A black and yellow sign with blue text

Description automatically generated

College of Engineering and Computing Science

**Project Report**

**Top Subscriber on YouTube**

NIRAJ RAJ

Term: Summer 2023

DTSC 701 – Introduction to Big Data

Professor:

Liangwen Wu

**Table of Content**

1. Introduction…………………………………………….………….3
2. Dataset overview……………………………………….………….3
3. Channel Category distribution, KNN and Confusion Matrix….….4
4. AWS spark cluster setup……………………………………..........5
5. SSH access…………………………………………………...........7
6. The final output of Spark……………………………………….…8
7. Conclusion………………………………………………………...9
8. **Introduction**

In the digital age, YouTube has emerged as a global platform where content creators, artists, and organizations showcase their creativity, share knowledge, and engage with audiences on an unprecedented scale. With millions of channels and billions of videos, YouTube represents a rich landscape of diverse content genres, captivating millions of viewers around the world. As part of this vibrant ecosystem, understanding the characteristics and dynamics of YouTube channels becomes pivotal.

This report embarks on an exploratory journey through the "Most Subscribed 1000 YouTube Channels" dataset, meticulously curated from the vast YouTube universe and sourced from Kaggle. With an unwavering focus on data-driven insights, the analysis encompasses two key pillars: the intricate world of channel categorization and the strategic implementation of machine learning techniques, specifically K-Nearest Neighbors (KNN) classification. Through the fusion of these analytical facets, we aim to illuminate the essence of YouTube content diversity and demonstrate the potential of intelligent algorithms in categorization tasks.

1. **Dataset Overview**

The dataset under investigation includes a broad view of the top 1000 YouTube channels with the most subscribers. Each entry in the dataset contains a wealth of channel-specific information, such as rank, title, subscriber count, video views, video count, categorical classification, and the channel's foundational year of creation. The foundation of our analysis is made up of this collection of characteristics.

Link: [https://www.kaggle.com/datasets/themrityunjaypathak/most-subscribed-1000-youtube-channels.](https://www.kaggle.com/datasets/themrityunjaypathak/most-subscribed-1000-youtube-channels" \t "_new)

1. **Channel Categories Distribution**

In the age of digital media, channel categorization serves as an illuminating compass, navigating users through the labyrinth of content offerings. The elucidation of category distribution is akin to unraveling the tapestry of human interests and preferences interwoven across the YouTube spectrum. By meticulously sifting through the dataset and deploying visualizations, we gleaned insights into the ebbs and flows of content genres that captivate the YouTube audience.

A screen shot of a computer

Description automatically generated

**K-Nearest Neighbors (KNN) Classification & Confusion Matrix**

In an era dominated by digital prowess, the application of machine learning algorithms breathes life into data, transforming it from mere numbers into predictive insights. The K-Nearest Neighbors (KNN) algorithm emerges as a torchbearer in the realm of classification. Our voyage into the application of KNN unfolds as we harness its power to predict channel categories based on features such as subscriber count, video views, and video count.

The ensuing bar chart orchestrates the dance of channel categories, echoing the rhythmic cadence of subscriber preferences. The orchestra is led by the crescendo of the "Music" category, harmonizing with the serenades of "Shows," "Entertainment," and "Education," each revealing a unique note in the symphony of YouTube.

A screen shot of a computer

Description automatically generated

Python code is on GitHub.

GitHub Link:

1. **AWS Spark setup**

The canvas grows beyond the confines of local processing as we go toward scalability and computational power. Utilizing cloud resources, the AWS Spark cluster spreads its wings and orchestrates the dance of data analysis. We entered the world of distributed computing, where data flows in torrents and insights come from the convergence of processing nodes, with the deployment of three instances and the magic of Spark setup.

1. Visit the AWS Website: Go to the official Amazon Web Services website at [https://aws.amazon.com/](https://aws.amazon.com/" \t "_new)
2. Sign Up: Click "Create an AWS Account" and follow the instructions. You'll need to give your email, password, and billing info.
3. Verify Identity: Complete identity verification, which might include using a phone number for a verification code.
4. Set Up Payment: Enter payment details to enable billing for your AWS account.
5. Accessing EMR: Access the "Amazon EMR" service through the AWS Management Console. Initiating Cluster Creation: Initiate the process by selecting the "Create cluster" button, which will lead you to the configuration of your EMR cluster.
6. Defining Cluster Settings: Assign a suitable name to your cluster and specify your preferred Spark application and version. Choosing Instance Types: Opt for 3 instance types for both master and core nodes.
7. Determining Instance Count: Indicate the desired number of core instances, in your case, select 3 instances. Configuring Security and Access: Set up security groups, opt for an EC2 key pair (if SSH access is needed), and configure other access parameters.
8. Commencing Cluster Launch: After reviewing your configurations, finalize by selecting the "Create cluster" button, initiating the launch of your Spark cluster.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **SSH Access**
2. SSH Key: If you configured an EC2 key pair during cluster creation, you can use it for SSH access.
3. Terminal: On your local machine, open a terminal window.

A screenshot of a computer

Description automatically generated

1. Navigate to Key Pair: If you created a new EC2 key pair, navigate to the directory where you saved the private key file (.pem) on your local machine.
2. Change Key Permissions: To ensure secure access, restrict the permissions of the private key file:

**chmod 400 mykeypair.pem**

1. SH Command: Use the SSH command to establish an SSH connection to one of the instances. Replace your-key-pair.pem with the actual path to your private key file and public-ip with the instance's public IP address:

[ssh -i ~/mykeypair.pemhadoop@ec2-3-15-222-135.us-east-2.compute.amazonaws.com](mailto:ssh -i ~/mykeypair.pemhadoop@ec2-3-15-222-135.us-east-2.compute.amazonaws.com)

A screenshot of a computer

Description automatically generated

1. Connected: You are now connected to the instance via SSH. You can use this terminal to interact with the instance, deploy Spark jobs, and manage the cluster.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Conclusion**

In this symphony of analysis, we traverse the expanse of channel categorization and machine learning classification through the lens of the "Most Subscribed 1000 YouTube Channels" dataset. With fervor, we unravel the intricacies of category distribution, witness the dance of KNN classification, and explore the integration of AWS Spark clusters and SSH access. Our analysis transcends numbers to unveil the narratives that shape the YouTube realm—a realm where data, intelligence, and creativity harmonize to orchestrate a digital odyssey.