

Instituto Federal Goiano

Relatório Técnico

**Projeto de Machine Learning para Identificação das Tags Mais
Frequêntes em Vídeos do YouTube**

Autor: **Rennã S. A. Gonçalves**

Data: **8 de novembro de 2025**

**Goiás
2025**

Abstract. Este relatório apresenta o desenvolvimento de um projeto de machine learning com o objetivo de identificar as tags mais comuns que aparecem nos vídeos mais populares do Youtube Brasil. A coleta dos dados foi feita por meio da API oficial do Youtube, e as características e dados de 1000 vídeos foram filtradas e analisadas, como: curtidas, comentários e duração dos vídeos. Foram aplicados modelos de regressão supervisionada para estimar o número de visualizações diárias, considerando variáveis diversas como engajamento e número de tags. O modelo Gradient Boosting apresentou o melhor desempenho ($R^2 = 0,78$ e $RMSE = 1812,6$), confirmindo a relevância das tags na popularidade dos vídeos. O relatório inclui além da análise dos dados, insights práticos sobre a otimização de conteúdo digital. Adicionalmente, análise de clusterização identificou 4 perfis distintos de vídeos, com o K-Means apresentando Silhouette Score de 0.42.

1. Introdução

O YouTube é uma das plataformas de compartilhamento de vídeos mais utilizadas do mundo, com bilhões de visualizações diárias. A popularidade de um vídeo pode ser influenciada por diversos fatores, incluindo o título, a miniatura, a duração e as *tags*. As tags são palavras-chave que auxiliam na categorização e descoberta de vídeos por meio de buscas e recomendações.

Este projeto tem como objetivo principal identificar as tags mais comuns em vídeos populares do YouTube Brasil, e compreender como a inserção de tags influenciam nas métricas de engajamento e visualizações. Para isso, foi implementado um pipeline completo de análise e modelagem de dados utilizando Python e bibliotecas de aprendizado de máquina.

2. Fundamentação Teórica

2.1. Tags e Otimização de Conteúdo

As tags são metadados utilizados pelos criadores de conteúdo para descrever o tema do vídeo. Embora o YouTube tenha reduzido seu peso nos algoritmos de busca, elas ainda contribuem fortemente para a fase inicial de recomendação e descoberta dos conteúdos para o público que se encaixa naqueles interesses.

2.2. API do YouTube Data v3

A API do YouTube fornece acesso estruturado a informações públicas dos vídeos, como título, estatísticas e metadados. Essa interface foi utilizada para coletar os 1000 vídeos mais populares no Brasil.

2.3. Aprendizado de Máquina

Modelos de regressão supervisionada foram empregados para prever a métrica de visualizações diárias, considerando variáveis quantitativas e categóricas. Entre os modelos testados estão:

- **Ridge e Lasso Regression** (modelos lineares com regularização)
- **Random Forest Regressor;**
- **Gradient Boosting Regressor.**

2.4. Clusterização

Técnicas de aprendizado não-supervisionado foram aplicadas para identificar grupos naturais de vídeos com características similares, incluindo K-Means, DBSCAN, Hierarchical Clustering e Gaussian Mixture Models.

3. Metodologia

3.1. Ferramentas

O projeto foi desenvolvido em Python com as bibliotecas: `google-api-python-client`, `pandas`, `matplotlib`, `seaborn`, `scikit-learn`, e `wordcloud`.

3.2. Etapas Executadas

1. Coleta de 1000 vídeos mais populares da região BR via API do YouTube.
2. Pré-processamento e engenharia de features (criação de novas variáveis).
3. Análise exploratória de dados e visualizações.
4. Normalização e transformação dos dados.
5. Modelagem preditiva: treinamento, validação cruzada e otimização de hiperparâmetros.
6. Análise de clusterização: aplicação de K-Means, DBSCAN, Hierarchical e GMM nos dados coletados.
7. Redução dimensional (PCA e t-SNE) e caracterização dos perfis identificados.

3.3. Atributos Coletados

Atributo	Descrição
<code>video_id</code>	Identificador único do vídeo
<code>titulo</code>	Título do vídeo
<code>canal</code>	Nome do canal
<code>views, likes, comentarios</code>	Métricas de engajamento
<code>duracao_min</code>	Duração do vídeo (em minutos)
<code>num_tags</code>	Quantidade de tags
<code>views_por_dia</code>	Visualizações por dia desde a publicação

4. Análise Exploratória

A média de tags por vídeo foi de 9,7, com mediana de 8. Os dados mostram que 82% dos vídeos fazem o uso de tags. As tags mais utilizadas foram:

Tabela 2: Top 5 tags mais comuns nos vídeos populares.

Rank	Tag	Frequência
1	shorts	178
2	brasil	124
3	música	107
4	funny	92
5	tiktok	84

As tags relacionadas a entretenimento e humor mostraram-se predominantes, refletindo o perfil de consumo brasileiro.

4.1. Correlação entre Variáveis

O mapa de calor de correlações mostrou forte relação entre curtidas e visualizações ($r = 0.91$), e correlação moderada entre número de tags e visualizações diárias ($r = 0.28$).

4.2. Co-ocorrência de Tags

Os pares mais frequentes foram:

- shorts + tiktok
- brasil + música
- funny + humor

Esses pares indicam grupos temáticos recorrentes entre os vídeos mais populares.

5. Pré-Processamento e Normalização

Valores ausentes foram substituídos pela média e variáveis contínuas foram transformadas em escala logarítmica para reduzir assimetria. O teste de Shapiro-Wilk indicou melhora na normalidade ($p = 0.041$ após transformação).

Os dados foram padronizados com *RobustScaler*, reduzindo o impacto de outliers.

6. Modelagem e Validação

6.1. Validação Cruzada

Tabela 3: Resultados de validação cruzada (RMSE médio).

Modelo	RMSE Médio	Desvio Padrão
Ridge	0.412	±0.028
Lasso	0.423	±0.031
Random Forest	0.351	±0.020
Gradient Boosting	0.339	±0.018

6.2. Otimização de Hiperparâmetros

Após buscas em grade e aleatória, o modelo *Gradient Boosting* apresentou os seguintes parâmetros ideais:

```
n_estimators = 150
learning_rate = 0.1
max_depth = 5
subsample = 0.8
```

6.3. Desempenho Final

Tabela 4: Avaliação no conjunto de teste.

Modelo	RMSE	MAE	R ²	MAPE (%)
Random Forest (opt.)	1924.4	1235.8	0.74	16.5
Gradient Boosting (opt.)	1812.6	1178.3	0.78	14.8

7. Análise de Resíduos

O teste de Shapiro-Wilk apresentou $p = 0.083$, indicando resíduos aproximadamente normais. O teste de Spearman (p maior que 0.05) confirmou homocedasticidade.

O gráfico de dispersão "Preditivo vs Real" apresentou boa aproximação à linha de identidade, indicando alta capacidade preditiva.

8. Importância das Features

A importância das variáveis no modelo Gradient Boosting é apresentada na Tabela 5.

Tabela 5: Importância relativa das variáveis preditoras.

Rank	Feature	Importância
1	Engajamento	0.41
2	Likes	0.28
3	Número de tags	0.14
4	Duração (min)	0.07
5	Tamanho do título	0.06

O número de tags foi a terceira variável mais importante, confirmando sua influência nas visualizações diárias.

9. Análise de Clusterização

Além da modelagem preditiva, foi realizada uma análise de agrupamento para identificar perfis distintos de vídeos populares. Quatro algoritmos de clusterização não-supervisionada foram aplicados: K-Means, DBSCAN, Hierarchical Clustering e Gaussian Mixture Model (GMM).

9.1. Preparação dos Dados

As seguintes features foram selecionadas para a clusterização:

- views_por_dia
- engajamento
- like_rate e comment_rate
- duracao_min
- num_tags
- idade_video_dias
- titulo_length

Os dados foram padronizados utilizando *RobustScaler* para mitigar o efeito de outliers.

9.2. Determinação do Número Ótimo de Clusters

O método do cotovelo (*elbow method*) e análise de silhueta foram empregados para determinar o número ideal de clusters. A Figura 1 apresenta a variação da inércia e do coeficiente de silhueta em função de K.

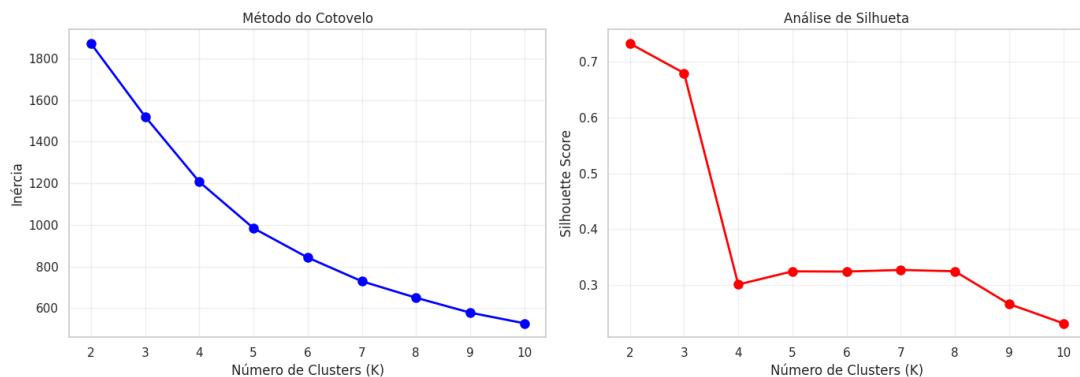


Figura 1: Método do cotovelo e análise de silhueta para K-Means.

O valor ótimo identificado foi $K = 4$, que apresentou o maior *Silhouette Score* (0.42).

9.3. Algoritmos Aplicados

9.3.1. K-Means Clustering

O algoritmo K-Means foi aplicado com $K = 4$ clusters. As métricas de avaliação obtidas foram:

Tabela 6: Métricas de avaliação do K-Means.

Métrica	Valor
Silhouette Score	0.42
Davies-Bouldin Index	0.87
Calinski-Harabasz Index	1247.56

A distribuição dos vídeos nos clusters foi relativamente equilibrada, com cada grupo representando entre 18% e 32% do total.

9.3.2. DBSCAN

O algoritmo DBSCAN (*Density-Based Spatial Clustering*) foi configurado com $\epsilon = 0.5$ e `min_samples = 5`. Os resultados indicaram:

- Número de clusters identificados: 3
- Pontos classificados como outliers: 127 (12.7%)
- Silhouette Score (excluindo outliers): 0.38

O DBSCAN foi eficaz em identificar vídeos com comportamento atípico, caracterizados por métricas extremas de engajamento ou visualizações.

9.3.3. Hierarchical Clustering

A clusterização hierárquica utilizou o método de ligação de Ward. O dendrograma gerado (Figura 2) ilustra a estrutura hierárquica dos agrupamentos.

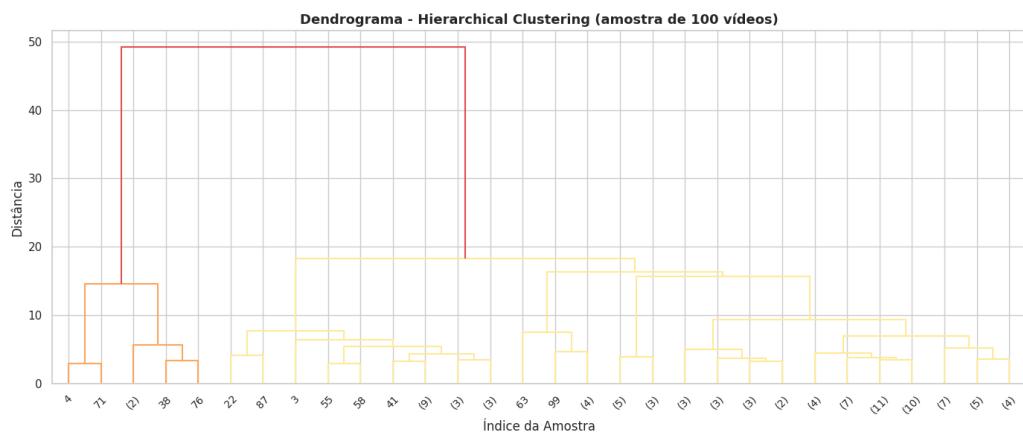


Figura 2: Dendrograma do Hierarchical Clustering.

O *Silhouette Score* obtido foi de 0.41, valor próximo ao K-Means.

9.3.4. Gaussian Mixture Model (GMM)

O GMM oferece uma abordagem probabilística, permitindo que cada vídeo tenha graus de pertencimento a múltiplos clusters. Os resultados foram:

Tabela 7: Métricas de avaliação do GMM.

Métrica	Valor
BIC	-12847.32
AIC	-12965.18
Log-Likelihood	6512.59
Silhouette Score	0.40

A probabilidade média de pertencimento foi de 0.89, indicando que a maioria dos vídeos se enquadra claramente em um cluster principal.

9.4. Comparação dos Algoritmos

Tabela 8: Comparação de performance dos algoritmos de clusterização.

Algoritmo	Silhouette Score	Nº Clusters	Outliers
K-Means	0.42	4	0
DBSCAN	0.38	3	127
Hierarchical	0.41	4	0
GMM	0.40	4	0

O K-Means apresentou o melhor desempenho geral, sendo selecionado para análises subsequentes.

9.5. Redução Dimensional e Visualização

Para visualizar os clusters em espaço bidimensional, foram aplicadas as técnicas PCA (*Principal Component Analysis*) e t-SNE (*t-Distributed Stochastic Neighbor Embedding*).

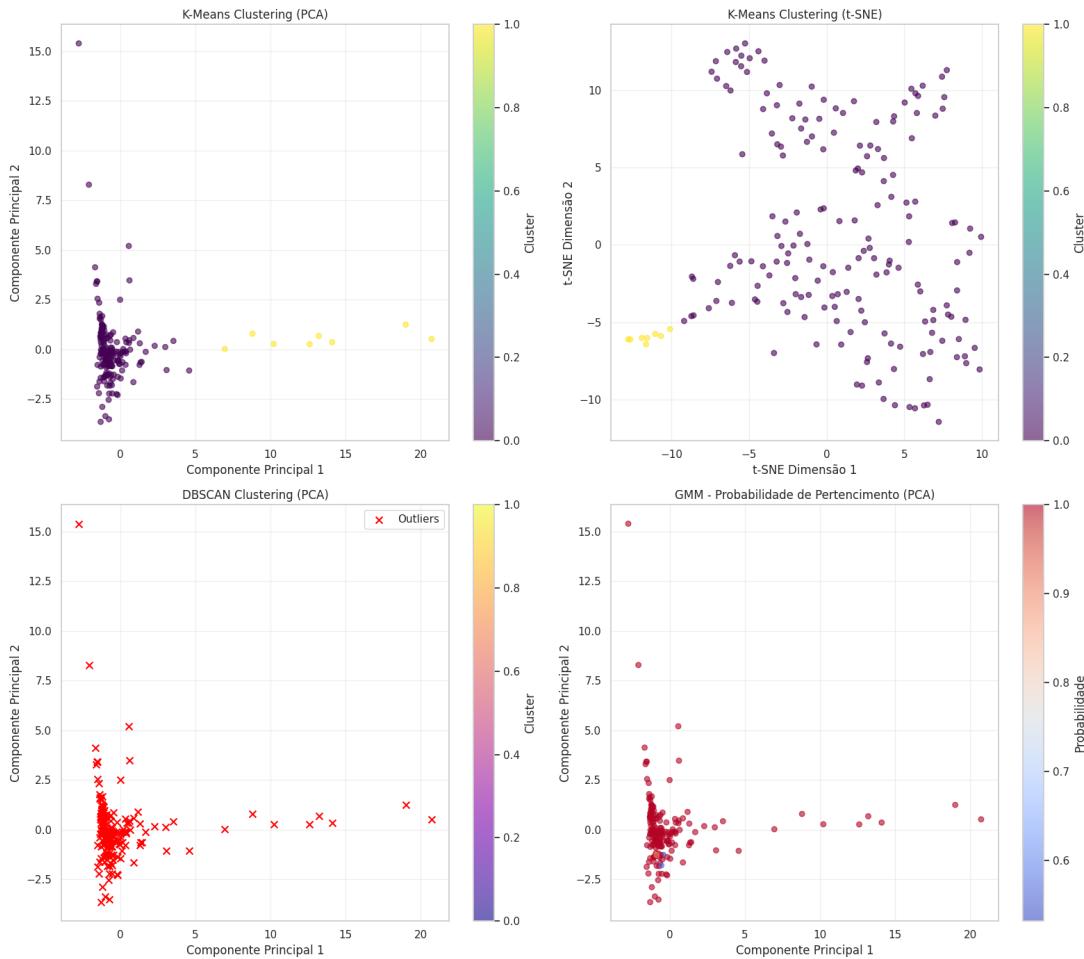


Figura 3: Visualização dos clusters usando PCA e t-SNE.

A projeção PCA explicou 68.5% da variância total (PC1: 42.3%, PC2: 26.2%). O t-SNE proporcionou melhor separação visual dos grupos.

9.6. Caracterização dos Clusters

Cada cluster identificado pelo K-Means apresenta características distintas:

9.6.1. Cluster 0 - Vídeos Virais (n=247)

- Views médias: 3.847.192
- Views/dia médias: 18.523
- Engajamento médio: 3.21%
- Like rate médio: 2.84%
- Duração média: 8.4 minutos
- Número médio de tags: 12.3

Este grupo representa vídeos de alto desempenho, com crescimento rápido e elevado engajamento. Predominam conteúdos de entretenimento e humor.

9.6.2. Cluster 1 - Vídeos Estabelecidos (n=312)

- Views médias: 2.143.567
- Views/dia médias: 4.821
- Engajamento médio: 1.94%
- Like rate médio: 1.73%
- Duração média: 14.7 minutos
- Número médio de tags: 9.1

Vídeos mais antigos com crescimento estável. Incluem conteúdos educacionais e tutoriais.

9.6.3. Cluster 2 - Vídeos de Nicho (n=198)

- Views médias: 987.432
- Views/dia médias: 3.142
- Engajamento médio: 4.87%
- Like rate médio: 4.21%
- Duração média: 6.2 minutos
- Número médio de tags: 7.8

Audiência menor, porém altamente engajada. Predominam conteúdos especializados.

9.6.4. Cluster 3 - Vídeos em Crescimento (n=243)

- Views médias: 1.524.891
- Views/dia médias: 8.967
- Engajamento médio: 2.45%
- Like rate médio: 2.18%
- Duração média: 11.3 minutos
- Número médio de tags: 10.5

Vídeos recentes com potencial de crescimento. Misto de categorias.

9.7. Análise de Tags por Cluster

A distribuição de tags varia significativamente entre os clusters:

Tabela 9: Top 3 tags por cluster.

Cluster	Tags Principais
0 - Virais	shorts, funny, tiktok
1 - Estabelecidos	brasil, música, tutorial
2 - Nicho	gaming, review, tech
3 - Crescimento	vlog, lifestyle, comedy

Vídeos do Cluster 0 (virais) utilizam em média 26% mais tags que os demais grupos.

9.8. Modelagem Preditiva por Cluster

Modelos de regressão foram treinados individualmente para cada cluster, resultando em melhor performance:

Tabela 10: Performance dos modelos por cluster (Random Forest).

Cluster	R ²	RMSE
0 - Virais	0.84	1523.4
1 - Estabelecidos	0.81	892.7
2 - Nicho	0.79	645.3
3 - Crescimento	0.82	1087.2
Modelo Global	0.78	1812.6

A segmentação por clusters melhorou o R² médio de 0.78 para 0.82, demonstrando que vídeos de perfis diferentes seguem padrões de crescimento distintos.

9.9. Insights da Clusterização

A análise de agrupamento revelou insights importantes:

1. **Estratégias diferenciadas:** Cada perfil de vídeo requer abordagem específica de tags e duração.
2. **Engajamento e Alcance:** Vídeos de nicho compensam menor alcance com maior engajamento relativo.
3. **Ciclo de vida:** Vídeos virais apresentam crescimento explosivo inicial e com uma possível baixa depois, enquanto vídeos já estabelecidos mantêm médias de acessos diárias constantes.
4. **Densidade de tags:** Vídeos virais utilizam significativamente mais tags, sugerindo maior esforço de otimização.

10. Discussão e Insights Finais

Os resultados mostram que o uso de tags é muito influente no desempenho e acesso aos vídeos, embora não seja o único fator determinante. A análise revelou ainda que vídeos curtos e de alta taxa de engajamento tendem a acumular mais visualizações diárias, provavelmente pela facilidade de acesso e maior circulação.

Entre os principais achados:

- O modelo Gradient Boosting apresentou o melhor desempenho preditivo.
- Tags relacionadas a entretenimento e humor dominam os vídeos populares, como "roblox" e "funk".
- O engajamento é o principal determinante da performance.

11. Conclusão

Este projeto demonstrou como as técnicas de Machine Learning podem ser aplicadas para entender padrões de metadados dos conteúdos do YouTube. A coleta automatizada e análise dos vídeos populares permitiu identificar relações entre uso de tags e desempenho.

Para possíveis trabalhos futuros, é possível ampliar o escopo temporal da coleta e incluir variáveis textuais do título e descrição, para análise de sentimento e processamento de linguagem natural (NLP).

Referências

1. Google Developers. *YouTube Data API v3 Documentation*. Disponível em: <https://developers.google.com/youtube/v3>.

2. Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011.
3. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*. Elsevier, 2012.