

# Análise de tendências com Machine Learning - YouTube

## Trabalho Prático Nº 2

José Francisco Fernandes  
Instituto Politécnico de Beja  
Beja, Portugal  
[jose20fernandes03@gmail.com](mailto:jose20fernandes03@gmail.com)

Tierri Ferreira  
Instituto Politécnico de Beja  
Beja, Portugal  
[tierrif@hotmail.com](mailto:tierrif@hotmail.com)

**Resumo** — No presente artigo temos como objetivo fazer a previsão de visualizações e número de gostos e de comentários nas tendências do YouTube, usando o *software* Orange Data Mining, com um modelo de *forecasting*.

**Palavras Chave** – *tendências; YouTube; modelos de previsão; séries temporais.*

**Abstract** — In this article our objective is to make a forecast on views and like and comment counts in YouTube's trends, using the Orange Data Mining software, with a forecasting model.

**Keywords** - *trends; YouTube; forecasting models; time series.*

### I. INTRODUÇÃO

O YouTube é uma rede social de partilha de vídeos extremamente utilizada nos dias de hoje. Com o passar dos anos, tem crescido exponencialmente e é usado para diversos fins, tais como educação, entretenimento, partilha de música, etc.

Nesta análise fazemos uma previsão de alguns dados estatísticos desta rede para se verificar como será o desempenho da mesma nos próximos meses. Tal será possível usando o *software* Orange Data Mining, que nos permite de usar diversos algoritmos de Machine Learning num *workflow* simples e sem código. [1]

O *dataset* utilizado é uma lista de vídeos que entraram nas tendências. Estes podem repetir-se devido à possível (e frequente) entrada nas tendências mais que uma vez em diferentes datas.

As previsões serão feitas em três métricas de vídeo diferentes: visualizações, gostos e comentários. Em relação aos desgostos, foi inicialmente experimentada uma previsão com esta métrica, mas devido à desativação do número público de desgostos no YouTube, estes passaram a ser 0, o que torna a previsão inútil nesse contexto.

O modelo escolhido foi o modelo ARIMA, para *Time Series*. Antes de escolher o mesmo, pensámos em outros modelos que não são para *Time Series*, como redes neuronais e *Random Forest*, mas imediatamente percebemos que estes não faziam sentido, pelo menos da forma que tentámos usá-los. Assim, focámo-nos em investigar séries temporais e os seus conceitos, assim como modelos que estivessem disponíveis no Orange.

### II. DESCRIÇÃO DO PROCESSO APLICADO

O processo aplicado neste trabalho foi o KDD.

#### A. Preparação dos Dados

Os dados foram preparados com o processo ETL feito no trabalho anterior com algumas alterações.

Foi usada a linguagem de programação Python para fazer os *scripts* de extração e de transformação do *dataset* utilizado. No primeiro ficheiro, extraíram-se todos os valores do ficheiro CSV original e foram filtrados os números de visualizações, de gostos e de comentários de cada vídeo.

Como o mesmo vídeo pode-se repetir nas tendências diversas vezes, foi necessário filtrar os vídeos pelo seu ID, ficando apenas a tendência deste com mais visualizações (que automaticamente significa a mais recente).

Após esta filtragem, dividem-se os três tipos de dados em três ficheiros CSV, exportados para uso no Orange.

Juntamente com estes ficheiros, outros três foram gerados, mas para o canal *MrBeast*, que é o canal que mais se destaca nas tendências. Assim, faz-se uma análise experimental a um canal só, para comparar resultados e verificar o desempenho do modelo.

#### B. Aplicar Algoritmo

O algoritmo utilizado é o modelo ARIMA, que faz parte da biblioteca Time Series no Orange. Este modelo é aplicado em problemas de *forecasting*, ou seja, para fazer previsões futuras de dados que estão diretamente ligados ao tempo. [2]

Inicialmente, pensámos em dividir os dados em 70% para treino do modelo e 30% para testes. No entanto, no modelo ARIMA, verificou-se que isto é desnecessário, dado que os resultados foram essencialmente os mesmos, no nosso caso.

No Orange, a biblioteca Time Series não está instalada por defeito, sendo necessário instalá-la no menu "Add-ons".

Na Fig. 1 é possível observar o exemplo de um *workflow* no Orange deste trabalho. Foi necessário repetir estes *workflows* para cada ficheiro.

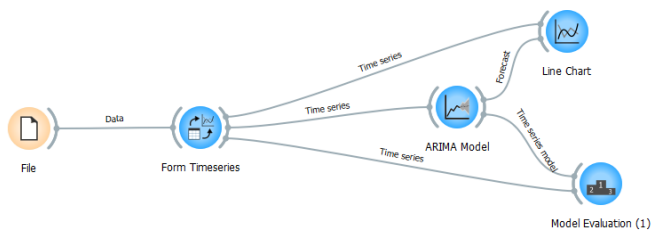


Figure 1. Workflow no Orange para o modelo ARIMA com visualização e avaliação.

### C. Avaliação dos Dados

A avaliação dos dados foi feita usando o widget “Model Evaluation” da biblioteca Time Series. Este permite-nos de avaliar o modelo ARIMA com as métricas aplicáveis em forecasting.

As métricas principais que serão usadas para avaliar o modelo são o RMSE (*root-mean-square deviation*), o MAE (*mean absolute error*) e o MAPE (*mean absolute percentage error*).

Para os melhores resultados, testámos o modelo ARIMA com diversos valores nos seus argumentos, e obtivemos os melhores resultados com os seguintes valores:

- Auto-regression order ( $p$ ): 0;
- Differencing degree ( $d$ ): 1;
- Moving average order ( $q$ ): 3.

O número de *steps* do *forecast* foi 3, considerando que, visualmente, fazia mais sentido. Quanto mais *steps* forem dados, menos precisão haverá na previsão.

Uma vez que o Orange não tem uma ferramenta que permite calcular os valores ótimos destas variáveis, estivemos a comparar resultados da avaliação do modelo com diferentes valores até chegar a este resultado, que considerámos a configuração ótima do modelo para este *dataset*.

### D. Visualização dos Resultados

Para a visualização dos resultados, foi usado o widget “Line Chart”, também pertencente à biblioteca Time Series, que nos permite de visualizar a previsão juntamente com os dados do passado. Na secção III serão apresentados os resultados.

## III. RESULTADOS

No total foram feitas 6 (seis) análises com o modelo ARIMA para todos os ficheiros CSV gerados. Na Fig. 2 demonstra-se o resultado da previsão do número total de visualizações das tendências do YouTube ao longo do tempo. Na zona azul, é-nos dada a previsão.

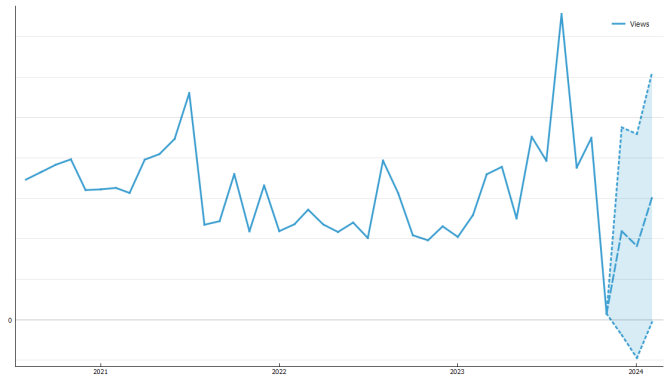


Figure 2. Previsão de visualizações totais nas tendências do YouTube.

De acordo com a mesma, o número de visualizações tende a crescer após uma queda do pico anterior. No futuro, apenas tende a subir, pelo menos até ao valor mínimo anterior neste gráfico.

Nas Fig. 3 e 4 o mesmo é feito, mas para número de gostos e de comentários, respetivamente.

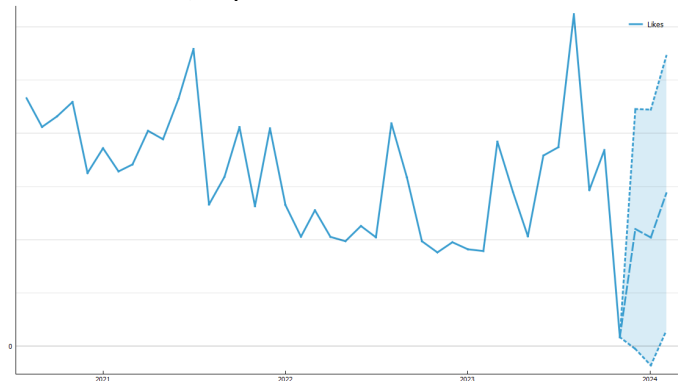


Figure 3. Previsão de gostos totais nas tendências do YouTube.

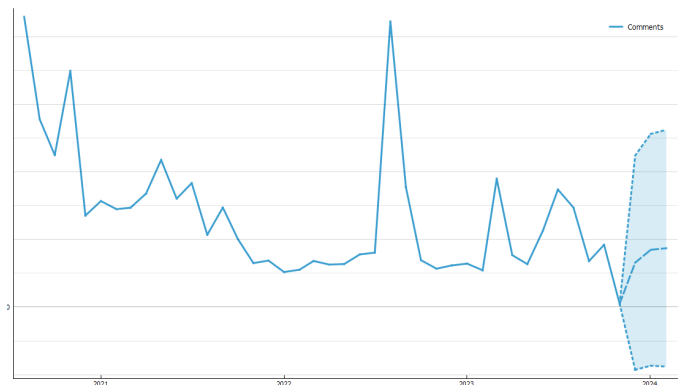


Figure 4. Previsão de comentários totais nas tendências do YouTube.

Podemos observar uma semelhança entre os gráficos das visualizações e dos gostos, mas uma diferença maior nos comentários, sendo a previsão igualmente diferente.

Agora, como experiência, foi usado o mesmo workflow para fazer uma previsão individual a um canal, tendo sido escolhido o canal MrBeast devido ao facto de ser o canal que tem mais visualizações nas tendências.

Nas Fig. 5, 6 e 7 mostram-se os resultados da previsão para este canal.

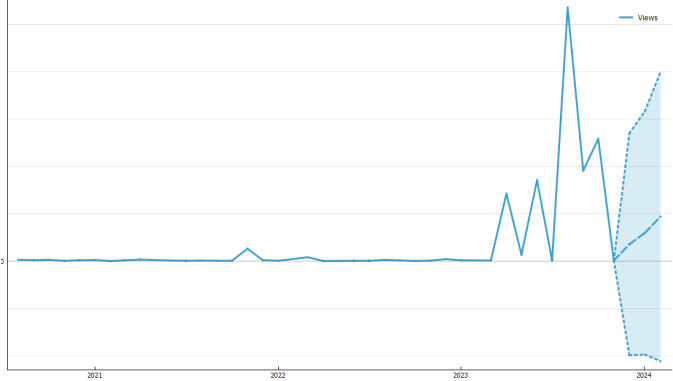


Figure 5. Previsão de visualizações do canal MrBeast nas tendências do YouTube.

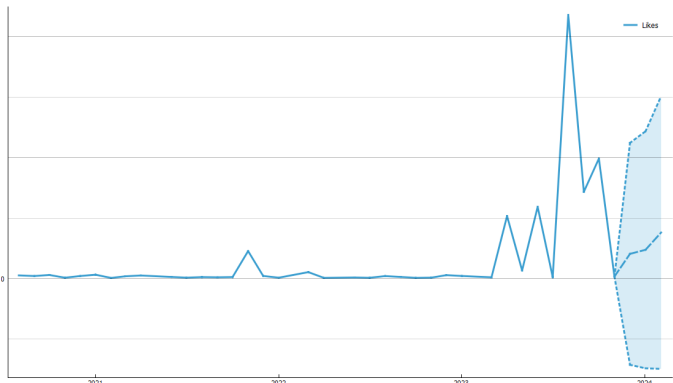


Figure 6. Previsão de gostos do canal MrBeast nas tendências do YouTube.

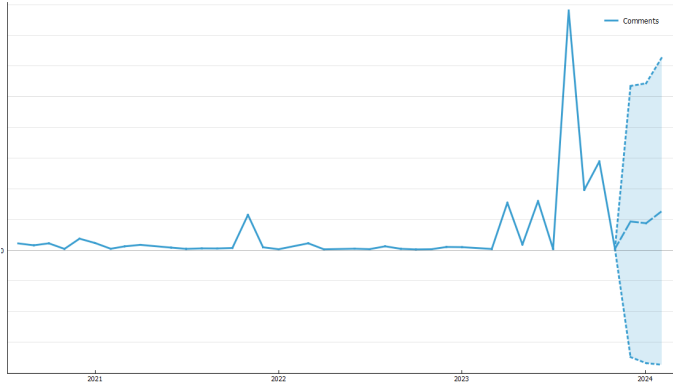


Figure 7. Previsão de comentários do canal MrBeast nas tendências do YouTube.

Devido à diferença muito grande entre valores do passado e mais recentes, a previsão pode ser menos precisa. Tal poderá ser visto na secção seguinte de avaliação do modelo.

#### IV. AVALIAÇÃO DO MODELO

Tal como mencionado na secção II, foram usadas três métricas para avaliar o modelo ARIMA com estes dados.

#### A. Modelos para Valores Totais

Na Fig. 8 mostra-se a avaliação do modelo ARIMA usado para as visualizações totais nas tendências do YouTube de acordo com o dataset usado.

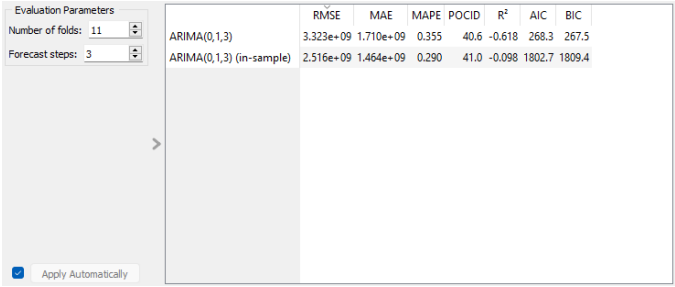


Figure 8. Avaliação do modelo ARIMA para visualizações totais.

Os valores do RMSE/MAE podem parecer elevados, mas efetivamente os dados trabalhados são números de grande dimensão.

Se olharmos para o MAE, o valor está abaixo de 0.5 (50%), significando que os valores são aceitáveis. No entanto, como vamos observar noutros resultados, estes valores podem piorar, dependendo do contexto.

Note-se que devemos ter mais em conta os valores da linha de cima (ARIMA(0, 1, 3)), e não a versão *in-sample*. Nas Fig. 9 e 10 podemos observar a avaliação dos modelos para o número de comentários e gostos.

RMSE	MAE	MAPE
1.473e+08	7.203e+07	0.351

Figure 9. Avaliação do modelo ARIMA para número de gostos totais.

RMSE	MAE	MAPE
2.410e+07	8.835e+06	0.638

Figure 10. Avaliação do modelo ARIMA para número de comentários totais.

Podemos observar que nos comentários houve uma subida no erro, possivelmente devido à inconsistência dos dados.

#### B. Modelos Específicos (Canal MrBeast)

Os modelos feitos especificamente para este canal tiveram resultados piores. Nas Fig. 11, 12 e 13 mostram-se os mesmos.

RMSE	MAE	MAPE
2.178e+09	5.321e+07	1.048

Figure 11. Avaliação do modelo ARIMA para número de visualizações do canal MrBeast

RMSE	MAE	MAPE
9.186e+07	4.111e+06	1.158

Figure 12. Avaliação do modelo ARIMA para número de gostos do canal MrBeast.

RMSE	MAE	MAPE
3.318e+06	3.088e+05	1.227

Figure 13. Avaliação do modelo ARIMA para número de gostos do canal MrBeast.

Como podemos observar nestes resultados, este modelo não funciona bem com dados pouco consistentes. A pouca quantidade de visualizações neste canal comparado com um pico de visualizações tornou o modelo extremamente impreciso.

#### V. COMPARAÇÃO DE RESULTADOS

No *dataset* usado, não encontramos nenhuma previsão como a nossa, pelo menos que tivesse sido partilhada. [5]

Isto torna este trabalho único o que impossibilita a comparação de resultados no seu estado atual.

#### VI. CONCLUSÕES

Podemos concluir que os resultados dos modelos mais gerais (visualizações e gostos, principalmente) para todos os vídeos nas tendências neste *dataset* foram positivos. No entanto, pudemos observar um erro maior quando os gráficos são menos

consistentes, ou seja, quando os valores mudam demasiado, o modelo tende a errar mais.

Este trabalho foi importante para os nossos conhecimentos no âmbito desta Unidade Curricular de Sistemas de Informação e na área de Data Science, que é uma área em constante expansão nos dias de hoje e muito relevante para o nosso conhecimento.

#### REFERÊNCIAS BIBLIOGRÁFICAS

- [1] N. Humphrey, "Time-Series Forecasting Part 7 - Time-Series Forecasting with Orange," 16 01 2021. [Online]. Available: <https://www.youtube.com/watch?v=bpPwDMmouyUJ>. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] O. Gerasymov, "Machine Learning For Time Series Forecasting," 04 01 2023. [Online]. Available: <https://codeit.us/blog/machine-learning-time-series-forecasting>.
- [3] Orange Data Mining, "Model Evaluation," Orange, [Online]. Available: [https://orangedatamining.com/widget-catalog/time-series/model\\_evaluation\\_w/](https://orangedatamining.com/widget-catalog/time-series/model_evaluation_w/). [Acedido em 27 12 2023].
- [4] Python, "Python 3.12.1 documentation," Python, [Online]. Available: <https://docs.python.org/3/>. [Acedido em 26 12 2023].
- [5] R. Sharma, "YouTube Trending Video Dataset (updated daily)," Kaggle, [Online]. Available: <https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/>. [Accessed 12 11 2023].