# DEPARTMENT OF INFORMATION TECHNOLOGY DELHI TECHNOLOGICAL UNIVERSITY

**IT-306: ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEM**

**PROJECT REPORT**

**ANALYSIS AND DETECTION OF FAKE REPLICAS OF YOUTUBE CHANNEL**

## SUBMITTED TO:

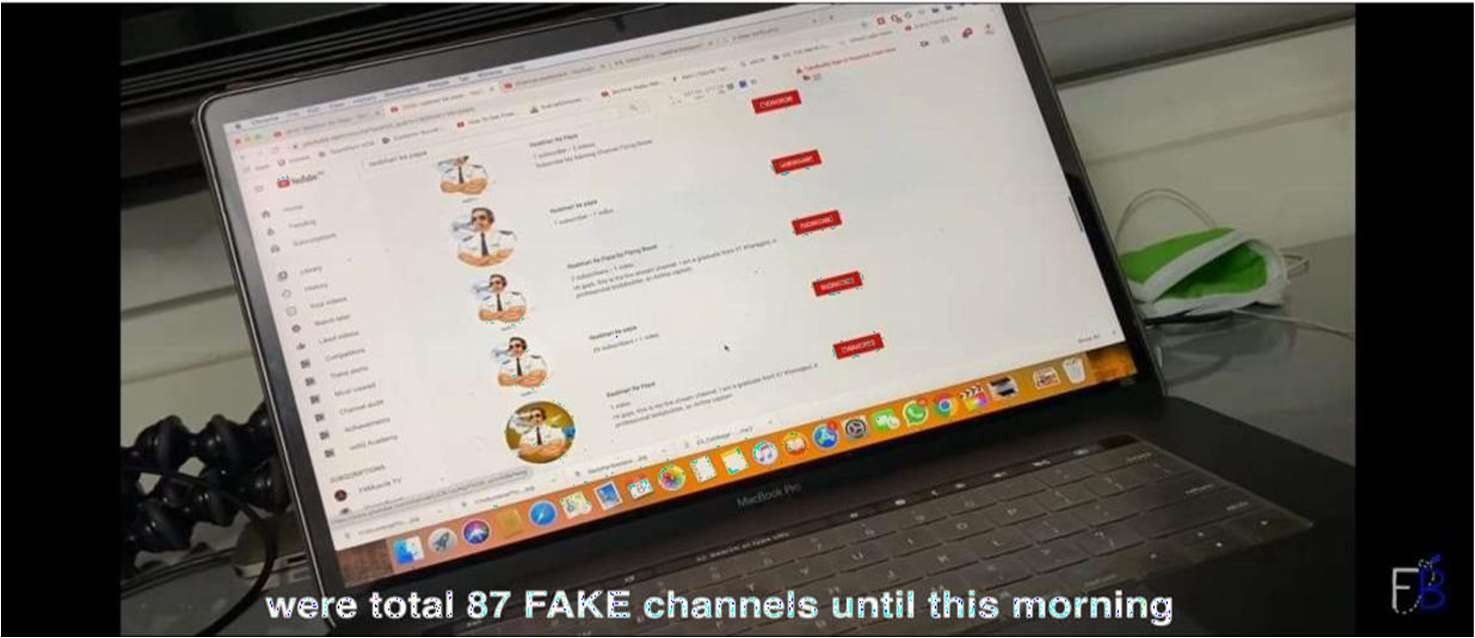


Ms. Reena Tripathi

## SUBMITTED BY:

Mrityunjay Mishra (2K19/IT/080) Naman Jain (2K19/IT/086)

# INTRODUCTION



The motivation to do this project came from the well-known YouTuber “Gaurav Taneja” whose channel names are “Flying Beast” and “Fit Muscle TV” respectively and currently holding around

6.93 million subscribers and counting.

In this Lockdown he launched his 3rd channel “Rasbhari Ke Papa”. This is a gaming channel; he launched this channel in afternoon and on next day he found around 87 fake replicas of his channel and it led to the subscriber in dilemma that which is the legit channel.

Instagram status of Gaurav Taneja

## So, we thought to make an analyzer which can analyze and detect fake replicas of YouTube channels. The main objectives were:

1. To explore the different attributes through which we can differentiate between the legit and fake channels using YouTube API.
2. To build a sustainable model using Machine Learning Models such as K Means Clustering and Artificial Neural Networks

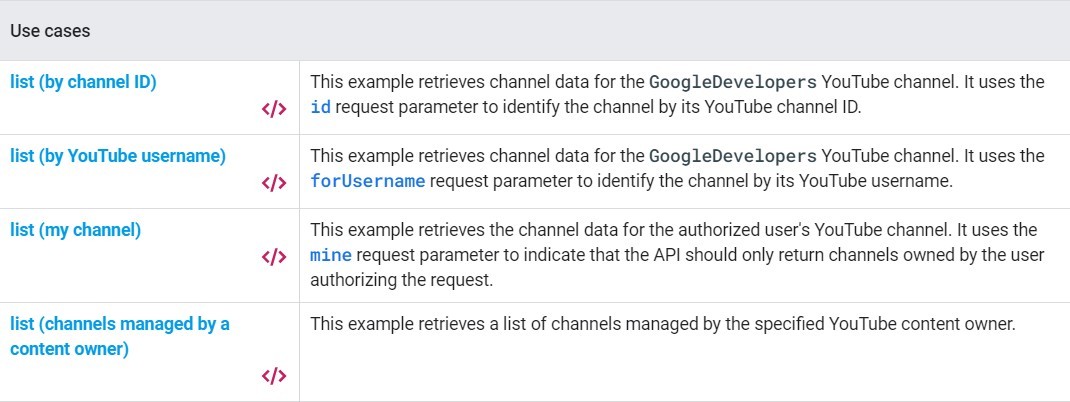
# IMPLEMENTATION DETAILS



## DATASETS

An informational index (or dataset) is an assortment of information. In this task we will make the dataset according to there is no dataset accessible on the web and the information is dynamic so we use YouTube API.

The YouTube Application Programming Interface (YouTube API) permits designers to get to video insights and YouTube channels information through two sorts of calls, REST and XML- RPC. Google portrays YouTube API Resources as "APIs and Tools that let you bring the YouTube experience to your website page, application, or gadget.



Channel details using YouTube API

## INDEPENDENT AND DEPENDENT VARIABLE

**INDEPENDENT VARIABLE** - Independent variables *(also referred to as feature)* are the input for a process that is being analyses.

**DEPENDENT VARIABLE -** Dependent variables are the output of the process.

|  |  |
| --- | --- |
| **INDEPENDENT VARIABLES** | **DEPENDENT VARIABLE** |
| 1. View Count 2. Subscribers Count 3. Videos Count 4. Date of Registration 5. Likes – Dislikes Ratio | 1. Label |



## PREPROCESSING AND DATA ANALYSIS TECHNIQUES

* Data cleansing to handle missing data by replacing them with the **mean value** of the particular attribute with respect to the whole dataset.

Encoding the dependent variable as:

Legit: 1

Fake: 0

Scaling the features using **normalization**.

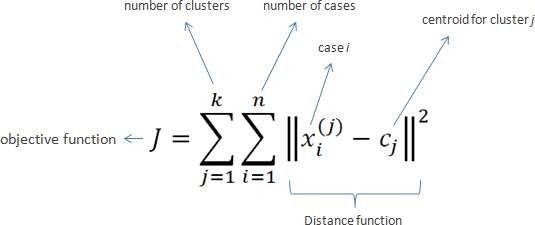
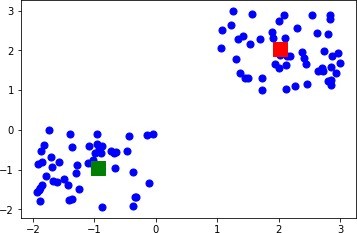
Data labelling using **K means clustering** and **neural network** for channel classification

## DATA PREPROCESSING TECHNIQUES

Steps involved in data pre-processing:

1. **Acquire the dataset**
2. **Import all the crucial libraries**
3. **Import the dataset**
4. **Identifying and handling the missing values**
5. **Encoding the categorical data**
6. **Feature scaling**

## DATA ANALYSIS TECHNIQUES



Our problem statement requires classification of data on basis of various features.

Since our data is not labelled hence, we have to use unsupervised learning technique such as K means clustering to label the data which will later be used for training our neural network.

## K means clustering

K-means clustering implies bunching is one of the least difficult and famous unaided AI calculations. Ordinarily, solo calculations make inductions from datasets utilizing just information vectors without alluding to known, or named, results.

To handle the learning information, the K- means clustering implies calculation in information mining begins with a first gathering of haphazardly chose centroids, which are utilized as the starting focuses for each bunch, and afterward performs iterative (tedious) estimations to enhance the places of the centroids It halts creating and optimizing clusters when either:

* + The centroids have stabilized-there is no change in their values because the clustering has been successful.
  + The defined number of iterations has been achieved.

## Euclidean distance + Centroid Calculation

**Artificial Neural Network**



Neural networks are mind-boggling models, which attempt to impersonate the manner in whichthe human cerebrum creates characterization rules. A neural net comprises a wide range of layers of neurons, with each layer accepting contributions from past layers, and passing yieldsto additional layers. The manner in which each layer yield turns into the contribution for the following layer relies upon the weight given to that particular connection, which relies upon the cost work, and the analyzer. The neural net emphasizes a foreordained number of cycles, called ages. After every age, the cost of work is broken down to see where the model could be improved. The advancing capacity at that point adjusts the inside mechanics of the organization, for example, the loads, and the predispositions, in light of the data gave by the cost work, until the cost work is limited.

We have utilized 1 shrouded layer and change the number of neurons present inside that layer to accomplish the greatest execution.

Our last layer has a solitary neuron comparing to the yield which relates to the validness of theinformation channel.

We have utilized sigmoid capacity as the initiation work for the yield of each layer.

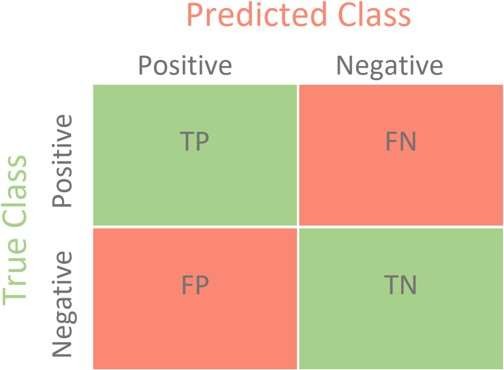
Bias is an additional parameter with an input value of 1 in the Neural Network which is used to adjust the output along with the weighted sum of the inputs to the neuron.

The cost function is the Cross-entropy function

Where y stands for the number we want to get from the network and y with a hat is the numberwe actually got by passing our example through the network.

Finally training the model on our dataset involves gradient descent of all the weights until the cost function is minimized.

# EXPERIMENTAL DETAILS

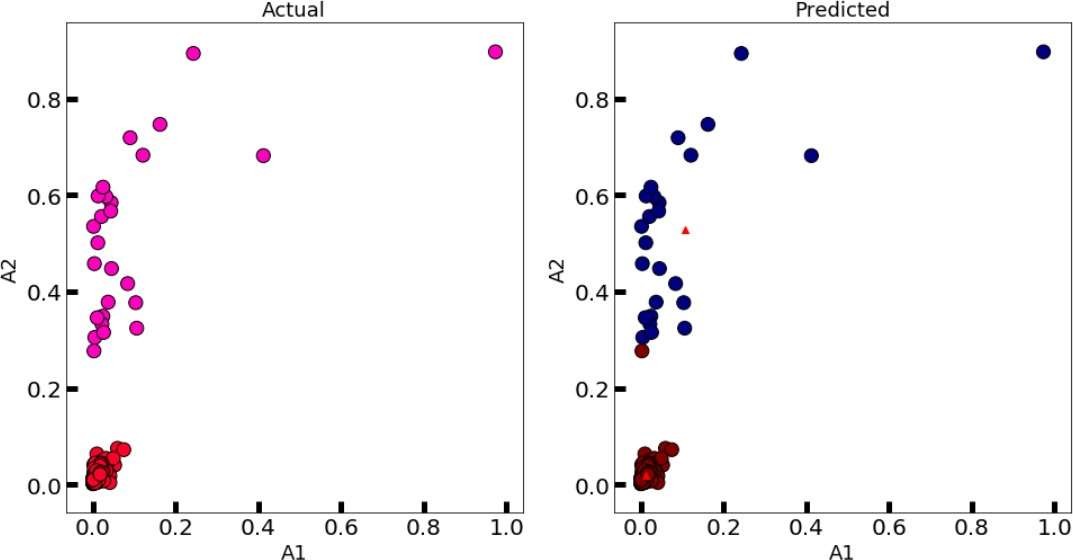


## PERFORMANCE MEASURE

We will use confusion matrix-based metrics which are mentioned below:

1. Accuracy:
2. Precision:
3. Sensitivity:
4. Specificity:
5. F measure:

## K-MEANS CLUSTERING:



**ARTIFICIAL NEURAL NETWORK:**

**For 200 epochs** which is the default value, the model got overfitted.Accuracy: 97.89%

Precision: 93.33%

Sensitivity: 93.33%

Specificity: 98.75%

F-measure: 0.933

**For 100 epochs** the results were satisfactory.Accuracy: 98.42%

Precision: 100%

Sensitivity: 90%

Specificity: 100%

F-measure: 0.937

**For 150 epochs** the performance measures are:Accuracy: 98.42%

Precision: 96.55%

Sensitivity: 93.33%

Specificity: 99.37%

F-measure: 0.949

Since F-measure for 150 epochs is greater than that for 100, hence we choose 150 as the number of epochs while training the model



## Confusion Matrix for Alpha = 0.01,

**Epochs = 150**

**CONFUSION MATRIX FOR NEURAL NETWORK:**



# DISCUSSION



We have successfully implemented our project with fairly high accuracy for both themodels i.e., K means clustering and the neural networks.

The only thing that can be added is more dataset as “more the dataset, better the model”. We have used around 200 rows because we need to add each data manually (it was the onlyoption with YouTube API) but if somehow, we manage to automate this process, we canget a large collection of data which can further increase our accuracy.

This model can also be used by YouTube to ban the fake channels which is a big problem as these fake channels earn a lot without any efforts by just copying the content and the actual video creator gets nothing.

# REFERENCES

## https://cse.iitkgp.ac.in/~pawang/papers/ecir19\_youtube.pdf

1. **https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation- methods-and-drawbacks-aa03e644b48a**
2. **https://wiki.pathmind.com/neural-network**