# Predictive analysis of YouTube trending videos using Machine Learning

Dissertation submitted in partial fulfilment of the requirements for the degree of

MSc Data Analytics

At Dublin Business School

Aakash Ashok Niture

Supervisor: Mr. Pierpaolo Dondio

MSc Data Analytics 2020 - 21

# DECLARATION

‘I declare that this applied project that I have submitted to Dublin Business School for the award of MSc Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.’

Signed: Aakash Ashok Niture

Student Number: 10525178

Date: 11/01/2021

# ACKNOWLEDGEMENTS

I would like to express my appreciation of gratitude towards Mr. Pierpaolo Dondio, my research supervisor for his valuable guidance from the beginning of the thesis with constructive and important suggestions provided throughout the designing, implementation, and betterment of dissertation. He also helped me with the thesis research, and his readiness for having meetings about discussing the progress is very much appreciated. I would also like to express my humble gratitude towards Dublin Business School for providing me with the platform to enhance my educational abilities.

Secondly, I would like to thank my family and friends for their support and encouragement throughout my studies.

# ABSTRACT

YouTube is a world-famous video sharing interactive platform which allows its users to rate, share, save, comment, and upload the content. Unlike popular videos which get number of likes and views by the time they are stated as popular, YouTube trending videos represents the content which is gaining viewership over a certain time period and has a potential to be popular. Despite their importance YouTube trending video's analysis have not been a well-researched area yet. This research proposes to analyse interactive features to determine correlation and importance of variables for the trendiness of a video. Study focuses on how interactive video features helps a video trend on YouTube. Research is based on YouTube trending video's viewership statistics of more than 40000 videos over a certain time period. Since trending video statistics consists of number of Views, Likes, Dislikes and Comment counts, the research performed Linear regression model of Machine Learning for predictive analysis of number of views for YouTube trending videos. In addition, the study performs a comparative analysis of a number of classification models namely Random Forest, SVM, Decision Tree, Logistic Regression and Gaussian Naïve Bayes, to determine which model suits better for predicting the number of days a video will take to get trending from its upload time and the number of days a video will trend on the trending list. Research achieved maximum accuracy of 62.53% for predicting YouTube's trending video's lifecycle. Cross Validation method have been used for statistical significance testing and the performance evaluation matrix has compared and determined the most useful classifiers. Furthermore, this research follows CRISP DM methodology design with correlational quantitative research method. Study will bring objectivity towards the popularity constraint of YouTube trending videos.

## Table of Contents

## Table of Figures and Tables

# CHAPTER 1 – INTRODUCTION

## 1.1 Background

YouTube is the largest online video sharing platform in the world. Founded in 2005 YouTube now has become a cosmos in its own. With over 4 billion views per day YouTube is the most suited platform for user generated content (Iman Barjasteh et al., 2014). YouTube offers interactive video features for public and content creators such as Views, which denotes the total number of viewership gathered by the particular video till date. Generally, the number of views determines the popularity of videos and it takes a certain amount of time for a video to become popular. There is always some content which catches attention of masses in a short period of time and such content falls under the YouTube's trending videos feature. Although trending videos cannot be called popular when featured under trending tab in YouTube but have the potential to be popular in the future. Despite their importance YouTube trending videos have not been a well-analysed area yet. Considering that over a billion unique users visit YouTube in a month and 72 hours of video is uploaded over YouTube per minute it has become one of the largest platforms for business (Iman Barjasteh et al., 2014). YouTube opened these new revenue streams in 2013 where brand management, advertising, promotion like businesses have been introduced. Whenever a video gains popularity, it is made available to the large number of viewers for free and it gains masses attention for a while. It is hard to keep track of which content might get trending in near future or might become popular, hence predictive analysis using Machine Learning is introduced.

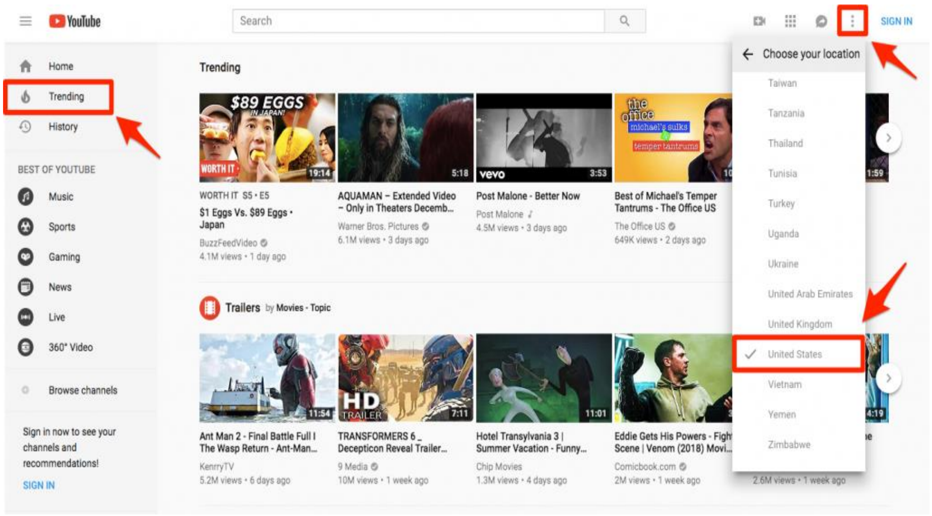


*Figure 1.1 YouTube Trending page*

Past studies focused on early observations of a video statistics to predict the popularity of the video. Few explored the random virality of video resulting in burst of views for the video. Current study follows CRISP DM methodology and it implements Machine Learning algorithms Random Forest, SVM (Support Vector Machine), Decision Tree, Logistic Regression and Gaussian Naïve Bayes classifier for classification method and Linear

Regression for regression approach in predictive analysis of YouTube trending video's life cycle.

This research helps YouTube to analyse and predict the lifecycle of a trending content so that it can build its business model around that content accordingly. Content creators or as they are called YouTubers also generate revenue from their videos. YouTube is the sole source of income for a lot of YouTubers and this study will help creators to analyse their contents life cycle and make improvements in required areas. Such as feedback from viewers is a very important aspect for YouTubers as they can understand that how their content is being received by people, and this study helps YouTube and YouTubers understand how the interactive features affect their videos performance on the social platform.

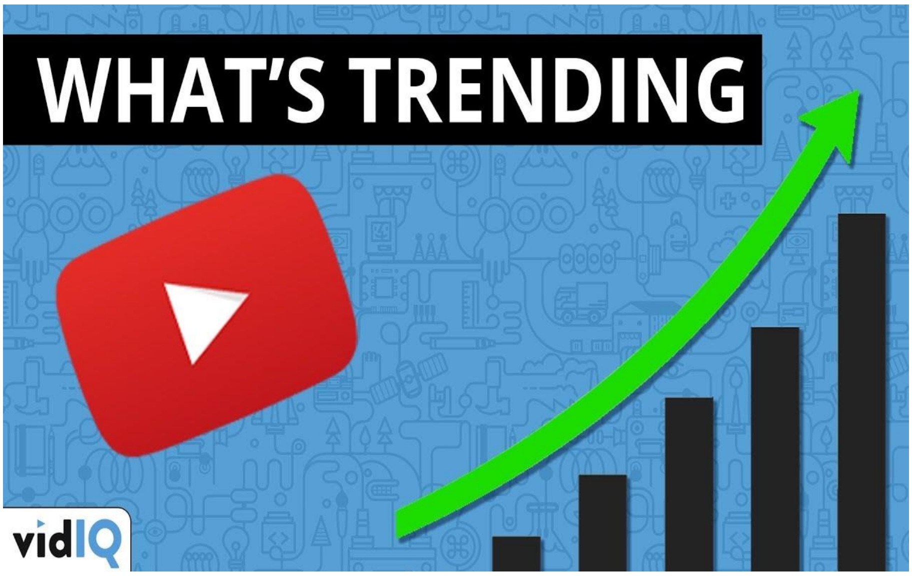


*Figure 1.2 YouTube trending*

## 1.2 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. The key factors of this study are as follows:

### 1.2.1 correlation between features

Basic statistical analysis performed over the dataset. Dataset is obtained from Kaggle (J, 2017). Dataset contains unsupervised data hence, to turn the dataset into supervised data two target variables (Labels) are created for respective predictive analysis problems. Dataset contains regular YouTube features such as number of Views, Likes, Dislikes, Comments count, Comments_disabled, Ratings_disabled, Category Id, Trending date e.g. Initial observations will be to determine, how interactive video features helps a video trend on YouTube? is having a large number of views required for a video to trend? how important is the Correlation between features?

### 1.2.2 Comparative analysis of ML classifiers for predicting YouTube trending video's lifecycle

Thesis implemented comparative analysis of Machine Learning algorithms specifically Random Forest, SVM (Support Vector Machine), Decision Tree, Logistic Regression and Gaussian Naïve Bayes classifier to determine which classifier is best suited for forecasting. YouTube trending video's life cycle is analysed into two parts –

- *Predicting the number of days video will take to get on trending list*

The first part of a trending video's life cycle is to rise above competition hence, Classifiers analyze and predict the number of days it will take for a video to get on the trending page on YouTube since its upload on the social platform.

- *Predicting the number of days video will trend on trending list*

The second part of life cycle starts after the video gets on the trending list on YouTube. Classifiers analyze and predict the number of days a video will trend on the trending page on YouTube since its first trending day on the social platform.

This research will help YouTube into key factors such as marketing, advertisements, and brand management.

### 1.2.3 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.

## 1.3 RESEARCH OBJECTIVES

*Aim –*

Assist YouTube and YouTubers with decisions relating to YouTube video's lifecycle forecasting using Machine Learning.

Objectives –

To analyse and compare different Machine Learning classification algorithms trained on YouTube's trending videos statistics data to predict the estimated lifecycle of a video. The objectives are:

- Model identification – best model selection through comparative study
- Model fitting – fitting all the models on trained data
- Model evaluation – evaluating results with significance testing

## 1.4 RESEARCH METHODOLOGIES

Thesis followed the traditional CRISP DM methodology as it is a very simple and go to process for Machine Learning and Data Analysis projects since decades. CRISP DM methodology is very efficient as it covers the business aspect of a project as well as its technical side. Hence, CRISP DM holds a dominant share in methodology domain even after decades. CRISP DM consists of 6 steps Business understanding, Data understanding, and Data preparation, Modeling, Evaluation and Deployment process for identifying, selecting, and assessing the most suitable model for projects business effectiveness. For this dissertation proposed steps are:

- Step 1: 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: Data Evaluation, Evaluate findings, Compare models
- Step 6: Deployment

Thesis have a business outlook as it helps the social media giants YouTube to understand and enhance the behaviour of trending video's life cycle and the importance of features in video's success, also study helps YouTubers to garner their videos accordingly. Hence, CRISP DM methodology is preferred because of its emphasis on business aspect of a project.

## 1.5 DISSERTATION OUTLINE

Research covered prediction models; the algorithms used to fit the model and optimization technique for obtaining better performance and accuracy. The performance evaluation matrix analysed and selected the better suited models. Various feature selection parameters help determining the best attributes that are positively correlated with the label for model fitting.

The proposed study followed CRISP DM methodology and the report roadmap is as follows,

- *Introduction* – This chapter includes background, research problem, aim and objectives and hypothesis to be tested.
- *Literature Survey* – This chapter includes existing study of YouTube's trending videos using journals and books which includes concepts and models for prediction
- *Methodology* – This chapter follows CRISP DM methodology to implement and conduct the thesis. The methodology is tailored according to research.
- *Implementation* – This chapter conducts the data analysis for research.
- *Results* – This chapter includes discussion and evaluation of models fitted in the research.
- *Scope and Limitations* – This chapter covers the future scope and limitations for the study.
- *Conclusion* – This chapter summarizes the findings of the research and comes to conclusion.

# CHAPTER 2 - LITERATURE REVIEW

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).

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).

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.

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 algorithms have the better performances. Study is limited to textual data due to social platforms limitation (Tejal Rathod and Mehul Barot, 2018).

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

# CHAPTER 3 – METHODOLOGY

The proposed research follows correlational quantitative research method and integrates the CRISP DM methodology as Model consists of 6 steps namely business understanding, data understanding, data preparation, model, evaluation, and deployment which covers the defecto process of a general Machine Learning project. Determining business objectives, project plan, data collection, data cleaning, model selection and building, result evaluation and deployment of project with review consideration all such necessary steps are covered by CRISP DM model hence making it the first choice of data domain projects (Shearer et al., 2000). There are different methodologies such as KDD, SEMMA, Two Crowds and many more but CRISP DM is more complete and other methodologies are inspired from KDD and CRISP DM (Umair Shafique and Haseeb Qaiser,2014)

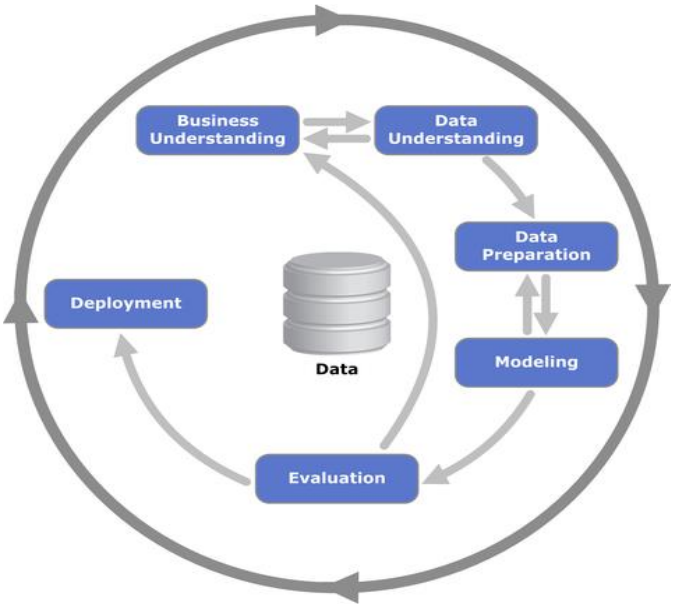


*Figure 3.1 CRISP DM MODEL (Shearer et al., 2000)*

## 3.1 Business understanding

YouTube is the booming platform for business such as advertising, brand management e.g. and it is hard to predict the life cycle of a trending video. This research can be commercialized for Google, Yahoo, and various other advertisement services, also for the YouTube and content creators. Project helps these three parties by helping them organise and regulate a stable income with calculated business decisions. Hence, there a lot of business opportunities and scope for the study.

YouTube's recommender system can also benefit from the study as recommendation system needs to distribute and recommend the trending videos to users. Study focuses on implementing and determining the best performing model for forecasting a video's life cycle and give better solution and a path for such commercial purpose.

## 3.2 data understanding

Dataset is obtained from Kaggle (J, 2017). Dataset contains regular YouTube features such as number of Views, Likes, Dislikes, Comments_count, Comments_disabled, Ratings_disabled, Category_Id, Trending_date, publish_date e.g. Study analysed the discrepancies in dataset, detected the null values. Also, research studied the distribution and correlation of features such as Views, Likes, Dislikes, Comments_count etc. using graphs and plots in python library to understand the data and its relationship with label.

Initial observations determined; the role of interactive video features in helping a video trend on YouTube, such as impact of having a large number of views for a video to trend.

## 3.3 data preparation

Dataset is in Unsupervised data format as it does not contain the dependent variable (label). Hence, research converted unsupervised dataset into supervised dataset. Study created two dependent variables (labels) for two classification problems respectively from the existing Unsupervised dataset transforming it into a Supervised dataset. Thesis detected and eliminated the null values from the data; also dropped the unnecessary columns from the dataset. Study encoded and normalised the categorical and numerical features for the ease of data fitting in model.

Feature selection methods such as RFE (Recursive Feature Elimination) and Chi square were used to select the most relevant attributes according to label for modelling in research. Study also implemented Under sampling method of data balancing to balance the number of classes in data. Trending videos can trend for a number of days on YouTube, hence only the first instance of a video on trending list is extracted for fitting in the model to reduce the bias in model.

Data is cleaned and pre-processed and visualized using python libraries and environment.

## 3.4 Modeling

Thesis Implemented six Machine Learning algorithms on the dataset to get the desired results. Cross Validation method is used to assess the Significance testing for comparison of classifier's results. Algorithms used in the research are:

### ALGORITHMS

- *Linear Regression*

Linear Regression is the most common algorithm in Machine Learning used for determining the relationship between variables and label. The most basic regression correlation is Linear one. In this case,

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

A line with intercept $\beta_0$ and slope $\beta_1$.

One can interpret this simple equation as *Y* has a relative distribution with mean $\mu(X)$. Here Y is the predicted variable and X is the predictor (Altman and Krzywinski, 2015). In Linear Regression value of Y increases with value of X respectively. Hence, in Machine Learning it is used to predict the dependent variable (label) from independent correlated variables (features) having normal distribution.

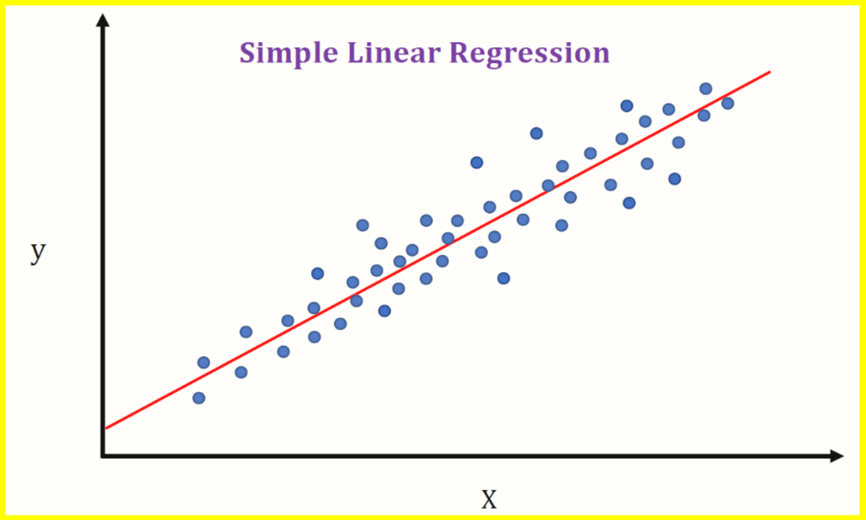


*Figure 3.2 Simple Linear Regression*

Study performed exploratory data analysis which confirmed the relationship between dependent variable (views) and independent variables (likes, dislikes, comment counts etc.) have a linear regression having a normal distribution. Hence, Linear Regression is performed to predict the number of views for trending videos.

- *Decision Tree*

Decision Tree algorithm is used for classification of labelled variable of a supervised dataset in Machine Learning.

Classifier is based on a simple tree structure where there is one root node which has the highest relevant attribute and branches denotes the supposed outcomes and leaf node represents further decision or variable. Its composition resembles that of a tree with categorical or numerical leaf nodes as a graph.

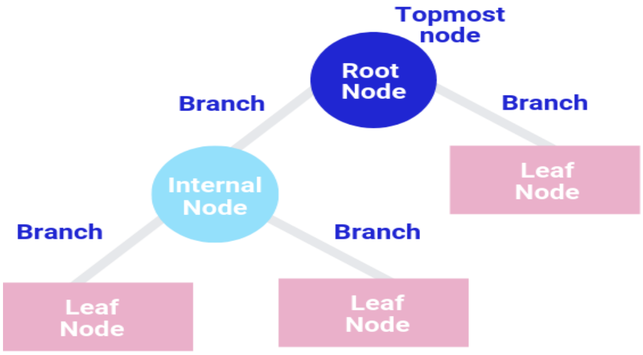


*Figure 3.3 Decision Tree*

Decision Tree consists of assessors namely Entropy, Gini Index, Gain ratio and Information Gain. The information gain is the amount of information derived about a variable from observing another independent variable. Entropy measures the uncertainty of an independent random variable.

Entropy – Entropy is regarded as a measure of uncertainty and randomness of output in machine learning models. If any action has no control over its outcome, such as coin toss or ball color in a ball, the randomness of the outcome is then estimated by the entropy of the model.

Information Gain – The information gain is the amount of information derived about a variable from observing another independent variable. Information Gain is an impurity-based criterion which uses the Entropy as an impurity-based measure (Lior Rokach and Oded Maimon, 2005). Information Gain reduces model's uncertainty level by decreasing the entropy levels. Higher the information gain value is lower the entropy resulting into better accuracy for the models. In decision tree, root node is the most relevant feature and provides information gain of next decision. Here is the equation for Information Gain,

$$InformationGain(a_i, S) = \\ Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{\left|\sigma_{a_i = v_{i,j}} S\right|}{|S|} \cdot Entropy(y, \sigma_{a_i = v_{i,j}} S)$$

where:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y = c_j} S\right|}{|S|} \cdot \log_2 \frac{\left|\sigma_{y = c_j} S\right|}{|S|}$$

*Figure 3.4 Information Gain and Entropy equation (Lior Rokach and Oded Maimon, 2005).*

Study Implemented Decision Tree classification model after transforming unsupervised dataset into supervised labelled output dataset.

- *Random Forest*

Leo Breiman developed the Machine Learning algorithm of Random Forest. It is a group of un-pruned classification or regression trees randomly selected from training set of the data (Ali et al., 2012). Features are selected on random basis in induction process. Prediction score is an aggregate of ensemble predictions such as majority votes in classification problems and averaging in regression problems.

Random Forest generally provides a significant performance than a single tree classifier (Ali et al., 2012). Random Forest algorithm is simply put an ensemble method which constructs multiple Decision Tree classifiers while training data in model and gives output as the mode of classes while classification or mean of classes for regression method. Random Forest is one of the most used algorithms as its simple, flexible, and diverse in nature.

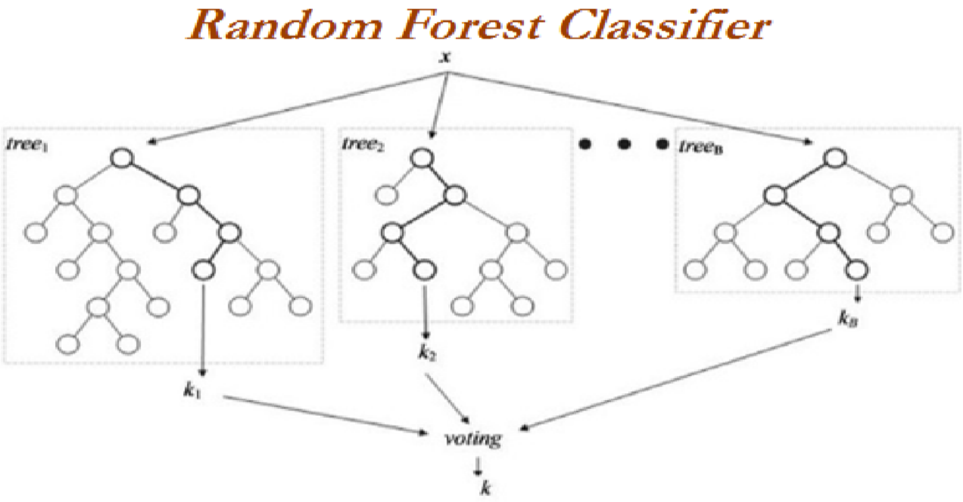


*Figure 3.5 Random Forest*

Research denotes Random Forest classifier as one of the finest classifiers as it edges other classifiers in performance for forecasting YouTube's trending video's life cycle.

- *SVM (Support Vector Machine)*

SVM (Support Vector Machine) algorithm of Machine Learning is one of the most reliable and popular algorithms as it works for both regression as well as classification methods. Support Vector Machine considers extreme features of the dataset and plot a decision boundary known as hyperplane. Support is supposed to be best at segregating two classes. Every data point in the dataset is a support vector in SVM.

An optimal decision boundary is established with help of the nearest positive and negative distance of support vectors D+ and D- respectively. This hyperplane provides a margin of distance (D+) + (D-) for better classification of dataset. Most basic form of SVM is a linear line separating the classes hence, it is called Linear Support Vector Machine (LSVM) (Theodoros Evgeniou and Massimiliano Pontil, 2001).

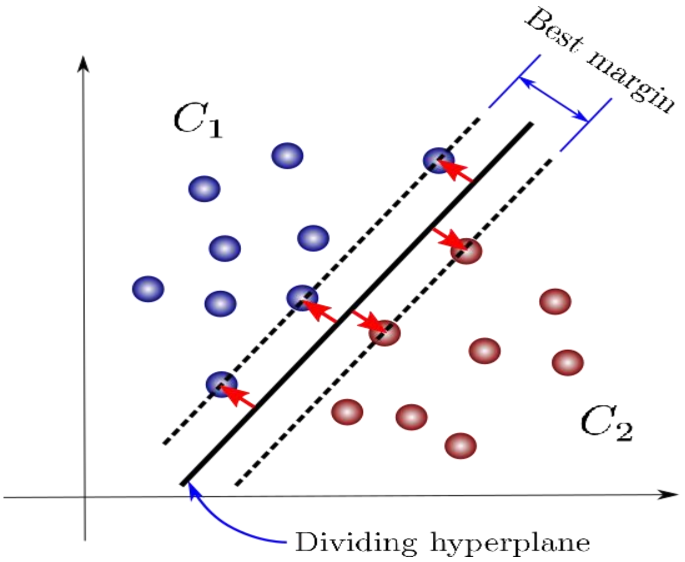


*Figure 3.6 Random Forest*

When the data is nonlinear Support Vector Machine works in high dimensional space efficiently with help of its kernels hence, it is a very versatile algorithm. For linear (LSVM) approach the formula for SVM is,

$$f(x) = w \cdot x + b$$

When the dataset is nonlinear and SVM uses its kernels to plot hyperplanes in the feature space induced by kernel K, the formula includes kernel domain K,

$$\left\{ f : \|f\|_K^2 < \infty \right\}$$

Advantage of Support Vector Machine are, SVM is effective in high dimensions and it is memory efficient. Support Vector Machine works better even when number of dimensions are greater than number of samples hence, R^N > Samples. Thesis have used Support Vector Machine as it is one of the best classifiers for binary classification.

- *Gaussian Naïve Bayes*

Naïve Bayes algorithm of Machine Learning is a classification method which follows Bayes theorem with strong naïve assumptions that are independent. Naïve Bayes assumes features are independent of given classes. It is called Naïve Bayes as algorithm makes some Naïve assumptions for classifying dataset. Naïve Bayes is used extensively as it has minimal storage and fast training properties (Pouria Kaviani and Sunita Dhotre, 2017).

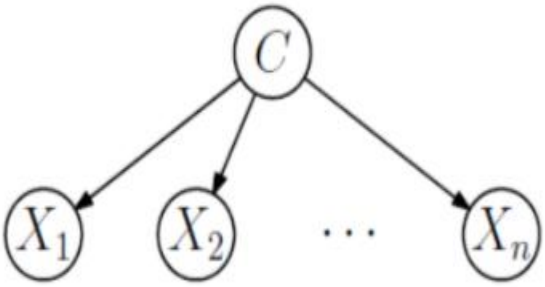


*Figure 3.7 Naïve Bayes Classifier (Pouria Kaviani and Sunita Dhotre, 2017).*

Gaussian Naïve Bayes is an extension of Naïve Bayes algorithm. Gaussian Naïve Bayes is considered when the dataset has a normal distribution. Gaussian Naïve Bayes is the simplest Naïve Bayes to implement as it estimates only the mean and standard deviation from training data. This algorithm has a probabilistic approach as it considers prior and posterior probability calculations of the given classes in training and test data set, respectively. It induces the distribution using,

$$Pr(c, x_1, \ldots, x_m) = Pr(c) \cdot \prod_{i=1}^{m} Pr(x_i \mid c)$$

*Figure 3.8 Naïve Bayes equation (Pouria Kaviani and Sunita Dhotre, 2017).*

Thesis have implemented Gaussian Naïve Bayes classifier for forecasting the life cycle of trending videos. Study found that it is the least performing classifier amongst others.

- *Logistic Regression*

In Machine Learning when the dataset has dependent variable or Label in numeric continuous distribution Linear Regression is used to prediction, Logistic Regression is used to classify Dependent Variable (Label) when the data is in the form of binary classes 1 and 0. Logistic Regression fits a S shape logistic function to the data points unlike a linear line in Linear Regression.

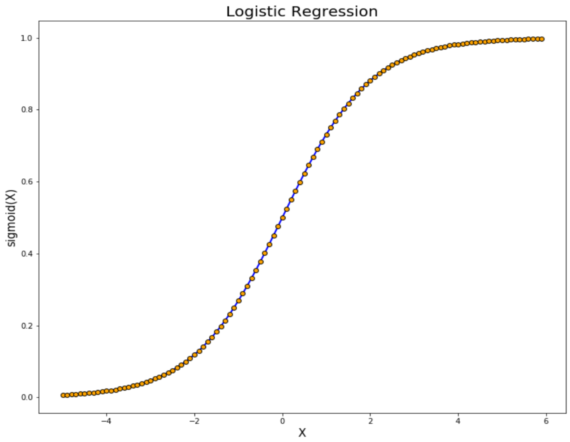


*Figure 3.9 Logistic Regression*

A simple Logistic Regression is used when there is a binary prediction variable in dataset as the study has. For more than two classes Multinomial Logistic Regression is used. Logistic Regression follows a simple LOGIT mathematical concept to plot the S shaped logistic function. LOGIT is a natural logarithmic function of an odds (binary) ration (Peng, Kuk Lida Lee and Ingersoll, 2002). A 2*2 contingency table is derived from the LOGIT function. The simple Logistic Regression form is,

$$\mathrm{logit}(Y) = \text{natural}\log(\text{odds}) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X.$$

*Figure 3.10 Logistic Regression formula (Peng, Kuk Lida Lee and Ingersoll, 2002).*

Usually, logistic regression is considered best for describing and testing hypotheses about correlation between a categorical dependent (label) variable and one or more categorical or continuous independent variables (features).

## 3.5 Evaluation

Evaluation phase in CRISP DM methodology evaluates the findings of models fitted on the dataset. Thesis have used performance evaluation matrix instead of t test for evaluating the various model’s performance as the method is more descriptive.

- *Cross Validation*

Thesis implemented K-fold Cross Validation method to compare the results obtained from classifiers. Cross Validation method provides readers with an explanatory performance evaluation matrix for classifiers results hence, making it easy to compare and determine the best performing classifier model on the dataset. Cross Validation method uses subsets of dataset such as no two subsets overlap with each other. K-fold determines the number of subsets to be made in the dataset.

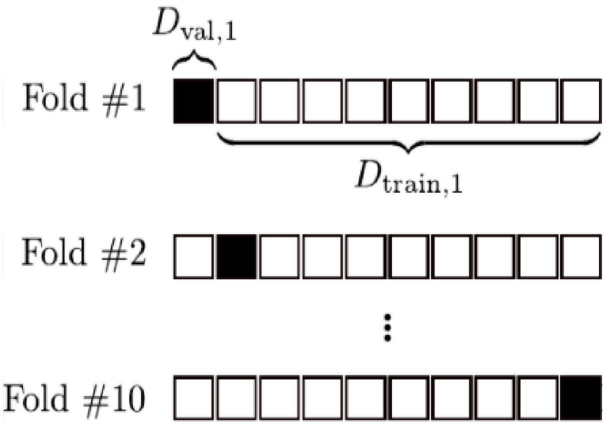
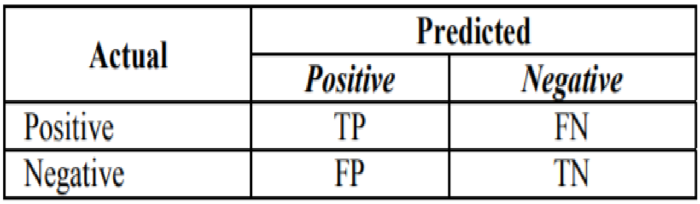


*Figure 3.11 10-fold Cross Validation (Berrar, 2018).*

Research preferred Cross Validation method instead of a simple t-test for significance testing as Cross Validation provides with multiple performance majors such as Accuracy, Precision, Recall, AUC and F1 score to assess and compare multiple classification models used in the process. Cross Validation method derives a confusion matrix from the following variables,

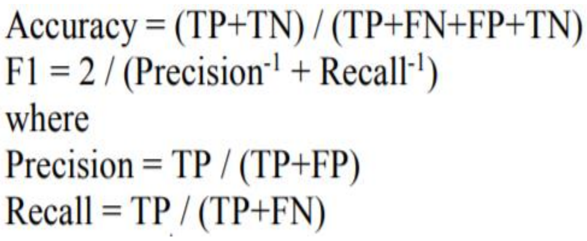
- True Positive: outcome correctly predicted as positive class
- True Negative: outcome correctly predicted as Negative class
- False Positive: outcome incorrectly predicted as positive class
- False Positive: outcome incorrectly predicted as positive class



| Actual | Predicted | |
|---|---|---|
| | *Positive* | *Negative* |
| Positive | TP | FN |
| Negative | FP | TN |

*Figure 3.12 Confusion Matrix for binary classification (Wardhani et al., 2019).*

Performance measures such as Accuracy, Precision, Recall, AUC and F1 score are derived from this confusion matrix,



$$\text{Accuracy} = (TP+TN) / (TP+FN+FP+TN)$$
$$F1 = 2 / (\text{Precision}^{-1} + \text{Recall}^{-1})$$
where
$$\text{Precision} = TP / (TP+FP)$$
$$\text{Recall} = TP / (TP+FN)$$

*Figure 3.13 Definitions for performance measures (Wardhani et al., 2019).*

Cross Validation method gives performance evaluation matrix which is simple to understand and explain and efficient to compare and select the most efficient model.

## 3.6 Deployment

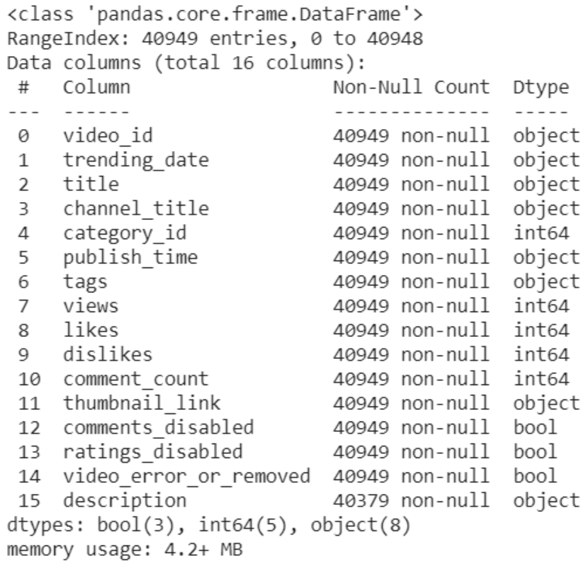
Deployment is where the projects findings pays off. In this last and crucial phase of CRISP DM methodology, it does not matter how accurate models' results are or how perfect the model fits the data if the model is not integrated to improve the business objectives.

In deployment phase research gives scope to build an interface or an application with the best performing model which will predict the life cycle of a YouTube video for users. An application where user can input the current status of video's features such as number of views, likes, dislikes etc. and with one click the bot will predict the video's performance analysis for next few days. YouTube and YouTubers can benefit from such an interface as it provides them answers with a forecasting outlook to understand when and what are the chances of a video to get on the trending list, and For how long can a video trend on YouTube.

# CHAPTER 4 – IMPLEMENTATION

Exploratory Data Analysis – EDA (Exploratory Data Analysis) is a method of exploring and understanding the different distribution and relationships between variables. Initial unsupervised dataset contains features that are views, likes, dislikes, comment_count, category_id, ratings_disabled, comments_disabled etc. Study has explored the data distribution and nature using visualization libraries in python.

Thesis analysed the entire unsupervised dataset using info() function. Function denotes the number of values per column, data type of a column and if the column contains any null values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40949 entries, 0 to 40948
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   video_id                40949 non-null  object
 1   trending_date           40949 non-null  object
 2   title                   40949 non-null  object
 3   channel_title           40949 non-null  object
 4   category_id             40949 non-null  int64 
 5   publish_time            40949 non-null  object
 6   tags                    40949 non-null  object
 7   views                   40949 non-null  int64 
 8   likes                   40949 non-null  int64 
 9   dislikes                40949 non-null  int64 
 10  comment_count           40949 non-null  int64 
 11  thumbnail_link          40949 non-null  object
 12  comments_disabled       40949 non-null  bool  
 13  ratings_disabled        40949 non-null  bool  
 14  video_error_or_removed  40949 non-null  bool  
 15  description             40379 non-null  object
dtypes: bool(3), int64(5), object(8)
memory usage: 4.2+ MB
```

*Figure 4.1 unsupervised dataset information*

Figure denotes that dataset contains 40949 samples in every column except column description. Description column contains some null values. Also, study analysed that dates columns trending_date and publish_time have object as their data type instead of a datetime format.

Research plotted histogram of certain variables to understand the distribution of data samples. Histogram shows the variables like views, likes, dislikes, comment_count have a very right skewed/positive graph as some of the values lie far right under the curve. Also, variables such as comments_disabled, ratings_disabled have one class significantly larger than other.

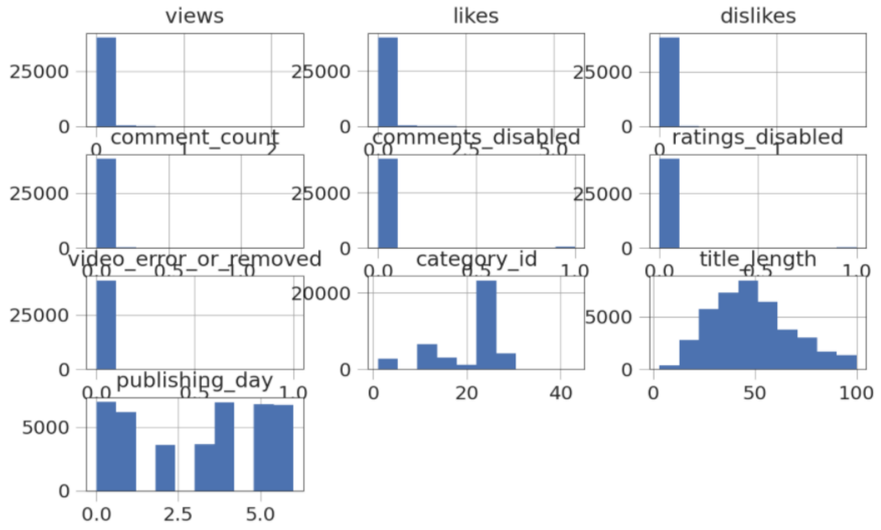


*Figure 4.2 Data distribution of variables*

Here is a histogram of variable views to understand the right skewed nature of the distribution better.

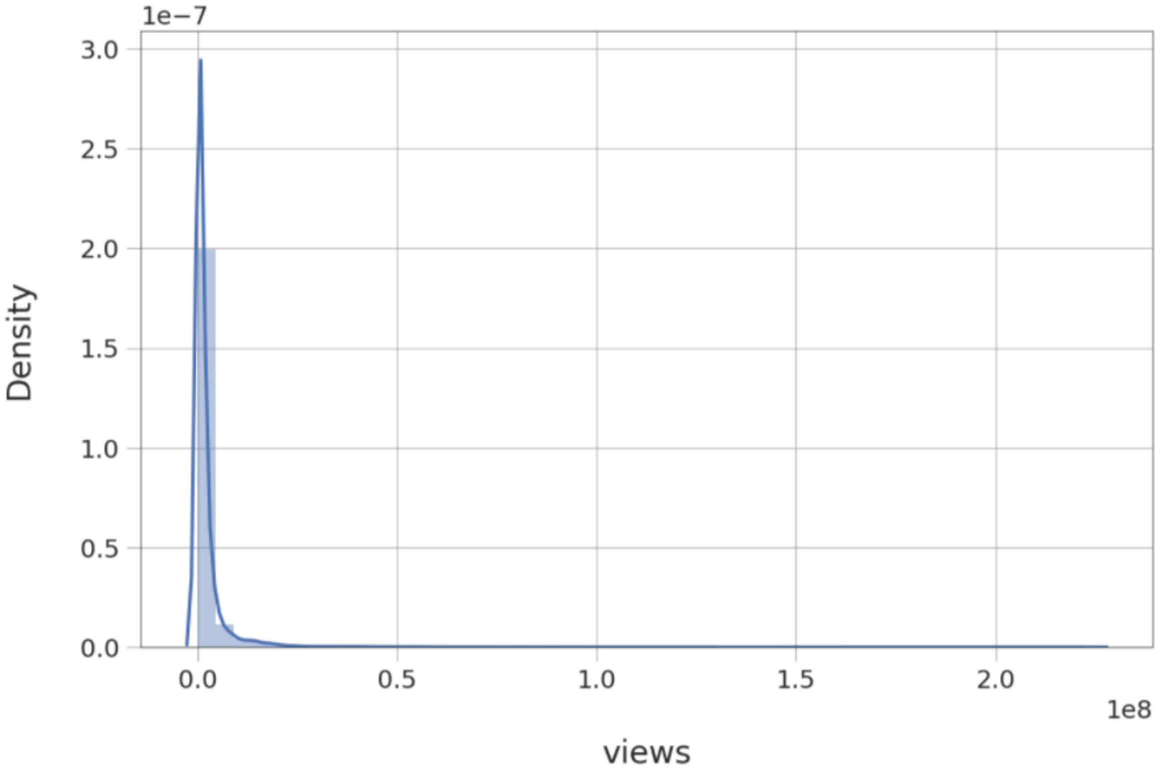


*Figure 4.3 Data distribution of views*

Some of the data points lie far right under the distribution curve as views on some videos are way more because of their popularity. study learns that the vast majority of trending videos have 5 million views or less. To understand the relationship between variables study explored the distribution of data between them using scatter plot.

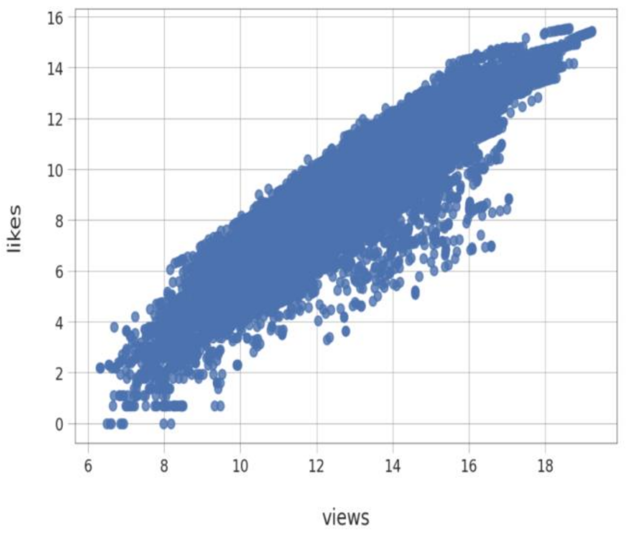


*Figure 4.4 correlation between views and likes*

Scatter plot shows the positive correlation between views and likes features. Number of likes increase with the number of views. Respectively there is a correlation between variables views, dislikes, comments_count etc.

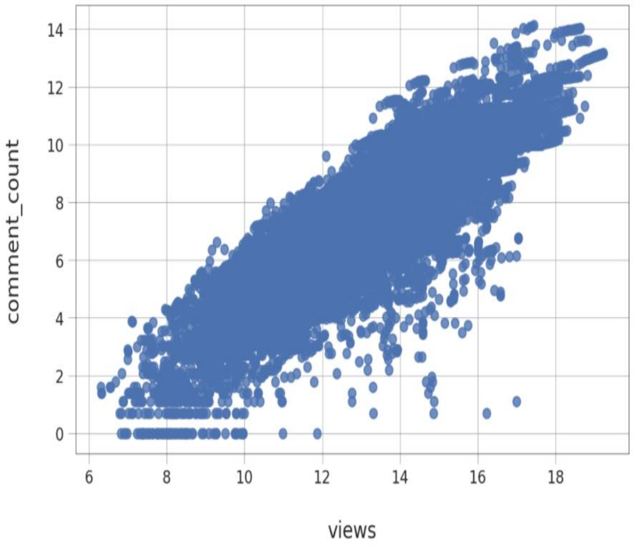


*Figure 4.5 correlation between views and comments_count*

Some of the comments are stagnant but study learns that most of the features have positive correlation with each other. Positive correlation helps in obtaining better accuracy to predict the lifecycle of trending videos.

Thesis explored non numerical columns of unsupervised dataset to explore their impact on making a video trend. Research added a new column using lambda function to the unsupervised dataset to denote the length of each video title, for plotting the histogram of title length to get an idea about the lengths of trending video titles.

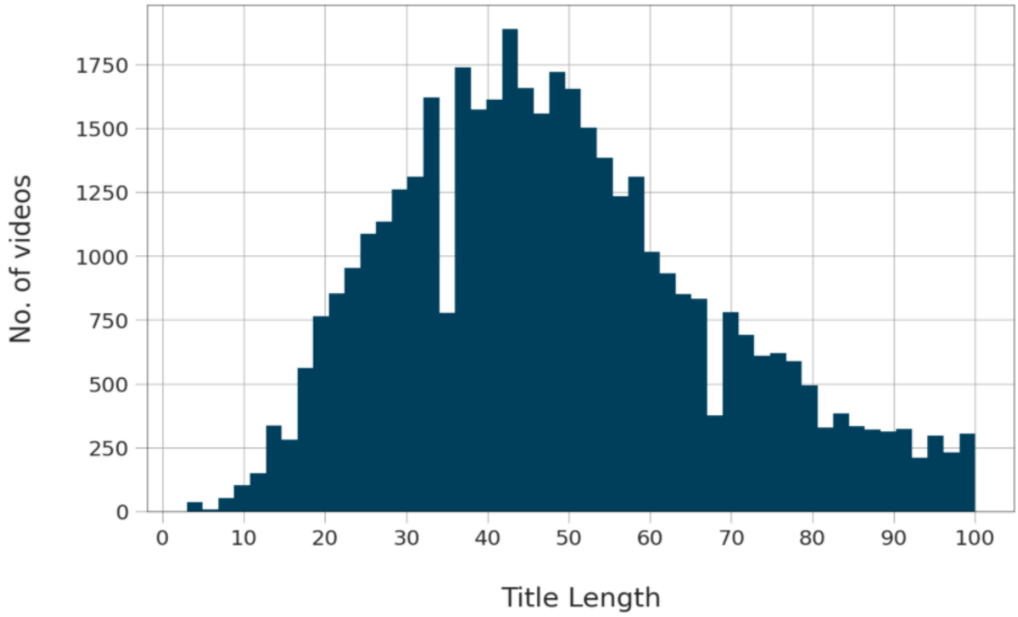


*Figure 4.6 Title Length of YouTube trending videos*

The Title Length data distribution shows a normal distribution, where most videos have Title Lengths between 40 and 60 character approximately. Study plotted a scatter plot between Title Length and number of views to understand the relationship between the two attributes.

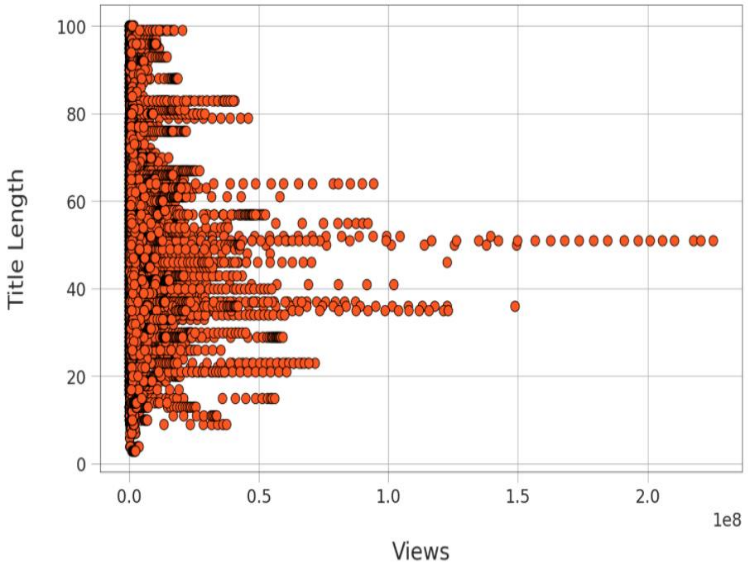


*Figure 4.7 Impact of Title Length on Views*

The scatter plot helps study learn that there is no relationship between the Title Length and the number of views. Hence, study rules out any impact of Title Length in making a video trend. however, plot shows videos that have 100,000,000 views and more have Title Length between 35 and 60 characters approximately.

Titles contains words, it is a marketing strategy in today’s world to use bold words to attract attention of viewers to the videos. Research explored the most common words occurred in the YouTube's trending videos titles, to understand the Impact of words on the success of a video.

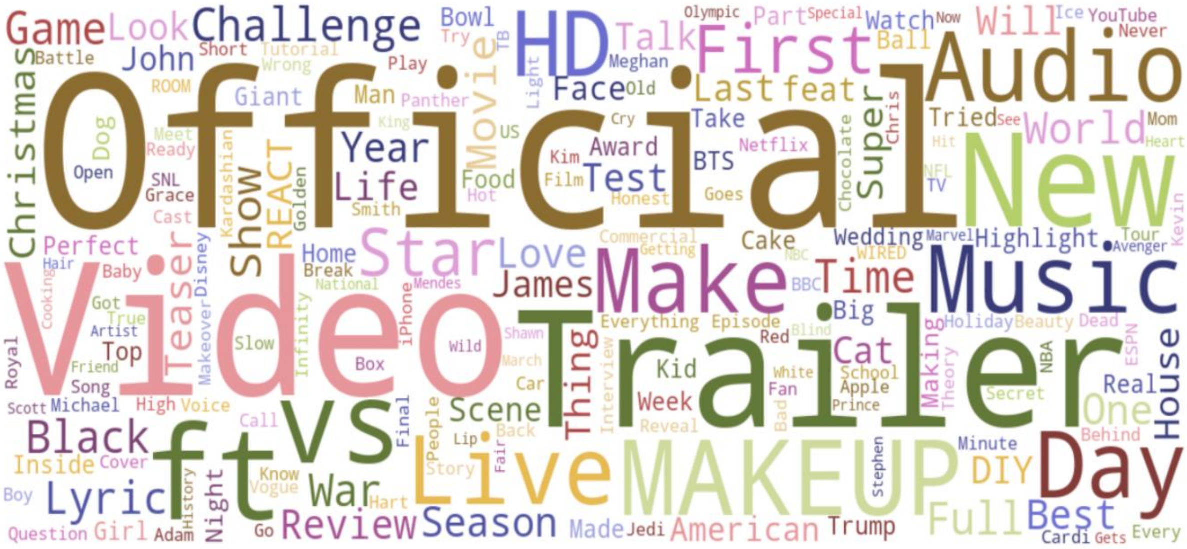


*Figure 4.8 Word cloud of most common words*

Trending videos contain words like Official, Trailer, Live, New, Challenge, HD etc. which helps attracting the viewer's attention.

Thesis extracted and added a new column from dataset for denoting the publishing day of the video to explore the impact a day can have in the trendiness of a video on YouTube. Publishing day column is extracted from publish_time column that denotes the uploading date and time of the video.

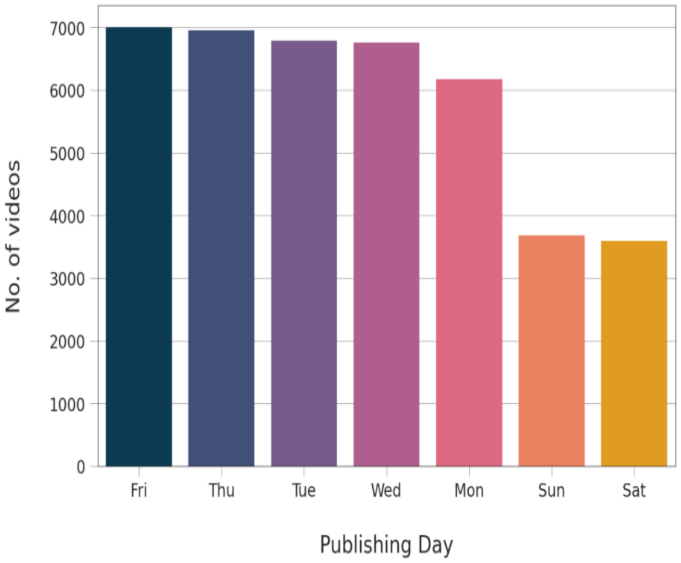


*Figure 4.9 publishing day of the videos*

Study learns that number of videos published on Friday, Thursday, Tuesday, and Wednesday trend more than the number of videos uploaded on the weekend.

Research also explored the impact of publishing time in the trendiness of a video. Thesis used these two variables in prediction analysis. Variable is extracted from publish_time column in the dataset.

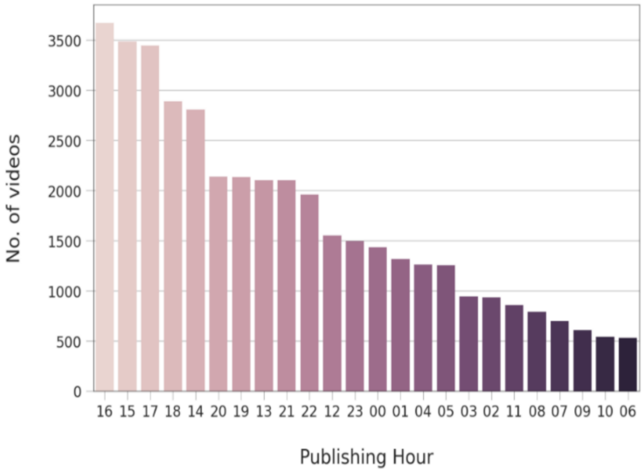
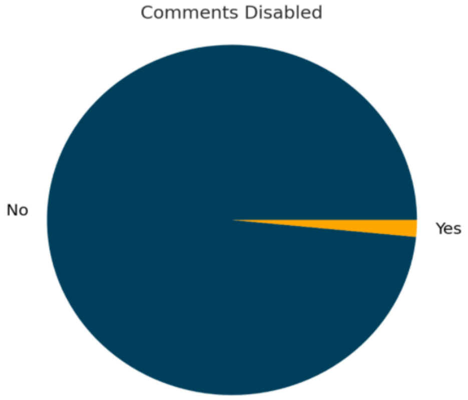


*Figure 4.10 publishing time of the videos*

Study learns that number of videos published at 4 pm, 3 pm, 5 pm and 6 pm trend more than the number of videos uploaded 8 am, 7 am, 6 am.

User interactive features such as likes, dislikes, comments etc. are very important as YouTube and content creators can understand the feedback they are receiving on the content. YouTube allows YouTubers to disable the user rating features hence, study analysed the impact of disabling comments and ratings into success of a video.



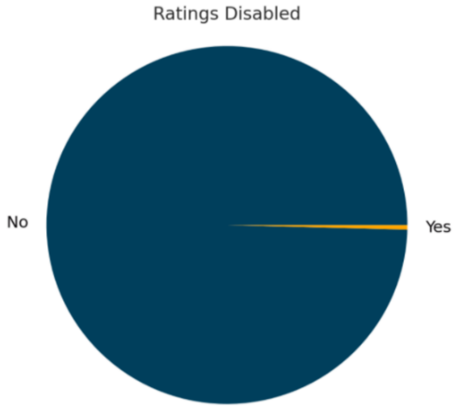


*Figure 4.11 comments disabled of the videos*

*Figure 4.12 ratings disabled of the videos*

Research denotes that YouTube's trending videos have insignificant number of videos which have their comments and ratings disabled. Hence, as a result public interactive feature analysis will help creators to understand the impact of closing rating features or comments of a video.

Thesis covers prediction of YouTube trending video’s life cycle in two parts and forecasting analysis of views of trending videos, therefore data analysis falls under three sections.

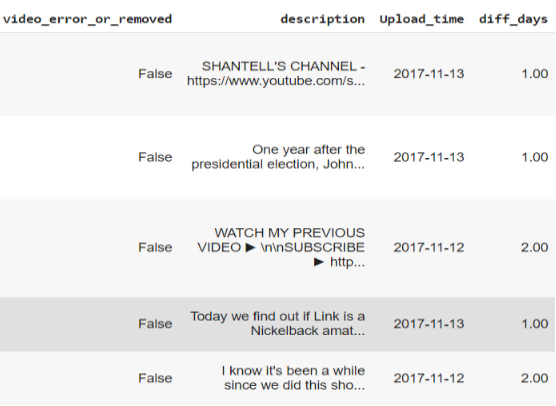
## Predicting the number of days video will take to get on trending list

Thesis implementation followed the CRISP DM methodology hence,

### Data Preparation –

dataset obtained from Kaggle was in unsupervised data format. Dataset did not have any dependent variable (label) to predict. Research transformed unsupervised data into supervised data for forecasting of YouTube trending videos life cycle.

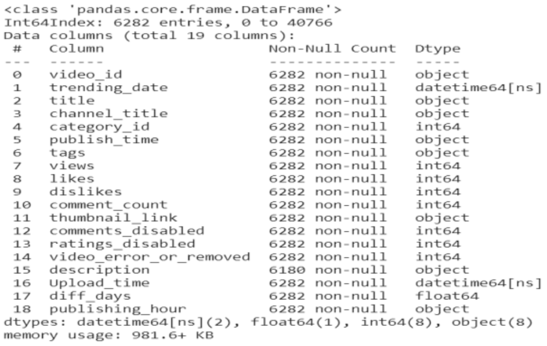
Thesis derived a new column diff_days, which denotes the number of days a video took to get on trending list from its publishing day. Study transformed trending_date and publishing_day column’s datatype from object to datetime using python. Upload_day column was extracted from publish_day column and thesis added diff_day column based on the difference between them.



| video_error_or_removed | description | Upload_time | diff_days |
|---|---|---|---|
| False | SHANTELL'S CHANNEL - https://www.youtube.com/s... | 2017-11-13 | 1.00 |
| False | One year after the presidential election, John... | 2017-11-13 | 1.00 |
| False | WATCH MY PREVIOUS VIDEO ▶ \n\nSUBSCRIBE ▶ http... | 2017-11-12 | 2.00 |
| False | Today we find out if Link is a Nickelback amat... | 2017-11-13 | 1.00 |
| False | I know it's been a while since we did this sho... | 2017-11-12 | 2.00 |

*Figure 4.13 diff_days column*

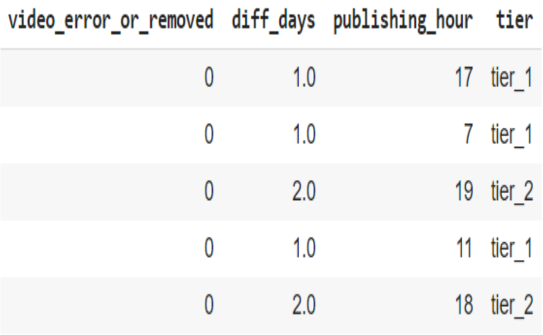
A trending video trends on YouTube for multiple days, thesis analysed only the first instance of a video on the trending list to avoid feeding bias to the model. Every day the features such as number of likes, dislikes, comments increase hence machine will get bias to predict the number of days a video will take to get on the trending list.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6282 entries, 0 to 40766
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                6282 non-null   object
 1   trending_date           6282 non-null   datetime64[ns]
 2   title                   6282 non-null   object
 3   channel_title           6282 non-null   object
 4   category_id             6282 non-null   int64
 5   publish_time            6282 non-null   object
 6   tags                    6282 non-null   object
 7   views                   6282 non-null   int64
 8   likes                   6282 non-null   int64
 9   dislikes                6282 non-null   int64
 10  comment_count           6282 non-null   int64
 11  thumbnail_link          6282 non-null   object
 12  comments_disabled       6282 non-null   int64
 13  ratings_disabled        6282 non-null   int64
 14  video_error_or_removed  6282 non-null   int64
 15  description             6180 non-null   object
 16  Upload_time             6282 non-null   datetime64[ns]
 17  diff_days               6282 non-null   float64
 18  publishing_hour         6282 non-null   object
dtypes: datetime64[ns](2), float64(1), int64(8), object(8)
memory usage: 981.6+ KB
```

*Figure 4.14 First instance data of YouTube videos*

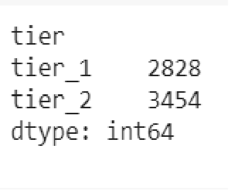
Thesis extracted 6282 samples of first instances of videos with help of video_id column which contains unique video ids for each video. For prediction using classification models of Machine Learning, label must be in categorical form. Thesis created the dependent variable (label) from diff_days column transforming unsupervised data to supervised format.



| video_error_or_removed | diff_days | publishing_hour | tier |
|---|---|---|---|
| 0 | 1.0 | 17 | tier_1 |
| 0 | 1.0 | 7 | tier_1 |
| 0 | 2.0 | 19 | tier_2 |
| 0 | 1.0 | 11 | tier_1 |
| 0 | 2.0 | 18 | tier_2 |

*Figure 4.15 Dependent variable tier created*

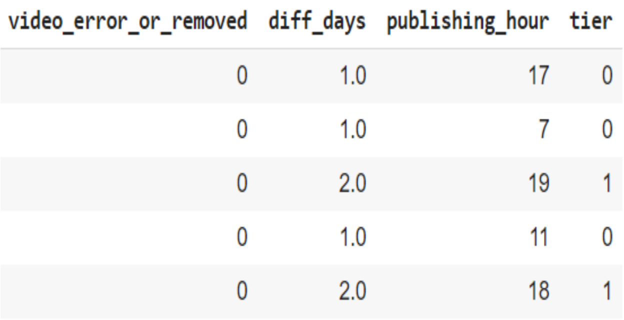
Binary classes were created tier_1 and tier_2 using conditions in python such as if number of days a video took to get on trending list (diff_days) is less than 1 then the class of that video is tier_1 and if the video takes more than a day to get on trending list it belongs to tier_2. The number days parameter as 1 here has been taken in this condition algorithm for getting a somewhat balanced class.



```
tier
tier_1    2828
tier_2    3454
dtype: int64
```

*Figure 4.16 tier classes*

Now, that the output variable is ready, study removed the unnecessary columns from dataset and eliminated null values. Categorical variable needs to be encoded before modelling the data in algorithms hence, research transformed categorical output variable into binary using Label encoder.

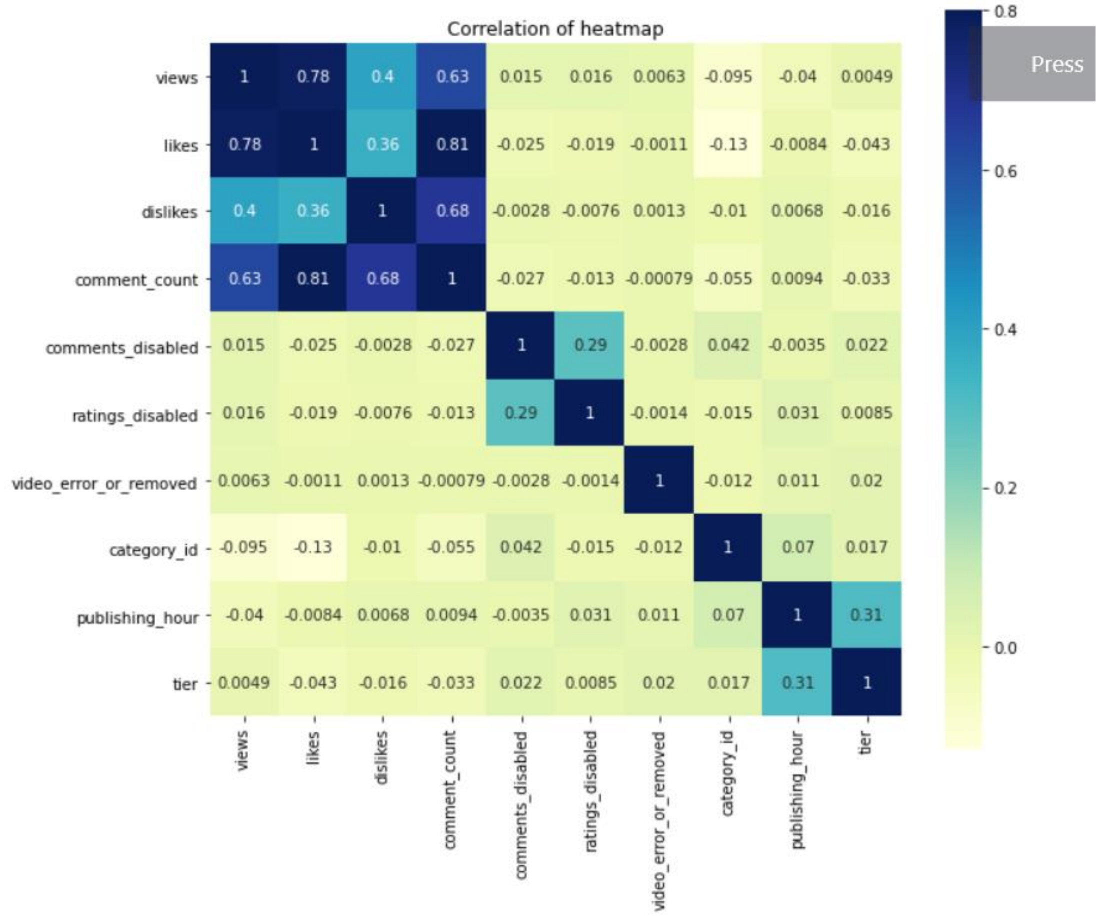


| video_error_or_removed | diff_days | publishing_hour | tier |
|---|---|---|---|
| 0 | 1.0 | 17 | 0 |
| 0 | 1.0 | 7 | 0 |
| 0 | 2.0 | 19 | 1 |
| 0 | 1.0 | 11 | 0 |
| 0 | 2.0 | 18 | 1 |

*Figure 4.17 encoded label output*

Thesis plotted Correlation between the variables to determine the best features for forecasting.

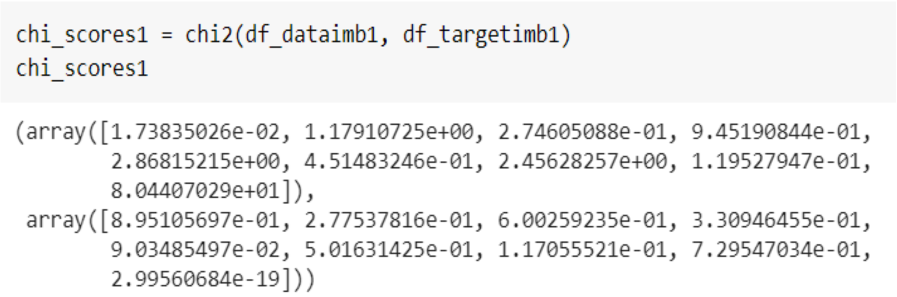
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2fd57410f0>
```



| | views | likes | dislikes | comment_count | comments_disabled | ratings_disabled | video_error_or_removed | category_id | publishing_hour | tier |
|---|---|---|---|---|---|---|---|---|---|---|
| views | 1 | 0.78 | 0.4 | 0.63 | 0.015 | 0.016 | 0.0063 | -0.095 | -0.04 | 0.0049 |
| likes | 0.78 | 1 | 0.36 | 0.81 | -0.025 | -0.019 | -0.0011 | -0.13 | -0.0084 | -0.043 |
| dislikes | 0.4 | 0.36 | 1 | 0.68 | -0.0028 | -0.0076 | 0.0013 | -0.01 | 0.0068 | -0.016 |
| comment_count | 0.63 | 0.81 | 0.68 | 1 | -0.027 | -0.013 | -0.00079 | -0.055 | 0.0094 | -0.033 |
| comments_disabled | 0.015 | -0.025 | -0.0028 | -0.027 | 1 | 0.29 | -0.0028 | 0.042 | -0.0035 | 0.022 |
| ratings_disabled | 0.016 | -0.019 | -0.0076 | -0.013 | 0.29 | 1 | -0.0014 | -0.015 | 0.031 | 0.0085 |
| video_error_or_removed | 0.0063 | -0.0011 | 0.0013 | -0.00079 | -0.0028 | -0.0014 | 1 | -0.012 | 0.011 | 0.02 |
| category_id | -0.095 | -0.13 | -0.01 | -0.055 | 0.042 | -0.015 | -0.012 | 1 | 0.07 | 0.017 |
| publishing_hour | -0.04 | -0.0084 | 0.0068 | 0.0094 | -0.0035 | 0.031 | 0.011 | 0.07 | 1 | 0.31 |
| tier | 0.0049 | -0.043 | -0.016 | -0.033 | 0.022 | 0.0085 | 0.02 | 0.017 | 0.31 | 1 |

*Figure 4.18 correlation between variables*

Correlation map denotes that variables such as views, likes, dislikes, comment_counts have positive correlation with each other. Correlation helps in feature selection but to improve the accuracy of model thesis performed chi square test on the variables.



```
chi_scores1 = chi2(df_dataimb1, df_targetimb1)
chi_scores1
```

```
(array([1.73835026e-02, 1.17910725e+00, 2.74605088e-01, 9.45190844e-01,
        2.86815215e+00, 4.51483246e-01, 2.45628257e+00, 1.19527947e-01,
        8.04407029e+01]),
 array([8.95105697e-01, 2.77537816e-01, 6.00259235e-01, 3.30946455e-01,
        9.03485497e-02, 5.01631425e-01, 1.17055521e-01, 7.29547034e-01,
        2.99560684e-19]))
```

*Figure 4.19 chi square test*

Chi square test denotes the coefficients for each variable which helps in feature selection for modeling, here first array denotes the estimate values and second array denotes the p-values. Hence, plotting p-values,

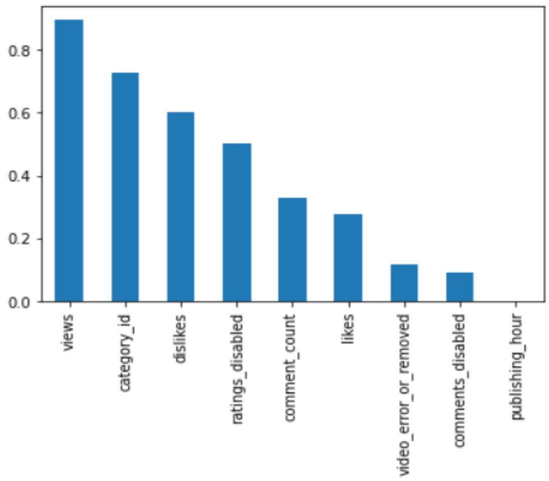


*Figure 4.20 p-value of variables*

Study selected variables such as comment_count, likes, video_error_or_removed, comments_disabled and publishing_hour features based on chi square test as their p-value is less than 0.5. also, study normalized the data for ease of machine's understanding of values.

Thesis tuned the Machine Learning models further using under sampling technique for balancing classes in label to achieve better prediction accuracy.

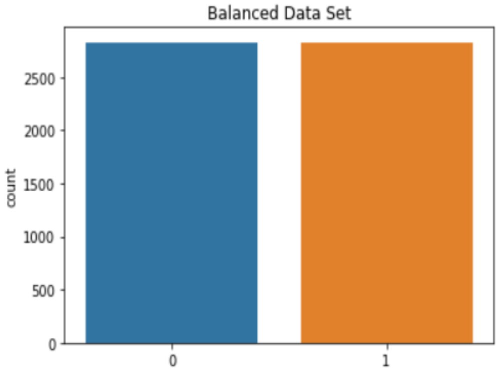
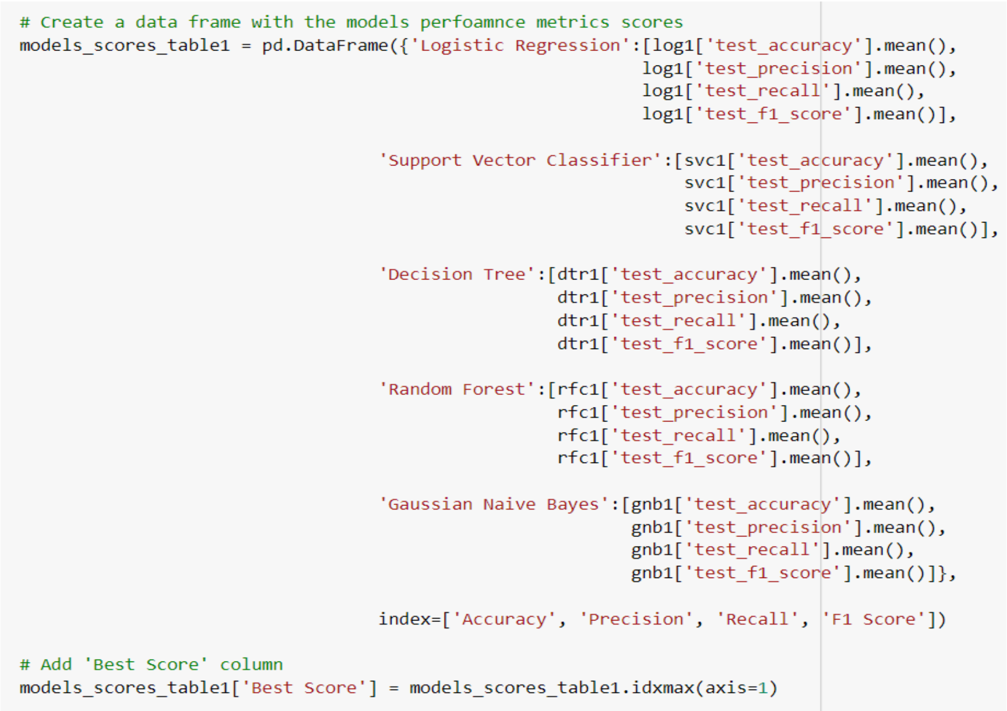


*Figure 4.21 Balanced tier label output*

Now, the study has prepared the data to model into classification algorithms.

### Modeling –

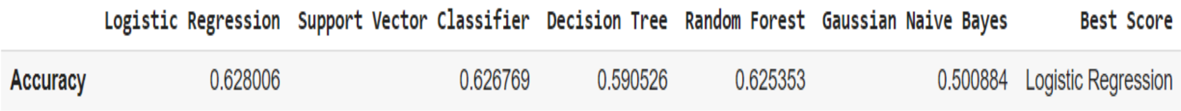
Research created a function using make scorer method in python, instead of modelling five classification models separately it was easy to create a cross validation function. Function performs Random Forest, SVM (Support Vector Machine), Decision Tree, Logistic Regression and Gaussian Naïve Bayes classifiers using k fold cross validation with 5 folds.



```
# Create a data frame with the models perfoamnce metrics scores
models_scores_table1 = pd.DataFrame({'Logistic Regression':[log1['test_accuracy'].mean(),
                                                           log1['test_precision'].mean(),
                                                           log1['test_recall'].mean(),
                                                           log1['test_f1_score'].mean()],

                                  'Support Vector Classifier':[svc1['test_accuracy'].mean(),
                                                              svc1['test_precision'].mean(),
                                                              svc1['test_recall'].mean(),
                                                              svc1['test_f1_score'].mean()],

                                  'Decision Tree':[dtr1['test_accuracy'].mean(),
                                                  dtr1['test_precision'].mean(),
                                                  dtr1['test_recall'].mean(),
                                                  dtr1['test_f1_score'].mean()],

                                  'Random Forest':[rfc1['test_accuracy'].mean(),
                                                  rfc1['test_precision'].mean(),
                                                  rfc1['test_recall'].mean(),
                                                  rfc1['test_f1_score'].mean()],

                                  'Gaussian Naive Bayes':[gnb1['test_accuracy'].mean(),
                                                         gnb1['test_precision'].mean(),
                                                         gnb1['test_recall'].mean(),
                                                         gnb1['test_f1_score'].mean()]},

                                  index=['Accuracy', 'Precision', 'Recall', 'F1 Score'])

# Add 'Best Score' column
models_scores_table1['Best Score'] = models_scores_table1.idxmax(axis=1)
```

*Figure 4.22 model evaluation function*

Model evaluation function gives various parameters of performance of classifiers such as Accuracy, Precision, Recall and F1 score.



| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| **Accuracy** | 0.628006 | 0.626769 | 0.590526 | 0.625353 | 0.500884 | Logistic Regression |

*Figure 4.23 Accuracy measure*

Results show that Logistic Regression, Support Vector Classifier and Random Forest classifier give better performance than Gaussian Naïve Bayes classifier for forecasting. Study implemented Performance evaluation matrix to compare and detect the best model in evaluation phase in results.

Logistic Regression classifier is comparatively better classification model for predicting the number of days a video will take to get on the trending list. Research covered the first part of YouTube’s trending video’s life cycle.

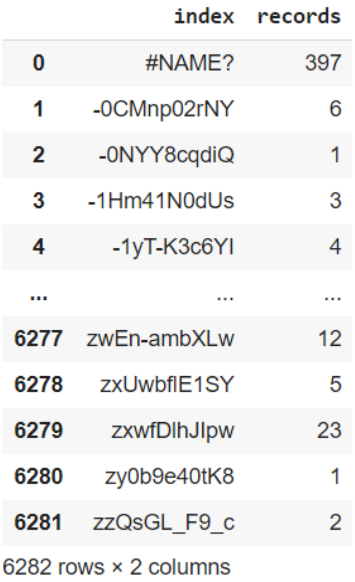
## Predicting the number of days video will trend on trending list

After getting on the trending list on You Tube, a video generally trends for several days and randomly goes off the list. To understand and predict the period a video will trend on You Tube trending page, study have implemented Machine Learning’s Classification models on the dataset.

Thesis implementation followed the CRISP DM methodology hence,

## Data Preparation –

Research extracted and created a new label from the initial unsupervised dataset with 49000 samples. Every video on You Tube contains a unique “video_id”, by counting the occurrences of specific video_ids in video_id column in dataset study obtained the number of days a video got trending.

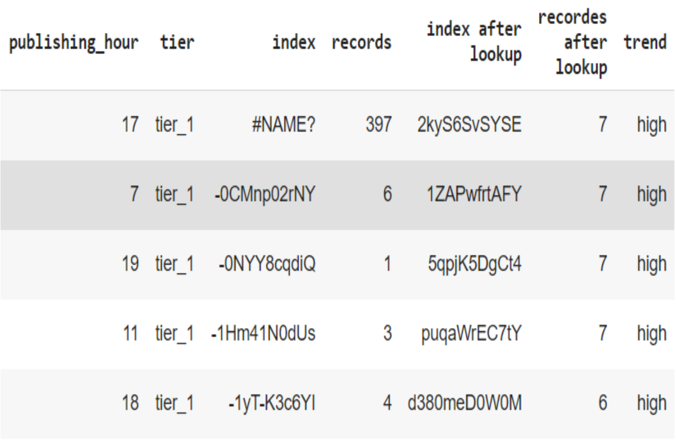


| | index | records |
|---|---|---|
| **0** | #NAME? | 397 |
| **1** | -0CMnp02rNY | 6 |
| **2** | -0NYY8cqdiQ | 1 |
| **3** | -1Hm41N0dUs | 3 |
| **4** | -1yT-K3c6YI | 4 |
| **...** | ... | ... |
| **6277** | zwEn-ambXLw | 12 |
| **6278** | zxUwbflE1SY | 5 |
| **6279** | zxwfDlhJlpw | 23 |
| **6280** | zy0b9e40tK8 | 1 |
| **6281** | zzQsGL_F9_c | 2 |

6282 rows × 2 columns

*Figure 4.24 Number of video frequency on trending page*

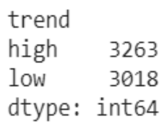
Thesis created a new data frame for this predictive analysis by using look up function in EXCEL to denote the number of records to each video based on their video_id. Study transformed unsupervised data into supervised data and to obtain the categorical variables into label output study transformed records column with a new binary label column trend.



| publishing_hour | tier | index | records | index after lookup | recordes after lookup | trend |
|---|---|---|---|---|---|---|
| 17 | tier_1 | #NAME? | 397 | 2kyS6SvSYSE | 7 | high |
| 7 | tier_1 | -0CMnp02rNY | 6 | 1ZAPwfrtAFY | 7 | high |
| 19 | tier_1 | -0NYY8cqdiQ | 1 | 5qpjK5DgCt4 | 7 | high |
| 11 | tier_1 | -1Hm41N0dUs | 3 | puqaWrEC7tY | 7 | high |
| 18 | tier_1 | -1yT-K3c6YI | 4 | d380meD0W0M | 6 | high |

*Figure 4.25 Dependent variable trend created*

Using conditions method in python study created the label output column trend such as if a video trends on YouTube less than 5 days it belongs to low class and if a video trends on YouTube for more than 5 days it belongs to High class.

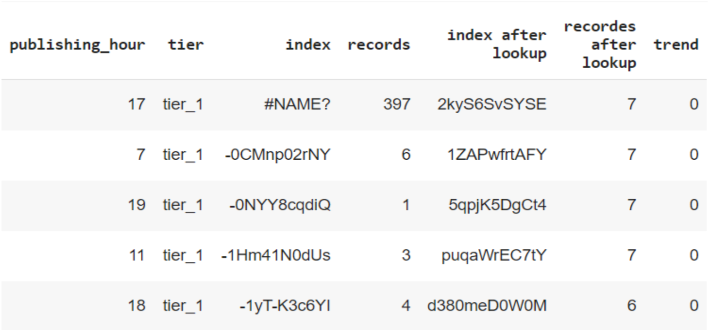


```
trend
high    3263
low     3018
dtype: int64
```

*Figure 4.26 trend classes*

The number days parameter as 5 here has been taken in this condition algorithm for getting a somewhat balanced class.

Now that the output variable is ready, study removed the unnecessary columns from dataset and eliminated null values. Categorical variable needs to be encoded before modelling the data in algorithms hence, research transformed categorical output variable into binary using Label encoder. Also, study normalised the data using scalar function for ease of understanding for the machine.

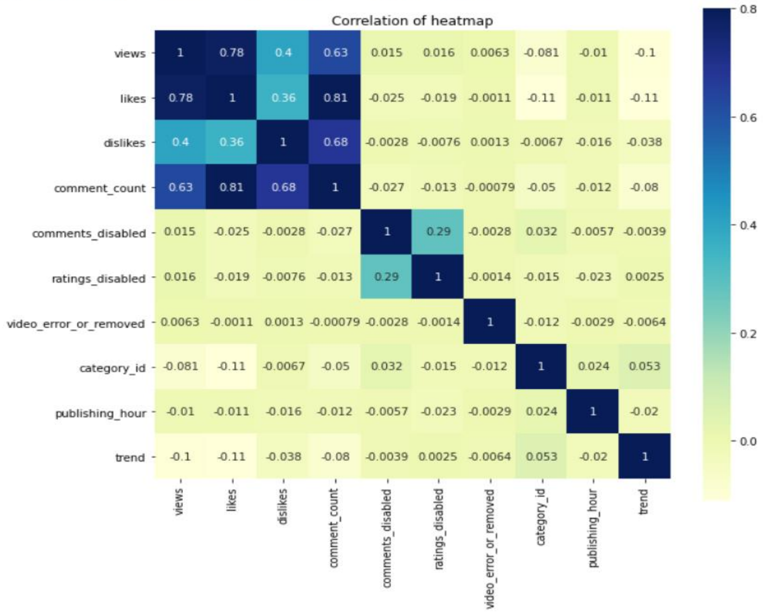


| publishing_hour | tier | index | records | index after lookup | recordes after lookup | trend |
|---|---|---|---|---|---|---|
| 17 | tier_1 | #NAME? | 397 | 2kyS6SvSYSE | 7 | 0 |
| 7 | tier_1 | -0CMnp02rNY | 6 | 1ZAPwfrtAFY | 7 | 0 |
| 19 | tier_1 | -0NYY8cqdiQ | 1 | 5qpjK5DgCt4 | 7 | 0 |
| 11 | tier_1 | -1Hm41N0dUs | 3 | puqaWrEC7tY | 7 | 0 |
| 18 | tier_1 | -1yT-K3c6YI | 4 | d380meD0W0M | 6 | 0 |

*Figure 4.27 encoded label output*

As the prediction is about the duration a video will trend on YouTube study used only the unique videos data. This gave a sample size of 6281 unique video count in dataset. Thesis plotted Correlation between the variables to determine the best features for forecasting.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2fe0a60e10>
```

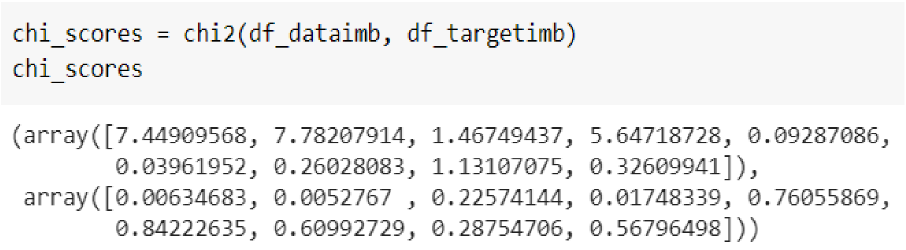


Correlation of heatmap

|  | views | likes | dislikes | comment_count | comments_disabled | ratings_disabled | video_error_or_removed | category_id | publishing_hour | trend |
|---|---|---|---|---|---|---|---|---|---|---|
| views | 1 | 0.78 | 0.4 | 0.63 | 0.015 | 0.016 | 0.0063 | -0.081 | -0.01 | -0.1 |
| likes | 0.78 | 1 | 0.36 | 0.81 | -0.025 | -0.019 | -0.0011 | -0.11 | -0.011 | -0.11 |
| dislikes | 0.4 | 0.36 | 1 | 0.68 | -0.0028 | -0.0076 | 0.0013 | -0.0067 | -0.016 | -0.038 |
| comment_count | 0.63 | 0.81 | 0.68 | 1 | -0.027 | -0.013 | -0.00079 | -0.05 | -0.012 | -0.08 |
| comments_disabled | 0.015 | -0.025 | -0.0028 | -0.027 | 1 | 0.29 | -0.0028 | 0.032 | -0.0057 | -0.0039 |
| ratings_disabled | 0.016 | -0.019 | -0.0076 | -0.013 | 0.29 | 1 | -0.0014 | -0.015 | -0.023 | 0.0025 |
| video_error_or_removed | 0.0063 | -0.0011 | 0.0013 | -0.00079 | -0.0028 | -0.0014 | 1 | -0.012 | -0.0029 | -0.0064 |
| category_id | -0.081 | -0.11 | -0.0067 | -0.05 | 0.032 | -0.015 | -0.012 | 1 | 0.024 | 0.053 |
| publishing_hour | -0.01 | -0.011 | -0.016 | -0.012 | -0.0057 | -0.023 | -0.0029 | 0.024 | 1 | -0.02 |
| trend | -0.1 | -0.11 | -0.038 | -0.08 | -0.0039 | 0.0025 | -0.0064 | 0.053 | -0.02 | 1 |

*Figure 4.28 correlation between features*

Correlation map denotes that variables such as views, likes, dislikes, comment_counts have positive correlation with each other. Correlation helps in feature selection but to improve the accuracy of model thesis performed chi square test on the variables.



```
chi_scores = chi2(df_dataimb, df_targetimb)
chi_scores
```

```
(array([7.44909568, 7.78207914, 1.46749437, 5.64718728, 0.09287086,
        0.03961952, 0.26028083, 1.13107075, 0.32609941]),
 array([0.00634683, 0.0052767 , 0.22574144, 0.01748339, 0.76055869,
        0.84222635, 0.60992729, 0.28754706, 0.56796498]))
```

*Figure 4.29 chi square test*

Chi square test denotes the coefficients for each variable which helps in feature selection for modeling, here first array denotes the estimate values and second array denotes the p-values. Hence, plotting p-values,

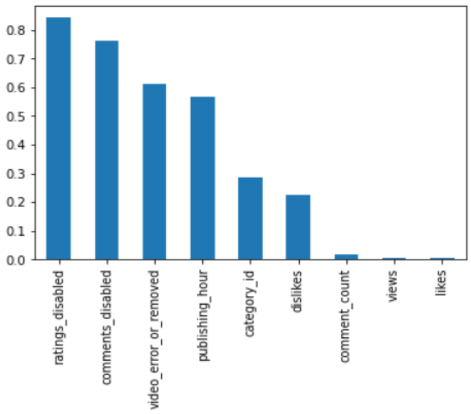


*Figure 4.30 p-value for variables*

Study selected variables such as comment_count, likes, dislikes, views, and category_id features based on chi square test as their p-value is less than 0.5. Thesis tuned the Machine Learning models further using under sampling technique for balancing classes in label to achieve better prediction accuracy.

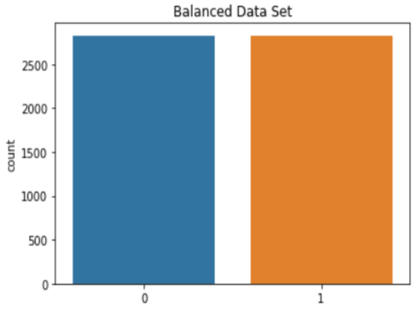
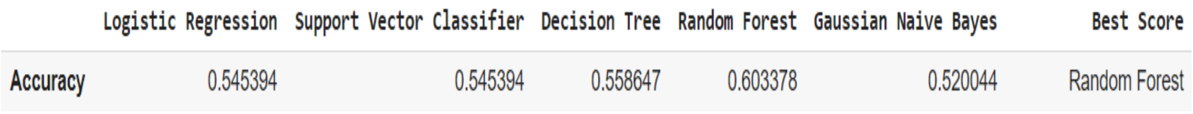


*Figure 4.31 balanced trend output*

Now, the study has prepared the data to model into classification algorithms.

### Modeling –

Study used the same make scorer model evaluation function for predicting the number of days a video will trend on the trending list.



| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| **Accuracy** | 0.545394 | 0.545394 | 0.558647 | 0.603378 | 0.520044 | Random Forest |

*Figure 4.32 accuracy measure for model 2*

Results show that Random Forest, Support Vector Classifier and Decision Tree classifier give better performance than Gaussian Naïve Bayes classifier for forecasting. Study implemented Performance evaluation matrix to compare and detect the best model in evaluation phase in results.

## Predictive analysis of number of views for YouTube's trending videos

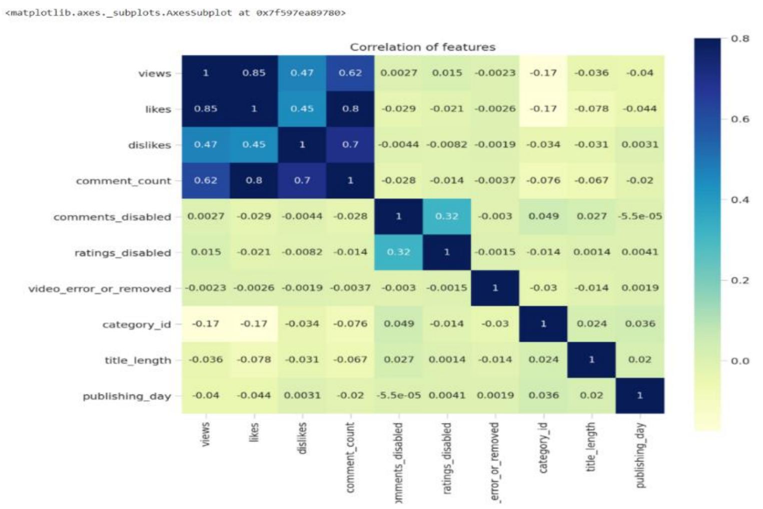
Study have performed Linear regression model of Machine Learning on target variable, which is "Views" for predicting the number of views of a video on trending list. YouTube's features such as Likes, Dislikes, Comments counts, category id, publishing hour, ratings disabled, comments disabled and video error or removed have been the variables for model.

Thesis implementation followed the CRISP DM methodology hence,

### Data Preparation –

Study determined that there is moderate correlation between features that are Likes, Dislikes, Comment counts, category id, comments disabled and the target variable views using correlation function.

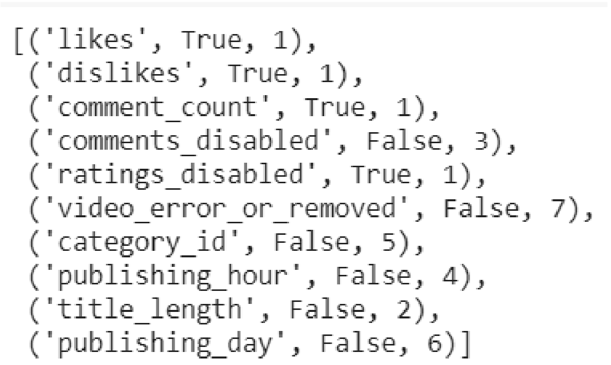
Correlation helps in feature selection but to improve the accuracy of model thesis performed various feature selection methods on the variables.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f597ea89780>
```

*Figure 4.33 correlation between variables*

Study used RFE (Recursive Feature Elimination) algorithm and cross validation induced Linear Regression method for obtaining the better R2 score.



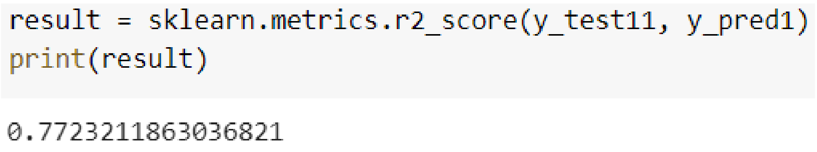
```
[('likes', True, 1),
 ('dislikes', True, 1),
 ('comment_count', True, 1),
 ('comments_disabled', False, 3),
 ('ratings_disabled', True, 1),
 ('video_error_or_removed', False, 7),
 ('category_id', False, 5),
 ('publishing_hour', False, 4),
 ('title_length', False, 2),
 ('publishing_day', False, 6)]
```

*Figure 4.34 RFE output*

Thesis selected features with RFE score not more than 1 which has True values. Hence, variables such as likes, dislikes, comment_count, and ratings_disabled have been modelled for predictive analysis. Study normalized the dataset for scaling the extreme data values between 0 and 1.

Modeling –

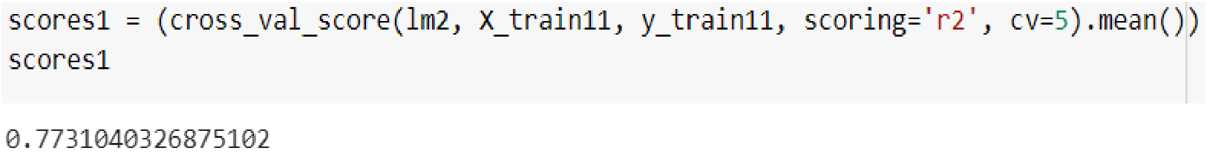
The dependent variable (label) and independent variables (features) shows normal distribution nature of data hence a simple logarithmic model is better to predict the output variable. Research modelled Linear Regression algorithm using 70% - 30% trainset and test set split function. Study tuned the Machine Learning algorithm using RFE and normalization methods for getting better accuracy.



```
result = sklearn.metrics.r2_score(y_test11, y_pred1)
print(result)

0.7723211863036821
```

*Figure 4.35 accuracy output for Linear Regression model*

Research also implemented k fold cross validation induced Linear Regression model to compare the accuracy output with the normal Linear Regression model.



```
scores1 = (cross_val_score(lm2, X_train11, y_train11, scoring='r2', cv=5).mean())
scores1

0.7731040326875102
```

*Figure 4.36 accuracy output for K-fold CV induced Linear Regression*

Hence, Linear Regression model of Machine Learning gives 77.3% accuracy for predicting the number of views of a YouTube trending video.

# CHAPTER 5 – RESULTS

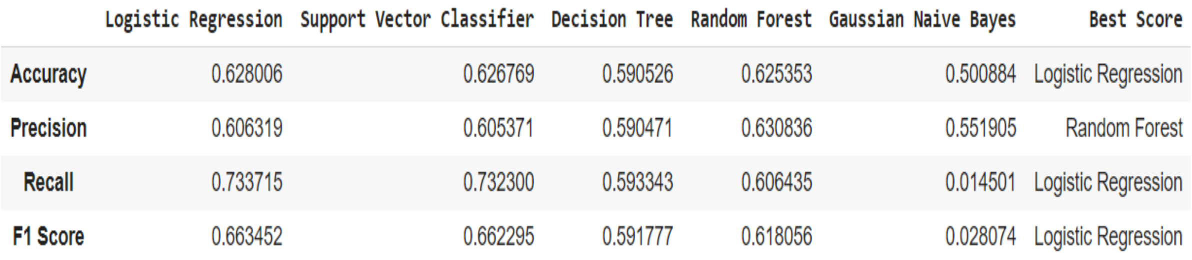
## Evaluation –

Research performed K-fold Cross Validation method for evaluating and comparing the multiple classifiers implemented on the dataset. Cross Validation method gives a Performance Evaluation Matrix with key performance measures such as Accuracy, Precision, Recall and F1 score. Matrix gives elaborated and knowledgeable results rather than a simple hypothesis test to determine the significance in classifiers findings.

## Predicting the number of days video will take to get on trending list with Classification:

A video takes certain amount of time to get trending since it is published on YouTube. Classification models such as Decision Tree, Random Forest, Naïve Bayes etc. were implemented on the target variable "tier" to predict the classes. Following is the Performance Evaluation Matrix for determining significance of results.

Evaluation of Classification models on "tier" –



| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| **Accuracy** | 0.628006 | 0.626769 | 0.590526 | 0.625353 | 0.500884 | Logistic Regression |
| **Precision** | 0.606319 | 0.605371 | 0.590471 | 0.630836 | 0.551905 | Random Forest |
| **Recall** | 0.733715 | 0.732300 | 0.593343 | 0.606435 | 0.014501 | Logistic Regression |
| **F1 Score** | 0.663452 | 0.662295 | 0.591777 | 0.618056 | 0.028074 | Logistic Regression |

*Figure 5.1 Performance Evaluation Matrix for "tier"*

Evaluation Matrix shows that Logistic Regression classifier gives 12% better accuracy than the least performing Gaussian Naïve Bayes classifier. Also, Logistic Regression have the 13% and 5% higher Precision and Recall than Random Forest classifier, respectively. Support Vector Classifier have the second-best significant performance for predicting the first phase of a YouTube's trending video's life cycle.

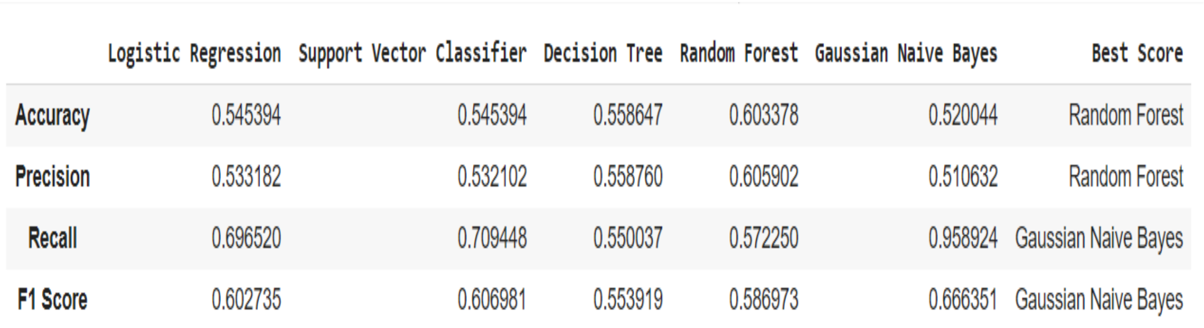
Hence, Performance Evaluation Matrix shows that for predicting the number of days a video will take to get on the trending list Logistic Regression, Support Vector Machine and Random Forest classifiers give the moderate performance.

## Predicting the number of days video will trend on trending list with Classification:

After getting on the trending list on You Tube, a video generally trends for several days and randomly goes off the list. To understand and predict the period a video will trend on You Tube trending page, Study have implemented Machine Learning's Classification models on the dataset.

Classification models such as Decision Tree, Random Forest, Naïve Bayes etc. were implemented on the target variable "trend" to predict the classes. Following is the Performance Evaluation Matrix for "trend".

Evaluation of Classification models on trend –



| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| Accuracy | 0.545394 | 0.545394 | 0.558647 | 0.603378 | 0.520044 | Random Forest |
| Precision | 0.533182 | 0.532102 | 0.558760 | 0.605902 | 0.510632 | Random Forest |
| Recall | 0.696520 | 0.709448 | 0.550037 | 0.572250 | 0.958924 | Gaussian Naive Bayes |
| F1 Score | 0.602735 | 0.606981 | 0.553919 | 0.586973 | 0.666351 | Gaussian Naive Bayes |

*Figure 5.2 Performance Evaluation Matrix for "trend"*

Evaluation Matrix shows that Random Forest classifier gives 8% better accuracy than the least performing Gaussian Naïve Bayes classifier. Also, Logistic Regression and Support Vector Machine have the 13% and 2% higher Recall and F1 score than Random Forest classifier, respectively. Also, Random Forest classifier has the highest Precision of 60.59%. Logistic Regression classifier have the second-best significant performance for predicting the second phase of a YouTube's trending video's life cycle.

Hence, Performance Evaluation Matrix shows that for predicting the number of days a video will trend on the trending list Random Forest, Logistic Regression and Support Vector Machine classifiers give the moderate performance. Also, Study finds that Gaussian Naïve Bayes classifier has the least significant performance to predict the YouTube's trending video's life cycle.

### Predicting the number of views of a video on trending list with Regression:

For predicting the number of views of a video on trending list. Thesis have performed Linear Regression models on target variable, which is "Views". YouTube's features such as Likes, Dislikes, Comments counts, category id, publishing_hour, ratings_disabled, and comments_disabled have been the attributes for model. There is moderate correlation between features that are Likes, Dislikes, Comment_counts, ratings_disabled, and the target variable views.

RFE (Recursive Feature Elimination) algorithm has been used for identifying the significant variables for modeling and obtaining the better R2 score. Thesis performed two Linear Regression methods, a simple Linear Regression and Linear Regression with Cross Validation where K-fold is 5.

Results of Linear Regression models on views –

*Table 5.1 Results for Linear Regression models*

| Model | Score for r2 |
|---|---|
| Linear Regression | 0.772 |
| Cross validation induced Linear Regression | 0.773 |

As we can see both the Linear Regression models has a moderate r2 score of 0.773.

Various regression models such as S-H model, ML model, MRBF model have been implemented on predicting the number of views for YouTube videos in the past. Most of the study focused on very initial statistics of a video to predict the future statistics, they explored the sudden burst in views of a video, virality of a content with real world changes. Current study lacks the proper supervised dataset for conducting such complex regression models, also features and label shows a normal positively skewed distribution nature in data hence, a simple Logarithmic Linear Regression model is performed to predict the number of views.

For classification problem Logistic Classifier stands out giving significant performance for both the trending video's lifecycle. Study conducts a binary classification problem for predicting the lifecycle of a trending videos. Logistic Regression is one of the best classifiers when it comes to classify a binary problem. Random Forest and Support Vector Machine classifiers are also very good binary classifiers hence they have the second-best performance in study. Whereas Gaussian Naïve Bayes classifier has the least significant performance in the study. Naïve Bayes classifier makes some naïve assumptions for classification and this classifier might work better if the dataset is a well maintained, labelled dataset.

Current dataset is a transformed supervised dataset from an unsupervised dataset, still the research managed a 62.52% decent accuracy for predicting the lifecycle of a trending video. Study can achieve better improved results with an ideal supervised data.

# CHAPTER 6 – SCOPE AND LIMITATIONS

## *SCOPE AND LIMITATION*

### Scope –

Analysis of unstructured data in the form of unsupervised dataset is challenging. However, Analysis of structured supervised data is much easier and more reliable. YouTube, Google's social media platform for videos, has over a billion users and averages billions of views per day.

Forecasting is a science of predicting the dependent variable (label) based on its independent correlated variables (features) using Machine Learning algorithms. Classification and Regression algorithms used in the study provides the social giants YouTube with useful business analysis necessary for making business decisions.

Many of forecasting's involves analysis of supervised or unsupervised data. Study predicts a YouTube's trending video's lifecycle with 62% accuracy by transforming unsupervised data into supervised data. Hence, there is huge scope for improvement which can be obtained by using much more complex and hybrid Machine Learning algorithms on a specific supervised data collected for studies research question. Furthermore, research can explore the non-numerical attributes such as Title, Description etc. contribution combined with current model in future.

There is scope in exploring a video's sudden decline from popularity. Study can further add and analyse a third lifecycle of a trending video which is its decline from the trending page.

### Limitation –

The main limitation in this research is the initial unsupervised dataset. Study could not implement unsupervised machine learning algorithms such as Deep Learning as the samples and features in the dataset are limited as well.

Good statistics of samples helps achieve better prediction results.

# CHAPTER 7 – CONCLUSION

Popularity is often considered a random act. Thesis analyzed YouTube’s trending video’s lifecycle forecasting aspect to understand the randomness of a video’s virality for academic and business perspective.

Thesis implemented five classification models namely Random Forest, SVM, Decision Tree, Logistic Regression and Gaussian Naïve Bayes to explore the research question that is ‘Comparative analysis of Machine Learning algorithms for YouTube trending video’s life cycle prediction by analysing YouTube’s trending videos statistics data’. Comparative studies analysis concludes that Logistic Regression, Random Forest, and Support Vector Machine (SVM) classifiers give better performance for predicting the lifecycle of YouTube’s trending Videos. K-fold Cross Validation method evaluates the model results and help research confirm the findings. Study also concludes the scope of better classification results by performing the existing methodology on a better supervised dataset.

Research implemented a simple Logarithmic Linear Regression algorithm to predict the number of views for a YouTube trending video. Study findings conclude that basic Linear Regression model gives a moderate accuracy percentage of 77.3% with help of limited features. Hence, there is scope of obtaining better accuracy by implementing hybrid S-H, MRBF model with the current study on a better supervised dataset with added numerical and non-numerical features.

Study successfully implemented the Machine Learning algorithms on an unsupervised dataset and analysed the findings to compare and select the best classification models for predicting YouTube’s trending video’s lifecycle.

# REFERENCES

Ali, J., Rehanullah Khan, Ahmad, N. and Imran Maqsood (2012). *Random Forests and Decision Trees*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees [Accessed 2 Jan. 2021].

Altman, N. and Krzywinski, M. (2015). Simple linear regression. *Nature Methods*, [online] 12(11), pp.999–1000. Available at: https://www.nature.com/articles/nmeth.3627 [Accessed 2 Jan. 2021].

Anders Brodersen, Scellato, S. and Mirjam Wattenhofer (2012). *YouTube around the world: Geographic popularity of videos*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/262290085_YouTube_around_the_world_Geographic_popularity_of_videos [Accessed 9 Jan. 2021].

Berrar, D. (2018). *Cross-Validation*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/324701535_Cross-Validation [Accessed 3 Jan. 2021].

Chelaru, S.V., Orellana-Rodriguez, C. and Altingovde, I.S. (2012). Can Social Features Help Learning to Rank YouTube Videos? *Web Information Systems Engineering - WISE 2012*, [online] pp.552–566. Available at: https://link.springer.com/chapter/10.1007/978-3-642-35063-4_40 [Accessed 9 Jan. 2021].

Davidson, J., Liebald, B., Junning Liu and Dasarathi Sampath (2010). The YouTube video recommendation system. [online] ResearchGate. Available at: https://www.researchgate.net/publication/221140967_The_YouTube_video_recommendation_system [Accessed 7 Aug. 2020].

Figueiredo, F., Almeida, J.M., Gonçalves, M.A. and Benevenuto, F. (2016). TrendLearner: Early prediction of popularity trends of user generated content. *Information Sciences*, 349–350, pp.172–187.

Flavio Figueiredo, F. Benevenuto and J. Almeida (2011). *The tube over time: characterizing popularity growth of youtube videos*. [online] undefined. Available at: https://www.semanticscholar.org/paper/The-tube-over-time%3A-characterizing-popularity-of-Figueiredo-Benevenuto/0b1e520fbca86e377678d82d9f6144bcf17f606e#extracted [Accessed 9 Jan. 2021].

Gábor Szabó and Huberman, B. (2008). *Predicting the Popularity of Online Content*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/23417017_Predicting_the_Popularity_of_Online_Content [Accessed 9 Jan. 2021].

Henrique Pinto, Jussara Almeida and Marcos André Gonçalves (2013). *Using early view patterns to predict the popularity of YouTube videos*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/266653405_Using_early_view_patterns_to_predict_the_popularity_of_YouTube_videos [Accessed 9 Jan. 2021].

Hoiles, W., Aprem, A. and Krishnamurthy, V. (2017). Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(7), pp.1426–1437.

Iman Barjasteh, Liu, Y. and Hayder Radha (2014). Trending Videos: Measurement and Analysis. [online] ResearchGate. Available at: https://www.researchgate.net/publication/266262149_Trending_Videos_Measurement_and_Analysis [Accessed 24 Jul. 2020].

J, M. (2017). Trending YouTube Video Statistics. [online] Kaggle.com. Available at: https://www.kaggle.com/datasnaek/youtubenew?select=USvideos.csv [Accessed 24 Jul. 2020].

Krishna, A., Zambreno, J. and Krishnan, S. (n.d.). *Polarity Trend Analysis of Public Sentiment on YouTube*. [online] Available at: http://www.rcl.ece.iastate.edu/sites/default/files/papers/KriZam13A.pdf.

Li, Y., Eng, K. and Zhang, L. (n.d.). *YouTube Videos Prediction: Will this video be popular?* [online] Available at: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647615.pdf.

Lior Rokach and Oded Maimon (2005). *Decision Trees*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/225237661_Decision_Trees [Accessed 2 Jan. 2021].

Ouyang, S., Li, C. and Li, X. (2016). A Peek Into the Future: Predicting the Popularity of Online Videos. *IEEE Access*, [online] 4, pp.3026–3033. Available at: https://ieeexplore.ieee.org/document/7491235?denied= [Accessed 9 Jan. 2021].

Peng, J., Kuk Lida Lee and Ingersoll, G.M. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. [online] ResearchGate. Available at:

https://www.researchgate.net/publication/242579096_An_Introduction_to_Logistic_Regression_Analysis_and_Reporting [Accessed 3 Jan. 2021].

Pouria Kaviani and Sunita Dhotre (2017). *Short Survey on Naive Bayes Algorithm*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/323946641_Short_Survey_on_Naive_Bayes_Algorithm [Accessed 3 Jan. 2021].

Prabha, G.M., Madhumitha, B., Ramya, R.P., 2019. Predicting the Popularity of Trending Videos in YouTube Using Sentimental Analysis 8, 6.

s. Amudha, Niveditha V.R, P S Raja Kumar and Radha Rammohan Shanthanam (2020). *Youtube Trending Video Metadata Analysis Using Machine Learning*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/342150876_Youtube_Trending_Video_Metadata_Analysis_Using_Machine_Learning [Accessed 9 Jan. 2021].

Shearer, C., Moss, L., Adelman, S., Herdlein, S.A., Fong, J., Wong, H.K. and Fong, A. (2000). The CRISPDM Model: The New Blueprint for Data Mining E-Business and the New Demands on Data Warehousing Technology: The New Demands E-Commerce Places on Data Warehousing Technology Katherine Hammer Turning the Corner from Data Warehousing to Electronic C. JOURNAL, [online] 5. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dmmodel-the-newblueprint-for-data-mining-shearer-colin.pdf.

Tejal Rathod and Mehul Barot (2018). *Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/323877617_Trend_Analysis_on_Twitter_for_Predicting_Public_Opinion_on_Ongoing_Events [Accessed 9 Jan. 2021].

Theodoros Evgeniou and Massimiliano Pontil (2001). *Support Vector Machines: Theory and Applications*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications [Accessed 3 Jan. 2021].

Trzcinski, T. and Rokita, P. (2017). Predicting Popularity of Online Videos Using Support Vector Regression. *IEEE Transactions on Multimedia*, [online] 19(11), pp.2561–2570. Available at: https://ieeexplore.ieee.org/document/7903630 [Accessed 9 Jan. 2021].

Umair Shafique and Haseeb Qaiser (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). [online] ResearchGate. Available at: https://www.researchgate.net/publication/268770881_A_Comparative_Stu dy_of_Data_Mining_Pro cess_Models_KDD_CRISP-DM_and_SEMMA [Accessed 16 Jun. 2020].

Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D. and Lestantyo, P. (2019). Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. [online] Available at: https://ieeexplore.ieee.org/document/8949568 [Accessed 3 Jan. 2021].

Zhou, R., Khemmarat, S., & Gao, L. (2010). The impact of YouTube recommendation system on video views. Proceedings of the 10th Annual Conference on Internet Measurement - IMC ’10. doi:10.1145/1879141.1879193