Project On Machine Learning

Topic Idea: Trending YouTube Video Statistics

Group Name: Data Pirates

Group Members: Shahrin Shafiq Simran - 20101358

Arafat Rahman - 20101121

Dilshad Jahan - 18101453

Tanzina Binte Azad - 20201217

Submitted to: Benjir Islam Alvee

Monirul Haque

Introduction:

YouTube trending videos reflect material that attracts viewers' attention over a very short period of time and has the potential to become popular, in contrast to popular videos, which would have already attained substantial viewership figures by the time they are labeled popular. YouTube popular videos have not been properly investigated or evaluated despite their significance and exposure. We measure, examine, and compare several important components of YouTube popular videos in this research. Our analysis is based on gathering and tracking high-resolution time-series data on over 8,000 YouTube videos for a total of nine months to determine how many people watched each video and other associated data. We are able to study popular videos' time-series across crucial domains because they are branded as trending within a few hours after they are uploaded. Youtube is a popular website that maintains a list of the top trending videos. These trending videos are categorized in terms of views, likes, dislikes and comments. In this machine learning project we chose this dataset and this dataset is a daily record of the top trending videos. The data in our dataset is divided into categories based on the number of views, likes, dislikes and comments. We will be able to analyze what factors affect how trending a video will be. We used a classifier to test the accuracy of classifying the rest test into the mentioned categories.

Methodology:

This dataset contains data on daily trending YouTube videos for several months (and counting) of data on daily trending YouTube videos. Data is provided for the US, GB, DE, CA and FR regions with up to 200 trending videos listed each day. Now includes data from the RU, MX, KR, JP and IN regions for the same time period. Each region's data is stored in a separate file. The dataset was collected using the YouTube API.

The process where most of the issues in the data are being cleaned and solved is the Data Pre-processing step. This has various methods. In this project, we tried to remove null values by data cleaning technique but there were no null values. Then removed outlets, did scale the data by Standard Scaler, then took the feature engineering approach to create better features. After that feature selection has been used to select the most important variables.

Data Preprocessing:

- Importing Libraies: Numpy, Pandas, Scikit Learn, Seaborn, Matplotlib
- Dropping Useless Columns
- Removing Null Values
- Visualize: Scatter Matrix
- Scaling: Standard Scaler
- Splitting Data: Train and Test Set

Here, we used 4 models.

1. **Naive Bayes Classifier:** The term "Naive bayes classifiers" refers to a set of classification algorithms based on Bayes' Theorem. It is a family of algorithms that all share a common principle, namely that every pair of features being classified is independent of each other.

2. **K- Neighbors Classifier:** It is an instant-based, non-parametric learning method. The classifier in this method learns from the instances in the training dataset and classifies new input using previously measured scores.

3. **Random Forest Classifier:** A random forest classifier is a supervised learning algorithm that can be used to solve regression and classification problems. Because of its high flexibility and ease of implementation, it is one of the most used ML algorithms.

4. **Neural Network Classifier:** Neural nets are ins[ired by the learning process that occurs in human brains. They are made up of an artificial network of functions known as parameters that allow the computer to learn and fine-tune itself by analyzing new data.

These mentioned approaches are followed for train and testing accuracy.

RESULT

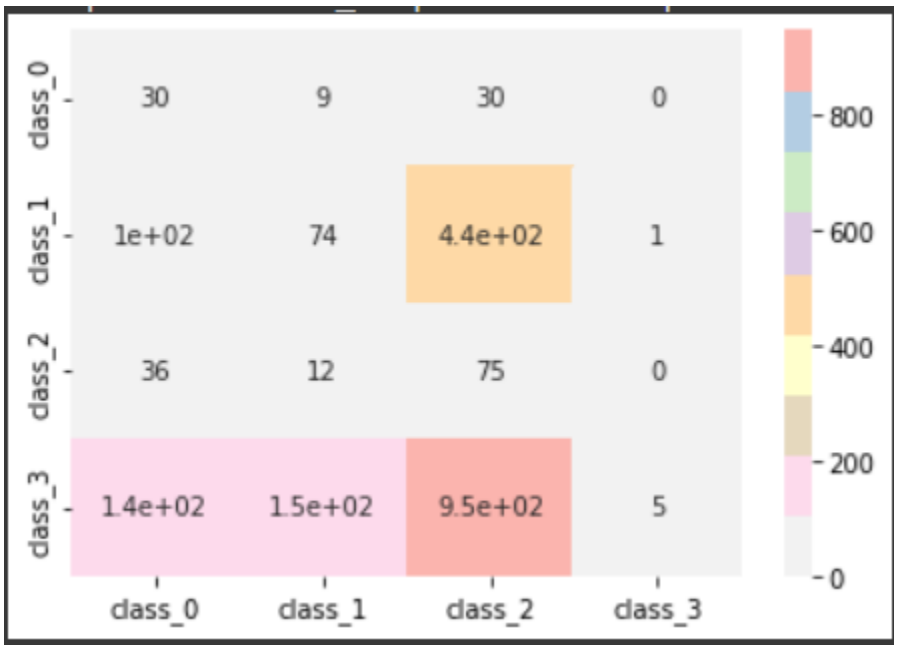
1. **Naive Bayes classifier:**

Training accuracy of model is 0.11
Testing accuracy of model is 0.15

Confusion matrix:

[38 3 41 0]]
[92 76 380 1]
[60 22 109 0]
117 144 956 5]

Heat map:



Classification Report:

	precision	recall	f1-score	support
1	0.1173	0.4138	0.1827	87
2	0.3020	0.1201	0.1719	616
3	0.0518	0.6364	0.0958	121
4	1.0000	0.0049	0.0098	1220
accuracy			0.0944	2044
macro avg	0.3678	0.2938	0.1151	2044
weighted avg	0.6960	0.0944	0.0711	2044

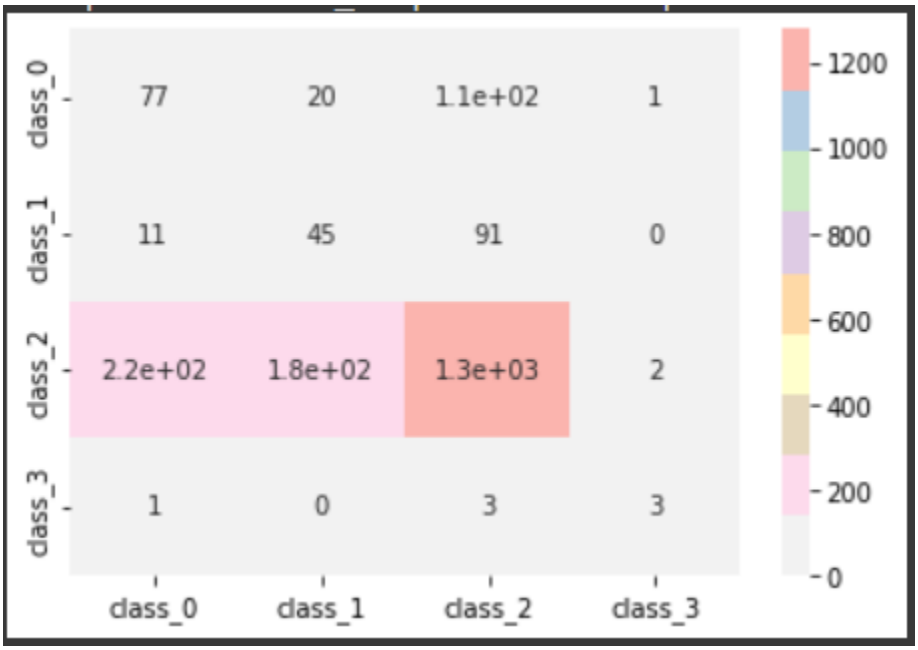
2. K-neighbors classifier:

Training accuracy of the model is 0.79
Testing accuracy of the model is 0.71

Confusion matrix:

```
[[89   13   112    1]
 [18   67   75     0]
 [200  165  1296   4]
 [0    0    3     1]
```

Heat map:



Classification Report:

	precision	recall	f1-score	support
1	0.2541	0.3920	0.3083	199
2	0.2327	0.3519	0.2801	162
3	0.8735	0.7717	0.8194	1682
4	0.0000	0.0000	0.0000	1
accuracy			0.7011	2044
macro avg	0.3401	0.3789	0.3520	2044
weighted avg	0.7620	0.7011	0.7265	2044

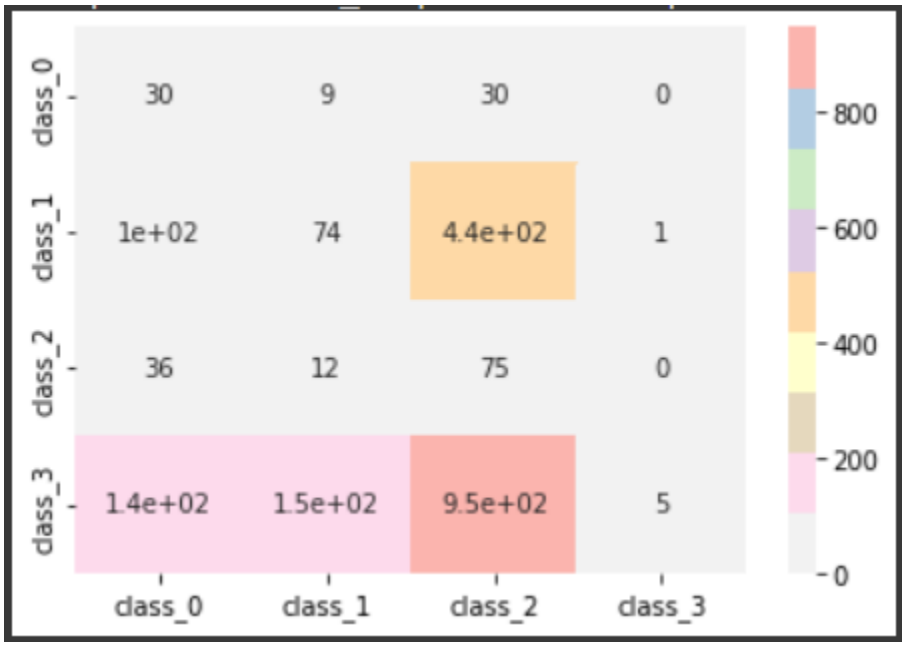
3. **Random forest Classifier:**

Training accuracy of the model is 1.00
Testing accuracy of the model is 0.70

Confusion matrix:

```
[[3      0     4      0]
 [1      0     2      0]
 [ 303   245   1480   6]
 [0      0     0      0]
```

Heat Map:



Classification Report:

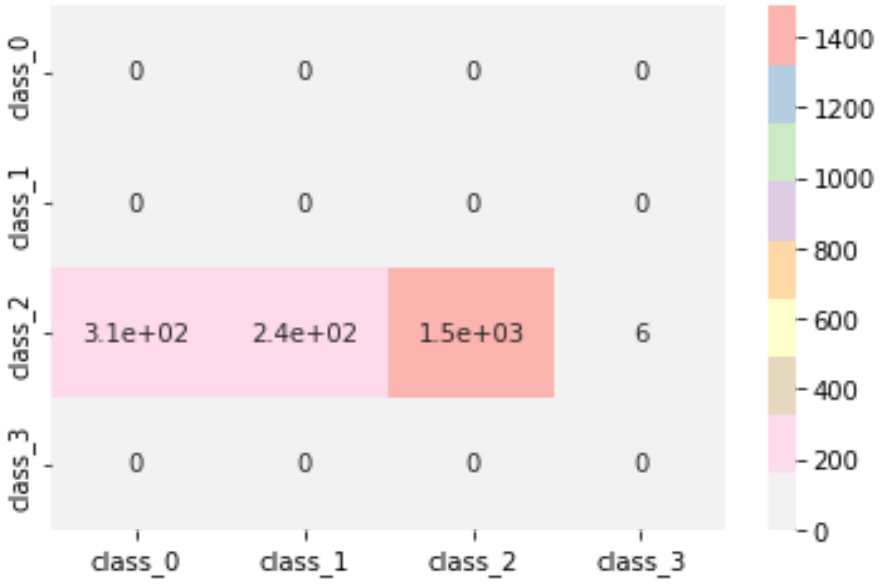
	precision	recall	f1-score	support
1	0.0261	0.1212	0.0429	66
2	0.0000	0.0000	0.0000	0
3	0.9724	0.7305	0.8343	1978
4	0.0000	0.0000	0.0000	0
accuracy			0.7109	2044
macro avg	0.2496	0.2129	0.2193	2044
weighted avg	0.9419	0.7109	0.8087	2044

4. Neural Network Classifier:

The Training accuracy of the model is 0.74
The Testing accuracy of the model is 0.74

Confusion matrix:

```
[[   0    0    0    0]
 [   0    0    0    0]
 [ 307  245 1486    6]
 [   0    0    0    0]]
```

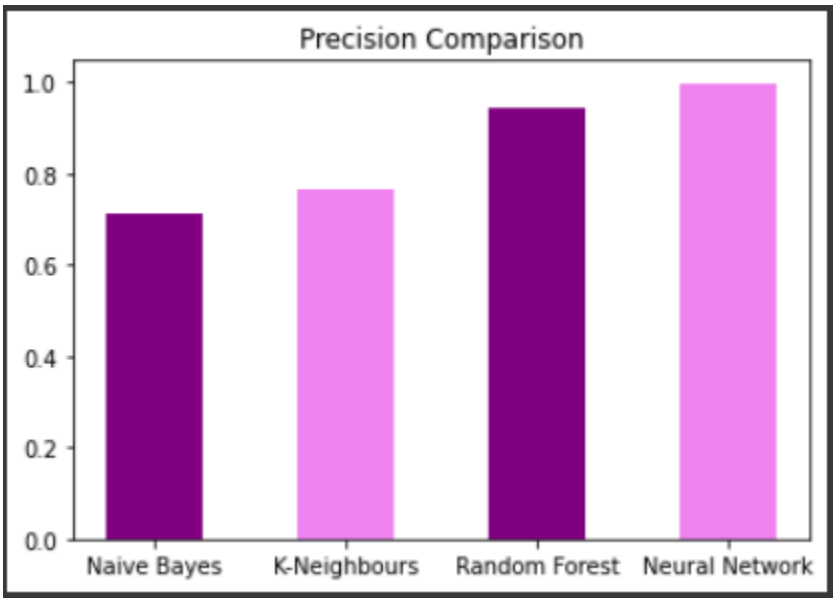


Classification Report:

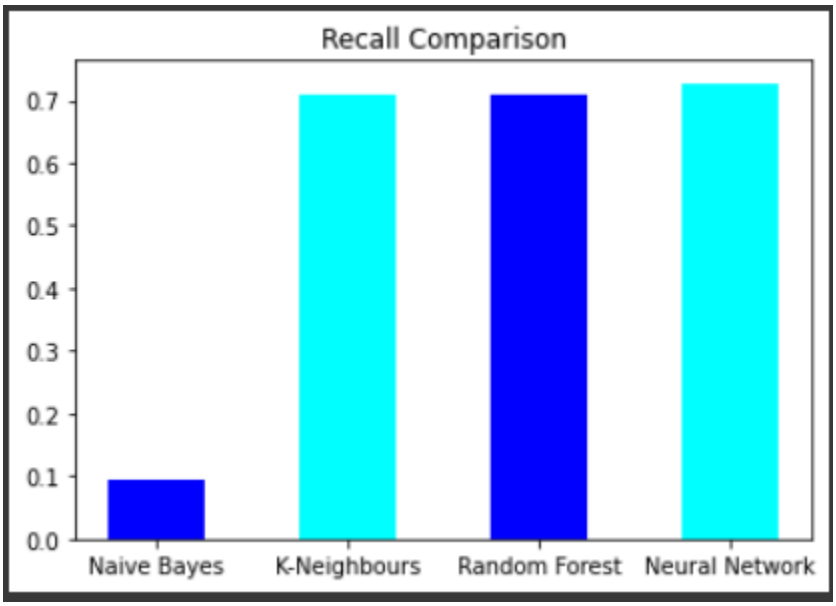
	precision	recall	f1-score	support
1	0.0000	0.0000	0.0000	0
2	0.0000	0.0000	0.0000	0
3	1.0000	0.7270	0.8419	2044
4	0.0000	0.0000	0.0000	0
accuracy			0.7270	2044
macro avg	0.2500	0.1818	0.2105	2044
weighted avg	1.0000	0.7270	0.8419	2044

ANALYSIS

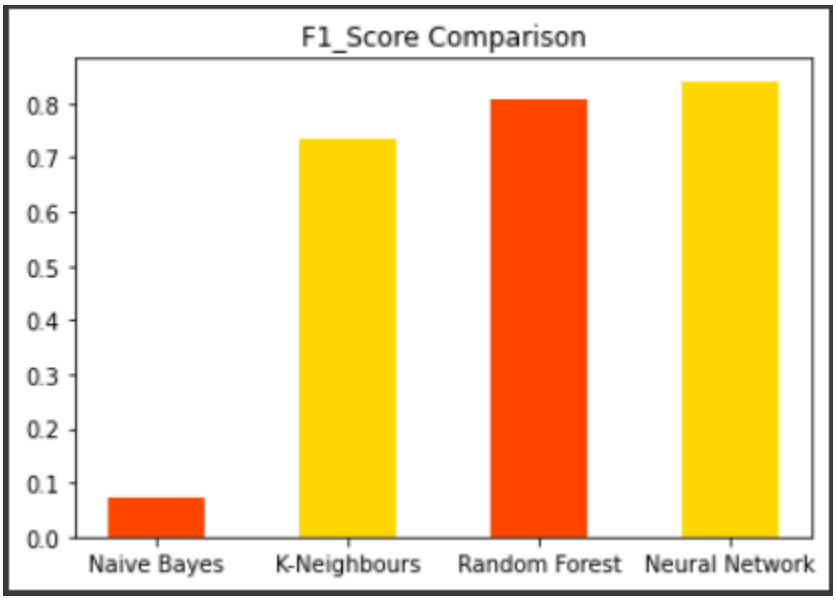
Precision: Precision, or the caliber of a successful prediction produced by the model, is one measure of the model's performance. Precision is calculated by dividing the total number of positive predictions by the proportion of genuine positives (i.e., the number of true positives plus the number of false positives). For instance, in a customer attrition model, accuracy is the ratio of the total number of consumers the model properly anticipated would unsubscribe to the number of customers who actually did so.



Recall: The recall is determined as the proportion of Positive samples that were properly identified as Positive to all Positive samples. The recall gauges how well the model can identify Positive samples. The more positive samples that are identified, the larger the recall. Only the classification of the positive samples is important to the recall. This is unrelated to the manner in which the negative samples are categorized, such as for accuracy. Even if all of the negative samples were mistakenly categorized as Positive, the recall will be 100% when the model labels all of the positive samples as Positive.



F1 Score: The F-score is used in statistical analysis of binary classification as a gauge of test precision. It is derived from the test's precision and recall, where precision is the proportion of true positive results to all positive results, and recall is the proportion of real positive findings to all samples that should be classified as positive. In binary classification, recall is sometimes referred to as sensitivity, while precision is also known as positive predictive value. The harmonic mean of the accuracy and recall is the F1 score. An F-score can have a maximum value of 1.0, which denotes perfect accuracy and recall, and a minimum value of 0, which occurs when either precision or recall are zero.



Reference:

1.Organization, 2019, June 3. Trending YouTube video statistics. https://www.kaggle.com/datasets/datasnae

2.*(PDF) trending videos: Measurement and analysis - researchgate*. (n.d.). Retrieved September 3, 2022, from https://www.researchgate.net/publication/266262149_Trending_Videos_Measurement_and_Analysis

3.Google. (n.d.). YouTube trending video view statistics - think with google. Google. Retrieved September 3, 2022, from https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/youtube-trending-video-views/

4.YouTube. (n.d.). *YouTube*. Retrieved September 3, 2022, from https://www.youtube.com/trends/records/.

5.*Analysis on YouTube trending videos - IRJET-international research ...* (n.d.). Retrieved September 3, 2022, from https://www.irjet.net/archives/V7/i8/IRJET-V7I8732.pdf