2022 Julho

MINIO

STACKLABS - FINANÇAS

NOSSA EQUIPE

ANDERSON FELIPE

Analista de dados

FELIPE PEREIRA

Engenheiro de dados

LENON BORGES

Cientista de dados

MARIO CARVALHO

Engenheiro de dados

WEBER GODOI

Cientista de dados & Gerente de projetos

TÓPICOS ABORDADOS

- Desafio escolhido
- Objetivos
- Coleta de dados
- Análise de dados
- Pré-Processamento
- Modelos NLP
- Feature Engineering
- Seleção de features
- Validação estatística
- Ensembles
- Engenharia de dados
- Deployment

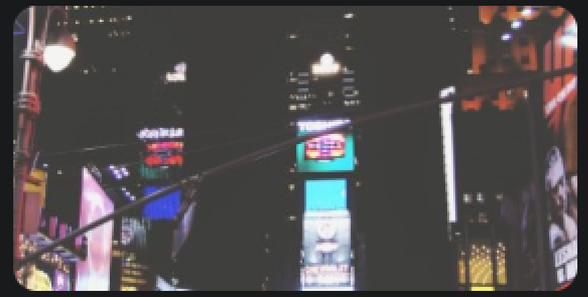
Jul/2022

DESAFIO ESCOLHIDO

AARON7SUN · UPDATED 3 YEARS AGO 1477 New Notebook Download (6 MiB)

Daily News for Stock Market Prediction

Using 8 years daily news headlines to predict stock market movement



- Previsão sobre o Índice de Dow Jones
- 25 principais notícias do dia
- Reddit WorldNews Channel

1º PROBLEMA

Exemplo:

A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours later.

```
dfred = df1 = pd.read_csv('/content/drive/MyDrive/Job/Stack Labs/archive (2)/RedditNews.csv')
dfred
```

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours' Not...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...
...	...	...
73603	2008-06-08	b'Man goes berzerk in Akihabara and stabs ever...
73604	2008-06-08	b'Threat of world AIDS pandemic among heterose...
73605	2008-06-08	b'Angst in Ankara: Turkey Steers into a Danger...
73606	2008-06-08	b"UK: Identity cards 'could be used to spy on ...
73607	2008-06-08	b'Marriage, they said, was reduced to the stat...

73608 rows × 2 columns

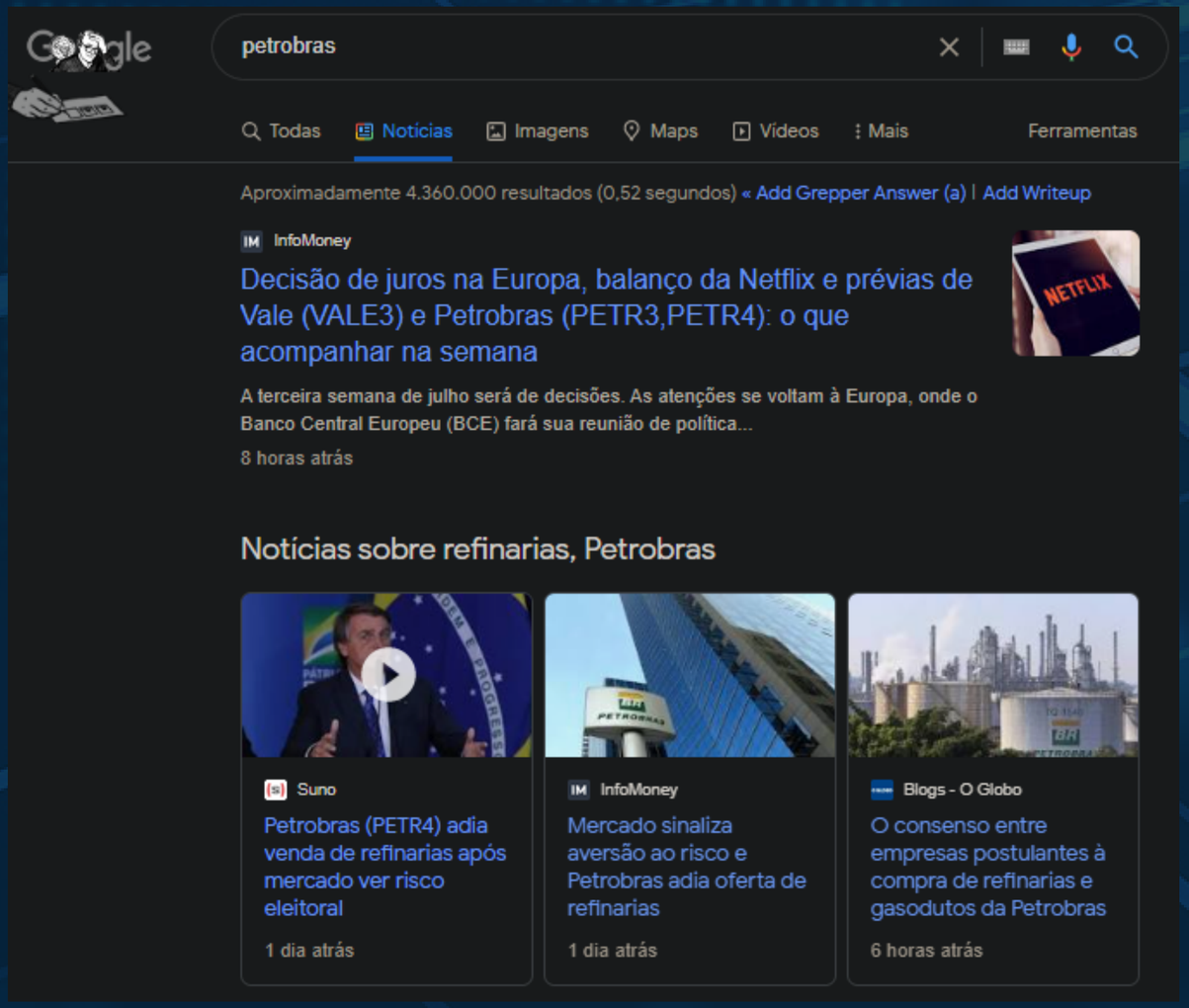
Precisamos de um dataset apropriado…

TURNING POINT

Por que não sobre ativos brasileiros?

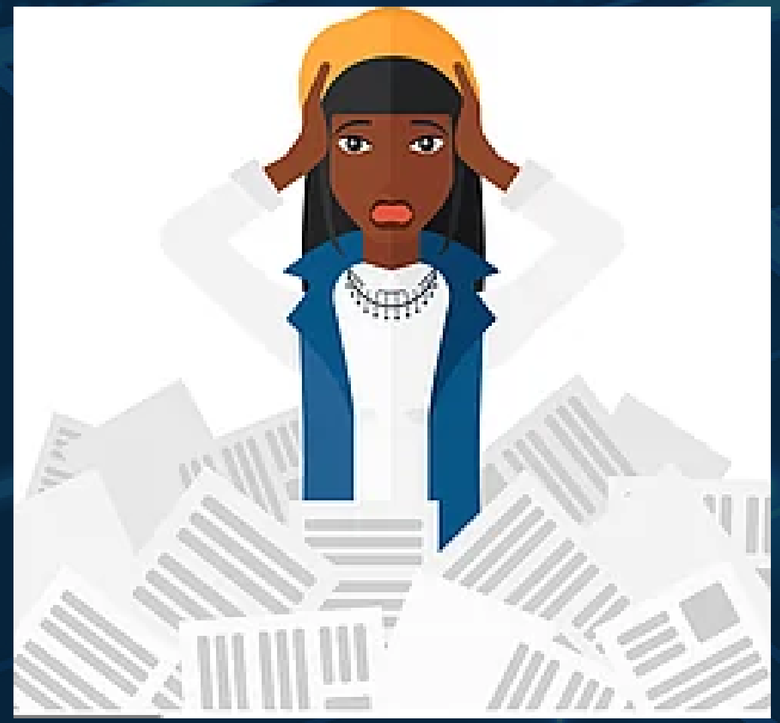
- Criação de Valor
- Inovação
- Mercado endereçável

5 Milhões de
CPF's na B3



OBJETIVOS

- Apoio à Tomada de Decisão
- Acompanhamento das notícias
- Tempestividade na informação



COLETA DE DADOS

Para notícias, GoogleNews Api:

- Pricing
- Tratamento
- Controle
- Fonte dos dados

Para série histórica de preços, Yahoo! Finance's API:

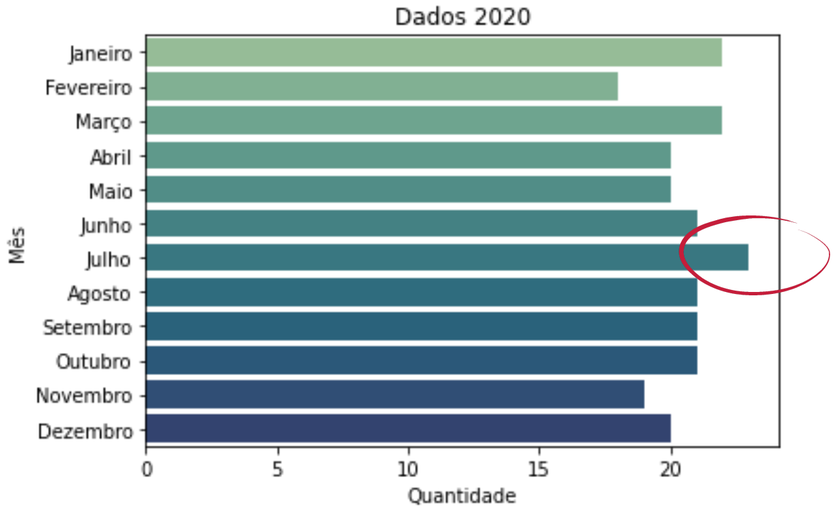
- Confiabilidade
- Preço de Fechamento Ajustado

ANÁLISE DE DADOS

Dos pregões da Bolsa:

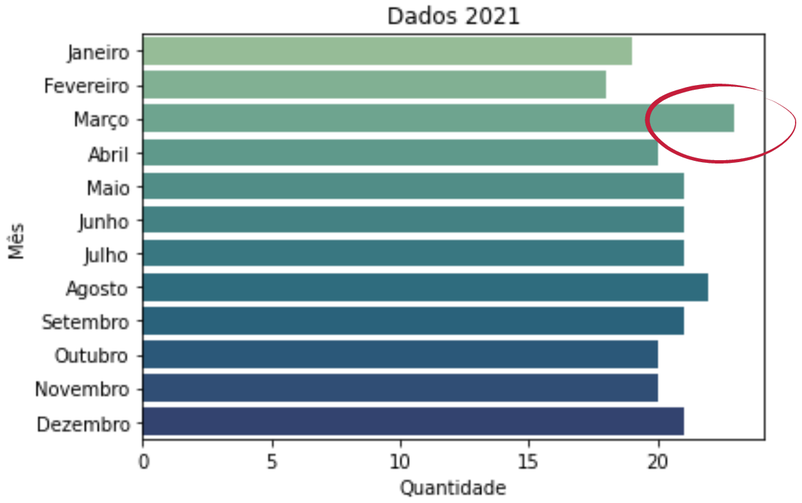
Analise 2020:

Total: **248 pregões**
Média: **20,6 p/mês**
Máximo: **Julho (23)**
Mínimo: **Fevereiro (18)**



Analise 2021:

Total: **247 pregões**
Média: **20,6 p/mês**
Máximo: **Março (23)**
Mínimo: **Fevereiro (18)**



ANÁLISE DE DADOS

Das notícias mineradas:

Mineração (ano):

2019 -> 883
2020 -> 3434
2021 -> 3026
2022 -> 1817

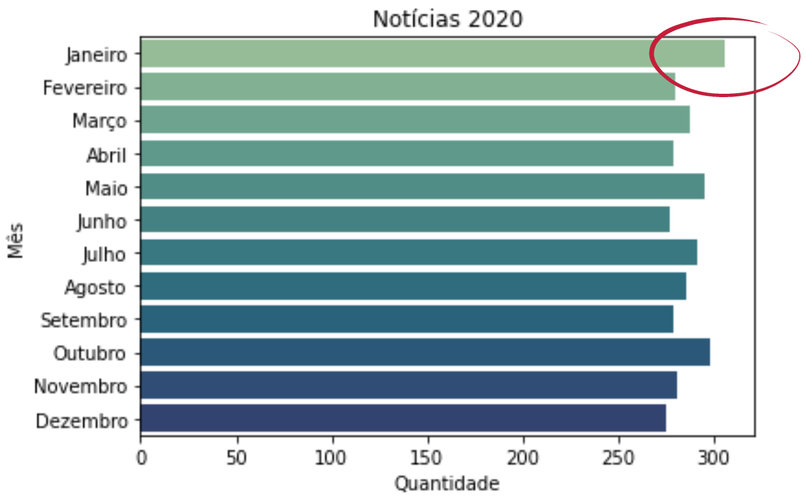
Total -> 9160

Média entre anos

Média -> 3230
Média/Mês -> 269,2
Média/Dia -> 8,85

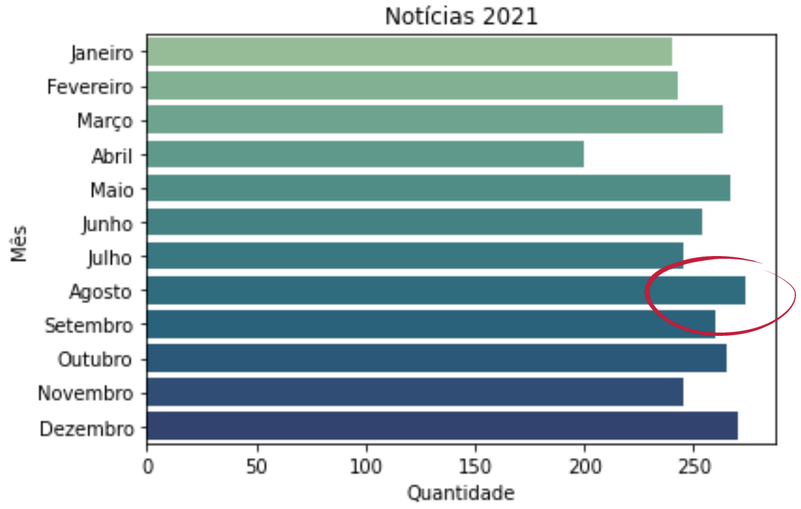
Analise 2020:

Total: **3434 news**
Média 1: **286,17 p/mês**
Média 2: **9,40 p/dia**
Máximo: **Janeiro (306)**



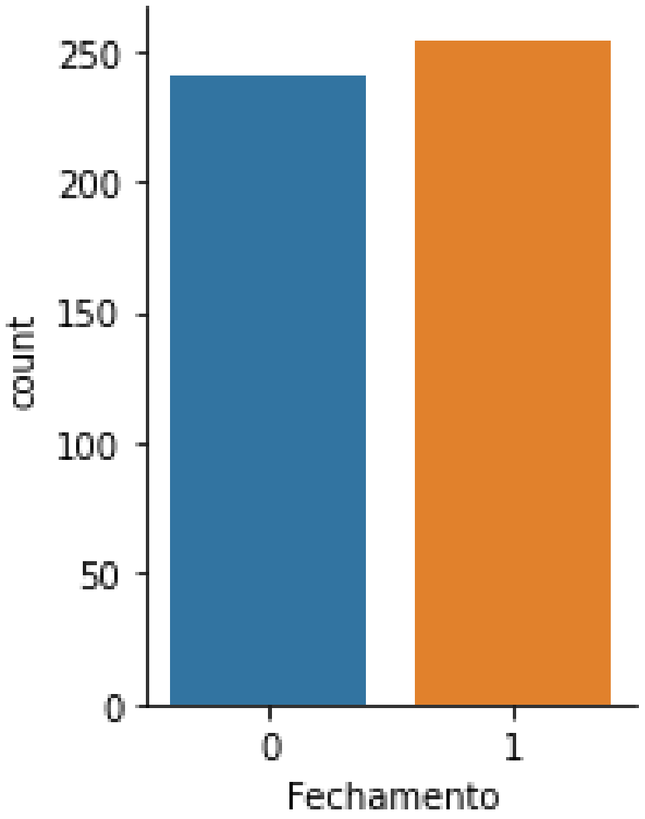
Analise 2021:

Total: **3026 news**
Média 1: **252,17 p/mês**
Média 2: **8,29 p/dia**
Máximo: **Agosto (274)**

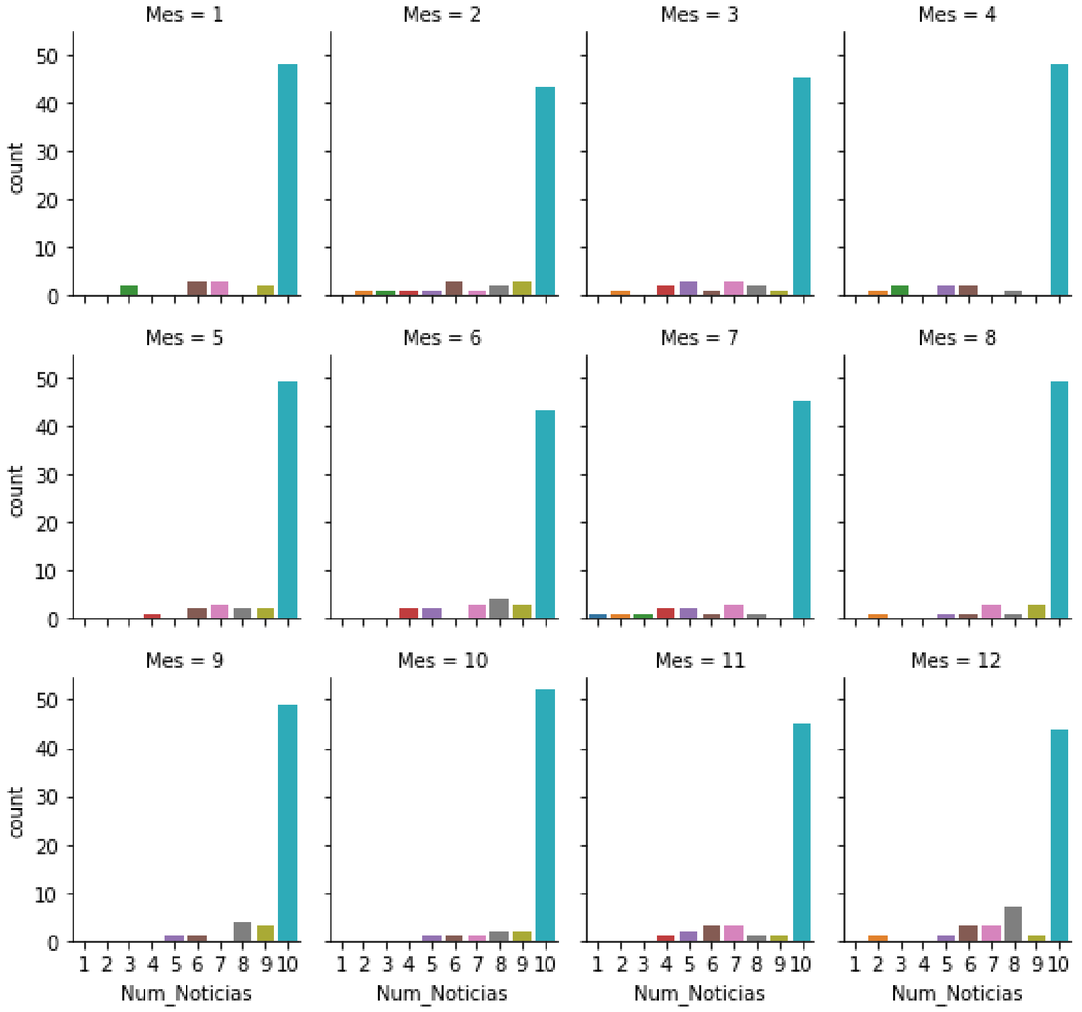


ANÁLISE DE DADOS

Fechamentos da Bolsa:



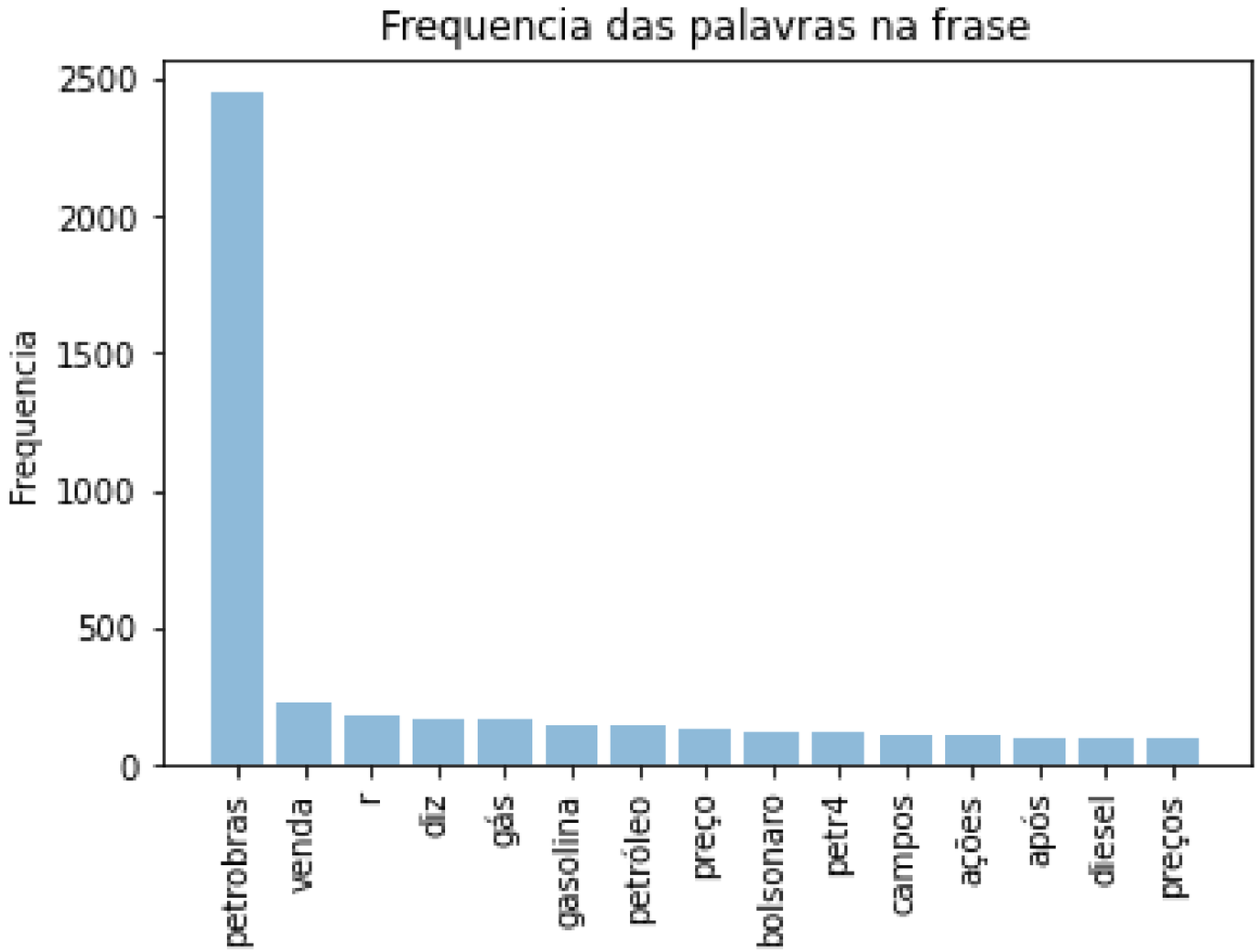
Distribuição das Notícias:



Mais de 80% dos dias com 10 News

ANÁLISE DE DADOS

Frequência de palavras:

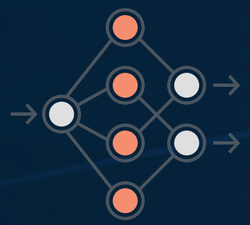


PRÉ-PROCESSAMENTO

- Limpeza: Remoção notícias não relacionadas à Petrobras
- Concatenação de título de noticias por dia
- Tratamento dias sem pregão: Concatenação ao próximo dia com pregão

MODELOS NLP

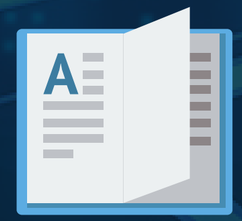
Supervisionado



roBERTa
- Negatividade
- Positividade
- Neutralidade

finBERT
- Negatividade
- Positividade
- Neutralidade

Não Supervisionado



Sentilex → Score

VaderSentiment
- Negatividade
- Positividade
- Neutralidade
- Composição

FEATURE ENGINEERING

- Notícias do dia atual influenciam no comportamento futuro do mercado de ações

11 Features dia atual

+

11 Features x 4 dias anteriores

=

55 Features

FEATURE ENGINEERING

- Frequencia de palavras: inviesamento
- Palavras inviesadas aumentam a performance na classificação

Inviesamento positivo

	palavra	qtde_baixa	qtde_alta	%_tendencia_alta
417	vale3	1	7	0.88
391	rlam	1	7	0.88
467	biodiesel	1	6	0.86
461	milhão	1	6	0.86
458	diretoria	1	6	0.86
449	df	1	6	0.86
483	verde	1	5	0.83
497	vale-gás	1	4	0.80
319	vídeo	2	8	0.80
234	defende	3	12	0.80

Inviesamento negativo

	palavra	qtde_baixa	qtde_alta	%_tendencia_alta
448	resultado	10	0	0.00
268	suspende	12	1	0.08
365	denuncia	9	1	0.10
372	analistas	7	1	0.12
395	saúde	5	1	0.17
310	fecham	10	2	0.17
407	13	10	2	0.17
277	nordeste	9	2	0.18
239	atividades	8	2	0.20
428	dilma	4	1	0.20

FEATURE ENGINEERING

TOP 500 PALAVRAS
VARIAÇÕES SENTILEX

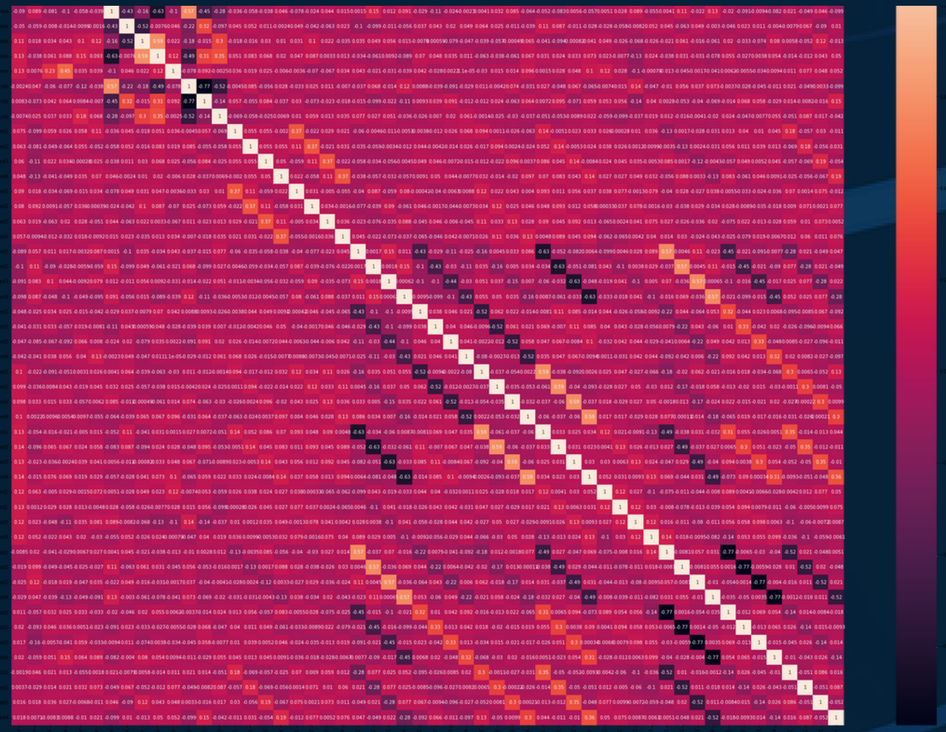
% Enviesamento	Polaridade
>= 65%	±1 á ±5
>= 70%	±1 á ±5
>= 75%	±1 á ±5
>= 80%	±1 á ±5

VIÉS 70%
INPUTAÇÃO ±3

Palavras 70% enviesadas - 3x mais importância

SELEÇÃO DE FEATURES

CORRELAÇÃO 55 FEATURES



BIBLIOTECA LAZYPREDICT

- Experimentos com normalização e PCA

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
LogisticRegression	0.727569	0.726983	0.726983	0.726944
CalibratedClassifierCV	0.725063	0.724544	0.724544	0.724493
LinearSVC	0.723810	0.723160	0.723160	0.723101
NearestCentroid	0.722055	0.721716	0.721716	0.721484
RidgeClassifierCV	0.722306	0.721678	0.721678	0.721600
LinearDiscriminantAnalysis	0.722055	0.721456	0.721456	0.721397
RidgeClassifier	0.721805	0.721210	0.721210	0.721142
SVC	0.697744	0.697443	0.697443	0.697179

1. **Sentilex** d-0: Score
2. **roBERTa** d-1: Neutralidade
3. **finBERT d-2: Negatividade**
4. **Sentilex d-3:** Score

VALIDAÇÃO ESTATÍSTICA

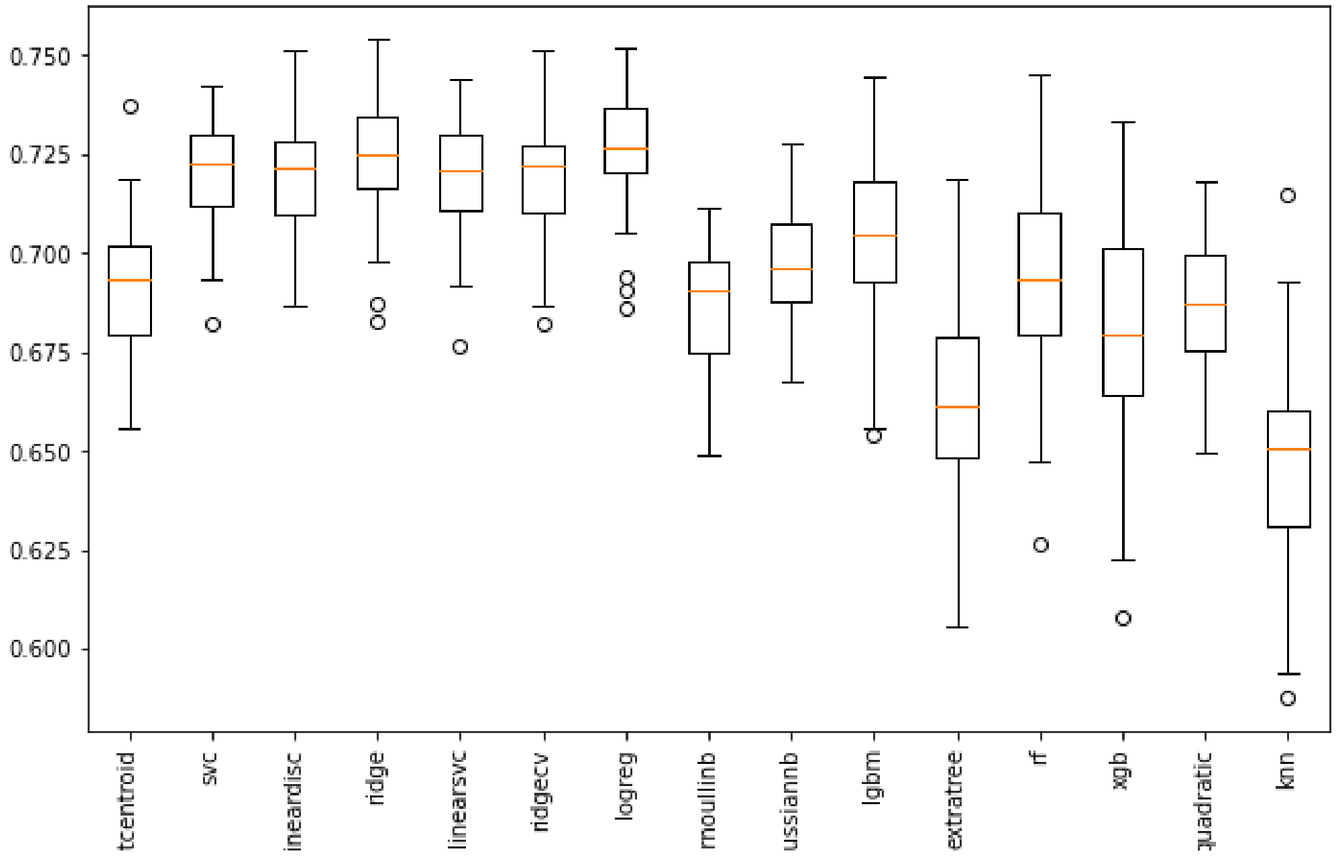
- Tunning 15 algorítmos
- Teorema do Limite Central:

30 Experimentos validação cruzada 10 kfolds

- Seleção candidatos:

SVC, Ridge, RidgeCV, Logistic Regression

- Teste de normalidade
- ANOVA
- Teste de Tukey



VALIDAÇÃO ESTATÍSTICA

H0 - Hipotese Nula: Não há diferença estatística entre os algorítmos

- ANOVA: p-valor 0.029
- Teste Tukey:

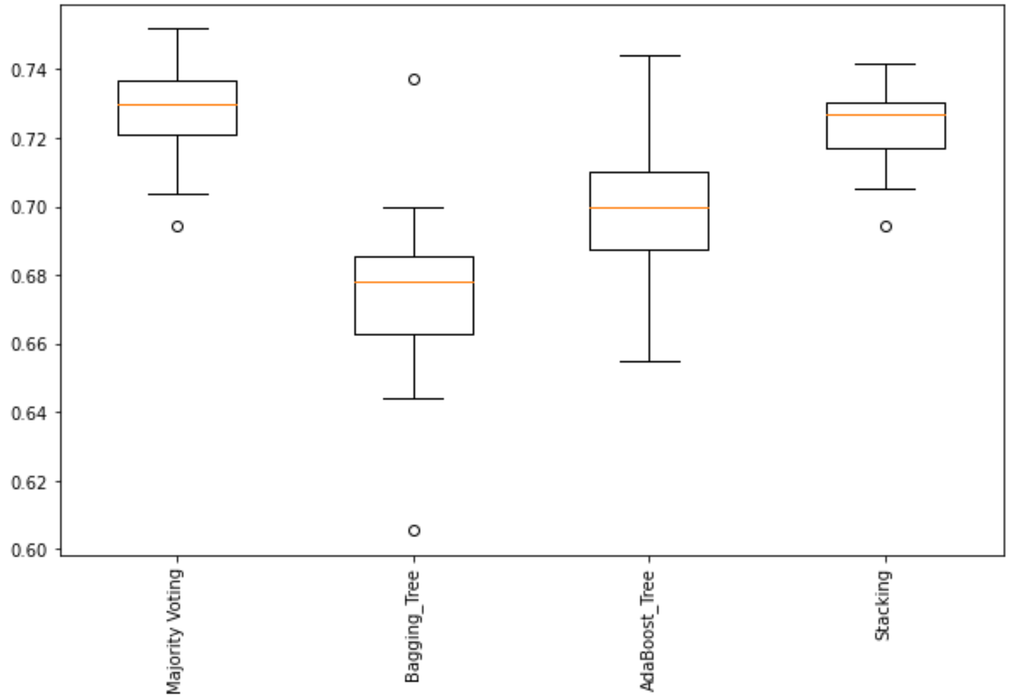
```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
 group1  group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
 logreg   ridge  -0.0028 0.6985 -0.0096  0.0041  False
 logreg ridgecv  -0.0072 0.0329 -0.0141 -0.0004   True
 logreg     svc   -0.006 0.1102 -0.0128  0.0009  False
  ridge ridgecv  -0.0045  0.327 -0.0113  0.0023  False
  ridge     svc  -0.0032 0.6038   -0.01  0.0036  False
ridgecv     svc   0.0013    0.9 -0.0055  0.0081  False
----------------------------------------------------
```

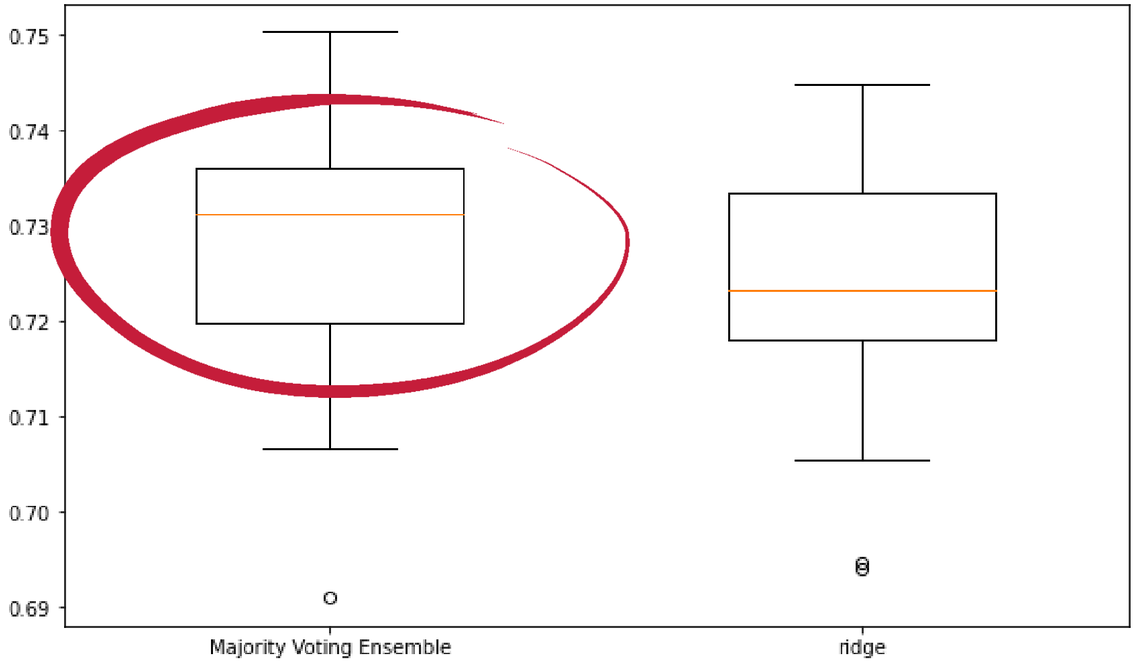
Modelo	Média Acc
LogReg	0.726056
Ridge	0.723297
SVC	0.720093
RidgeCV	0.718820

ENSEMBLES

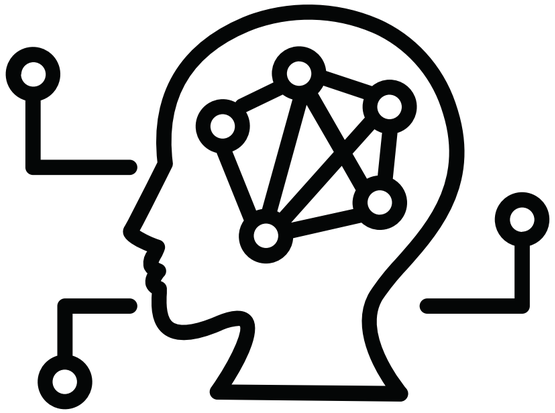
- Ensembles:

Majority Voting, Bagging, Adaboost, Stacking



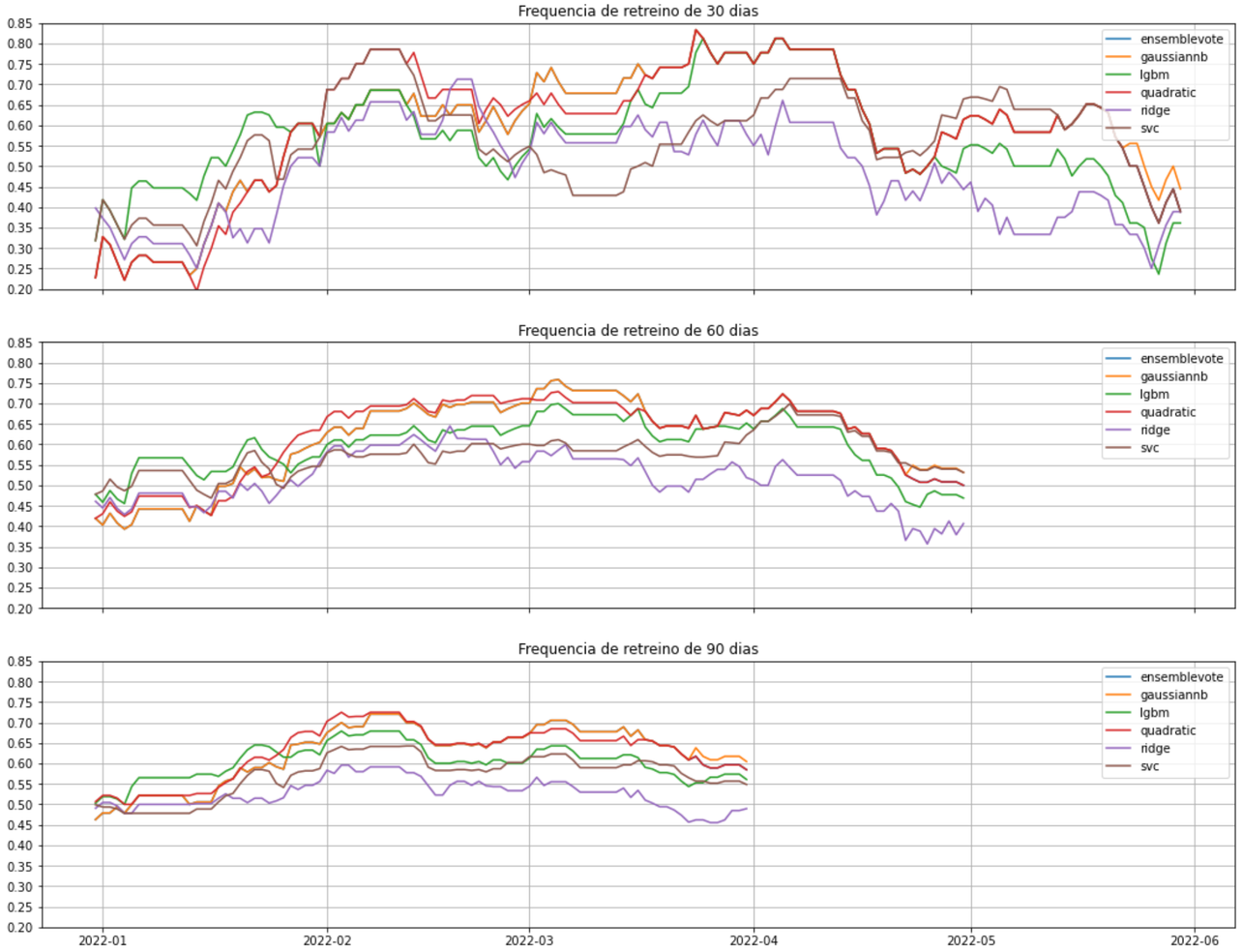


PREVISÃO 2022

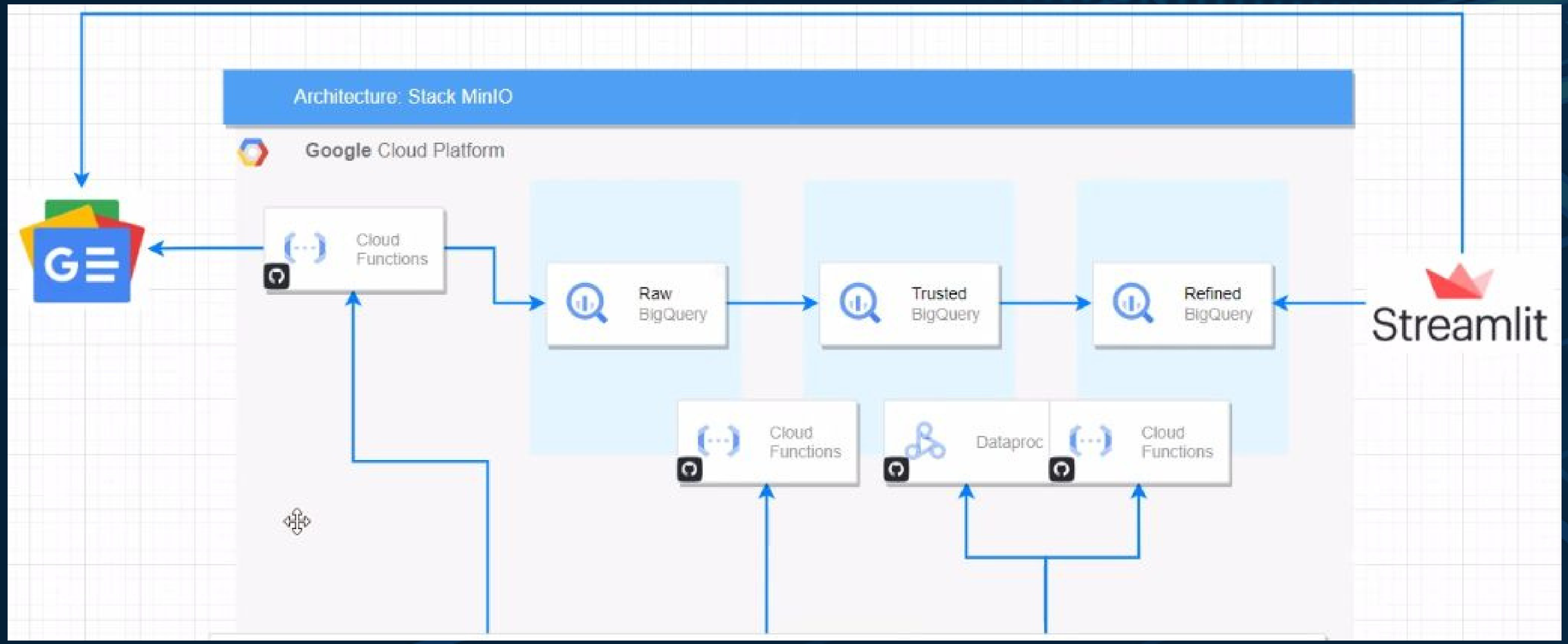


- Acompanhamento modelo em 2022
  - Baixa acurácia no inicio do ano

- Experimentos frequencia de retreino
  - 30, 60 e 90 dias



ENGENHARIA DE DADOS



DEPLOYMENT

SCAN ME

https://mariusss21-teste-googlenews-app-zfhhgo.streamlitapp.com/

OBRIGADO!!