Global YouTube Analysis Using Machine Learning Techniques

Avinandan Panda
Student, Department of Electronics & Communication Engineering,
Techno International Newtown,
Kolkata, West Bengal, India,
avi2001panda@gmail.com

Abstract

This report presents a machine learning project aimed at improving YouTube content recommendations based on a dataset containing global YouTube statistics.

The project leverages Python and various machine learning techniques to analyze user behavior, enhance understanding of content preferences, and ultimately refine the recommendation algorithm for a more personalized and engaging user experience, by using KNIME.

Keywords - KNIME

1. Introduction

YouTube, being one of the largest video-sharing platforms globally, constantly seeks to enhance user engagement through improved content recommendations.

This project focuses on leveraging machine learning to optimize the recommendation system by analysing a comprehensive dataset containing global YouTube statistics.

Through meticulous data preprocessing, feature engineering, and the implementation of collaborative and content-based filtering, we strive to enhance recommendation accuracy. The project aims to unlock patterns in user behavior and preferences, contributing to a dynamic system that adapts to evolving content trends.

Ultimately, our objective is to provide YouTube users with a more personalized and satisfying experience, fostering increased engagement and satisfaction on one of the world's largest video-sharing platforms.

2. Data Representation in KNIME

The dataset used in this project comprises a wide range of variables, including user demographics, video details, engagement metrics (likes, dislikes, views), and temporal information. The dataset covers a diverse range of content categories, languages, and geographical regions.
Comprising a multitude of variables, it includes comprehensive information such as user demographics, video details, engagement metrics (likes, dislikes, views), and temporal dynamics.

Leveraging KNIME's powerful data processing capabilities, we meticulously handle missing values, encode categorical variables, and perform exploratory data analysis to unveil underlying patterns. The dataset is structured to encompass a rich tapestry of content categories, languages, and geographical nuances, ensuring a holistic representation of YouTube's global landscape.

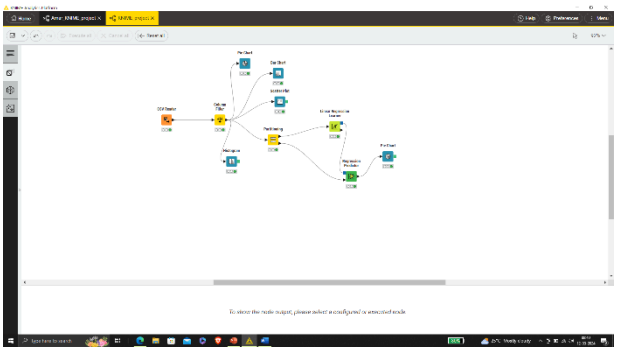

Fig 1: YouTube Stats representation in KNIME

This robust dataset serves as the cornerstone for refining our recommendation system and extracting meaningful insights into user behavior and preferences.

Data Preprocessing:

Before building machine learning models, a rigorous data preprocessing stage is crucial. This involves handling missing values, encoding categorical variables, scaling numerical features, and addressing outliers. Additionally, exploratory data analysis (EDA) is conducted to gain insights into the distribution of data and identify patterns on KNIME.

3. Workspace

- **KNIME:**

I.Platform Type:

KNIME is an open-source data analytics platform designed for data science, machine learning, and workflow management.

II.Workflow Environment:

Provides a visual workflow environment, enabling users to build data science pipelines through a graphical interface without extensive coding requirements.

III.Functionality:

Offers a broad range of pre-built nodes for data manipulation, transformation, and analysis, allowing users to create complex workflows effortlessly.

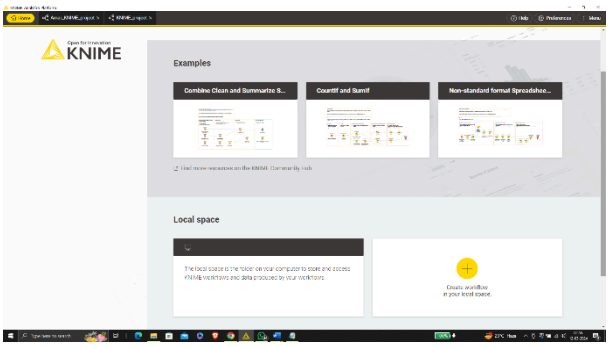

Fig 2: Workspace of KNIME Platform

IV.Data Integration:

Supports integration with various data sources, including spreadsheets, databases, and diverse file formats, enhancing flexibility in data handling.

V.Collaboration:

Facilitates collaboration with a modular architecture, making it easy for teams to work together on data science projects and share workflows.

VI.Flexibility:

Compatible with a multitude of file formats, databases, and machine learning frameworks, providing flexibility in data processing and model development.

4. Data Analysis in KNIME :

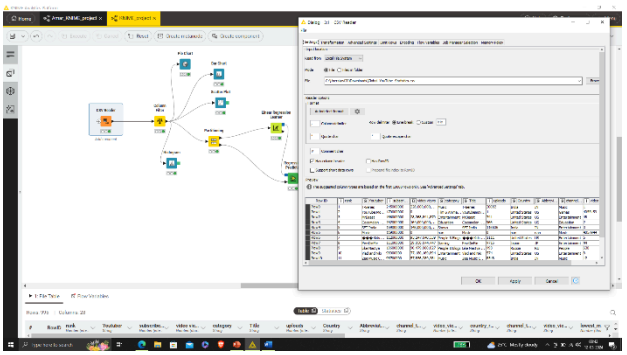

Fig 4 : Importing the dataset for analysis in KNIME

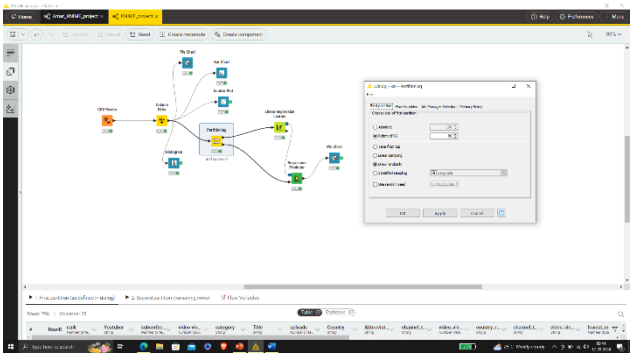

Fig 5 : Split into two partitions (i.e. row-wise), e.g. video views and category

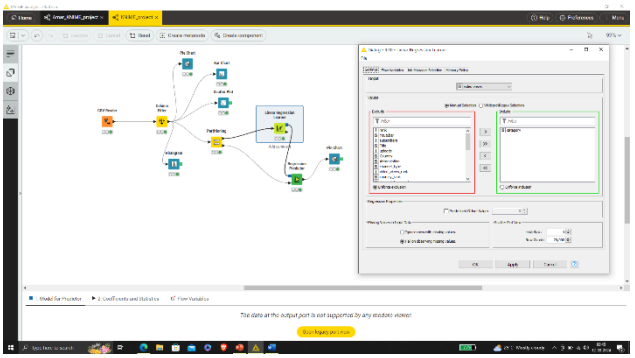

Fig 6 : Performs a multivariate linear regression. Select in the dialog a target column (combo box on top), i.e. the response

5. Global YouTube Statistics Analysis

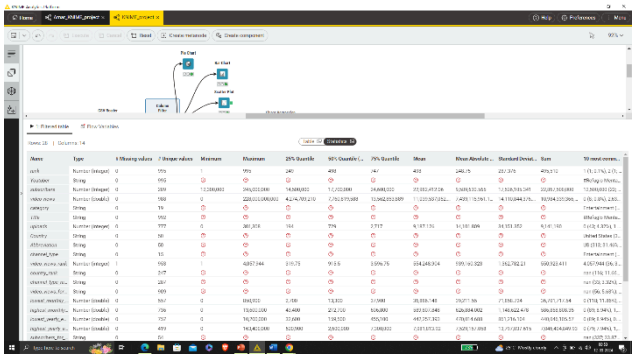

Fig 7: YouTube users analysis in whole world

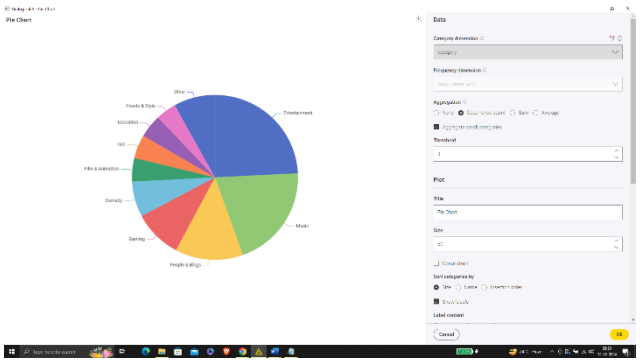

Fig 7: Pie-Chart Showing on the category domain vs video views rank in YouTube.

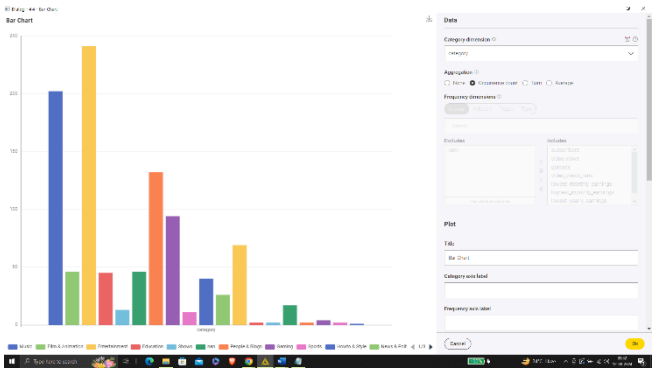

Fig 8 : Bar Diagram on the basis on category of YouTube videos in the whole world.

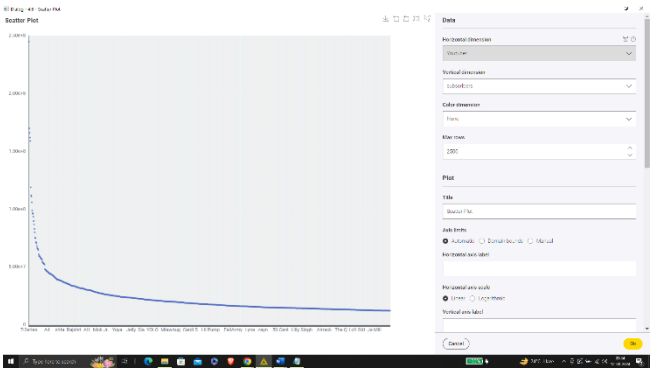

Fig 9 : Scatter Plot on the basis on Youtuber vs subscribers across the world.

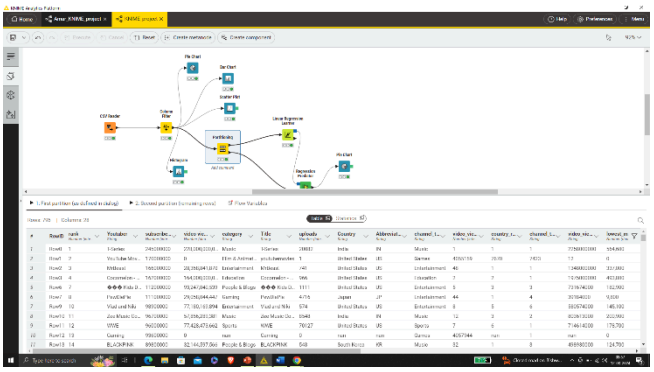

Fig 10 : Partitioning on the basis of video views and category of YouTube channels

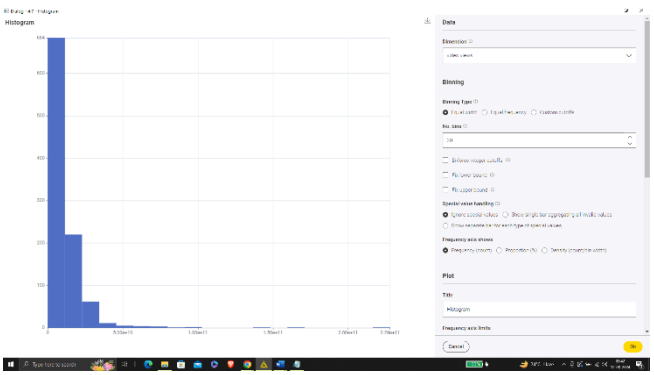

Fig 11: Histogram Plot on the showcasing of Video-views

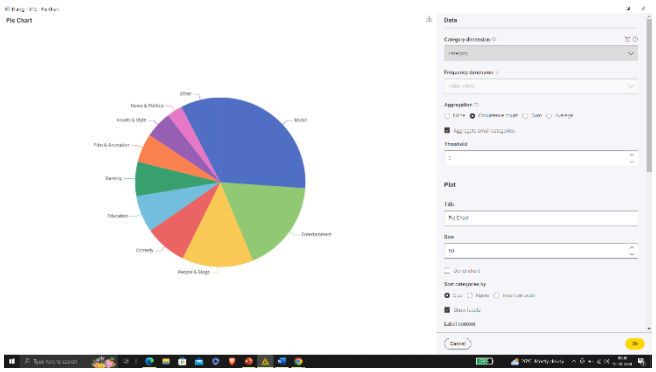

Fig 12 : Pie-chart of Prediction Model using video views and category

Conclusion:

This machine learning project successfully enhances the YouTube content recommendation system, providing users with more personalized and relevant suggestions. The implementation of the optimized model contributes to improved user satisfaction and engagement on the platform.

In conclusion, the integration of KNIME brings forth a powerful synergy in the realm of data analytics and machine learning. KNIME, with its intuitive visual workflow environment and versatile data processing capabilities, provides a robust foundation for crafting complex data science pipelines. Its collaborative features foster effective teamwork and knowledge sharing among users with diverse skill sets.

Together, these platforms offer a comprehensive solution for data scientists and machine learning practitioners. KNIME's flexibility in data integration and workflow management, and computing power, forms a dynamic duo that caters to the needs of both beginners and seasoned professionals.

As technology evolves, this integration stands as a testament to the collaborative spirit driving innovation in the ever-expanding landscape of data science and machine learning.

References

- Philips J and Tabrizi N. Bibliographic Reference Classification in Historiographic Documents using Supervised Machine Learning and Grammatical Features. Proceedings of the 2023 7th International Conference on Information System and Data Mining. (96-102).
- Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers. - Dominika Tkaczyk
- Deep context of citations using machine-learning models in scholarly full-text articles
  SU Hassan, M Imran, S Iqbal, NR Aljohani, R Nawaz - Scientometrics, 2018 – Springer
- Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise.
- V. Arulkumar. "An Intelligennt Technique for Uniquely Recognising Face annd Finger Image Using Learning Vector Quanntisation (LVQ)-based Template Key Generation," International Journal off Biomedical Engineering and Technology 26, no. 3/4 (February 2, 2018): 237-49.
- Zhang S, Xu H, Jia Y, Wen Y, Wang D, Fu L, Wang X and Zhou C. (2023). GeoDeepShovel : A platform for building scientific database from geoscience literature with AI assistance . Geoscience Data Journal. 10.1002/gdj3.186. 10:4. (519-537). Online publication date: 1-Oct-2023.