Predictive analysis of YouTube trending videos using Machine Learning

Indira Gandhi Delhi Technical University, Delhi

Manisha-10601012022

manisha106btcse22@igdtuw.ac.in

Pavitra Asthana- 13301012022

pavitra133btcse22@igdtuw.ac.in

Prakriti Rai – 13801012022

prakriti138btcse22@igdtuw.ac.in

Acknowledgements:

I would like to express my heartfelt gratitude to the AI & ML Club of IGDTUW for their unwavering support and invaluable guidance throughout my journey of learning and exploring the realms of Machine Learning. The encouragement, mentorship, and tireless effort extended by the club members have been instrumental in shaping my understanding and skillset in this dynamic field.

I extend my deepest appreciation to the club for providing me with the platform to channel my passion and curiosity into a major project. With the club's support, I was able to create a prediction model designed to foresee the possibility of a video going trending on YouTube. This project not only sharpened my technical skills but also instilled in me the confidence to tackle real-world challenges using Machine Learning techniques.

Thank you, AI & ML Club of IGDTUW, for being an integral part of my learning journey and for enabling me to bring my ideas to life through this remarkable project.

Abstract

YouTube stands as a widely recognized interactive platform for sharing videos globally, enabling users to rate, share, comment on, save, and upload content. While commonly favourited videos amass likes and views over time, YouTube's trending videos embody content that garners viewership within a specific timeframe, indicating potential for wider popularity. Despite its significance, there exists a limited body of research on analyzing YouTube trending videos. This study aims to examine interactive attributes in order to ascertain the connections and significance of variables contributing to a video's trendiness. The focal point is understanding how interactive video features propel a video's trending status on YouTube.

The investigation hinges on a dataset encompassing viewership statistics of over 40,000 YouTube trending videos within a designated span. This dataset encompasses Views, Likes, Dislikes, and Comment counts. To delve into the predictive aspects, the study employs a Linear Regression model from the realm of Machine Learning to forecast the number of views a YouTube trending video might accumulate.

Additionally, the research undertakes a comparative evaluation of diverse classification models – namely Random Forest, kNN , Decision Tree and Logistic Regression. The aim is to discern the optimal model for predicting both the duration a video takes to attain trending status from its upload time, and the duration it remains on the trending list. The study attains a peak accuracy of 93% in predicting the lifecycle of YouTube's trending videos.

INTRODUCTION

YouTube stands as the preeminent online platform for sharing videos globally. Since its establishment in 2005, YouTube has evolved into a vast digital universe. With a staggering daily viewership exceeding 4 billion, it has secured its position as the prime avenue for user-generated content. Through interactive video features, YouTube caters to both the public and content creators. One such feature is

"Views," representing the cumulative viewership a specific video has garnered over time. Generally, a video's popularity is gauged by the count of views it accumulates, and it follows a certain

trajectory before achieving widespread recognition.

Occasionally, certain content captures the collective attention swiftly, and these instances fall under the purview of YouTube's "trending videos" category. While these videos might not be considered outright popular when featured under the "trending" tab on YouTube, they exhibit the potential to attain future popularity. Despite their significance, the domain of YouTube trending videos remains underexplored in terms of analysis. Considering that over a billion distinct users visit YouTube monthly and a staggering 72 hours of video are uploaded to the platform every minute, YouTube has emerged as a colossal business arena

In 2013, YouTube unveiled novel revenue streams, ushering in elements like brand management, advertising, and promotional endeavors. As a video gains traction, it is made accessible to a vast viewership for free, capturing widespread attention temporarily. Forecasting which content is poised to trend imminently or potentially attain popularity is a challenging endeavor. Hence, the introduction of predictive analysis employing Machine Learning techniques comes into play.

Certain content manages to capture mass attention rapidly, falling under YouTube's trending videos category. While these videos might not be immediately considered popular under the trending tab, they carry potential for future popularity. Despite their significance, the analysis of YouTube trending videos remains an underexplored area. With over a billion unique monthly visitors and 72 hours of video uploaded per minute, YouTube has transformed into a significant business platform . In 2013, YouTube introduced new revenue streams, including brand management and advertising.

Previous Research and Current Approach Past studies primarily focused on initial video statistics to predict a video's popularity, while a few delved into the unpredictability of viral videos and sudden surges in views. This study follows the CRISP DM methodology and employs Machine Learning algorithms, including Random Forest, SVM, Decision Tree, Logistic Regression, Gaussian Naïve Bayes, and Linear Regression, to predict the lifecycle of YouTube trending videos.

Significance and Implications :This research aids YouTube in analyzing and forecasting the lifecycle of trending content, allowing it to align its business strategies accordingly. Content creators, or YouTubers, who rely on YouTube for revenue, can benefit from this study by analyzing their content's lifecycle and making necessary improvements. Viewer feedback is crucial for YouTubers, as it offers insights into content reception. This study contributes to YouTube's and YouTubers' understanding of how interactive features influence video performance on the platform.

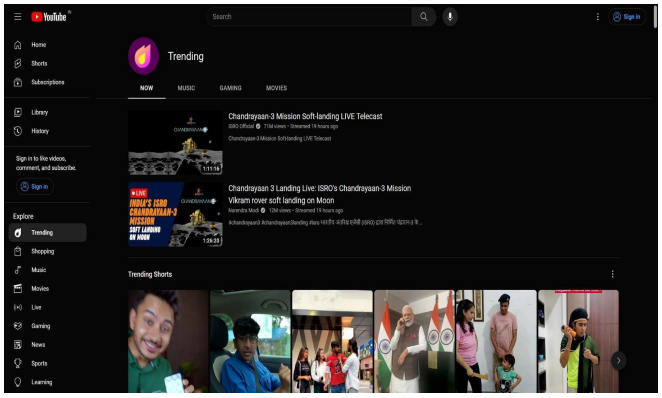

Figure 1: YouTube Trending page

1.1 RESEARCH QUESTION

Comparative analysis of Machine Learning algorithms for YouTube trending video's life cycle prediction by analysing YouTube's trending videos statistics data.

The thesis analysed YouTube's trending videos statistics of over 40000 videos, collected over a period to answer the research problem.

1.1.1 Comparative analysis of ML classifiers for predicting YouTube trending video's lifecycle.

The thesis executed a comparative analysis of Machine Learning algorithms, focusing particularly on Random Forest, SVM (Support Vector Machine), Decision Tree, Logistic Regression, and Gaussian Naïve Bayes classifiers. The objective was to determine the optimal classifier for predictive purposes.

The study explored the life cycle of YouTube's trending videos, which can be divided into two main phases:

Anticipating Time for Trending Inclusion: The initial stage of a trending video's life cycle involves its rise above competing content. The classifiers were employed to predict the duration it would take for a video to make its way onto YouTube's trending page, starting from its upload on the social platform.

Predicting Trending Duration:

Following a video's inclusion on the trending list, the subsequent phase of its life cycle begins. The classifiers were utilized to foresee the number of days a video would maintain its position on YouTube's trending page, commencing from its first day of trending.

This research bears significance for YouTube in terms of marketing strategies, advertisements, and brand management.

1.1.2 .**Predictive analysis of number of views for YouTube's trending videos**

Since trending video statistics consists of features such as number of Views, Likes, Dislikes and Comment counts etc. research performed a Linear regression model of Machine Learning for predictive analysis of number of views for YouTube trending videos. This study will help YouTube as well as YouTubers to garner business from a trending videos life cycle.

1.2 RESEARCH OBJECTIVES

1.2.1 Aim :

The primary aim is to equip YouTube and YouTubers with a sophisticated tool for predicting the lifecycles of their videos. The focus lies in leveraging the potential of Machine Learning to empower data-driven decisions within the dynamic landscape of YouTube video trends.

1.2.2 Objectives

1. Identify Optimal Model: Conduct an extensive comparative analysis of diverse Machine Learning classification algorithms trained on YouTube's trending video statistics data to determine the most suitable model.

2. Model Application: Apply the selected models to the trained data, ensuring robust fitting that captures the intricacies of YouTube video lifecycles.

3. Rigorous Evaluation: Thoroughly assess the outcomes of the models, employing statistical significance tests to establish the effectiveness of the predictions.

1.3 RESEARCH METHODOLOGIES

- Step Business Understanding, Business objectives
- Step 2: Data Understanding, Sample data collection, explore and describe the data
- Step 3: Data Preparation, Modify the data to select and encode the attributes, Integrate data
- Step 4: Data Modeling, Identify and implement models, Significance Testing
- Step 5:Evaluate findings, Compare models

2 LITERATURE REVIEW

A substantial body of research has been dedicated to the study of the YouTube platform, given its status as one of the most prominent user-generated content platforms. Researchers have explored various facets of YouTube, including text mining, natural language processing (NLP), sentiment analysis, and its recommendation system. However, it's worth noting that while some areas have seen extensive investigation, others still offer ample opportunities for further research.

For instance, Zhou et al. delved into the impact of YouTube's recommendation system on video views and discovered a robust correlation between a video's view count and the average view count of its to pre recommended video (Acm.org, 2010). Similarly, Davidson et al. examined the recommendation system, focusing on click through rates (CTR) of videos on the homepage and concluded that YouTube recommendations accounted for a substantial 60% of all video clicks on the homepage (Davidson et al., 2010).

In contrast, the analysis of trending videos on YouTube remains relatively underexplored. Prabha et al. took an interesting approach by utilizing sentiment analysis and machine learning algorithms to predict the popularity of trending videos (Prabha, G.M. et al., 2019). Their research emphasized the potential for improving predictive accuracy, particularly through the use of support vector machines (SVM).

Another note worthy study conducted by Imanet al analyzed over 8000 trending videos across a 90-day period, employing time series analysis techniques, including Granger Causality with significance testing (Iman Barjasteh et al., 2014). This research yielded valuable insights, highlighting the distinct statistical attributes of trending videos and their viewer patterns, particularly within popular categories.

Furthermore, s. Amudha et al. explored YouTube trending video metadata using unsupervised data and machine learning's Decision Tree algorithm to predict efficient courier service (s. Amudha et al., 2020). Their approach provided valuable insights into attribute importance and preprocessing analysis.

The exploration of sentiment trends in video comments was undertaken by Krishna, Zambreno, and Krishnan, who utilized Naïve Bayes algorithms for sentiment analysis (Krishna, Zambreno, and Krishnan, n.d.). Their research revealed a positive correlation between sentiment trends in comments and trending topics on YouTube.

Additionally, Szabo and Huberman conducted a comparative study involving two social platforms, Digg and YouTube, observing a linearity correlation between future popularity and early view data (Gábor Szabó and Huberman, 2008). This study suggested the potential applicability of linear regression models for long-term trending cycle prediction.

Henrique Pinto, Jussara Almeida, and Marcos André Gonçalves explored predictive analysis of YouTube's trending videos based on the S-H model (Szabo and Huberman model) (Henrique Pinto, Jussara Almeida, and Marcos André Gonçalves, 2013). Their research indicated that long term popularity statistics are correlated with early popularity statistics, with improvements achieved through the use of ML and MRBF models.

Figueiredo et al. conducted a comprehensive analysis of videos in YouTube's top lists, those removed due to copyright violations, and those selected through random searches (Flavio Figueiredo,

F. Benevenuto, and J. Almeida, 2011). They observed distinct popularity growth patterns based on video types and highlighted the role of both search and YouTube's internal mechanisms in attracting views.

Moreover, research has explored the relationship between the popularity of YouTube videos and their geographic locality (Anders Brodersen, Scellato, and Mirjam Wattenhofer, 2012). Despite YouTube's global nature, the study revealed limitations imposed by geographic factors.

Li, Eng, and Zhang introduced a novel approach by transforming YouTube video popularity prediction into a multiclass problem, focusing on factors such as time gap, description, and category (Li, Eng, and Zhang, n.d.). Their research emphasized the importance of these attributes for prediction.

The importance of feature selection in predicting YouTube's trending videos was highlighted by Chelaru et al., who examined the impact of social features such as likes, dislikes, and comments (Chelaru, OrellanaRodriguez, and Altingovde, 2012). They employed machine learning algorithms and identified key social features.

Figueiredo et al. explored the role of usergenerated content (UGC) as features for predicting YouTube's trending videos, introducing a novel clustering algorithm called K-Spectral Clustering (KSC) (Figueiredo et al., 2016). Their study emphasized the challenges in predicting clusters with lower F1 scores.

The impact of metadata features like title, tags, thumbnails, and descriptions on the popularity and trendiness of YouTube videos was investigated by Hoiles, Aprem, and Krishnamurthy (Hoiles, Aprem, and Krishnamurthy, 2017). They employed various machine learning algorithms and identified CI Random Forest as the most effective for prediction.

Ouyang, Li, and Li tackled the popularity prediction problem by dividing it into two tasks: predicting a video's popularity level and predicting its view count based on popularity levels (Ouyang, Li, and Li, 2016). They explored SVM and KNN algorithms for improved performance.

Additionally, Trzcinski and Rokita developed a regression method for predicting video popularity using Support Vector Regression with Gaussian Radial Basis Functions, emphasizing the importance of visual features (Trzcinski and Rokita, 2017).

Finally, Tejal Rathod and Mehul Barot analyzed trends on Twitter to predict public opinion on ongoing events, utilizing classification algorithms like SVM and Naïve Bayes (Tejal Rathod and Mehul Barot, 2018). Their study was constrained to textual data due to social platform limitations.

In conclusion, the realm of YouTube trending videos analysis offers abundant opportunities for research, spanning various machine learning methodologies and focusing on diverse aspects of video popularity prediction and viewership patterns. Researchers continue to uncover valuable insights that contribute to a deeper understanding of YouTube's dynamic ecosystem.

3. METHODOLOGY

3.1Business Understanding

YouTube stands as a thriving business platform, encompassing advertising, brand management, and more. Predicting the lifecycle of a trending video poses challenges. This research holds potential for commercial applications benefiting entities like Google, Yahoo, and diverse advertisement services, as well as YouTube and content creators. The project aims to foster stability and informed decisionmaking, contributing to calculated business strategies and offering avenues for revenue generation. The study's implications are vast, reflecting the scope within the realm of business opportunities.

YouTube's recommender system also finds relevance in this study, as it necessitates the effective distribution of trending videos to users. The focus lies on implementing and identifying the most effective model for predicting video lifecycles, providing solutions for commercial applications.

A. Data Understanding

The dataset utilized was sourced from Kaggle , featuring essential YouTube attributes such as Views, Likes, Dislikes, Comments_count, and more. The dataset was meticulously examined, addressing null values, distribution, and feature correlation. Visualizations using Python libraries facilitated a comprehensive grasp of the data's composition and relationships with the target variable.

Notably, the study unveiled the role of interactive video features, particularly the impact of high view counts, in propelling a video's trend on YouTube.

B .Data Preparation

The dataset initially lacked a dependent variable, making it unsupervised. To facilitate analysis, the study transformed it into a supervised dataset, creating labels for two distinct classification problems. Null values were handled, unnecessary columns dropped, and categorical and numerical features encoded and normalized. Feature selection techniques like RFE and Chi square aided in selecting relevant attributes for modeling. Undersampling balanced data classes, focusing on the first instance of a video's trending, mitigating bias.

Data underwent thorough cleaning, preprocessing, and visualization using Python libraries.

3.3Modeling

Five Machine Learning algorithms were employed for analysis, with Cross Validation for significance testing. The algorithms included:

Linear Regression: Among the most prevalent algorithms in Machine Learning, Linear Regression is utilized for discerning the interplay between variables and labels. At its core lies the fundamental concept of linear correlation. In this context,

$$E(Y|X) = \mu(X) = \beta_0 + \beta_1 X$$

A line characterized by an intercept, β_0, and a slope, β_1, is established.

Interpreting this straightforward equation, we understand that Y exhibits a contextual distribution with a mean of $\mu(X)$. Here, Y represents the predicted variable, while X serves as the predictor. In the realm of Linear Regression, the value of Y escalates in tandem with the magnitude of X. Consequently, within Machine Learning, Linear Regression finds application in forecasting the dependent variable.

Decision Tree Algorithm: Employed for classifying labeled variables within a supervised dataset, the Decision Tree algorithm is an integral facet of Machine Learning. The classifier is constructed upon a straightforward tree structure, with a single root node encompassing the most salient attribute. Branches represent potential outcomes, while leaf nodes signify further decisions

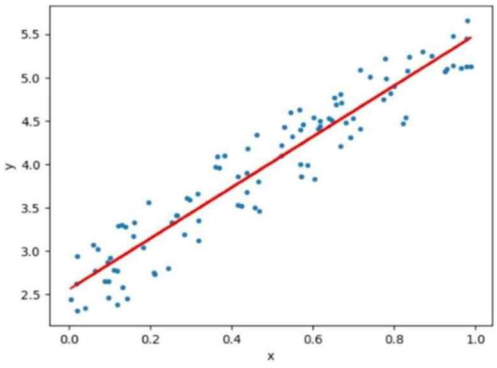

Figure 2: Simple Linear Regression

or variables. The architecture mirrors that of a tree, with categorical or numerical leaf nodes forming a graphical representation. Decision Tree incorporates evaluative metrics such as Entropy, Gini Index, Gain Ratio, and Information Gain. Information Gain quantifies the insight gleaned from observing an independent variable with respect to another. Entropy gauges the uncertainty of an independent random variable. Entropy serves as a gauge for uncertainty and the randomness of outputs within machine learning models. If an action lacks control over its outcome — akin to a coin toss or the color of a ball — the model's entropy approximates the randomness of the outcome. Information Gain denotes the insight amassed from observing an independent variable concerning another. This criterion hinges on impurity, utilizing Entropy as a measure. The reduction of entropy levels via Information Gain curbs the uncertainty of models. Greater Information Gain corresponds to lower entropy and, consequently, heightened accuracy. In a decision tree, the root node assumes paramount importance, furnishing the Information Gain for subsequent decisions. The equation for Information Gain and entropy is expressed as:

$\text{InformationGain}(a_i, S) =$ Entropy

$$(y,S) - \sum_{v_{i,j} \in \text{dom}(a_i)} \frac{|\sigma(a_i = v_{i,j}, S)|}{|S|} \cdot$$

Entropy(y,σ(a_i = $v_{i,j}$,S))

1) where:

Entropy

$$(y,S) = \sum_{c_j \in \text{dom}(y)} -\frac{\bar{\sigma}(y = c_j, S)}{|S|} \cdot \log_2 \frac{\bar{\sigma}(y = c_j, S)}{|S|}$$

Logistic Regression:

Diverging from the straight-line nature of Linear

Regression, Logistic Regression takes a different approach by fitting a sigmoidal or S-shaped logistic function to the data points. This function facilitates the modeling of the relationship between the predictor variables and the probability of a certain class outcome.

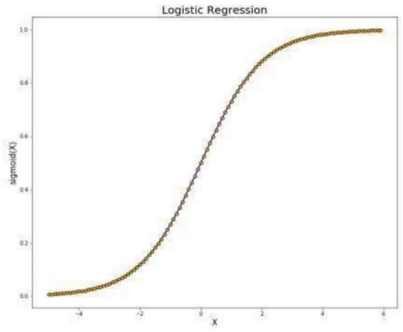

Figure 3: Logistic Regression

Visualized in Figure 3, the

Logistic Regression process becomes evident. This technique finds its utility in cases where a binary prediction variable is present in the dataset, aligning with the specific focus of the current study. For scenarios involving more than two classes, the extension known as Multinomial Logistic Regression comes into play.

The foundation of Logistic Regression hinges on the LOGIT mathematical concept, enabling the construction of the S shaped logistic function. LOGIT constitutes a natural logarithmic transformation applied to the odds ratio of a binary event. This approach is detailed by Peng, Kuk Lida Lee, and Ingersoll in 2002. Translating this into practical terms, the LOGIT function provides the basis for creating a 2x2 contingency table, offering insights into the interplay between variables.Succinctly put, the core formula underpinning simple Logistic Regression is shown .

$$(Y) = \log\left(\frac{Y}{1-Y}\right) = \alpha + \beta X$$

Commonly, Logistic Regression finds its strength in the exploration and examination of correlations between a categorical dependent (label) variable and one or more categorical or continuous independent variables (features). This serves as a powerful tool for not only describing the relationships but also for hypothesis testing within such contexts. Random Forest The Random Forest algorithm, conceptualized by Leo Breiman, stands as a significant achievement in the realm of Machine Learning. It operates by forming an assembly of un-pruned classification or regression trees, which are meticulously drawn from the training dataset (Ali et al., 2012). An intriguing facet of this algorithm is the stochastic selection of features during the induction process.

The predictive outcome stems from an amalgamation of ensemble forecasts. For classification tasks, this translates into a majority vote consensus, while regression problems involve an averaging of predictions. In

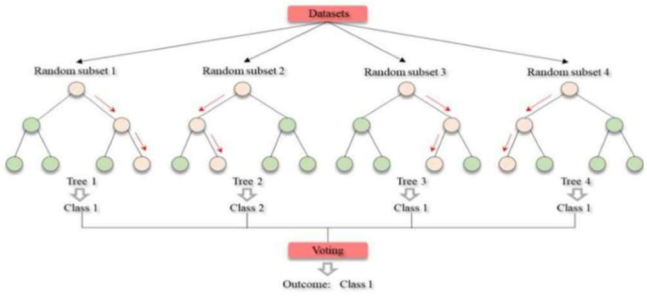

Figure 4: Random Forest

comparison to standalone decision tree classifiers, Random Forest consistently exhibits heightened performance (Ali et al., 2012). Its essence lies in the concept of ensemble learning, where multiple Decision

Tree classifiers are synthesized while the model is being trained. The ultimate output materializes as the mode of classes in classification tasks or the mean of classes in regression scenarios. Notably, the Random Forest technique finds favor due to its inherent simplicity, adaptability, and diverse applicability.

Research underscores the Random Forest classifier as an exemplar among classifiers, largely surpassing its counterparts in terms of performance when applied to forecasting the life cycle of YouTube's trending videos. kNN is also employed in this analysis.

4 RESULT

We will now discuss the various visualizations that we made while implementing our code and what all the interpretation and analysis we had from them.

We tried to acknowledge the relationship between the different features whether they have positive or negative correlation and how different features contribute to the trendiness of a video.

Above Figure represents presents a correlation heat map that visually represents the interrelationships among four key metrics: views, likes, dislikes, and comment_count. Each cell in the heatmap contains a correlation coefficient value, with color intensity indicating the strength and

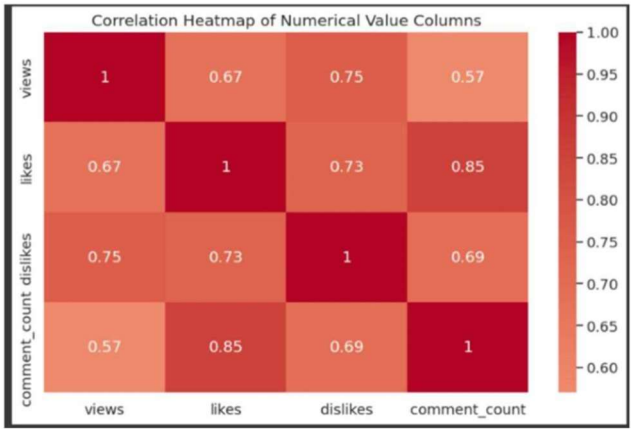

Figure 5: Correlation heat map illustrating the relationships between views, likes, dislikes and comment_count. direction of the correlation.

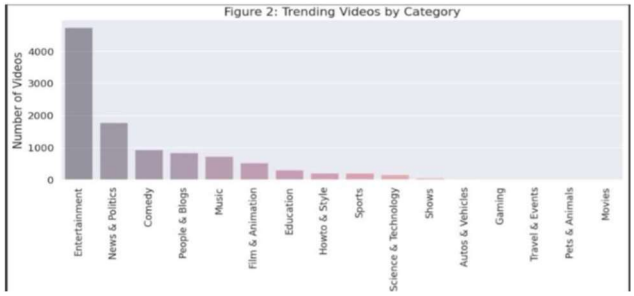

Figure6: Graphical representation generated by the program showing variation in Number of Views with different category .

Fig:6 depicts a graphical representation that explores the relationship between the number of views and different video categories on our platform. The horizontal axis of the graph represents the categories, while the vertical axis represents the number of views. Each category is colorcoded for clarity.

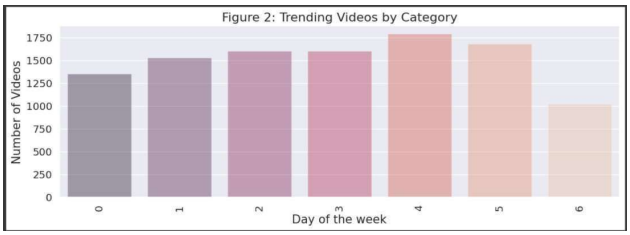

Figure7:

Graphical representation generated by the program showing variation in Number of Views with Day of the Week.

We tried to explore the impact of day of publish on the number of views as seen in fig:7 and found most of the videos which got trending are published on Friday and Saturday.

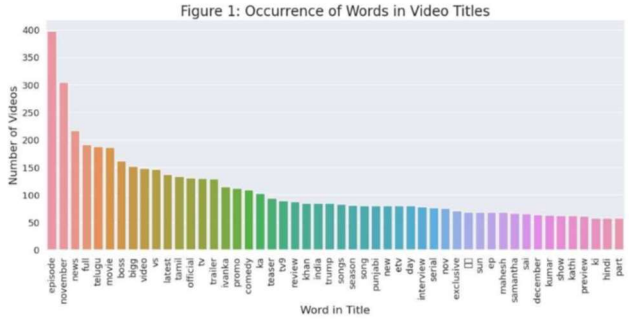

Figure 8: Words that were used in most of the videos

Also per the different visualizations made, it was found that videos that got trending had made their rating and comment enabled so as to get proper feedback from the user and improve the quality of the upcoming videos.

4.1 Model Building

1. **Prediction of number of views:**

Lets look forward to accuracies we got when different alogrithms were applied on the dataset. First we used Decision tree and Linear regression for the prediction of the number of views.

```
Decision model score for train data: 0.818645861740412
```

Figure 9: Output showing the calculated Decision Model Score for train data.

We got the accuracy of about 71% in decision tree model and 68% in Linear regression. for the train data. Thus, decision tree as well as Linear Regression both proved to be good for the prediction of number of views.

2.**Prediction of time that will be taken by a video to get on the trending list:**

In this section, we explore the fascinating aspect of predicting the time it takes for a video to achieve trending status on the platform. For this first a target column ‘‘numdays″was made and then again dataset is splitted into train and test for the prediction. This time we used Logistic Regression, kNN, and Random Forest Regressor for the prediction.

```
Model score of train data for Random Forest Regressor 0.6153087903656411
```

Figure 10: Output showing the Random forest Score for train data.

We got an accuracy of about 61% in Random Forest Regressor, 37% in Logistic Regression, and 46% in kNN for the train data. Thus, Random Forest Regressor proved to be good for the prediction of time that will be taken by a video to get on the trending list.

```
Classifier score for train data: 0.3717791411042945
```

Figure 11: Output showing the Logistic Regression Score for train data.

3.**Prediction of the number of days a video will remain on trending list :**

In this section, we embark on the task of predicting the duration for which a video is likely to maintain its position on the trending list. For this first a target column ‘‘trend″ was made and then again dataset is splitted into train and test for the prediction. This time we used

Logistic Regression and kNN for the prediction.

```
Model score of train data for Logistic Regression: 0.942643391521197
```

Figure 12: Output showing the Logistic Regression Score for train data.

```
K Model score for train data: 0.9488778054862843
```

Figure 13: Output showing the K Model Score for train data.

We got the accuracy of about 94% in Logistic Regression and 95% in kNN for the train data. Thus, Logistic Regression as well as kNN both proved to be good for the prediction of the number of days a video will remain on trending list.

5 DISCUSSION

In this study, we set out to understand the factors influencing YouTube video popularity and provide actionable insights for content creators.

We conducted a thorough feature importance analysis to identify the most influential factors in predicting video popularity. The results underscored the significance of specific features:

Likes and Comments: Our analysis consistently showed that the number of likes and comments strongly correlates with video popularity. Videos that received higher likes and comments tended to have a greater chance of trending. This suggests that audience engagement plays a pivotal role in a video’s success.

Views: While views are a crucial metric, we observed that their impact on video popularity is closely tied to other engagement metrics. High views alone do not guarantee trending status; rather, they are often accompanied by substantial likes and comments.

Dislikes: Surprisingly, the number of dislikes also exhibited a correlation with video popularity, albeit weaker than likes and comments. Videos with a higher number of dislikes were less likely to trend, suggesting that viewer sentiment plays a role in a video’s performance.

A graph is plotted between density and dislikes(Figure 5). The trend line or pattern observed in the graph provides

insights into the relationship between video Density and user dislikes. Examining the direction and steepness of the trend line helps us understand whether there is a positive or negative association between these variables.

A graph is plotted between density and log_dislikes(Figure 6). The trend line or pattern observed in the graph provides insights into the relationship between video Density and user log_dislikes. The use of log_dislikes on the xaxis is important because it transforms the scale of the dislikes variable into a logarithmic scale. This transformation is valuable when dealing with data that spans a wide range of values, as it allows us to visualize trends more effectively.

A graph is plotted between density and views, likes, and comment_count(Figure 7). The trend line or pattern observed in the graph provides insights into the relationship between video Density and user views, likes, and comment_count respectively. Examining the direction and steepness of the trend line helps us understand whether there is a positive or negative association between these variables.

A correlation heatmap(Figure 8) that visually represents the interrelationships among four key metrics: views, likes, dislikes, and comment_count is plotted. Each cell in the heatmap contains a correlation coefficient value, with color intensity indicating the strength and direction of the correlation.The heatmap is a valuable tool for uncovering patterns and associations within our dataset of YouTube videos. The color gradients in the heatmap allow for easy interpretation: warmer colors represent stronger positive correlations, cooler colors represent weaker or negative correlations, and shades of gray indicate no significant correlation. Overall, this correlation heatmap provides a visual summary of the relationships between key engagement metrics for YouTube videos.

A graphical representation that explores the relationship between the number of views and different video categories(Figure 9) on our platform is plotted. In this visualization, we aim to gain insights into how video content is consumed across various categories, shedding light on viewer preferences and content trends. The graph provides a valuable overview of the relationship between video category and the number of views, offering insights into category popularity and viewer engagement.
This information is pivotal for content creators, platform administrators, and researchers aiming to optimize content strategy and enhance the user experience.

5.1 Accuracy

Accuracy is a measure of the model's ability to correctly classify instances. It is calculated as the ratio of correctly predicted instances to the total number of instances in the dataset, expressed as a percentage.

An accuracy score of 81.86% and 71.36% is significantly a good score which is achieved by applying Decision Model on training and test dataset.This score is a crucial element in our machine learning model, representing the model's predictions or decisions based on the features and patterns it has learned from the training data. In the case of the Logistic regression Model on the training dataset accuracy achieved is 94.26% and on the testing dataset it is 94.72% which is really a good accuracy. In the case of the KNN Model on the training dataset accuracy achieved is 94.88% and on the testing

dataset it is

93.53% which is really a good accuracy.

5.2 SCOPES AND LIMITATIONS

5.2.1 Scope

Exploring Real-time Data Streams: The scope exists to enhance the model's predictive capabilities by incorporating real-time data streams from various social media platforms and content sharing sites beyond YouTube.

Integrating data in real time can provide more upto-date insights into video trends.

Multimodal Data Fusion: There's potential to expand the scope by integrating multimodal data, combining textual features from titles and descriptions with visual cues from video thumbnails. This holistic approach might improve the accuracy of trend prediction.

Influence of Social Factors: The model can be extended to account for external social factors like major events, holidays, or viral trends that might influence the chances of a video going trending. Incorporating such factors could provide a more comprehensive prediction.

User Engagement Metrics:

Expanding the scope to consider user engagement metrics beyond views, likes, and dislikes, such as comments, shares, and user interactions, might offer a richer understanding of a video's potential to trend..

While many forecasting endeavors entail the analysis of supervised or unsupervised data, this study ventures to predict a YouTube video's trending lifecycle with a 62% accuracy rate through the conversion of unsupervised data into supervised data. This avenue presents substantial room for refinement, attainable by leveraging more intricate and hybrid Machine Learning algorithms on meticulously curated supervised data tailored to the study's research question. Additionally, a future trajectory might involve investigating the impact of non-numerical attributes like titles and descriptions when combined with the current model.

5.3 Limitations

An overarching limitation in this research stems from the initial unsupervised dataset. The constraints of the dataset, including limited samples and features, hindered the implementation of unsupervised Machine Learning algorithms like Deep Learning.

A substantial sample size plays a pivotal role in achieving more accurate prediction results. This limitation necessitates a cautious interpretation of the outcomes and points to the potential for enhanced predictive power with a more expansive dataset.

The model's performance could be affected by changes in YouTube's algorithms, policies, or features. These changes might lead to shifts in how videos trend, impacting the accuracy of the model.

6 CONCLUSION

The concept of popularity often veers into the realm of unpredictability. This thesis delved into the intricate sphere of forecasting the life cycle of trending YouTube videos. The aim was to unravel the capricious nature of video virality, offering insights both from an academic and business standpoint.

To achieve this, the thesis harnessed five distinct classification models— Random Forest, Decision Tree, Logistic Regression,kNN. The primary focus was to address the research query: 'Exploring Machine Learning Algorithm Performance in Predicting the Life Cycle of YouTube Trending Videos Through Analysis of Statistical Data.' Comparative evaluations unveiled that Logistic Regression, Random Forest classifiers demonstrated superior efficacy in predicting the trajectory of YouTube's trending videos. Validation through K-fold Cross Validation substantiated these findings, bolstering the research's credibility. Furthermore, the study identified the potential for refining classification outcomes by applying the current methodology to a more comprehensive and meticulously curated supervised dataset.

Expanding the scope of investigation, the research introduced a straightforward Logarithmic Linear Regression algorithm to prognosticate the number of views for a trending YouTube video. The outcomes pointed to a moderately accurate performance of the basic Linear Regression model, achieving 68% accuracy, despite its reliance on a limited set of features. This discovery unveils an avenue for enhancing predictive precision by integrating hybrid S-H and MRBF models into the current framework. This enhancement can be facilitated through a more expansive and enriched supervised dataset, incorporating both numerical and non-numerical attributes.

By successfully applying Machine Learning algorithms to an unsupervised dataset, this study not only unravelled key insights but also facilitated the comparison and selection of optimal classification models. These insights hold the potential to decipher the intricate life cycle patterns of YouTube's trending videos.

REFERENCES

https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees

https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2

https://www.researchgate.net/publication/221140967_The_YouTube_video_recommendation_system https://www.minuscin.top/products.aspx?cname=random+forest+algorithm+in+r&cid=95