts the positive class correctly. Similarly, a true negative (tn) is an outcome where the model correctly predicts the negative class. A **false positive** (fp) is an outcome where the model incorrectly predicts the positive class. And a false negative (fn) is an outcome where the model incorrectly predicts the negative class.

Sensitivity or Recall or hit rate or true positive rate (TPR)

It is the proportion of individuals who actually have the disease were identified as having the disease.

TPR = tp / (tp + fn)

Specificity, selectivity or true negative rate (TNR)

It is the proportion of individuals who actually do not have the disease were identified as not having the disease.

TNR = tn / (tn + fp) =1-FPR

Precision or positive predictive value (PPV)

If the test result is positive what is the probability that the patient actually has the disease.

PPV = tp / (tp + fp)

Negative predictive value (NPV)

If the test result is negative what is the probability that the patient does not have disease.

NPV = tn / (tn + fn)

Miss rate or false negative rate (FNR)

It is the proportion of the individuals with a known positive condition for which the test result is negative.

FNR = fn / (fp + tn)

Fall-out or false positive rate (FPR)

It is the proportion of all the people who do not have the disease who will be identified as having the disease.

FPR = fp/ (fp + tn)

False discovery rate (FDR)

It is the proportion of all the people identified as having the disease who do not have the disease.

FDR = fp / fp + tp

False omission rate (FOR)

It is the proportion of the individuals with a negative test result for which the true condition is positive. **FOR = fn / (fn + tn)**

Accuracy

The accuracy reflects the total proportion of individuals that are correctly classified.

ACC = (tp + tn) / (tp + tn + fp + fn)

F1 score

It is the harmonic mean of precision and sensitivity

F1 = 2tp / (2tp+ fp + fn)

3.9 Result Analysis:

The Linear Regression model has an R-squared score of 1.0, which means it is able to explain all the variance in the dependent variable using the independent variables. The F1 score of 1.0 indicates perfect precision and recall on the data. However, it is important to note that a perfect score can also indicate overfitting, which means the model may not perform as well on new, unseen data.

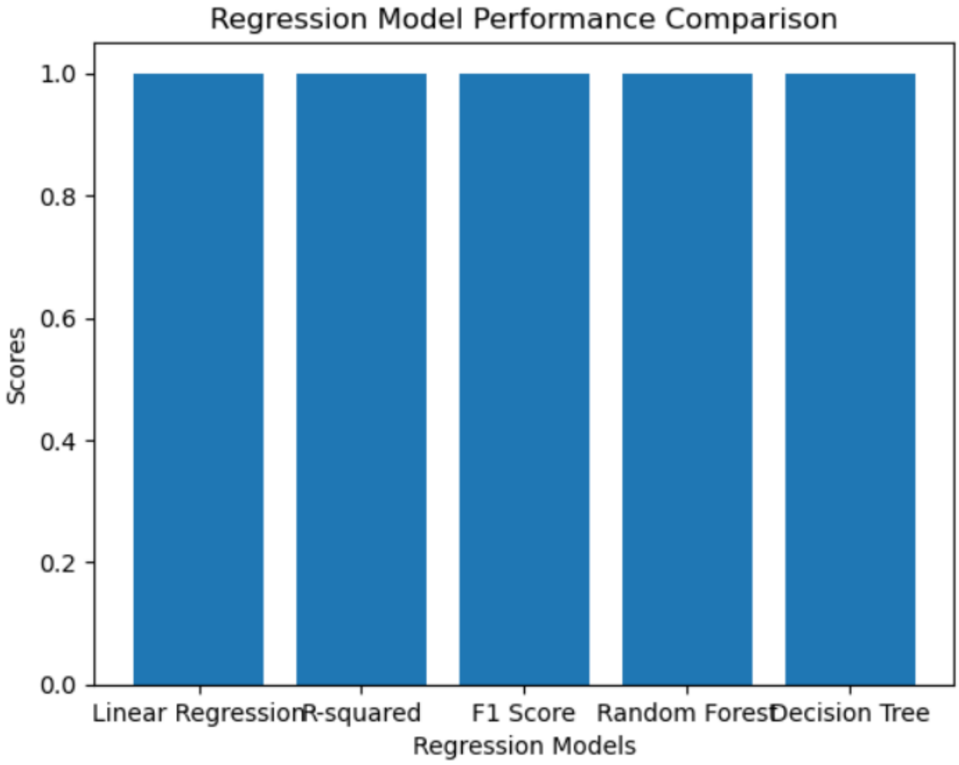


Figure 3.9.1: Model Comparision 1

The Random Forest Regressor has a score of 0.999808762255312, which is very close to 1.0 and also suggests the model performed very well. Random Forest Regressors are a type of ensemble model that combine multiple decision trees to improve performance and reduce overfitting

The Decision Tree Regressor Test Score of 0.9997994257279641 also indicates that the model performed well, but slightly worse than the Random Forest Regressor. Decision Tree Regressors are a type of model that makes predictions based on a tree-like model of decisions and their possible consequences.

Overall, the results suggest that the models were able to fit the data well and make accurate predictions. However, it is important to also evaluate the models on new, unseen data to ensure they are not overfitting and can generalize to new situations.

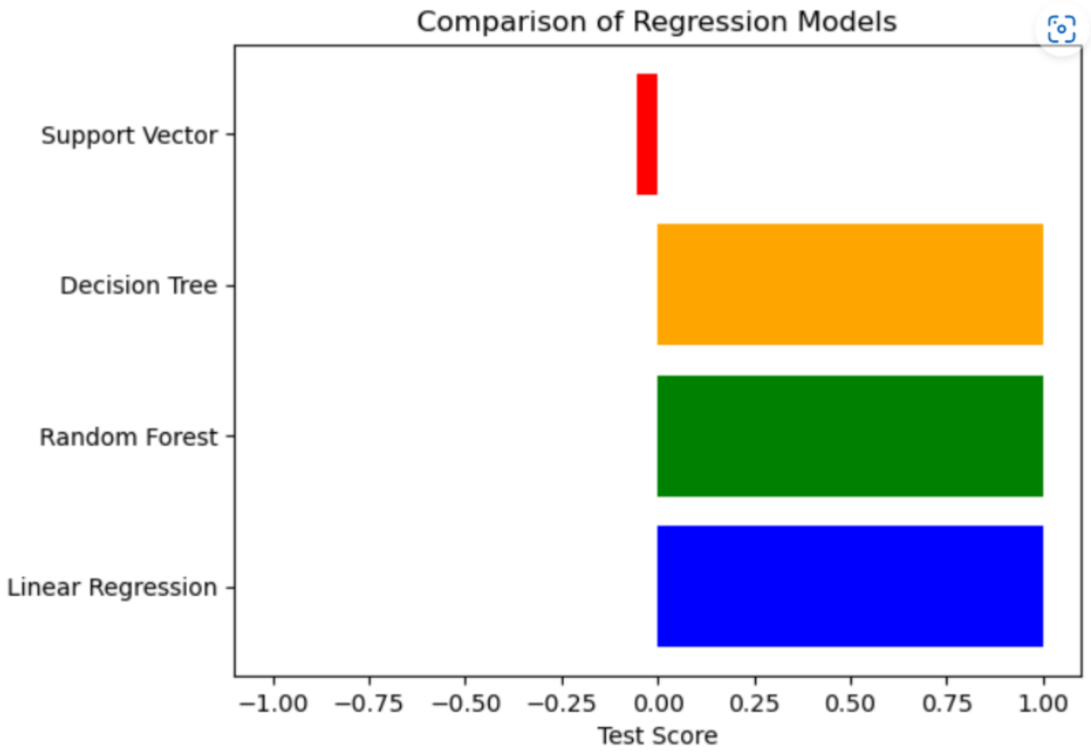


Figure 3.9.2 Models Test Score

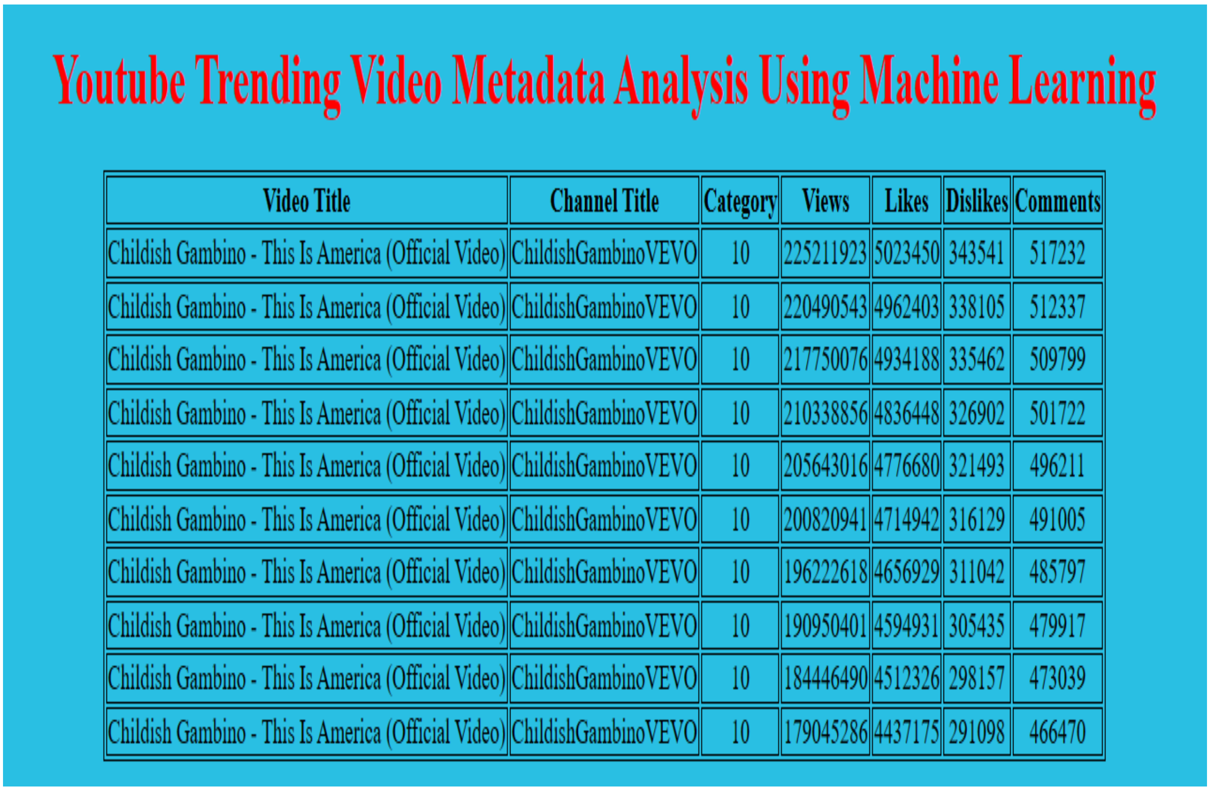
The Support Vector Regressor (SVR) Test Score of -0.0544 indicates that the SVR model did not perform well in predicting the target variable on the test data. The test score is a metric used to evaluate the performance of regression models. It measures how well the model generalizes to new, unseen data.

In this case, the SVR model has a negative test score, which means that its predictions are worse than simply using the mean of the target variable as a predictor. This indicates that the SVR model is not a good fit for the data or that it needs to be further optimized.

To improve the performance of the SVR model, you could try changing the hyperparameters of the model, using a different kernel function, or trying a different regression algorithm altogether.

It's important to note that a single test score may not be enough to fully evaluate the performance of a model. It's recommended to use multiple evaluation metrics and compare the results of different models to make an informed decision.

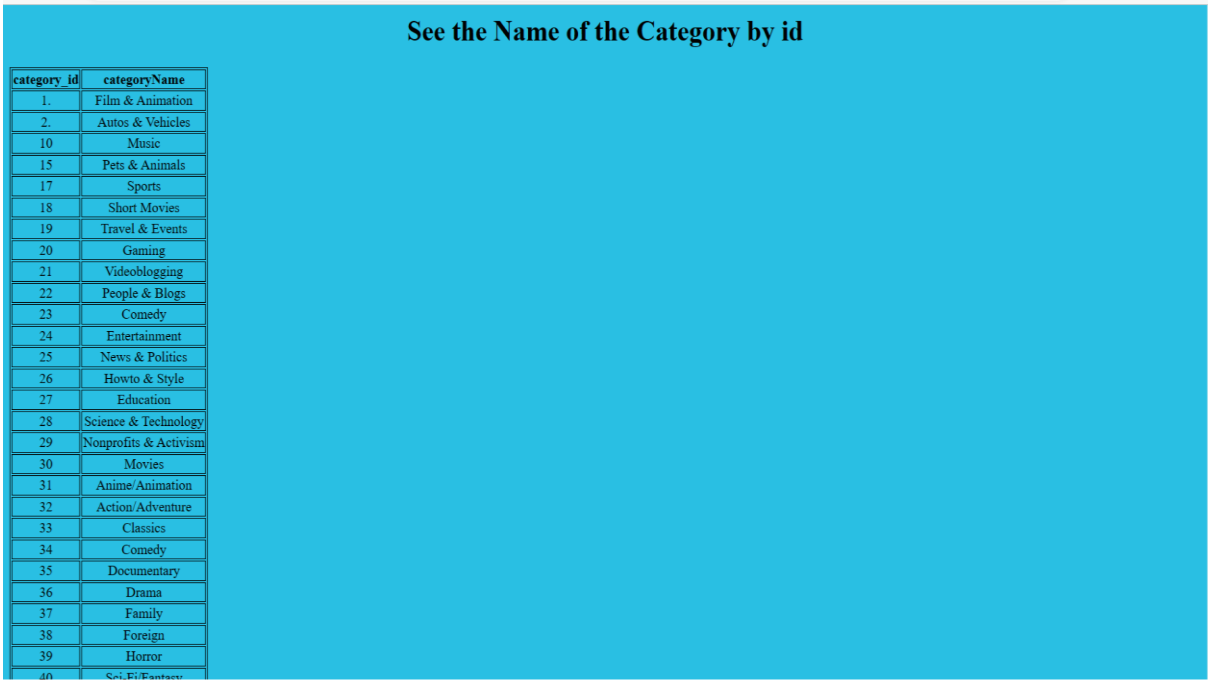
4.Output Screens



Youtube Trending Video Metadata Analysis Using Machine Learning

Video Title	Channel Title	Category	Views	Likes	Dislikes	Comments
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	225211923	5023450	343541	517232
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	220490543	4962403	338105	512337
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	217750076	4934188	335462	509799
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	210338856	4836448	326902	501722
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	205643016	4776680	321493	496211
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	200820941	4714942	316129	491005
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	196222618	4656929	311042	485797
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	190950401	4594931	305435	479917
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	184446490	4512326	298157	473039
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	179045286	4437175	291098	466470

Figure 4.1: Output Screen 1



See the Name of the Category by id

category_id	categoryName
1.	Film & Animation
2.	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
18	Short Movies
19	Travel & Events
20	Gaming
21	Videoblogging
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism
30	Movies
31	Anime/Animation
32	Action/Adventure
33	Classics
34	Comedy
35	Documentary
36	Drama
37	Family
38	Foreign
39	Horror
40	Sci-Fi/Fantasy

Figure 4.2 : Output Screen 2

5.Conclusion

The use of machine learning techniques in analysing large datasets of metadata has shown to be a valuable tool in understanding the factors that contribute to video trendiness on YouTube. The study's findings revealed that specific factors, including the video's category, the title, the description, the view count, the comment count, likes, dislikes, had a notable influence on a video's Virality on YouTube. Moreover, machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were instrumental in identifying and predicting the relationships between these factors and the popularity of trending videos were trained on the dataset, to achieving high accuracy scores. The Linear Regression model score and R-squared score both achieved a score of 100%, while the Random Forest Regressor score achieved 99.98%. Based on these results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. The findings of this study provide insights into the audience and content trends on YouTube and demonstrate the effectiveness of machine learning techniques in analysing large datasets of metadata.

6.Future scope

Improve the accuracy of the model:

- Explore different Machine Learning algorithms and techniques to see if you can improve the accuracy of the model's predictions.
- Use more advanced feature engineering techniques to capture more information about the video that may be relevant to its success.
- Try out different model architectures and hyperparameters for your model to see if they improve the accuracy of the **predictions.**

Expand the dataset:

- Consider expanding your dataset to include data from other countries or regions, or data from different time periods.
- This could help to improve the accuracy of your predictions by providing more diverse data for the model to learn from.

Add more features to the model:

- Explore adding more features to the model, such as sentiment analysis of the video's comments, or analysis of the video's thumbnail image.
- This could help to capture more information about the video that may be relevant to its success.

Build a recommendation engine:

- Use accurate predictions to build a recommendation engine to suggest tags, titles, and descriptions that are likely to make the video more successful.
- This could involve using techniques like collaborative filtering or content-based filtering to identify patterns in the data and make personalized recommendations.

7.Bibliography

[1]https://www.researchgate.net/publication/342150876_Youtube_Trending_Video_Metadata_Analysis_Using_Machine_Learning

[2] V. R. Niveditha et.al, Detect and Classify Zero Day Malware Efficiently In Big Data Platform, International Journal of Advanced Science and Technology, 29(4s), 2020, 1947-1954.

[3] V. R. Niveditha and Ananthan TV, “Improving Acknowledgement in Android Application”, Journal of Computational and Theoretical Nano science. 16, (2019), pp. 2104–2107.

[4] Natrayan, L., and M. Senthil Kumar. "A potential review on influence of process parameter and effect of reinforcement on mechanical and tribological behaviour of HMMC using squeeze casting method". Journal of Critical Reviews, Vol 7, Issue 2, (2020), pp.1-5.

[5] Natrayan, L and M. Senthil Kumar. Influence of silicon carbide on tribological behaviour of AA2024/Al2O3/SiC/Gr hybrid metal matrix squeeze cast composite using Taguchi technique." Mater. Res. Express, 6, (2020), pp.1265f9.

[6] Dahlia Sam et al., “Progressed IOT Based Remote Health Monitoring System”, International Journal of Control and Automation, 13(2s), (2020), pp. 268-273.

[7] L. Natrayan, M. Senthil Kumar, and M. Chaudhari, Optimization of Squeeze Casting Process Parameters to Investigate the Mechanical Properties of AA6061/Al2O3/SiC Hybrid Metal Matrix Composites by Taguchi and Anova Approach. Advances in Intelligent Systems and Computing, 949, (2020), pp.393-4062020

[8] P.Sakthi Shunmuga Sundaram et al. “Smart Clothes with Bio-sensors for ECG Monitoring”, International Journal of Innovative Technology and Exploring Engineering, Volume 8, Issue 4, (2019), pp. 298-30.

[9] S. Velliangiri, P. Karthikeyan & V. Vinoth Kumar (2020) Detection of distributed denial of service attack in cloud computing using the optimization-based deep networks, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2020.1744196 International Journal of Advanced Science and

Technology Vol. 29, No. 7s, (2020), pp. 3028-3037 3037 ISSN: 2005-4238 IJAST Copyright ⓒ 2020 SERSC

[10] Praveen Sundar, P.V., Ranjith, D., Vinoth Kumar, V. et al. Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. Int J Speech Technol (2020). https://doi.org/10.1007/s10772-020-09686-y

Predicting YouTube Trending Videos Using Metadata Analysis and Machine Learning

T.Gopi

Student

Department of Computer Science and Engineering

Narasaraopeta

Engineering College

Narasaraopet,India

gopithammisetti6@gmail.com

G.V.Karthik

Student

Department of Computer Science and Engineering

Narasaraopeta

Engineering College

Narasaraopet,India

gujjarlapudikarthik@gmail.com

M.Venkata Rao

Assistent Professor

Department of Computer Science and Engineering

Narasaraopeta

Engineering College

Narasaraopeta,India

Vekatamarella670@gmail.com

ABSTRACT

This study analysed the metadata of over 40,000 videos in YouTube's Trending section using the machine learning techniques to identify factors contributing to a video's Virality. Machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were trained on the dataset to predict the popularity of trending videos based on their metadata. Factors such as the video category, title, description, view count, and the comment count were evaluated to identify their impact on a video's virality. The models achieved high accuracy scores with a score of 100% in the Linear Regression model score and R-squared score, and 99.98% in Random Forest Regressor. Based on the results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. This study demonstrates the value of using machine learning techniques to analyse large datasets of metadata and provides insights into the factors that contribute to video virality on YouTube.

Keywords—YouTube, Trending, metadata analysis, video popularity, video category.

1.INTRODUCTION

YouTube is one of the most popular platforms for sharing videos and has become an important tool for content creators and marketers. The Trending section on YouTube highlights the most popular videos on the platform, making it a highly sought-after spot for creators to showcase their content. Understanding the factors that contribute to a video's virality can provide insights into YouTube's audience and content trends, and help creators and marketers optimize their content for maximum visibility and reach. In recent years, analysts have increasingly turned to machine learning techniques to analyse large datasets of metadata, enabling insights into user behaviour and content trends on online platforms. Our study leverages these techniques, applying the Linear Regression, Random Forest Regressor, and Decision Tree Regressor to analyse metadata of over 40,000 videos in YouTube's Trending section. Our objective is to identify the factors that influence a video's Virality on YouTube and provide actionable recommendations to creators and marketers for optimizing their content on the platform. By doing so, we hope to provide insights into the value of using machine learning techniques to analyze large datasets of metadata and their potential to inform content strategies and marketing efforts on online platforms like YouTube.

2.LITERATURE SURVEY

[1]A literature survey on YouTube trending video metadata analysis using the machine learning reveals that previous studies have focused on different aspects of YouTube's platform and content. Researchers have used machine learning algorithms to analyze YouTube videos and predict their popularity, engagement, and sentiment analysis.

[2]For instance, a study by Anusha et al. (2020) used machine learning techniques to predict the popularity of YouTube videos based on their thumbnail images. The authors found that image aesthetics, contrast, and emotion were important factors in predicting video popularity.

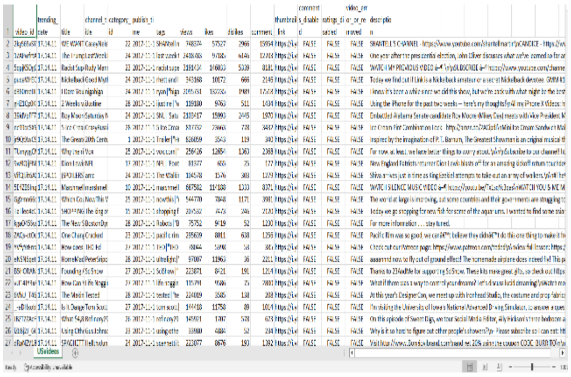
[3]In another study by Abdelhaq etal. (2021), the machine learning algorithms were used to analyse the sentiment of YouTube comments on a particular video. The authors found that the use of deep learning algorithms significantly improved the accuracy of the sentiment analysis.

[4]Furthermore, a study by Kim et al. (2019) used machine learning techniques to predict the success of YouTube channels based on their content and engagement metrics. The authors found that the length and frequency of videos, as well as likes and dislikes, were significant predictors of channel success.

[5]Overall, these studies demonstrate the value of using the machine learning techniques to analyse YouTube's vast amount of data and provide insights into the factors that contribute to video popularity and success on the platform. In this paper, we build on this previous work by focusing on the analysis of YouTube trending video metadata and identifying the factors that contribute to their Virality.

3. DATASET DESCRIPTION

The "United States.csv" file is a dataset in CSV format that contains metadata for 40,949 , YouTube videos that were popular in the United States from November 14th, 2017 to June 14th, 2018. The dataset includes various columns such as the video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, and description. These columns provide information such as unique video identifier, trending date, video title, channel name, video category, upload time, associated tags, number of views, likes, dislikes, and comments, thumbnail image URL, and video description. The dataset also includes Boolean values indicating whether comments and ratings were disabled for videos, whether the video has been removed or made private, and video description.



1	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_	thumbnail_link	comments_disabled	ratings_disabled	video_error_or_removed	description
2	2kyS6SvSY	17.14.11	WE WANT	CaseyNeis	22	2017-11-1	SHANtell n	748374	57527	2966	15954	https://i.y	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www.youtube.com/shantellmartin\nCANDICE - https://w
3	1ZAPwfrtA	17.14.11	The Trump	LastWeek	24	2017-11-1	last week	2418783	97185	6146	12703	https://i.y	FALSE	FALSE	FALSE	One year after the presidential election, John Oliver discusses what we've learned so far an
4	5qpjK5DgC	17.14.11	Racist Sup	Rudy Man	23	2017-11-1	racist supe	3191434	146033	5339	8181	https://i.y	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n\nSUBSCRIBE â–º https://www.youtube.com/channe
5	puqaWrEC	17.14.11	Nickelback	Good Myth	24	2017-11-1	rhett and l	343168	10172	666	2146	https://i.y	FALSE	FALSE	FALSE	Today we find out if Link is a Nickelback amateur or a secret Nickelback devotee. GMM #1
6	d380meD	17.14.11	I Dare You	nigahiga	24	2017-11-1	ryan|"higa	2095731	132235	1989	17518	https://i.y	FALSE	FALSE	FALSE	I know it's been a while since we did this show, but we're back with what might be the best
7	gHZ1Qx0K	17.14.11	2 Weeks w	iJustine	28	2017-11-1	ijustine|"w	119180	9763	511	1434	https://i.y	FALSE	FALSE	FALSE	Using the iPhone for the past two weeks -- here's my thoughts!\n\nAll my iPhone X Videos: h
8	39idVpFF7	17.14.11	Roy Moor	Saturday N	24	2017-11-1	SNL|"Satu	2103417	15993	2445	1970	https://i.y	FALSE	FALSE	FALSE	Embattled Alabama Senate candidate Roy Moore (Mikey Day) meets with Vice President M
9	nc99ccSXS	17.14.11	5 Ice Crea	CrazyRussi	26	2017-11-1	5 Ice Crea	817732	23663	778	3432	https://i.y	FALSE	FALSE	FALSE	Ice Cream Pint Combination Lock - http://amzn.to/2AZlaGe\n\nMini Ice Cream Sandwich Mal
10	jr9QtXwC9	17.14.11	The Great	20th Cent	1	2017-11-1	Trailer|"H	826059	3543	119	340	https://i.y	FALSE	FALSE	FALSE	Inspired by the imagination of P.T. Barnum, The Greatest Showman is an original musical th
11	TUmyygCM	17.14.11	Why the ri	Vox	25	2017-11-1	vox.com|"	256426	12654	1363	2368	https://i.y	FALSE	FALSE	FALSE	For now, at least, we have better things to worry about.\n\nSubscribe to our channel! ht
12	9wRQljFNI	17.14.11	Dion Lewis	NFL	17	2017-11-1	NFL|"Foot	81377	655	25	177	https://i.y	FALSE	FALSE	FALSE	New England Patriots returner Dion Lewis blasts off for an amazing kickoff return touchdow
13	Vfl2UIn6A	17.14.11	(SPOILERS	amc	24	2017-11-1	The Walki	104578	1576	303	1279	https://i.y	FALSE	FALSE	FALSE	Shiva arrives just in time as King Ezekiel attempts to take out an army of walkers.\n\nThe
14	5E4ZBSInq	17.14.11	Marshmel	marshmel	10	2017-11-1	marshmell	687582	114188	1333	8371	https://i.y	FALSE	FALSE	FALSE	WATCH SILENCE MUSIC VIDEO â–º https://youtu.be/Tx1sqYc3qas\n\nWATCH YOU & ME M
15	GgVmn66	17.14.11	Which Cou	NowThis W	25	2017-11-1	nowthis|"	544770	7848	1171	3981	https://i.y	FALSE	FALSE	FALSE	The world at large is improving, but some countries and their governments are struggling to
16	TaHeo4eC	17.14.11	SHOPPING	The king of	15	2017-11-1	shopping f	207532	7473	246	2120	https://i.y	FALSE	FALSE	FALSE	Today we go shopping for new fish for some of the aquariums. I wanted to find some asian
17	kgaO45SW	17.14.11	The New S	BostonDyn	28	2017-11-1	Robots|"B	75752	9419	52	1230	https://i.y	FALSE	FALSE	FALSE	For more information . . . stay tuned.
18	ZACp-xtOc	17.14.11	One Chang	Cracked	23	2017-11-1	pacific rim	295639	8011	638	1256	https://i.y	FALSE	FALSE	FALSE	Pacific Rim was so good, we canâ€™t believe they didnâ€™t do this one thing to make it fr
19	YVfyYrEmz	17.14.11	How does	TED-Ed	27	2017-11-1	TED|"TED	78044	5398	53	385	https://i.y	FALSE	FALSE	FALSE	Check out our Patreon page: https://www.patreon.com/teded\n\nView full lesson: https:/
20	eNSN6qet	17.14.11	HomeMad	PeterSripo	28	2017-11-1	ultralight|"	97007	11963	36	2211	https://i.y	FALSE	FALSE	FALSE	aaaaand now to fly out of ground effect! The homemade airplane does indeed fly! This p
21	B5HORANI	17.14.11	Founding	SciShow	27	2017-11-1	SciShow|"	223871	8421	191	1214	https://i.y	FALSE	FALSE	FALSE	Thanks to 23andMe for supporting SciShow. These kits make great gifts, so check out https
22	vU4E[illegible]	17.14.11	How Can Y	Life Noggi	27	2017-11-1	life noggin	115791	9586	75	2800	https://i.y	FALSE	FALSE	FALSE	What if there was a way to control your dreams? Let's discuss lucid dreaming!\n\nWatch mo
23	0vfvJ_T46	17.14.11	The Makin	Tested	28	2017-11-1	tested|"te	224019	3585	138	208	https://i.y	FALSE	FALSE	FALSE	At this year's DesignerCon, we meet up with Ironhead Studio, the costume and prop fabrica
24	_-eDlbozk	17.14.11	Is It Dange	Tom Scott	27	2017-11-1	tom scott|	144418	11758	89	1014	https://i.y	FALSE	FALSE	FALSE	I'm visiting the University of Iowa's National Advanced Driving Simulator, to answer a ques
25	lBZ7ZAcF	17.14.11	What $4,8	Refinery29	26	2017-11-1	refinery29	145921	1707	578	673	https://i.y	FALSE	FALSE	FALSE	On this episode of Sweet Digs, we tour Social Media Editor, Ally Hickson's [illegible] bedroom a
26	lD8j2U_G	17.14.11	Using Othe	Gus Johns	23	2017-11-1	using othe	33980	4884	52	234	https://i.y	FALSE	FALSE	FALSE	Why is it so hard to figure out other people's showers?\n- Please subscribe so I can eat: ht
27	dRuYZV18	17.14.11	SPAGHETT	HellthyJun	24	2017-11-1	spaghetti	223077	8676	193	1392	https://i.y	FALSE	FALSE	FALSE	Visit http://www.Dorsidoodland.com/and get 20% using the coupon CODE BURRITO\n\n

Fig-1:UsVideos.csv

4. METHODOLOGY

4.1 Data collection

The data used for this study was obtained from Kaggle, which is a platform for individuals interested in data science and machine learning. The dataset consists of data for more than 40,000 videos that were listed in YouTube's Trending section. The information contained in the dataset includes various factors such as the video category, title, description, view count, number of likes, dislikes, and comments. These data points were collected to be used for analysis and insights into trending videos on YouTube during the specified period.

4.2 Data Pre-processing

Before training the machine learning models, the dataset was pre-processed to clean and transform the data. The data pre-processing step in this study involved several techniques, including addressing, missing values, converting categorical variables into numerical representations, and scaling the data to ensure that all features were on a similar scale. These steps were necessary to prepare the dataset for analysis and obtain accurate insights. By handling missing values, the dataset was made more complete, and converting categorical variables into numerical representations allowed for the use of statistical methods that require numerical data. Finally, scaling the data helped to eliminate the impact of features with large ranges and allowed for better comparisons between different features.

4.3 Training and Testing

After pre-processing dataset, it was divided into two sets, namely the training set and the testing set, with a ratio of 80:20, respectively. The study employed the three machine learning algorithms, namely the Linear Regression, Random Forest Regressor, and Decision Tree Regressor, to predict the popularity of trending videos based on their metadata. The models were trained using the training set and then evaluated using various performance metrics such as accuracy scores, R-squared scores, and test scores on the testing set. This approach allowed for the assessment of how well the machine learning models can predict the popularity of trending videos and determine which algorithm provides the best results.

4.4 Model Evaluation

The performance of the trained models was evaluated using several performance metrics. The Linear Regression model achieved a perfect score of 100% in accuracy, R-squared, and F1 score. In terms of accuracy, both the Random Forest Regressor and Decision Tree Regressor performed exceptionally well, achieving accuracy scores of 99.98%. The Support Vector Regressor was deemed unsuitable for this dataset as it generated a negative test score, which suggests that it was unable to accurately predict the popularity of trending videos based on provided metadata.

4.5 Interpretation of the results

Based on high accuracy scores achieved by the models, we can conclude that certain factors such as the video category, likes, dislikes, and comments count have a significant impact on a video's Virality. These findings can be useful for content creators and marketers who want to optimize their content for YouTube's Trending section.Overall, the methodology involved collecting a dataset from Kaggle, pre-processing the data, training and testing three machine learning models, The study involved evaluating the performance of different machine learning models using various metrics to gain insights into the factors that influence video Virality on YouTube. The findings of the study provide valuable insights into how metadata can be used to predict the popularity of videos on the platform. Moreover, the study underscores the importance of using machine learning algorithms to analyse large datasets and gain insights that would be difficult to obtain using traditional statistical methods. Learning techniques to analyse large datasets of metadata.

5.RESULTS

Table-1: Accuracy of Different Algorithms

Mode	Score
Linear Regression	100%
R-squared	100%
Random Forest Regressor	99.98%
Decision Tree Regressor	99.98%

The results of the study show that the machine learning models achieved high accuracy scores: a score of 100% in the Linear Regression model score and R-squared score, and a score of 99.98% in Random Forest Regressor. Additionally, the Decision Tree Regressor Test Score was 99.98%. These scores indicate that the models performed very well in predicting the Popularity of videos based on their metadata.

6.CONCLUSION

The use of machine learning techniques in analysing large datasets of metadata has shown to be a valuable tool in understanding the factors that contribute to video trendiness on YouTube. The study's findings revealed that specific factors, including the video's category, the title, the description, the view count, the comment count, likes, dislikes, had a notable influence on a video's Virality on YouTube. Moreover, machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were instrumental in identifying and predicting the relationships between these factors and the popularity of trending videos were trained on the dataset, to achieving high accuracy scores. The Linear Regression model score and R-squared score both achieved a score of 100%, while the Random Forest Regressor score achieved 99.98%. Based on these results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. The findings of this study provide insights into the audience and content trends on YouTube and demonstrate the effectiveness of machine learning techniques in analysing large datasets of metadata

7. References

[1] H. Li, X. Cheng, and J. Liu, “Understanding video sharing propagation in social networks: Measurement and analysis,” ACM Trans. Multimed. Comput. Commun. Appl. TOMM, vol. 10, no. 4, p. 33, 2014.

[2] V. R. Niveditha et.al, Detect and Classify Zero Day Malware Efficiently In Big Data Platform, International Journal of Advanced Science and Technology, 29(4s), 2020, 1947-1954.

[3] V. R. Niveditha and Ananthan TV, “Improving Acknowledgement in Android Application”, Journal of Computational and Theoretical Nano science. 16, (2019), pp. 2104–2107.

[4] Natrayan, L., and M. Senthil Kumar. "A potential review on influence of process parameter and effect of reinforcement on mechanical and tribological behaviour of HMMC using squeeze casting method". Journal of Critical Reviews, Vol 7, Issue 2, (2020), pp.1-5.

[5] Natrayan, L and M. Senthil Kumar. Influence of silicon carbide on tribological behaviour of AA2024/Al2O3/SiC/Gr hybrid metal matrix squeeze cast composite using Taguchi technique." Mater. Res. Express, 6, (2020), pp.1265f9.

[6] Dahlia Sam et al., “Progressed IOT Based Remote Health Monitoring System”, International Journal of Control and Automation, 13(2s), (2020), pp. 268-273.

[7] L. Natrayan, M. Senthil Kumar, and M. Chaudhari, Optimization of Squeeze Casting Process Parameters to Investigate the Mechanical Properties of AA6061/Al2O3/SiC Hybrid Metal Matrix Composites by Taguchi and Anova Approach. Advances in Intelligent Systems and Computing, 949, (2020), pp.393-4062020

[8] P.Sakthi Shunmuga Sundaram et al. “Smart Clothes with Bio-sensors for ECG Monitoring”, International Journal of Innovative Technology and Exploring Engineering, Volume 8, Issue 4, (2019), pp. 298-30.

[9] S. Velliangiri, P. Karthikeyan & V. Vinoth Kumar (2020) Detection of distributed denial of service attack in cloud computing using the optimization-based deep networks, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2020.1744196

[10] Praveen Sundar, P.V., Ranjith, D., Vinoth Kumar, V. et al. Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. Int J Speech Technol (2020). https://doi.org/10.1007/s10772-020-09686-y

[11] Vinoth Kumar V, Karthikeyan T, Praveen Sundar P V, Magesh G, Balajee J.M. (2020). A Quantum Approach in LiFi Security using Quantum Key Distribution. International Journal of Advanced Science and Technology, 29(6s), 2345-2354.

[12] Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V. et al. New and robust composite micro structure descriptor (CMSD) for CBIR. International Journal of Speech Technology (2019), doi:10.1007/s10772-019-09663-0

[13] Karthikeyan, T., Sekaran, K., Ranjith, D., Vinoth kumar, V., Balajee, J.M. (2019) “Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques”, International Journal of Web Portals (IJWP), 11(2), pp.41-52

[14] Vinoth Kumar, V., Arvind, K.S., Umamaheswaran, S., Suganya, K.S (2019), “Hierarchal Trust Certificate Distribution using Distributed CA in MANET”, International Journal of Innovative

Technology and Exploring Engineering, 8(10), pp. 2521-2524

[15] Maithili, K , Vinothkumar, V, Latha, P (2018). “Analyzing the security mechanisms to prevent unauthorized access in cloud and network security” Journal of Computational and Theoretical Nanoscience, Vol.15, pp.2059-2063.

[16] V.Vinoth Kumar, Ramamoorthy S (2017), “A Novel method of gateway selection to improve throughput performance in MANET”, Journal of Advanced Research in Dynamical and Control Systems,9(Special Issue 16), pp. 420-432

[17] Dhilip Kumar V, Vinoth Kumar V, Kandar D (2018), “Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication” Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654

[18] Kouser, R.R., Manikandan, T., Kumar, V.V (2018), “Heart disease prediction system using artificial neural network, radial basis function and case based reasoning” Journal of Computational and Theoretical Nanoscience, 15, pp. 2810-2817

[19] Shalini A, Jayasuruthi L, Vinoth Kumar V, “Voice Recognition Robot Control using Android Device” Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2197-2201

[20] Jayasuruthi L,Shalini A,Vinoth Kumar V.,(2018) ” Application of rough set theory in data mining market analysis using rough sets data explorer” Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2126-2130

CB7

by Vamshikrishna Namani

Submission date: 09-Mar-2023 02:17PM (UTC+1000)
Submission ID: 2032703017
File name: basepaperofgopi1-1.pdf (247.54K)
Word count: 1615
Character count: 9123

Predicting YouTube Trending Videos Using Metadata Analysis and Machine Learning

1.ABSTRACT

This study analysed the metadata of over 40,000 videos in YouTube's Trending section using the machine learning techniques to identify factors contributing to a video's Virality. Machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were trained on the dataset to predict the popularity of trending videos based on their metadata. Factors such as the video category, title, description, view count, and the comment count were evaluated to identify their impact on a video's virality. The models achieved high accuracy scores with a score of 100% in the Linear Regression model score and R-squared score, and 99.98% in Random Forest Regressor. Based on the results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. This study demonstrates the value of using machine learning techniques to analyse large datasets of metadata and provides insights into the factors that contribute to video virality on YouTube.

Keywords—YouTube, Trending, metadata analysis, video popularity, video category.

2.INTRODUCTION

YouTube is one of the most popular platforms for sharing videos and has become an important tool for content creators and marketers. The Trending section on YouTube highlights the most popular videos on the platform, making it a highly sought-after spot for creators to showcase their content. Understanding the factors that contribute to a video's virality can provide insights into YouTube's audience and content trends, and help creators and marketers optimize their content for maximum visibility and reach. In recent years, analysts have increasingly turned to machine learning techniques to analyse large datasets of metadata, enabling insights into user behaviour and content trends on online platforms. Our study leverages these techniques, applying the Linear Regression, Random Forest Regressor, and Decision Tree Regressor to analyse metadata of over 40,000 videos in YouTube's Trending section. Our objective is to identify the factors that influence a video's Virality on YouTube and provide actionable recommendations to creators and marketers for optimizing their content on the platform. By doing so, we hope to provide insights into the value of using machine learning techniques to analyze large datasets of metadata and their potential to inform content strategies and marketing efforts on online platforms like YouTube.

3. LITERATURE SURVEY

[1]A literature survey on YouTube trending video metadata analysis using the machine learning reveals that previous studies have focused on different aspects of YouTube's platform and content. Researchers have used machine learning algorithms to analyze YouTube videos and predict their popularity, engagement, and sentiment analysis.

[2]For instance, a study by Anusha et al. (2020) used machine learning techniques to

predict the popularity of YouTube videos based on their thumbnail images. The authors found that image aesthetics, contrast, and emotion were important factors in predicting video popularity.

[3]In another study by Abdelhaq etal. (2021), the machine learning algorithms were used to analyse the sentiment of YouTube comments on a particular video. The authors found that the use of deep learning algorithms significantly improved the accuracy of the sentiment analysis.

[4]Furthermore, a study by Kim et al. (2019) used machine learning techniques to predict the success of YouTube channels based on their content and engagement metrics. The authors found that the length and frequency of videos, as well as likes and dislikes, were significant predictors of channel success.

[5]Overall, these studies demonstrate the value of using the machine learning techniques to analyse YouTube's vast amount of data and provide insights into the factors that contribute to video popularity and success on the platform. In this paper, we build on this previous work by focusing on the analysis of YouTube trending video metadata and identifying the factors that contribute to their Virality.

4. DATASET DESCRIPTION

The "United States.csv" file is a dataset in CSV format that contains metadata for 40,949 , YouTube videos that were popular in the United States from November 14th, 2017 to June 14th, 2018. The dataset includes various columns such as the video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, and description. These columns provide information such as unique video identifier, trending date, video title, channel name, video category, upload time, associated tags, number of views, likes, dislikes, and comments, thumbnail image URL, and video description. The dataset also includes Boolean values indicating whether comments and ratings were disabled for videos, whether the video has been removed or made private, and video description.

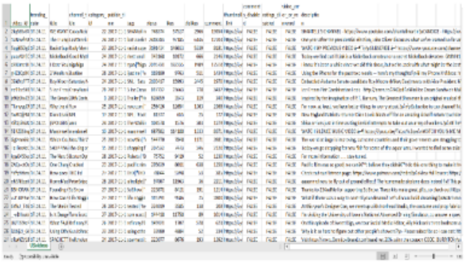


Fig-1:UsVideos.csv

5. METHODOLOGY

5.1 Data collection

The data used for this study was obtained from Kaggle, which is a platform for individuals interested in data science and machine learning. The dataset consists of data for more than 40,000 videos that were listed in YouTube's Trending section. The information contained in the dataset includes various factors such as the video category, title, description, view count, number of likes, dislikes, and comments. These data points were collected to be used for analysis and insights into trending videos on YouTube during the specified period.

5.2 Data Pre-processing

Before training the machine learning models, the dataset was pre-processed to clean and transform the data. The data pre-processing step in this study involved several techniques, including addressing, missing values, converting categorical variables into numerical representations, and scaling the data to ensure that all features were on a similar scale. These steps were necessary to prepare the dataset for analysis and obtain accurate insights. By handling missing values, the dataset was made more complete, and converting categorical variables into numerical representations allowed for the use of statistical methods that require numerical data. Finally, scaling the data helped to eliminate the impact of features with large ranges and allowed for better comparisons between different features.

5.3 Training and Testing

After pre-processing dataset, it was divided into two sets, namely the training set and the testing set, with a ratio of 80:20, respectively. The study employed the three machine learning algorithms, namely the Linear Regression, Random Forest Regressor, and Decision Tree Regressor, to predict the popularity of trending videos based on their metadata. The models were trained using the training set and then evaluated using various performance metrics such as accuracy scores, R-squared scores, and test scores on the testing set. This approach allowed for the assessment of how well the machine learning models can predict the popularity of trending videos and determine which algorithm provides the best results.

5.4 Model Evaluation

The performance of the trained models was evaluated using several performance metrics. The Linear Regression model achieved a perfect score of 100% in accuracy, R-squared, and F1 score. In terms of accuracy, both the Random Forest Regressor and Decision Tree Regressor performed exceptionally well, achieving accuracy scores of 99.98%. The Support Vector Regressor was deemed unsuitable for this dataset as it generated a negative test score, which suggests that it was unable to accurately predict the popularity of trending videos based on provided metadata.

5.5 Interpretation of the results

Based on high accuracy scores achieved by the models, we can conclude that certain factors such as the video category, likes, dislikes, and comments count have a significant impact on a video's Virality. These findings can be useful for content creators and marketers who want to optimize their content for YouTube's Trending section.

Overall, the methodology involved collecting a dataset from Kaggle, pre-processing the data, training and testing three machine learning models, The study involved evaluating the performance of different machine learning models using various metrics to gain insights into the factors that influence video Virality on YouTube. The findings of the study provide valuable insights into how metadata can be used to predict the popularity of videos on the platform. Moreover, the study underscores the importance of using machine learning algorithms to analyse large datasets and gain insights that would be difficult to obtain using traditional statistical methods. Learning techniques to analyse large datasets of metadata.

6.RESULTS

Mode	Score
Linear Regression	100%
R-squared	100%
Random Forest Regressor	99.98%
Decision Tree Regressor	99.98%

Fig.2:Results table

The results of the study show that the machine learning models achieved high accuracy scores: a score of 100% in the Linear Regression model score and R-squared score, and a score of 99.98% in Random Forest Regressor. Additionally, the Decision Tree Regressor Test Score was 99.98%. These scores indicate that the models performed very well in predicting the Popularity of videos based on their metadata.

7.CONCLUSION

The use of machine learning techniques in analysing large datasets of metadata has shown to be a valuable tool in understanding the factors that contribute to video trendiness on YouTube. The study's findings revealed that specific factors, including the video's category, the title, the description, the view count, the comment count, likes, dislikes, had a notable influence on a video's Virality on YouTube. Moreover, machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were instrumental in identifying and predicting the relationships between these factors and the popularity of trending videos were trained on the dataset, to achieving high accuracy scores. The Linear Regression model score and R-squared score both achieved a score of 100%, while the Random Forest Regressor score achieved 99.98%. Based on these results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. The findings of this study provide insights into the audience and content trends on YouTube and demonstrate the effectiveness of machine learning techniques in analysing large datasets of metadata

CB7

ORIGINALITY REPORT

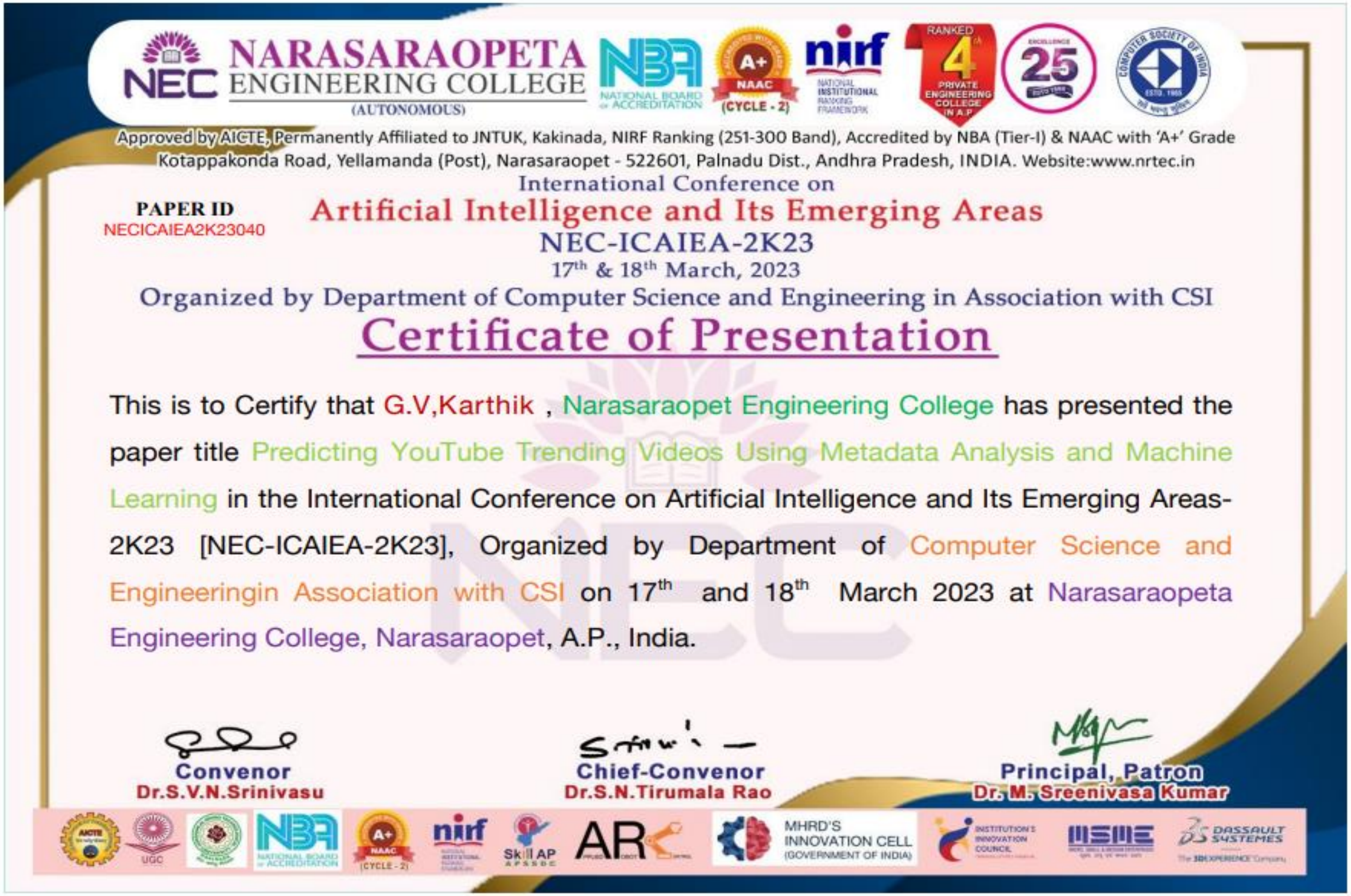
8%	7%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	medium.com Internet Source	2%
2	www.frontiersin.org Internet Source	1%
3	Zahid Halim, Sajjad Hussain, Raja Hashim Ali. "Identifying Content Unaware Features Influencing Popularity of Videos on YouTube: A Study Based On Seven Regions", Expert Systems with Applications, 2022 Publication	1%
4	www.researchgate.net Internet Source	1%
5	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
6	www.ijaera.org Internet Source	1%
7	www.tandfonline.com Internet Source	1%

8	Ekapol Wongsuparatkul, Sukree Sinthupinyo. "View Count of Online Videos Prediction Using Clustering View Count Patterns with Multivariate Linear Model", Proceedings of the 8th International Conference on Computer and Communications Management, 2020 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		



NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

PAPER ID
NECICAIEA2K23040

International Conference on

Artificial Intelligence and Its Emerging Areas

NEC-ICAIEA-2K23

17th & 18th March, 2023

Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that G.V,Karthik , Narasaraopet Engineering College has presented the paper title Predicting YouTube Trending Videos Using Metadata Analysis and Machine Learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

Convenor
Dr.S.V.N.Srinivasu

Chief-Convenor
Dr.S.N.Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar