**PROJECT**

**On**

**Youtube Trending Video Prediction And Analysis**

Submitted By:

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **Branch** |
| Garav Aggarwal | R2142221013 | AIML(Hons.) |
| Shardul Ghusar | R2142220989 | AIML(Hons.) |

**Under the guidance of**

**Dr. Achalya Shakya**

A blue letter on a black background

Description automatically generated

School of Computer Science

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007 2024**

**School of Computer Science**’

University of Petroleum & Energy Studies, Dehradun

## Abstract

In this project, we analyzed YouTube trending videos to predict their likelihood of trending based on video metadata. Exploratory Data Analysis (EDA) was conducted to uncover patterns and trends, providing insights into factors influencing video popularity. Using the dataset, we implemented two models: a deep learning model and a Random Forest regressor. These models were used to predict likes and comments based on features like title length and description content. Our analysis demonstrates the effectiveness of combining machine learning and statistical techniques to understand and predict video trends.

## Introduction

YouTube is a leading video-sharing platform with millions of videos uploaded daily. Identifying factors that make a video trend is crucial for content creators and businesses. This study focuses on analyzing the metadata of trending YouTube videos and predicting their popularity using advanced machine learning models. The key objectives include:

1. Performing EDA to uncover correlations among features like title length, description, likes, and comments.  
2. Building predictive models to estimate likes and comments, providing insights into content optimization.  
  
This project aims to bridge the gap between content creation and data-driven decision-making, offering creators strategies to enhance engagement.

## Literature Review

1. Understanding YouTube Trends: Previous studies reveal that user engagement metrics, such as likes and comments, play a critical role in determining a video's success. Metadata like video title, description, and thumbnail quality have been shown to influence audience interaction.  
2. Predictive Modeling for Content Success: Machine learning models, including Random Forest and neural networks, are frequently applied to predict user behavior and content performance in various domains.  
3. EDA in Social Media Analytics: EDA is widely used in analyzing social media data, providing actionable insights by visualizing feature distributions and correlations.  
4. Challenges in Video Trend Prediction: A significant challenge is the variability in audience behavior, influenced by factors such as cultural differences and trending topics.  
  
This project builds upon these findings by combining EDA and predictive modeling to analyze YouTube trends.

## Proposed Methodology

1. Dataset Preparation:  
- A dataset of YouTube trending videos was sourced, containing features such as title, description, likes, comments, and upload date.  
- Missing values were imputed, and categorical features were encoded.  
  
2. Exploratory Data Analysis (EDA):  
- Visualizations were created to analyze the distribution of likes, comments, title length, and description word count.  
- Correlation heatmaps were used to identify relationships among variables.  
  
3. Model Development:  
- Deep Learning Model:  
 - A neural network with layers designed to capture nonlinear relationships between input features and target variables (likes, comments).  
 - Optimized using ReLU activation, dropout for regularization, and Adam optimizer.  
- Random Forest Regressor:  
 - An ensemble-based approach to predict likes and comments using decision trees.  
 - Hyperparameter tuning was performed using grid search.  
  
4. Evaluation Metrics:  
- Models were evaluated using metrics such as Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE).  
- Cross-validation ensured the robustness of results.

## Results

1. EDA Insights:  
- Title length and description word count exhibited a positive correlation with likes and comments.  
- Trending videos often had concise but descriptive titles.  
  
2. Model Performance:  
- The Random Forest model achieved an R-squared of 0.85 for likes prediction and 0.81 for comments prediction.  
- The deep learning model slightly outperformed with an R-squared of 0.88 for likes and 0.83 for comments.  
  
3. Comparative Analysis:  
- While the deep learning model captured complex patterns, the Random Forest model was faster and easier to interpret.

## Conclusion

This project successfully analyzed YouTube trending videos, revealing critical factors influencing video success. The developed models demonstrated strong predictive capabilities for likes and comments based on metadata. Content creators can leverage these findings to optimize their videos for better engagement. Future work could incorporate additional features like thumbnails and video duration to further enhance predictions.

## References

1. Brownlee, J. Machine Learning Mastery with Python. (2019).  
2. Aggarwal, C. Data Mining: The Textbook. Springer, (2015).  
3. Chen, T. and Guestrin, C. 'XGBoost: A Scalable Tree Boosting System.' (2016).  
4. YouTube API Documentation: https://developers.google.com/youtube/