



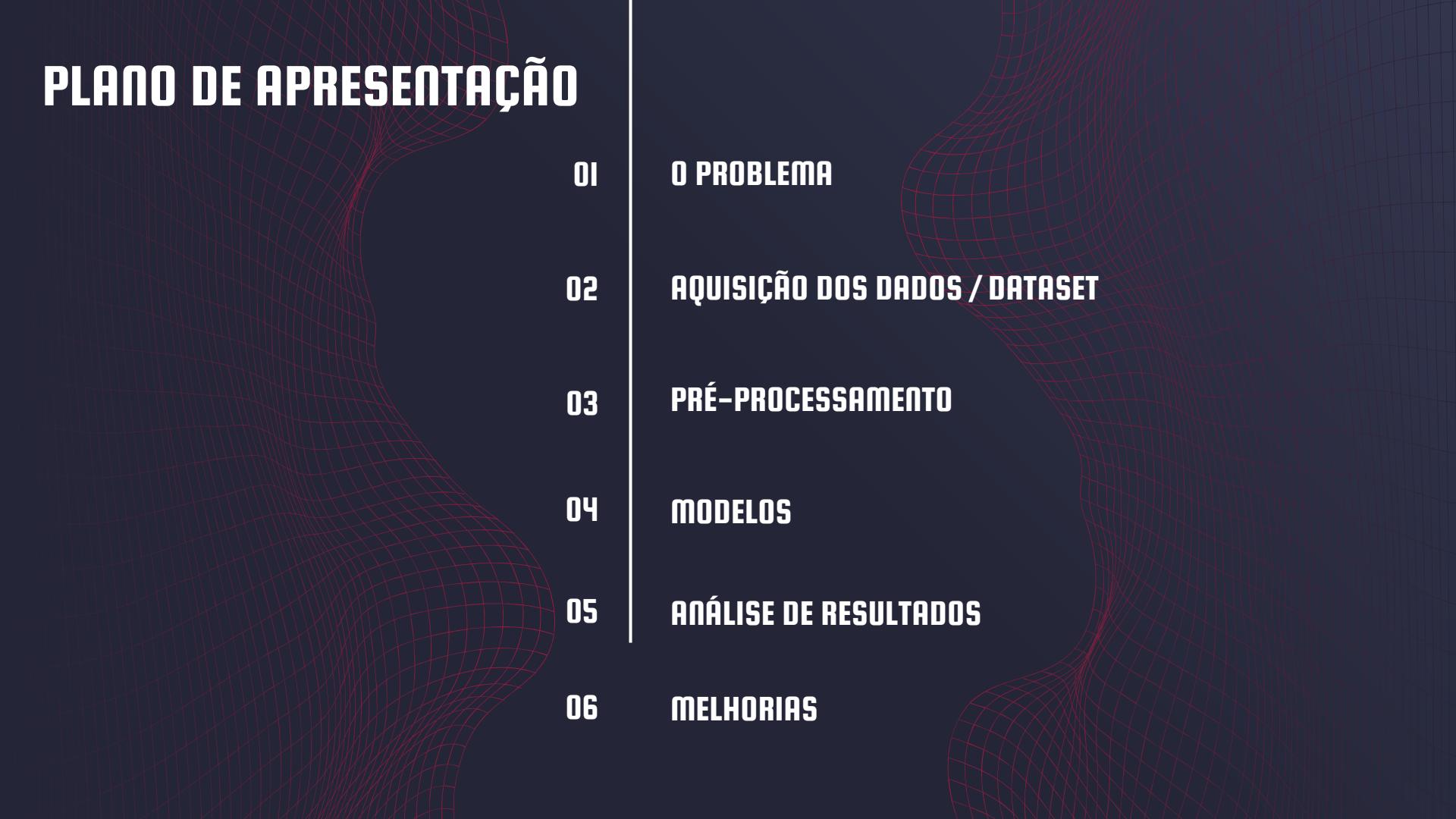
CLASSIFICAÇÃO DE PRODUTOS EM CATEGORIAS DE E-COMMERCE

ANTONIO MOREIRA - 9779242

LEONARDO MEIRELES - 4182085

JOAO PEDRO HANNAUER - 9390486

PLANO DE APRESENTAÇÃO

- 
- 01 O PROBLEMA
 - 02 AQUISIÇÃO DOS DADOS / DATASET
 - 03 PRÉ-PROCESSAMENTO
 - 04 MODELOS
 - 05 ANÁLISE DE RESULTADOS
 - 06 MELHORIAS

OI. O PROBLEMA

O PROBLEMA

magalu [Voltar ao topo](#)

Smartphone Samsung Galaxy A20 32GB Azul 4G
★★★★★ (9)
de R\$ 1.299,00 por
R\$ 999,00
10x de R\$ 99,90 sem juros

Smartphone Samsung Galaxy A20 32GB Azul 4...
★★★★★ (1)
por
R\$ 959,00
8x de R\$ 119,88 sem juros

Smartphone Samsung Galaxy A50 128GB Azul 4G
de R\$ 1.899,00 por
R\$ 1.439,10 à vista
ou R\$ 1.599,00
10x de R\$ 159,90 sem juros

Smartphone Motorola One Action 128GB Aqua Marine
★★★★★ (7)
de R\$ 1.794,00 por
R\$ 1.349,10 à vista
ou R\$ 1.499,00
10x de R\$ 149,90 sem juros

[<](#) [>](#)

mais vistos na semana

iPhone XR Apple 64GB Branco 4G Tela 6,1" Retina
★★★★★ (6)
de R\$ 5.994,00 por
R\$ 3.119,10 à vista

iPhone XR Apple 64GB Preto 4G Tela 6,1" Retina
★★★★★ (5)
de R\$ 5.994,00 por
R\$ 3.119,10 à vista

Smartphone Motorola G7 Play Edição Especial 32GB
★★★★★ (7)
de R\$ 1.099,00 por
R\$ 665,10 à vista

iPhone 8 Plus Apple 64GB Cinza Espacial 4G
★★★★★ (6)
de R\$ 4.599,00 por
R\$ 3.239,10 à vista

[<](#) [>](#)

O PROBLEMA

Informações do produto

Smartphone Samsung Galaxy A20 32GB Azul 4G - 3GB RAM 6,4" Câm. Dupla + Câm. Selfie 8MP

O Samsung Galaxy A20 azul será o seu novo Smartphone possui tela infinita de 6,4", câmera frontal de 8MP e traseira dupla de 13MP + 5MP com flash LED que, o torna ótimo para tirar fotos fantásticas e gravar vídeos em alta definição com resolução FHD (1920x1080). Dual Chip, tem também tecnologia 3G/4G que permite a transferência de dados e excelente navegação na internet, além de conectividade Wi-Fi e GPS. Conta ainda com bateria recarregável de 4000mAh, processador Octa-Core, 3GB de RAM e 32GB de armazenamento interno com possibilidade de expansão até 512GB através de cartão Micro SD. Além de tudo isso, ele oferece proteção personalizada com a tecnologia de reconhecimento facial. Desbloqueie e accese seu celular de forma fácil e segura. Basta você segurar o celular na frente do seu rosto que ele reconhecerá você. Simples assim.

Informações técnicas

Marca

Samsung

Referência

SM-A205GZBRZTO

Modelo

A20

Linha

Galaxy

Smartphone Samsung Galaxy A20 32GB Azul 4G - 3GB RAM 6,4" Câm. Dupla + Câm. Selfie 8MP

Código 155552800 | [Ver descrição completa](#) | [Samsung](#)



★★★★★ 4,7 (20) [Avaliar produto](#)

Cor:



Vendido e entregue por [magazineluiza.com](#)

de R\$ 1.299,00

por R\$ **999,00**

em 10x de R\$ 99,90 sem juros

[Mais formas de pagamento](#)

[Incluir garantia estendida e proteção roubo e furto](#)

[Adicionar à sacola](#)

[Retire na loja com frete grátis!](#) [Sobre mais](#)

[Consultar prazo e valor do frete](#)

00000-000 Ok Não sei o CEP

Azul

32GB

A memória disponível para uso do consumidor pode sofrer variações, conforme sistema operacional, aplicativos e/ou outros fatores

Micro SD

02. AQUISIÇÃO DE DADOS / DATASET

AQUISIÇÃO DE DADOS / DATASET

O processo de obtenção de dados se deu da seguinte maneira:

- Uso do framework **Scrapy** para a criação do crawler principal.
- Site escolhido para *crawlear* os dados foi o Magazine Luiza
- Extraímos o título do produto junto com sua descrição e categoria

AQUISIÇÃO DE DADOS / DATASET

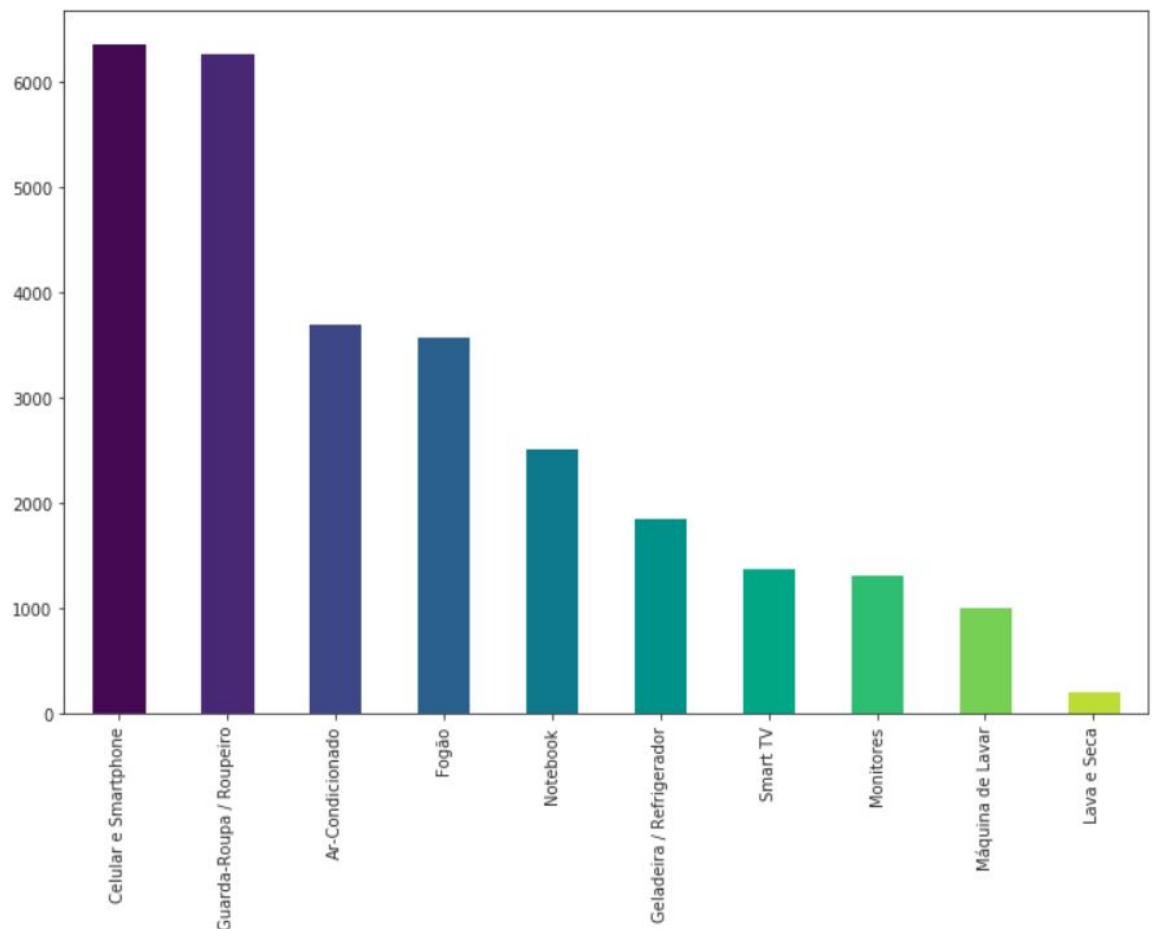
	title	category	description
0	Notebook Acer Aspire 5 A515-52-56A8 Intel Core...	Notebook	Cor Prata (partes A e C). Preto (partes B e D)...
1	Notebook Acer Aspire 5 A515-52G-522Z Intel Cor...	Notebook	Aspire 5 A515-52G-522Z - Processador Intel® Co...
2	Tela P/ Notebook Acer Aspire Es1-431-P0v7 14"...	Notebook	Enviamos Sua Tela Para Notebook Acer Aspire E...
3	Notebook Lenovo B330s-15ikbr Core I5 8250u 8gb...	Notebook	Lenovo B330s acessível e confiável. Acabament...
4	Notebook Lenovo Ideapad 330 - Tela 15.6" HD, ...	Notebook	
...
28315	Película Flexível 5D - Samsung Galaxy J2 Prime...	Celular e Smartphone	...
28316	Capa Carteira Flip Cover Asus Zenfone Max Shot...	Celular e Smartphone	01 Capa Carteira Flip Cover Asus Zenfone Max S...
28317	Carregador Portátil Pineng 10.000mah compativel...	Celular e Smartphone	Carregador Portátil Pineng 10000 A Original P...
28318	Smartphone Asus Zenfone Max Pro (M1) 32GB Dual...	Celular e Smartphone	O ZenFone Max Pro (M1) traz o melhor dos dois...
28319	Smartphone / Huawei / Mate 20 Pro / 128GB / Te...	Celular e Smartphone	Garantia : 3 meses contra defeitos de fab...

CARACTERÍSTICAS DO DATASET

10
CATEGORIAS

28330
PRODUTOS

NÃO-BALANCEADO



03. PRÉ-PROCESSAMENTO

PRÉ-PROCESSAMENTO

Etapas:

- Remoção de espaços em branco
- Tornando o texto *lowercase*
- Remoção de termos chaves (redundantes)
- Concatenação TITLE+DESCRIPTION=TEXT

PRÉ-PROCESSAMENTO

Etapas:

- Remoção de *stop words*
- Normalização (*latin_1*)
- *Tokenizer*
- *TF/IDF: Term Freq./Inverse Document Freq.*

PRÉ-PROCESSAMENTO

	category	text
0	Notebook	notebook aspire intel ageracao memoria windows...
1	Notebook	notebook aspire a-g-z intel geracao geforce as...
2	Notebook	notebook aspire neide notebook enviamos notebo...
3	Notebook	notebook lenovo bs-ikbr radeon windows preto l...
4	Notebook	notebook lenovo ideapad intel intel graphics s...
...
28315	Celular e Smartphone	pelicula flexivel samsung galaxy prime grand p...
28316	Celular e Smartphone	carteira cover zenfone maston carteira cover z...
28317	Celular e Smartphone	carregador portatil pineng compativel motorola...
28318	Celular e Smartphone	smartphone zenfone qualcomm snapdragon camera ...
28319	Celular e Smartphone	smartphone huawei camera wi-fi contra defeitos...

PRÉ-PROCESSAMENTO

NÚMERO DE PALAVRAS:

1.975.812

ANTES DO
PRÉ-PROCESSAMENTO

982.309

DEPOIS DO
PRÉ-PROCESSAMENTO

04. MODELOS

MULTINOMIAL NAIVE BAYES:

- A DISTRIBUIÇÃO MULTINOMIAL É BOA NA PRÁTICA PARA DADOS QUE PODEM SER TRANSFORMADOS EM FREQUÊNCIA/CONTAGEM
- CALCULA A PROBABILIDADE DE UMA CLASSE DADO QUE DOCUMENTO OU INSTÂNCIA DE DADO ACONTECEU
- BASTANTE UTILIZADO EM MODELAGEM DE DADOS TEXTUAIS

MULTILAYER PERCEPTRON:

- **Baseada nos conceitos Feedforward e Backpropagation**
- **Tende a apresentar bons resultados em problemas em dados com formato textual**
- **Resultados relevantes em problemas de classificação com classes bem definidas**

SUPPORT VECTOR MACHINE:

- **PORQUÊ SVM?**
 - **ESPAÇO DE ALTA DIMENSIONALIDADE;**
 - **POUCAS CARACTERÍSTICAS IRRELEVANTES;**
 - **O ESPAÇO É ESPARSO;**
 - **MUITOS PROBLEMAS DE CLASSIFICAÇÃO DE TEXTO SÃO LINEARMENTE SEPARÁVEIS.**
- **LINEAR SVM**
- **TREINAMENTO ‘OVR’ – ONE-VS-REST**

05. RESULTADOS

TÉCNICAS:

VALIDAÇÃO DO MULTINOMIAL NAIVE BAYES

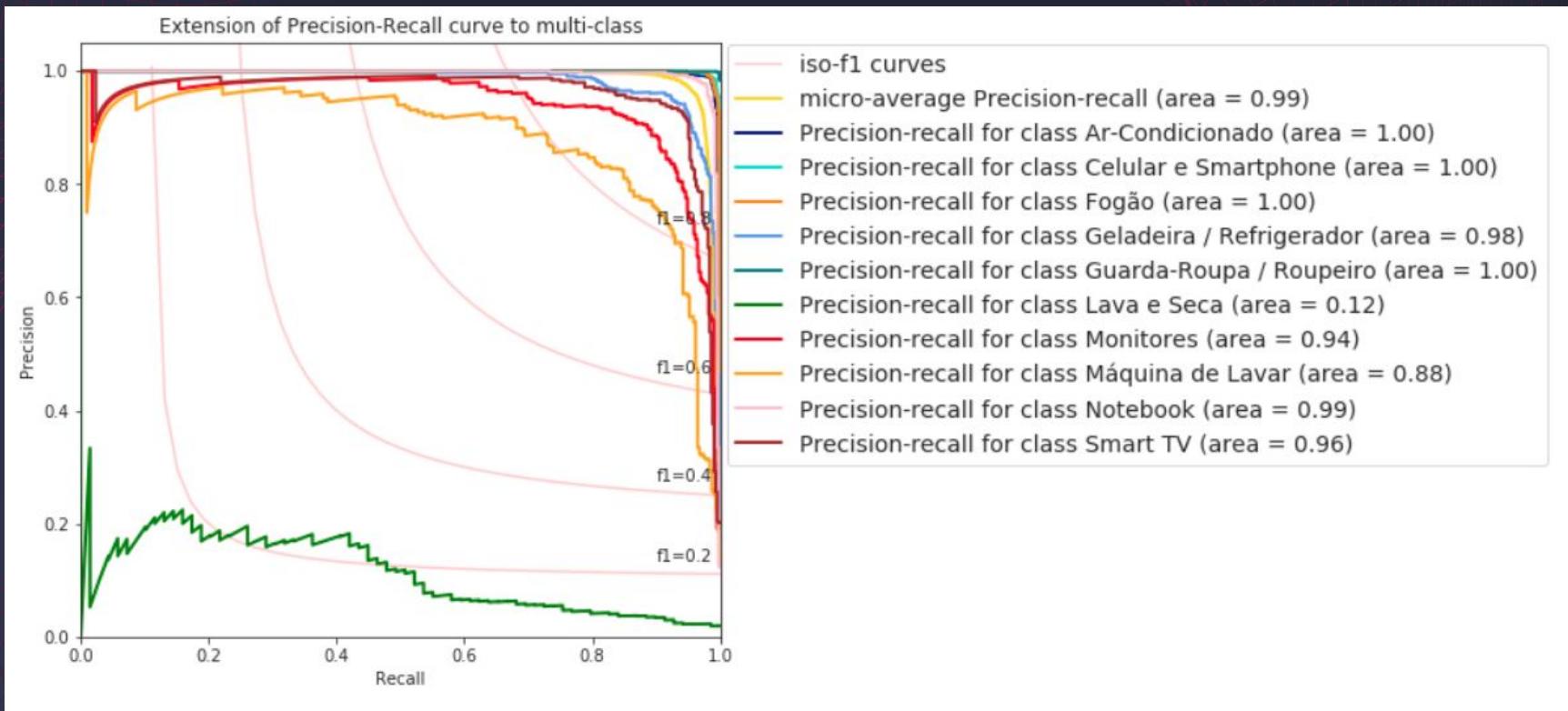
- ANÁLISE DAS CURVAS DE *PRECISION-RECALL*
- CLASSIFICADOR BINÁRIO P/ UMA CLASSE DE BAIXA OCORRÊNCIA NO DATASET (*LAVA E SECA*)
- *K-FOLD CROSS VALIDATION*

ANÁLISE DAS CURVAS PR

accuracy 0.9117543240381222

	precision	recall	f1-score	support
Ar-Condicionado	0.99	0.97	0.98	1110
Celular e Smartphone	0.99	0.98	0.99	1924
Fogão	0.99	0.98	0.99	1082
Geladeira / Refrigerador	0.97	0.83	0.89	546
Guarda-Roupa / Roupeiro	0.99	1.00	1.00	1895
Lava e Seca	0.33	0.01	0.03	69
Monitores	0.98	0.58	0.73	390
Máquina de Lavar	0.92	0.54	0.68	311
Notebook	0.99	0.95	0.97	747
Smart TV	0.99	0.71	0.83	425
micro avg	0.99	0.91	0.95	8499
macro avg	0.92	0.76	0.81	8499
weighted avg	0.98	0.91	0.94	8499
samples avg	0.91	0.91	0.91	8499

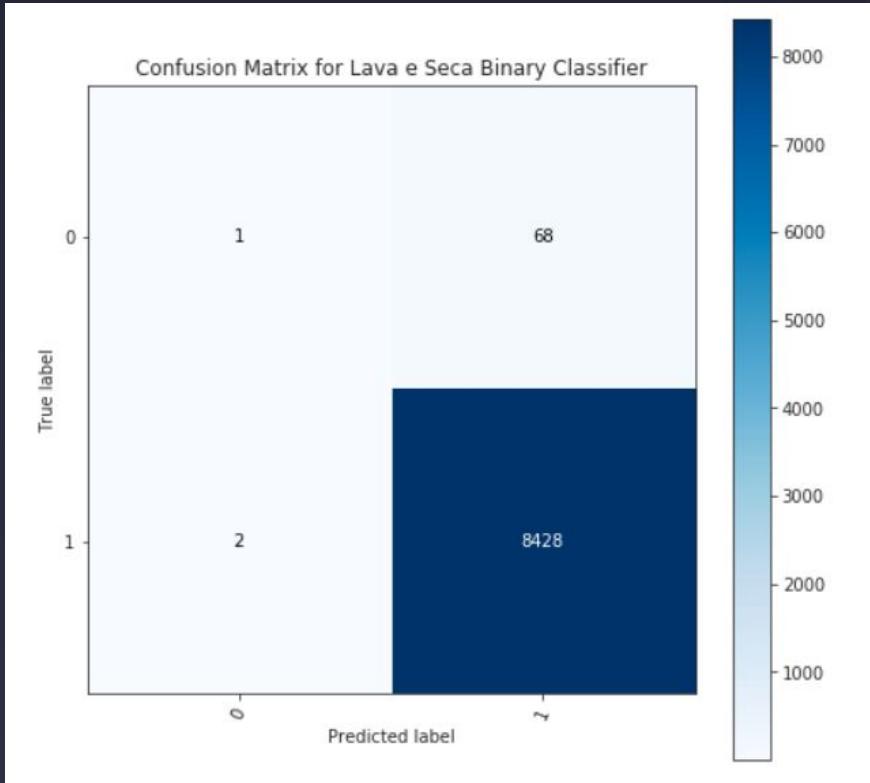
ANÁLISE DAS CURVAS PR



CLASSIFICADOR BINÁRIO PARA LAVA E SECA

	precision	recall	f1-score	support
0	0.33	0.01	0.03	69
1	0.99	1.00	1.00	8430
accuracy			0.99	8499
macro avg	0.66	0.51	0.51	8499
weighted avg	0.99	0.99	0.99	8499

CLASSIFICADOR BINÁRIO PARA LAVA E SECA



K-FOLD CROSS VALIDATION

K = 10

Accuracy Mean Score: 0.970
Precision Mean Score: 0.873
Recall Mean Score: 0.858

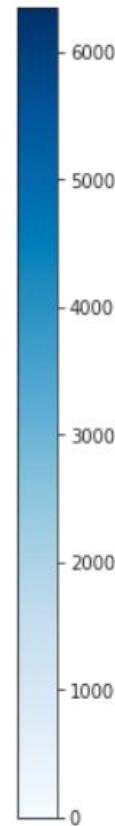
K-FOLD CROSS

VALIDATION

K = 10

Confusion Matrix of K-Fold validation

	Ar-Condicionado	Celular e Smartphone	Fogão	Geladeira / Refrigerador	Guarda-Roupa / Roupeiro	Lava e Seca	Monitores	Máquina de Lavar	Notebook	Smart TV	
Ar-Condicionado	3679	15	9	2	8	0	0	0	0	0	0
Celular e Smartphone	4	6355	5	0	1	0	0	0	5	0	0
Fogão	1	4	3578	6	7	0	0	0	0	0	0
Geladeira / Refrigerador	27	13	46	1766	20	0	0	0	0	0	0
Guarda-Roupa / Roupeiro	0	0	0	0	6286	0	0	0	0	0	0
Lava e Seca	44	35	18	2	7	0	0	102	1	6	
Monitores	4	104	16	0	23	0	1168	0	5	6	
Máquina de Lavar	28	13	65	11	7	0	0	897	0	0	
Notebook	1	39	10	0	4	0	1	0	2479	0	
Smart TV	6	81	4	1	21	0	2	0	0	1282	



VALIDAÇÃO DO MULTILAYER PERCEPTRON

- Accuracy Score analysis
- Análise de curvas Precision-Recall

ACCURACY SCORE

Training...

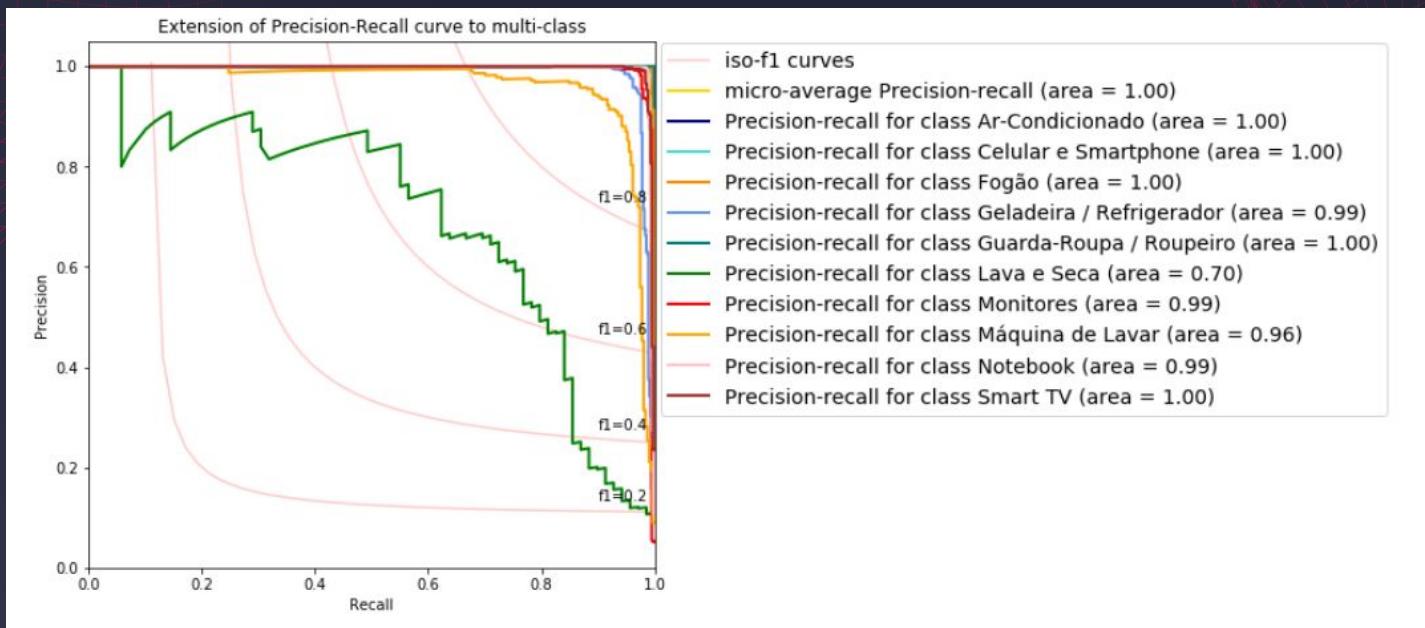
accuracy 0.9765854806447818

	precision	recall	f1-score	support
Ar-Condicionado	0.99	0.99	0.99	1110
Celular e Smartphone	1.00	0.99	0.99	1924
Fogão	1.00	0.99	1.00	1082
Geladeira / Refrigerador	0.99	0.94	0.97	546
Guarda-Roupa / Roupeiro	1.00	1.00	1.00	1895
Lava e Seca	0.68	0.62	0.65	69
Monitores	0.99	0.96	0.98	390
Máquina de Lavar	0.96	0.89	0.92	311
Notebook	1.00	0.98	0.99	747
Smart TV	1.00	0.97	0.98	425
micro avg	0.99	0.98	0.99	8499
macro avg	0.96	0.93	0.95	8499
weighted avg	0.99	0.98	0.99	8499
samples avg	0.98	0.98	0.98	8499

CPU times: user 2min 40s, sys: 1min 12s, total: 3min 52s

Wall time: 59.1 s

PRECISION-RECALL CURVES



VALIDAÇÃO DA SVM

TÉCNICAS:

- ONE VS REST
- VARIAÇÃO DE *MAX_FEATURES* NO *TF/IDF*
 - MELHOR N° DE FEATURES
- *K-FOLD CROSS VALIDATION* PARA MELHOR ESPAÇO ENCONTRADO

ENCONTRANDO MAX_FEATURES

Len space: 22549

Feature Names:

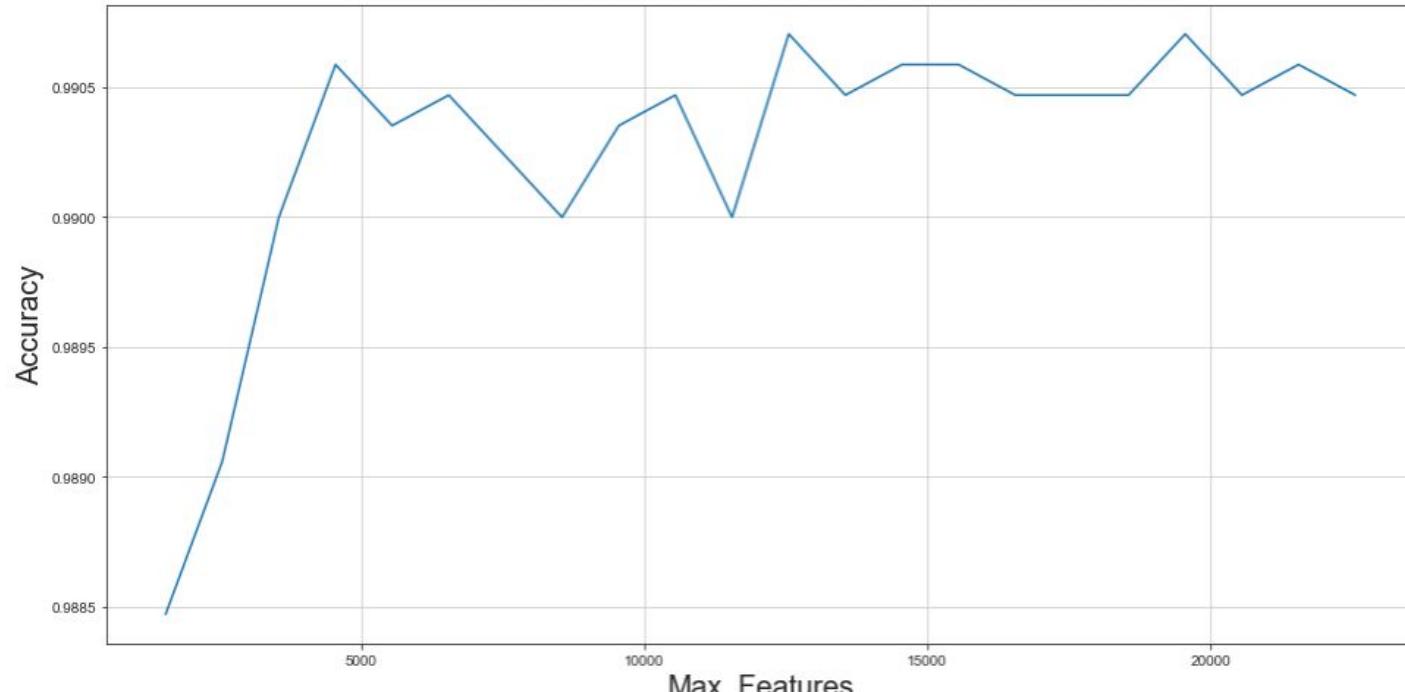
```
['abriu', 'abs', 'absdimensoes', 'absolut', 'absoluta', 'absolutamente', 'absolute', 'absoluto', 'absolutos', 'absorcao', 'absorva', 'absorve', 'absorvem', 'absorvendo', 'absorver', 'absorvicao', 'absorvidos', 'abstract', 'abstrato', 'abst', 'ratos', 'abusa', 'abusar', 'abuse', 'abwegs', 'abwfbrs', 'abwgbrs', 'ac', 'acaba', 'acabado', 'acabam', 'acabamento', 'acabamentos', 'acabando', 'acabar', 'acabaram', 'acabe', 'acabou', 'acacia', 'academia', 'acalcp']
```

Accuracy 0.9904694669961172

	precision	recall	f1-score	support
Ar-Condicionado	0.99	1.00	1.00	1110
Celular e Smartphone	0.99	1.00	0.99	1924
Fogão	1.00	1.00	1.00	1082
Geladeira / Refrigerador	0.98	0.97	0.98	546
Guarda-Roupa / Roupeiro	1.00	1.00	1.00	1895
Lava e Seca	0.82	0.71	0.76	69
Monitores	0.99	0.99	0.99	390
Máquina de Lavar	0.95	0.95	0.95	311
Notebook	1.00	0.99	0.99	747
Smart TV	1.00	0.98	0.99	425
accuracy			0.99	8499
macro avg	0.97	0.96	0.96	8499
weighted avg	0.99	0.99	0.99	8499

VARIANDO MAX_FEATURES 22549 → 1549

Impact of the number of features on the SVM classifier
as the result of the TF/IDF trasformation



MELHOR MAX_FEATURES: 1549

Len space: 1549

Feature Names:

```
['armazenar', 'armyshield', 'arquivos', 'arranhoes', 'artes', 'artificial', 'as', 'aspire', 'assim', 'assista', 'assistencia', 'assistir', 'atencao', 'atender', 'atendidos', 'atendimento', 'atente', 'aterramento', 'atinge', 'ativado', 'ativa', 'atividades', 'atlas', 'atoxico', 'atraente', 'atraves', 'atualizacao', 'atualle', 'audio', 'aumenta', 'aumentando', 'aumentar', 'auto', 'autolimpante', 'automatica', 'automaticamente', 'automatico', 'autonomia', 'auxiliar', 'avancado']
```

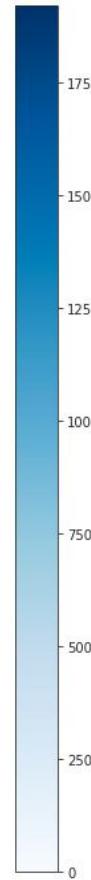
Accuracy 0.9884692316743147

	precision	recall	f1-score	support
Ar-Condicionado	0.99	0.99	0.99	1110
Celular e Smartphone	0.99	1.00	1.00	1924
Fogão	0.99	1.00	0.99	1082
Geladeira / Refrigerador	0.98	0.97	0.97	546
Guarda-Roupa / Roupeiro	1.00	1.00	1.00	1895
Lava e Seca	0.79	0.64	0.70	69
Monitores	0.98	0.98	0.98	390
Máquina de Lavar	0.93	0.95	0.94	311
Notebook	1.00	0.99	0.99	747
Smart TV	0.99	0.98	0.99	425
accuracy			0.99	8499
macro avg	0.96	0.95	0.96	8499
weighted avg	0.99	0.99	0.99	8499

CONFUSION MATRIX

Confusion Matrix for SVM Classifier

	Ar-Condicionado	2	2	2	0	0	0	0	0	1
Ar-Condicionado	1103	2	2	2	0	0	0	0	0	1
Celular e Smartphone	0	1920	1	1	0	0	1	0	0	1
Fogão	0	0	1077	2	2	0	0	1	0	0
Geladeira / Refrigerador	4	3	5	528	1	1	1	2	1	0
Guarda-Roupa / Roupeiro	0	0	0	0	1895	0	0	0	0	0
Lava e Seca	3	1	1	2	0	44	0	18	0	0
Monitores	1	2	0	0	0	0	384	0	2	1
Máquina de Lavar	1	1	1	3	0	11	0	294	0	0
Notebook	0	5	1	1	0	0	1	0	739	0
Smart TV	2	1	1	0	1	0	3	0	0	417



10-FOLD CROSS VALIDATION

0.9895

ACURÁCIA MÉDIA

06. MELHORIAS

MELHORIAS

NA AQUISIÇÃO DE DADOS:

- CRAWLING DE OUTRAS PLATAFORMAS
- BUSCAR DESCRIÇÕES DO MESMO PRODUTO EM PLATAFORMAS DIFERENTES

MELHORIAS

NA PRÉ-PROCESSAMENTO:

- AUMENTAR AS KEYWORDS DO DOMÍNIO
- UTILIZAR OUTRO MÉTODO ALÉM DO TF-IDF COMO EMBEDDING (GLOVE, WORD2VEC)

OBRIGADO!

Perguntas?