THE DEVELOPER'S
CONFERENCE

Trilha – Machine Learning

Fabíola S. F. Pereira

De pai pra filho! A Tarefa de Classificação Hierárquica aplicada em um Case Real Sobre Dados Textuais

Fabíola S. F. Pereira

Data Scientist @ ZUP IT

Olá! Muito prazer!

- Uberlândia/MG
- [Indústria] Cientista de Dados @ Zup IT
- [Pesquisa] Posdoc @ ICMC USP

O CASE DE ML

Interação com o Cliente

Modelagem do problema

DeSenvolvimento com ML

ReSultadoS

1º ATO

A CONVERSA COM O CLIENTE

Temos uma tabela Excel

Temos uma tabela Excel

nessa tabela tem textos

Temos uma tabela Excel

nessa tabela tem textos

atualmente categorizamos esses textos manualmente

Temos uma tabela Excel

nessa tabela tem textos

atualmente categorizamos esses textos manualmente

queremos categorizar automaticamente!

Quais textos vocês têm?

“cliente reclama que a Internet não está funcionando”

“o modem está com as luzes piscando, mas não consegue conectar. Já tentou resetar e nada aconteceu”

“ontem a Internet estava normal e hoje não está funcionando mais. Cliente solicitou aumento da velocidade de conexão”

“Cliente relata que teve uma chuva forte ontem e depois da chuva teve queda de energia e tudo parou de funcionar”

Quais categorias vocês têm?

1. *Bug no sistema (TI)*
2. *Bug no sistema (TI) - exceção*
3. *Problema de Rede*
4. *Cabeamento danificado*
5. *Cabeamento danificado – vandalismo*
6. *Cabeamento antigo*
7. *Erro de configuração para o perfil do cliente*
8. *Erro de configuração para clientes de um grupo*
9. *Erro de configuração para clientes ADSL*
10. *Erro de configuração para clientes Fibra*

...

96.(são 96!)

Então vocês já têm uma massa de dados categorizada para utilizarmos como exemplo?

Então vocês já têm uma massa de dados categorizada para utilizarmos como exemplo?

Quantos dados de exemplo vocês têm de cada categoria?

Então vocês já têm uma massa de dados categorizada para utilizarmos como exemplo?

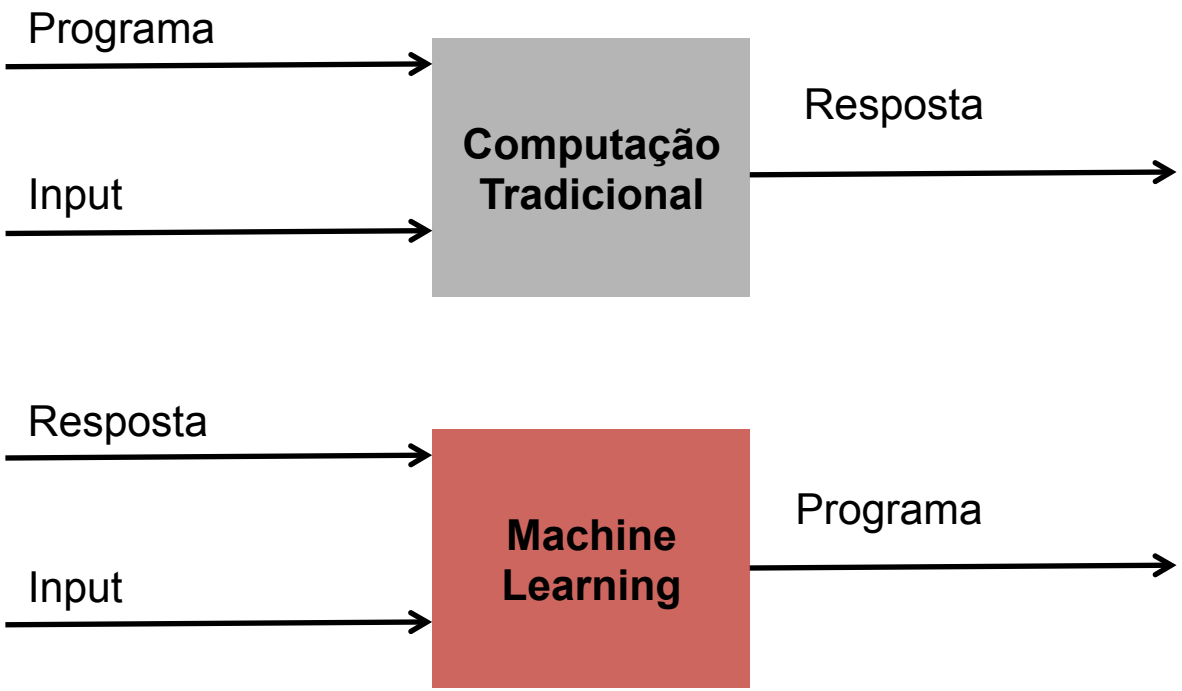
Quantos dados de exemplo vocês têm de cada categoria?

2º ATO

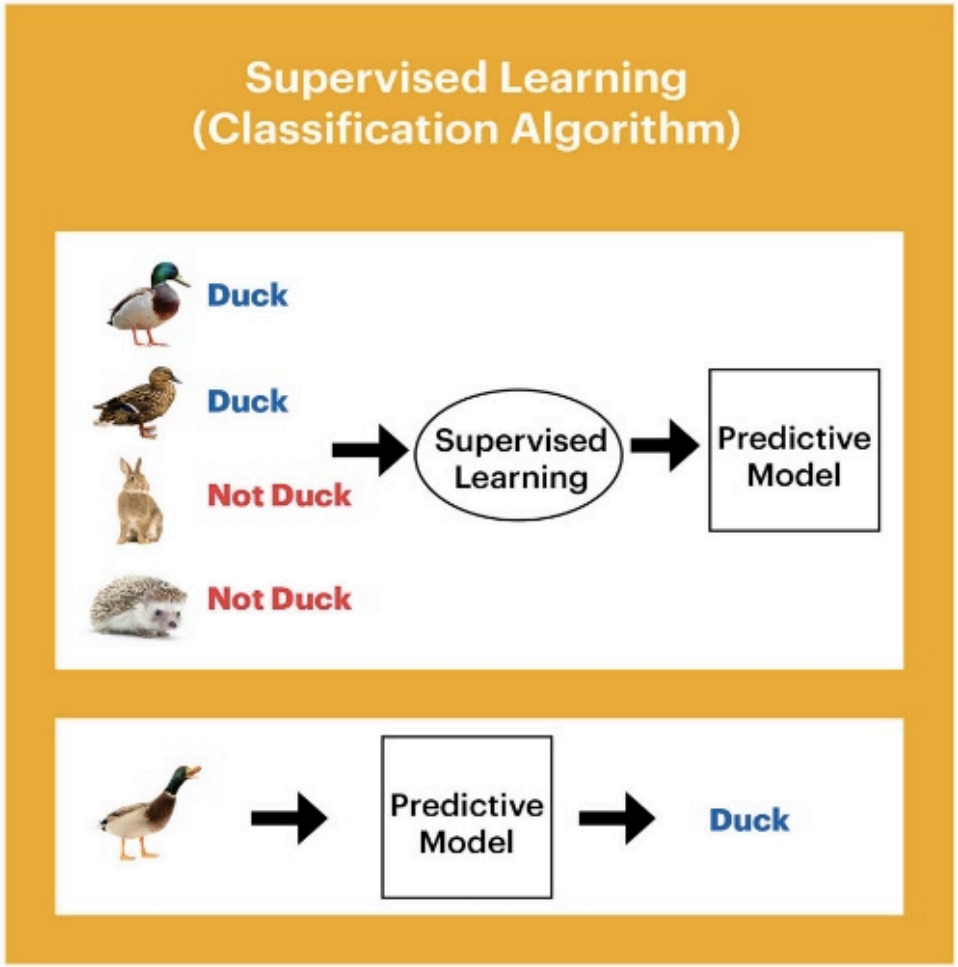
MODELAGEM DO PROBLEMA

Machine Learning é...

Machine Learning é...

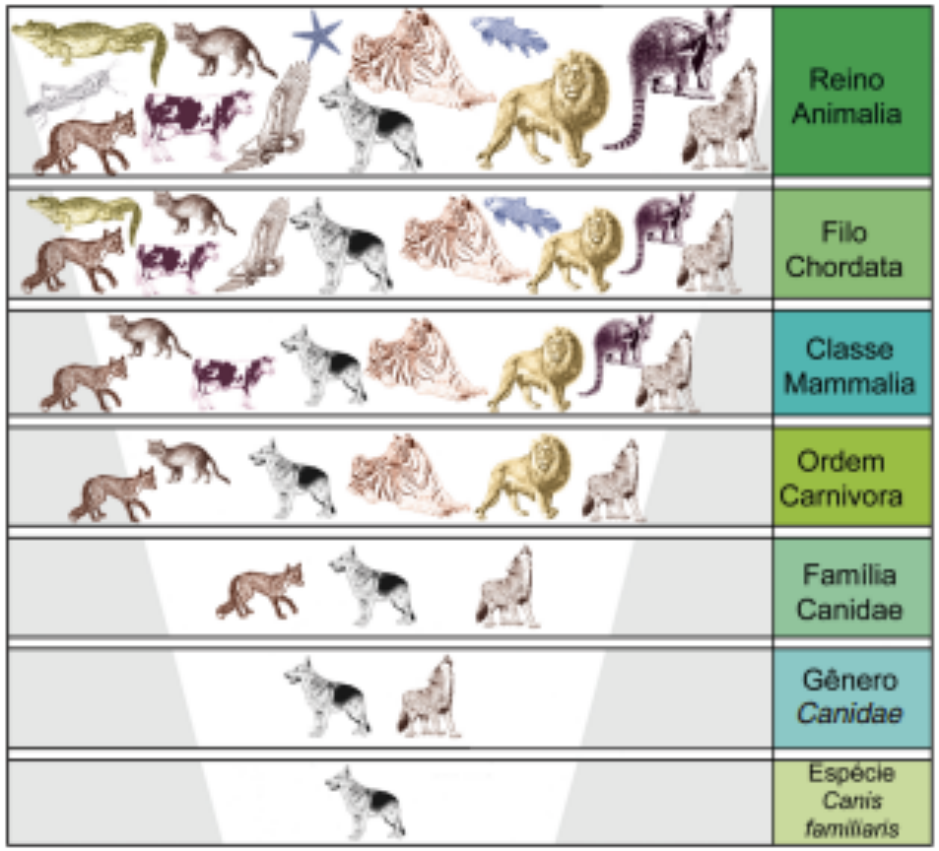


A tarefa de classificação



https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline/

A tarefa de classificação hierárquica



https://descomplica.com.br/blog/biologia/por-que-e-importante-classificar-os-seres-vivos/

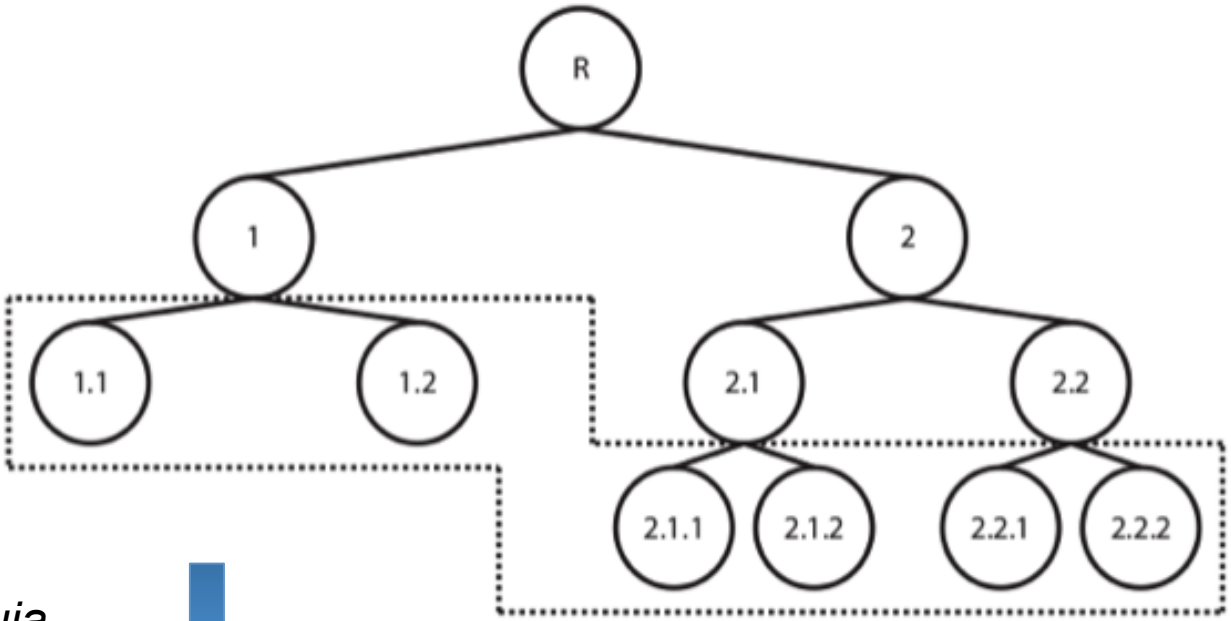
A tarefa de classificação hierárquica

abordagem flat

abordagem por hierarquia

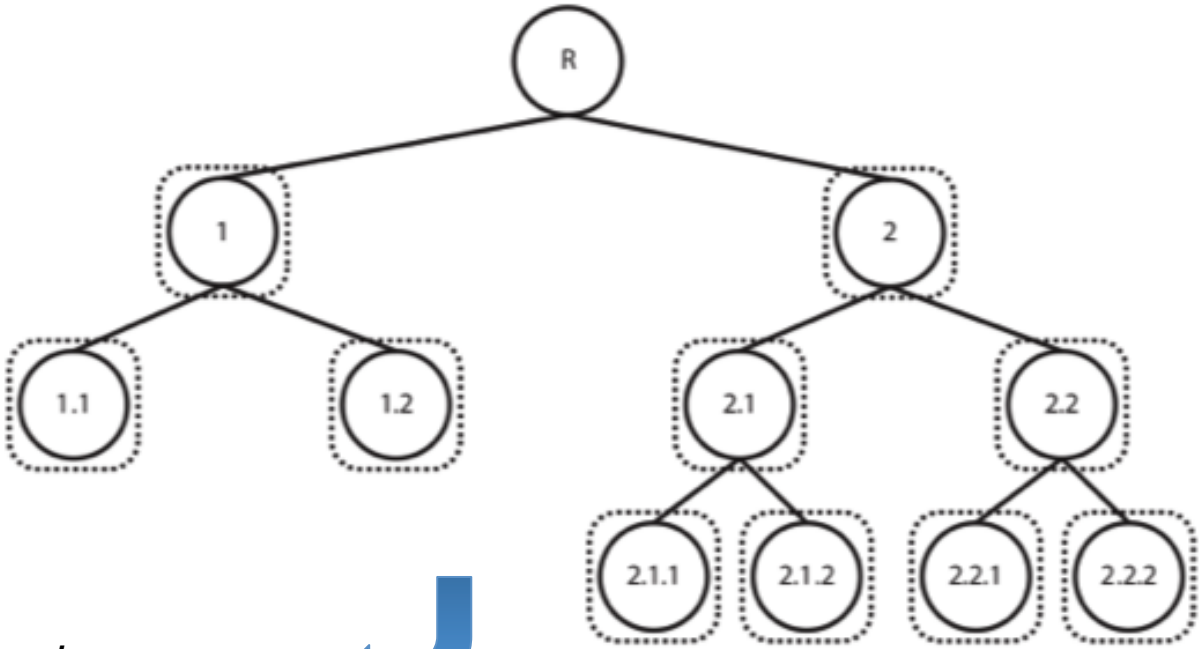
A tarefa de classificação hierárquica

FLAT



não existe uma hierarquia entre as categorias a serem preditas

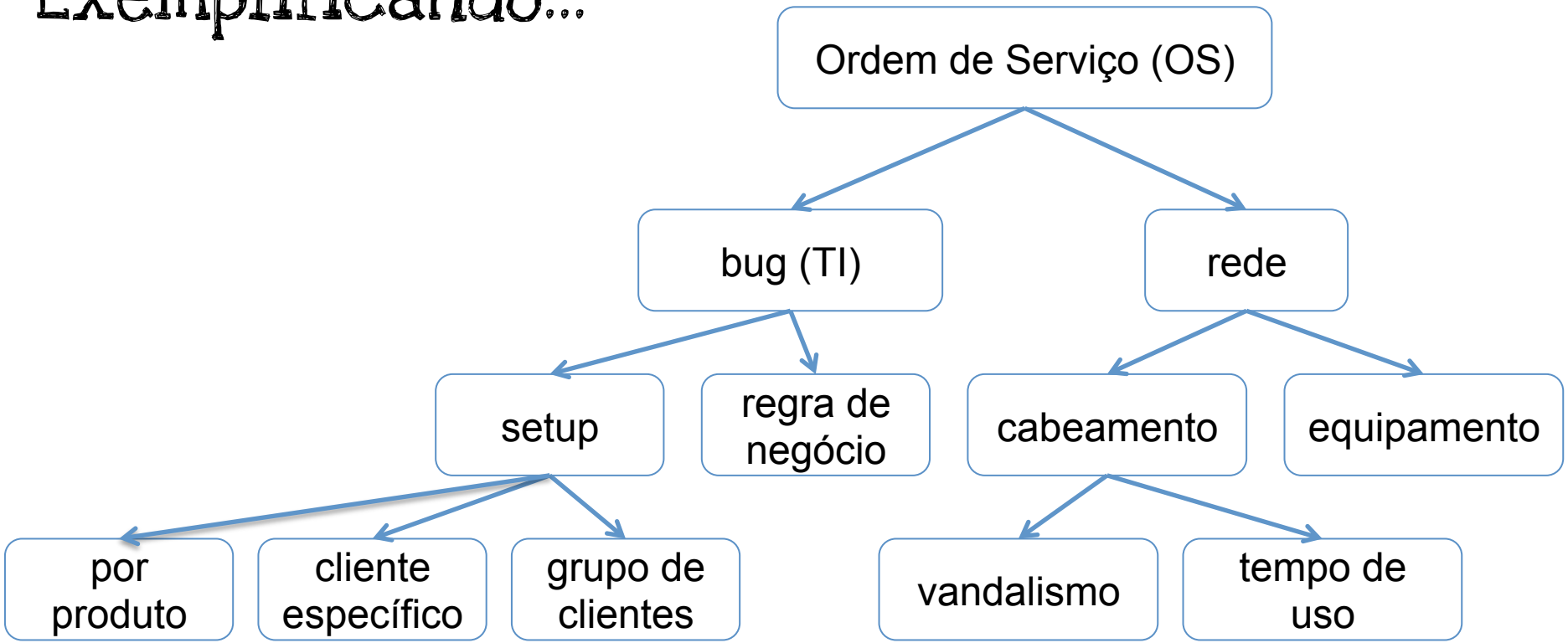
A tarefa de classificação hierárquica



HIERÁRQUICA

cada categoria a ser predita deve ser levada em consideração, sendo tais categorias organizadas em uma hierarquia

Exemplificando...



Decisão de modelagem	Cliente
Aprendizado supervisionado	é importante obter todas as categorias
Classificação hierárquica	fornecerá uma hierarquia dessas categorias
Abordagem hierárquica	fornecerá amostras de exemplo suficientes para cada categoria

Decisão de modelagem

Cliente

Aprendizado supe

obter todas as categorias

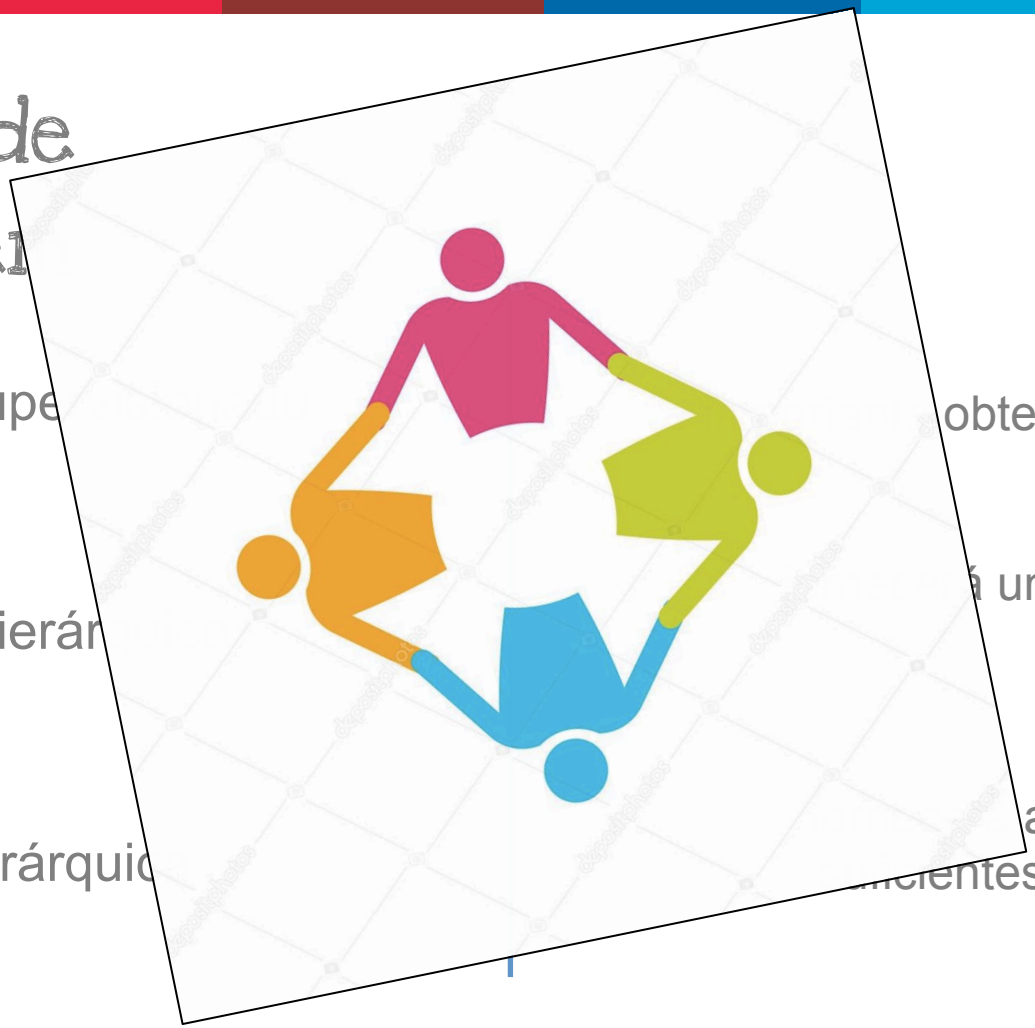
Classificação hierár

á uma hierarquia dessas categorias

Abordagem hierárquic

amostras de exemplo

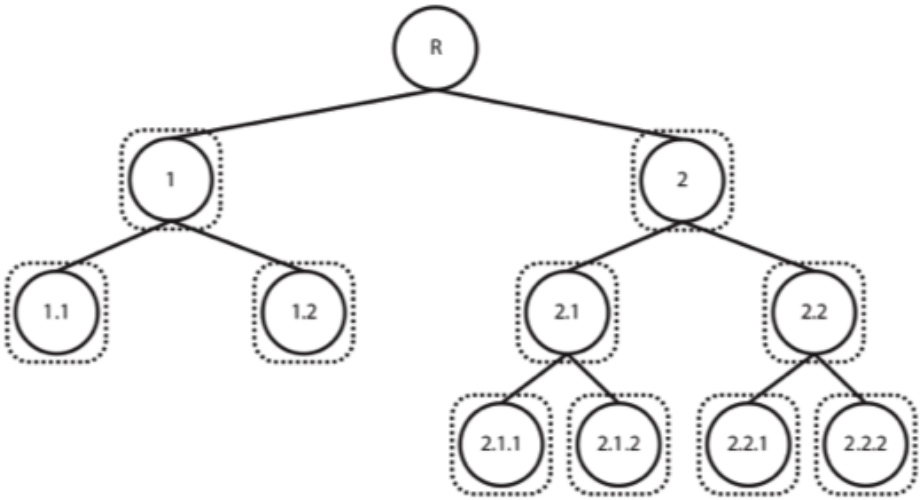
clientes para cada categoria



3º ATO

DESENVOLVIMENTO DA SOLUÇÃO

Para cada nó da árvore



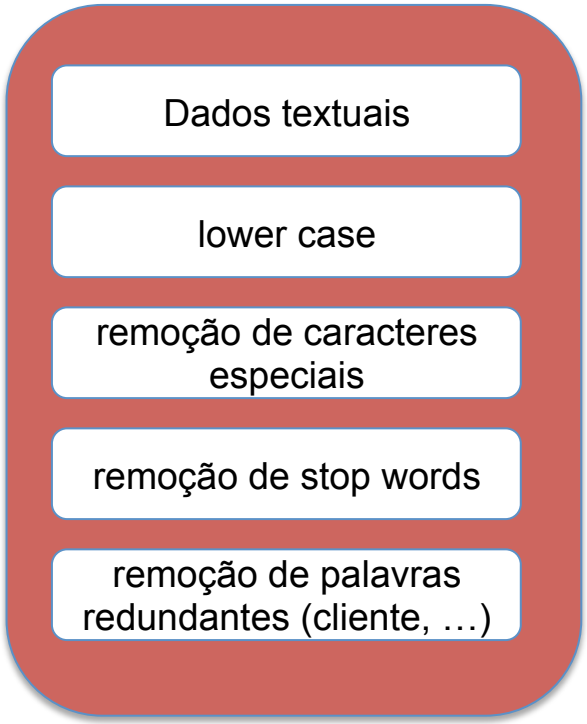
Pipeline de NLP

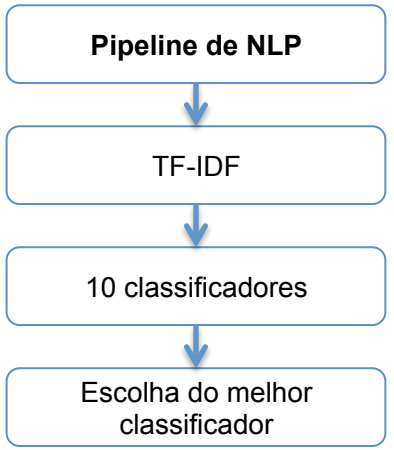
TF-IDF

10 classificadores

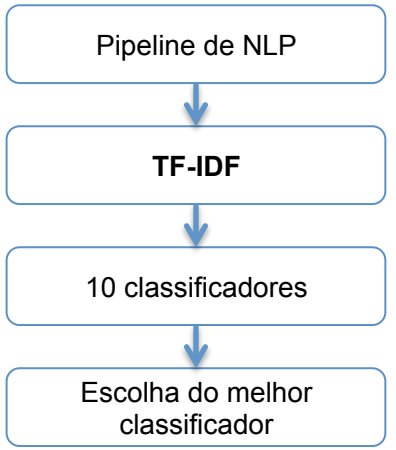
Escolha do melhor classificador

Pipeline de NLP





TF-IDF

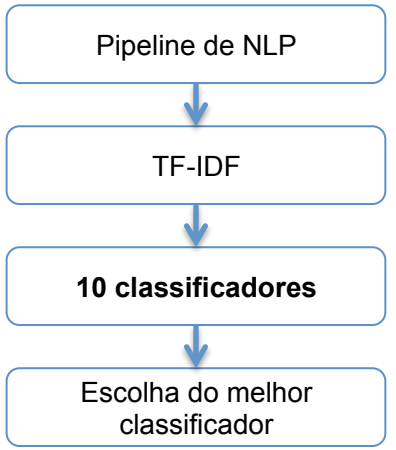


words =	aceit	alug	cart	credito	de	Mastercard	moto	qua	que	tem	tipo	vcs
bag1	1	0	1	1	1	0	0	0	0	0	0	1
bag2	1	0	0	0	0	1	0	0	0	0	0	1
bag3	1	0	1	0	0	0	0	0	0	0	0	0
bag4	0	0	0	0	0	0	1	1	0	1	0	1
bag5	0	0	0	0	0	0	1	0	1	1	1	1
bag6	0	1	0	0	0	0	0	0	1	0	0	1

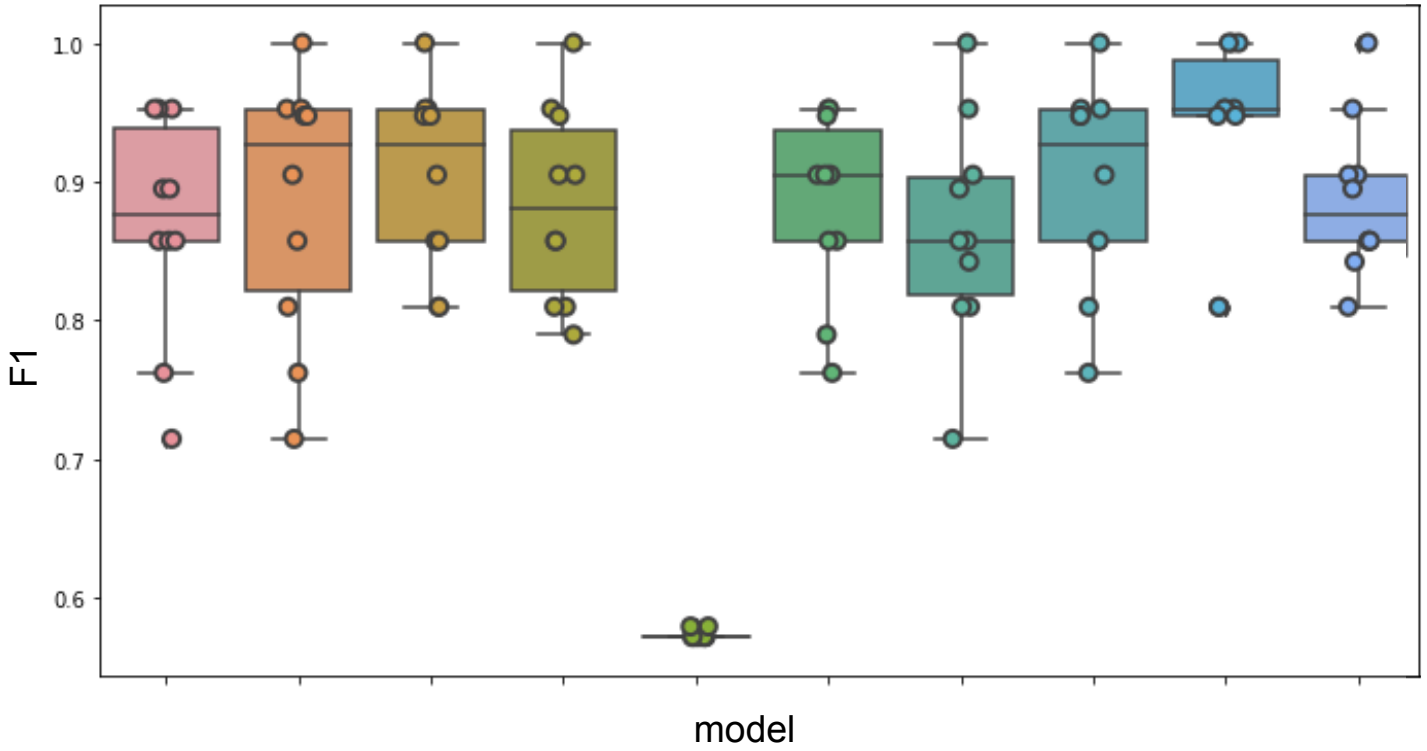
representando cada texto (ordem de serviço) em função das palavras que o formam

10 Classificadores (binários ou multi-classes)

NaiveBayes	básico
LinearSVC	básico
RandomForest	básico
MultiLayerPerceptron	básico
KNN	básico
Voting	ensemble
Extratrees	ensemble
Bagging	ensemble
Adaboost	ensemble
GradientBoosting	ensemble



Escolha do melhor classificador (F1)



Pipeline de NLP

↓

TF-IDF

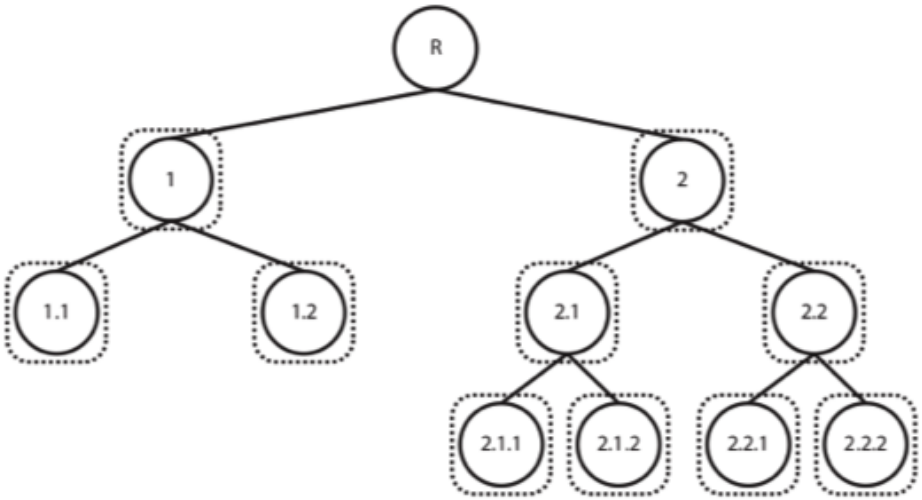
↓

10 classificadores

↓

Escolha do melhor classificador

Para cada nó da árvore



Pipeline de NLP

TF-IDF

10 classificadores

Escolha do melhor classificador

Ficha técnica

- Python – Scikit-learn, nltk, numpy
- Construção de modelos: Jupyter Notebook
- Automação das predições: PyCharm
- Desenvolvimento de uma peça de front para dashboards: Dash (python)


Time de deSenvolvedoreS

- 1 cientista de dados
- 1 dev python senior
- 4 meses de projeto

4º ATO

A ENTREGA

Conseguimos 100% de assertividade?

- Árvore final com 20 nós
- Média entre todos os classificadores
  - F1: 0.73
  - Acurácia: 75%
- Tempo de execução:
  - Antes era feito manualmente (400 registros/mês)
  - Com nossa abordagem: 4min (120000 registros/mês)

Percepção do Cliente

Foi bastante trabalhoso fornecer os dados de exemplo com amostras suficientes

Atingimos uma assertividade muito satisfatória!

Percepção dos Desenvolvedores

Dá pra melhorar a assertividade, porém precisamos de mais tempo

Difícil gerenciar a expectativa do cliente em um projeto como esse

De pai pra filho!
A Tarefa de Classificação Hierárquica aplicada em um Case Real Sobre Dados Textuais

Fabíola S. F. Pereira
Data Scientist @ ZUP IT

sites.google.com/site/fabiolasfpereira

THE DEVELOPER'S CONFERENCE