Objective

To analyze, categorize, and gain insights into YouTube channels to enhance strategic decision-making for marketing, content strategy, and partnerships.

Market Segmentation:

Classify YouTube channels based on content nature and viewer engagement to tailor marketing strategies effectively.

Content Optimization:

Identify content themes and trends across different clusters of channels to guide creators on potential areas for content diversification or specialization.

Partnership Identification:

Determine suitable channels for brand partnerships, sponsorships, and collaborations based on their content focus and audience reach.

.

Inputs

Rrank: Position of the YouTube channel based on the number of subscribers
Youtuber: Name of the YouTube channel
Channel_contents: Succinct description of the channel contents.
Primary_language: primary language of the channel
subscribers: Number of subscribers to the channel
video views: Total views across all videos on the channel
category: Category or niche of the channel
Title: Title of the YouTube channel
uploads: Total number of videos uploaded on the channel
Country: Country where the YouTube channel originates
Abbreviation: Abbreviation of the country
channel_type: Type of the YouTube channel (e.g., individual, brand)
video_views_rank: Ranking of the channel based on total video views
country_rank: Ranking of the channel based on the number of subscribers within its country
channel_type_rank: Ranking of the channel based on its type (individual or brand)
video_views_for_the_last_30_days: Total video views in the last 30 days
lowest_monthly_earnings: Lowest estimated monthly earnings from the channel
highest_monthly_earnings: Highest estimated monthly earnings from the channel
lowest_yearly_earnings: Lowest estimated yearly earnings from the channel
highest_yearly_earnings: Highest estimated yearly earnings from the channel
subscribers_for_last_30_days: Number of new subscribers gained in the last 30 days
created_year: Year when the YouTube channel was created
created_month: Month when the YouTube channel was created
created_date: Exact date of the YouTube channel's creation
Gross tertiary education enrollment (%): Percentage of the population enrolled in tertiary education in the country
Population: Total population of the country
Unemployment rate: Unemployment rate in the country
Urban_population: Percentage of the population living in urban areas
Latitude: Latitude coordinate of the country's location
Longitude: Longitude coordinate of the country's location

Outputs

Cluster Labels: Each channel is categorized into a cluster based on content.
Cluster Insights via Word CloudsPredictive
Category: Predicted categorization of channels into 'Popular', 'Influential', and 'Elite' based on subscriber counts.

Use

The use of the model is primarily for a feasibility check; it will then be applied to other similar datasets in a second stage.
Given the confidential nature of my own data—a list of customers with access to marketing content—I have opted to use surrogate data to model these interactions. This dataset detailing the 'first 1000 YouTube channels by subscribers,' sourced from Kaggle, serves as an excellent proxy for my analysis..

Limitations

Language Dependency:
The model's performance dependent on the quality and variety of the text in the channel descriptions, which may vary significantly across languages.

Dynamic Content Changes
YouTube channels often pivot their content, which can render static clusters less accurate over time.

Scalability Issues: While the model handles the current dataset effectively, larger datasets might require more computational resources or a simplified model to maintain efficiency.

Trade-offs Complexity vs. Performance:
The use of RandomForest and K-Means ensures robust performance but in some cases, the model may prioritize one over the other based on the classification threshold set, affecting its sensitivity to false positives or false negatives.

Generalization vs. Specificity:
The model is tuned to generalize across various types of channels, which might reduce its accuracy in identifying nuances of smaller niche channels.

Ethical

Bias and Fairness
Machine learning models, like yours, risk bias leading to unfair outcomes if data isn't diverse across languages, content types, and creator demographics. This could favour certain YouTube channels, perpetuating stereotypes and reinforcing visibility and monetization disparities.

Misuse of Model Outputs
Despite being for educational purposes, misuse of your model could result in harmful decisions. For instance, using it for real-world marketing without considering its limitations could harm niche creators by misclassifying or overlooking their content.

Model description

The model deployed in this project combines machine learning techniques like clustering and classification to analyse and categorize YouTube channels:

1. K-Means clustering to identify distinct groups of channels based on similarities in their content descriptions and viewer metrics. This method is ideal for segmenting large datasets into manageable groups that share common characteristics, making it easier to understand overarching trends and niche content areas.

   Latent Dirichlet Allocation (LDA) was used also as a more sophisticated approach for associating topics. The use of bi-grams and tri-grams (groupings of two and three words) in the TF-IDF vectorizer allows the model to capture phrases, which can be more expressive than single words. This can lead to more meaningful topics, as phrases often hold more specific meaning than individual words.

2. For predicting the category ('Popular’, 'Influential’, 'Elite’) of channels based on their subscriber counts, a RandomForestClassifier is employed. This model was chosen for its robustness and effectiveness in handling diverse datasets with a mix of numerical and categorical data. RandomForest is particularly advantageous due to its ability to manage overfitting, making it reliable for our purposes where the accuracy of prediction is key. The combination of these models provides a comprehensive approach to understanding YouTube channels, aiding in strategic decision-making for content creators and marketers alike.

The model is created in a Jupyter environment.

Hyperparameters optimization

Hyperparameter optimization was applied to improve the performance of the machine learning models, particularly:

a) For the *K-Means clustering* used in data segmentation, the main hyperparameter was the Number of Clusters (k).The Herarchiral Method was used here to identify the optimal k value.

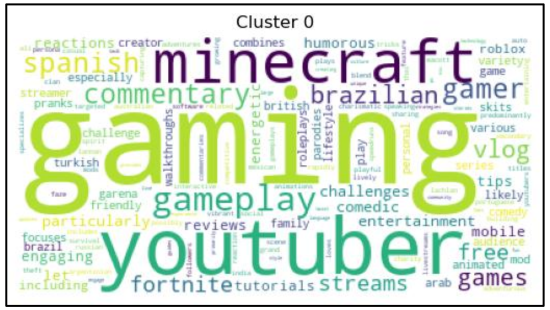
b) for the *RandomForestClassifier,* the main hyperparameters optimized for the RandomForestClassifier model include:

1. Number of Trees (n_estimators): The number of trees in the forest.
2. Maximum Depth of Trees (max_depth): Controls the maximum depth of each tree.
3. Minimum Samples Split (min_samples_split): The minimum number of samples required to split an internal node.
4. Minimum Samples Leaf (min_samples_leaf): The minimum number of samples required to be at a leaf node.

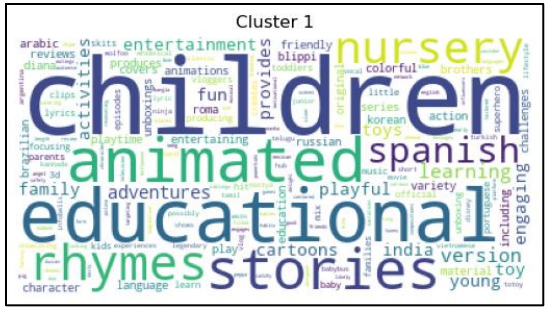
Optimization Methodology employed the use of the GridSearchCV, which systematically constructs and evaluates a model for each combination of algorithm parameters specified in a grid.
The process involved:

1. Defining a parameter grid for RandomForest parameters such as n_estimators, max_depth, min_samples_split, and min_samples_leaf.
2. Running GridSearchCV with cross-validation to ensure that the model's performance estimates are reliable and the model does not just perform well on a subset of the data.

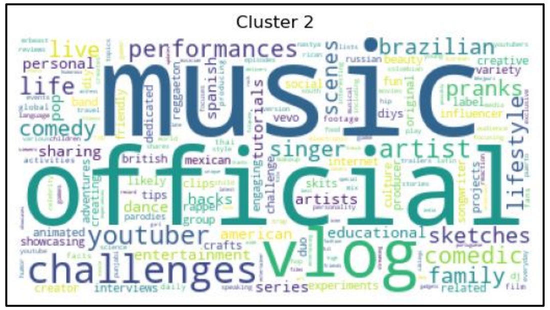
Results: K-means clustering for Channel Contents


Cluster 0

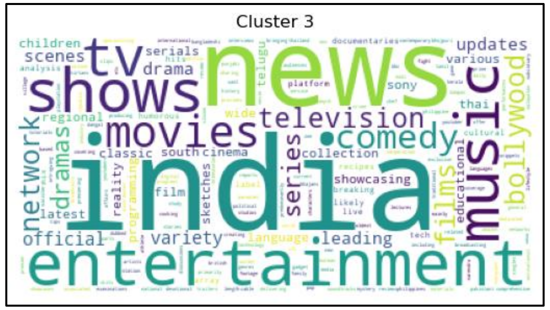
Cluster 0 focus on gaming content, as indicated by words like "Minecraft," "gameplay," "gamer," "Fortnite," and "streams." There's a strong emphasis on gaming platforms and terms, suggesting these channels are dedicated to video game content, reviews, and possibly live streaming of gameplay.


Cluster 1

Cluster 1 is centered around educational and child-friendly content, with terms such as "nursery," "rhymes," "educational," "cartoons," and "toys" standing out. This cluster includes channels that produce videos for children, like animated stories, learning materials, and music videos intended for young audiences.


Cluster 2

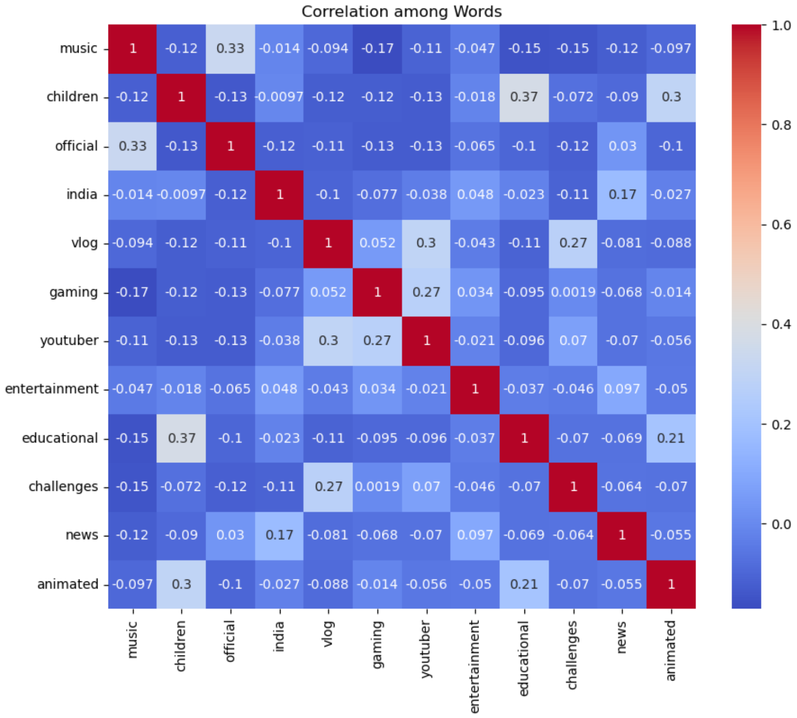
Cluster 2 has a mix of personal vlogging and entertainment, with words like "challenges," "comedy," "sketches," "artist," and "lifestyle." This cluster appears to include a variety of content creators, from comedians and musicians to lifestyle vloggers and entertainers.


Cluster 3

Cluster 3 focuses on more traditional forms of entertainment such as "shows," "movies," "drama," "television," and "Bollywood." Channels in this cluster might be offering streaming of television shows, movie reviews, celebrity news, and perhaps even network or studio-produced content.

The variety within each cluster point to the multiple facets of content that creators within similar genres explore. It also highlights the potential for targeting different audience demographics, from gamers and families to those interested in lifestyle and culture.

Results: Analysis correlation words



Distinct Content Themes

"Gaming" has a positive correlation with "youtuber," suggesting that gaming content is often associated with individual creators or 'YouTubers'.

"Educational" content shows a moderate positive correlation with "children," indicating that educational themes are often targeted towards young audiences.

"News" does not show strong correlation with most other content types, suggesting it stands as a distinct category.

Potential Content Overlaps:

"Challenges" show some correlation with "entertainment," which could indicate a trend where entertaining content often includes challenge videos.

"India" has a slight positive correlation with "official," possibly indicating that many official channels or content may be associated with Indian themes or produced in India

Distinct Topics identified:

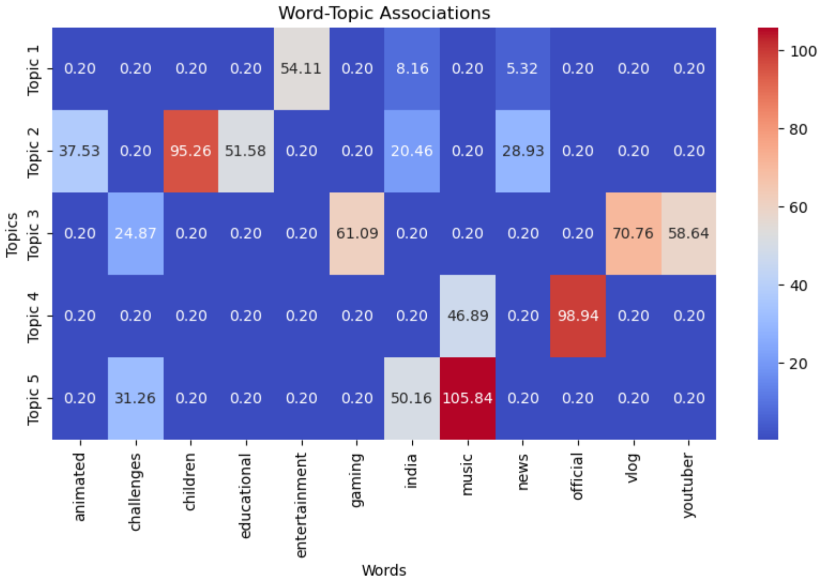
Topic 1 heavily associate with words like "vlog," "youtuber," and "gaming," which suggest channels focused on personal vlogging and gaming content.

Topic 2 emphasizes "entertainment," "gaming," and "news," hinting at channels that cover current events in entertainment and gaming.

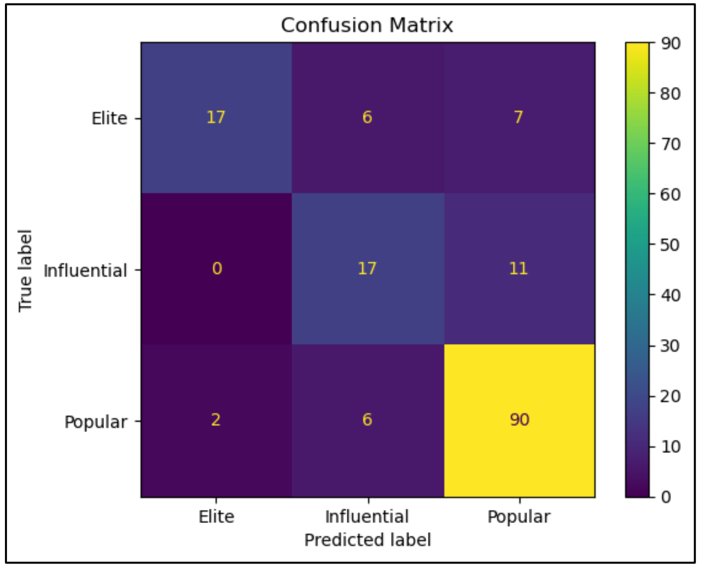
Topic 3 highlights "music," "india," and "children," which point to channels with content related to music for children or music from India.

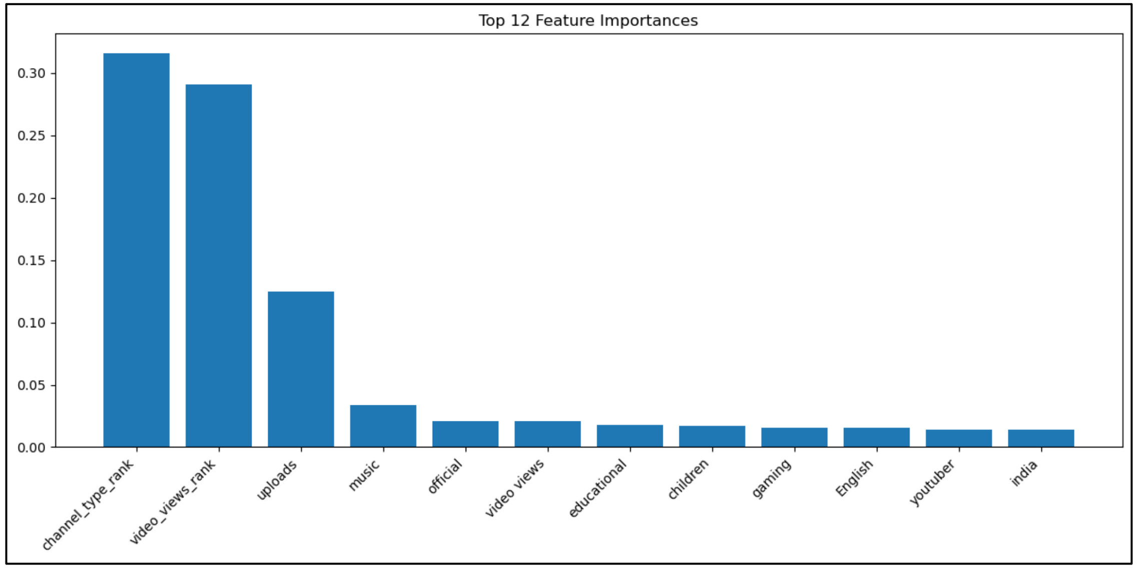
Topic 4 has a strong association with "official," which indicate official channels of creators, artists, or brands.

Topic 5 includes "children," "educational," and "animated," suggesting channels that produce educational content for children, possibly through animation.



Results: RandomForestClassifier to predict channel categories





The **RandomForestClassifier** model achieved a cross-validation score of **81.01%** and an independent test score of **79.49%**, indicating a reliable performance in categorizing YouTube channels into:

'Popular',
'Influential',
'Elite'

categories based on their content and viewership metrics.

The project results show a good level of accuracy in classifying YouTube channel categories. However, the confusion matrix indicates that the classification of 'Elite' channels is less precise compared to 'Popular' and 'Influential' ones. Therefore, with additional data or by obtaining a more balanced and comprehensive dataset, there is potential to further refine the model, enhance its predictive capabilities, and reduce misclassification rates.