



## Introdução à Machine Learning com Orange

Data: Agosto/2023

Versão 01

# APRESENTAÇÃO

- Professora na área de Tecnologia no Centro Paula Souza (ETEC Prof. Maria Cristina Medeiros)
- Atuo no Projeto “O Cubo” no CPS – disseminando de forma extra curricular o ensino de Python Básico e Inteligência Artificial (Machine Learn e Ciências de Dados) com alunos de diferentes áreas.
- Empreendedora da Startup CRIA (Corretor de Redações de Inteligência Artificial), sou líder da equipe que está desenvolvendo o projeto. A Startup incubada pela empresa Tecnologia Única, parceira no projeto. <http://cria.net.br>
- Avaliadora da Fundação Estudar, selecionei alunos que querem estudar fora do Brasil, em diferentes universidades. Avalio a capacidade técnica na área de Inteligência Artificial e Ciências de dados. Estudantes ganham bolsa de estudos pela <https://www.estudar.org.br/>



**Mestre em Mestre em Informática e Gestão do Conhecimento**

- Graduada em Informática para gestão de Negócios.
- Licenciatura em Matemática
- Especialização em Docência e Gestão na Educação a Distância.

# APRESENTAÇÃO

- ❑ Professor na área de Tecnologia no Centro Paula Souza (ETEC Prof. Maria Cristina Medeiros e FATEC Mauá)
- ❑ Atuo no Projeto “O Cubo” no CPS – disseminando de forma extra curricular o ensino de Python Básico, Inteligência Artificial (Machine Learning e Ciências de Dados) e IoT com alunos de diferentes áreas.
- ❑ Avaliador da Fundação Estudar, seleciono alunos que querem estudar fora do Brasil, em diferentes universidades. Avalio a capacidade técnica na área de Inteligência Artificial e Ciências de dados. Estudantes ganham bolsa de estudos pela <https://www.estudar.org.br/>



**Mestre em Mestre em Informática e Gestão do Conhecimento**

- Graduado em Ciência da Computação.
- Licenciatura em Informática
- Especialização em Banco de Dados.

# Ementa

1. Introdução à Inteligência Artificial
2. Papel da Ciência de Dados na IA;
3. Tipos de Aprendizado de Maquina;
4. Explorando uma base de dados jurídica com Orange;

# 1. Introdução à Inteligência Artificial

## Inteligência

***“faculdade de aprender, compreender e adaptar-se”***

## Aspectos da Inteligência

- **teológica**: “dom divino que nos torna semelhantes ao Criador”;
- **filosófica**: “princípio abstrato que é a fonte de toda a intelectualidade”;
- **psicológica**: “capacidade de resolver problemas novos com rapidez e êxito”.

# Inteligência Artificial

Definir precisamente o que é inteligência artificial é uma tarefa, se não impossível, pelo menos extremamente difícil.

Entretanto, podemos definir Inteligência Artificial (IA), enquanto **disciplina do conhecimento humano**. Segundo Russell & Norvig, as definições de IA, encontradas na literatura científica, podem ser agrupadas em quatro categorias principais:

- a) sistemas que pensam como humanos
- b) sistemas que agem como humanos
- c) sistemas que pensam logicamente
- d) sistemas que agem logicamente

# Algumas definições sobre IA

- IA é o ramo da computação preocupada com a **automação do comportamento inteligente** (Luger e Stubblefield).
- IA é o estudo da computação que torna possível perceber **raciocinar e agir**. Idéias que permitem que o computador seja inteligente (Winston).
- IA é a parte da ciência da computação voltada para o **desenvolvimento de sistemas inteligentes** (Feigenbaum).
- **Inteligência Artificial** é a capacidade de um sistema computacional ou mecanismos para realizar tarefas que normalmente exigem **inteligência humana**, como: raciocínio, aprendizado e adaptação a novas situações. Data Science Academy (2023)

# Áreas de aplicação da Inteligência Artificial

- jogos e brinquedos eletrônicos
- robótica e automação industrial
- verificação automática de software
- otimização e controle de processos
- processamento de linguagem natural
- bancos de dados dedutivos e mineração de dados
- aprendizagem, planejamento e escalonamento de tarefas
- reconhecimento de faces, de voz, de cheiros e de sabores

# Tipos de IA

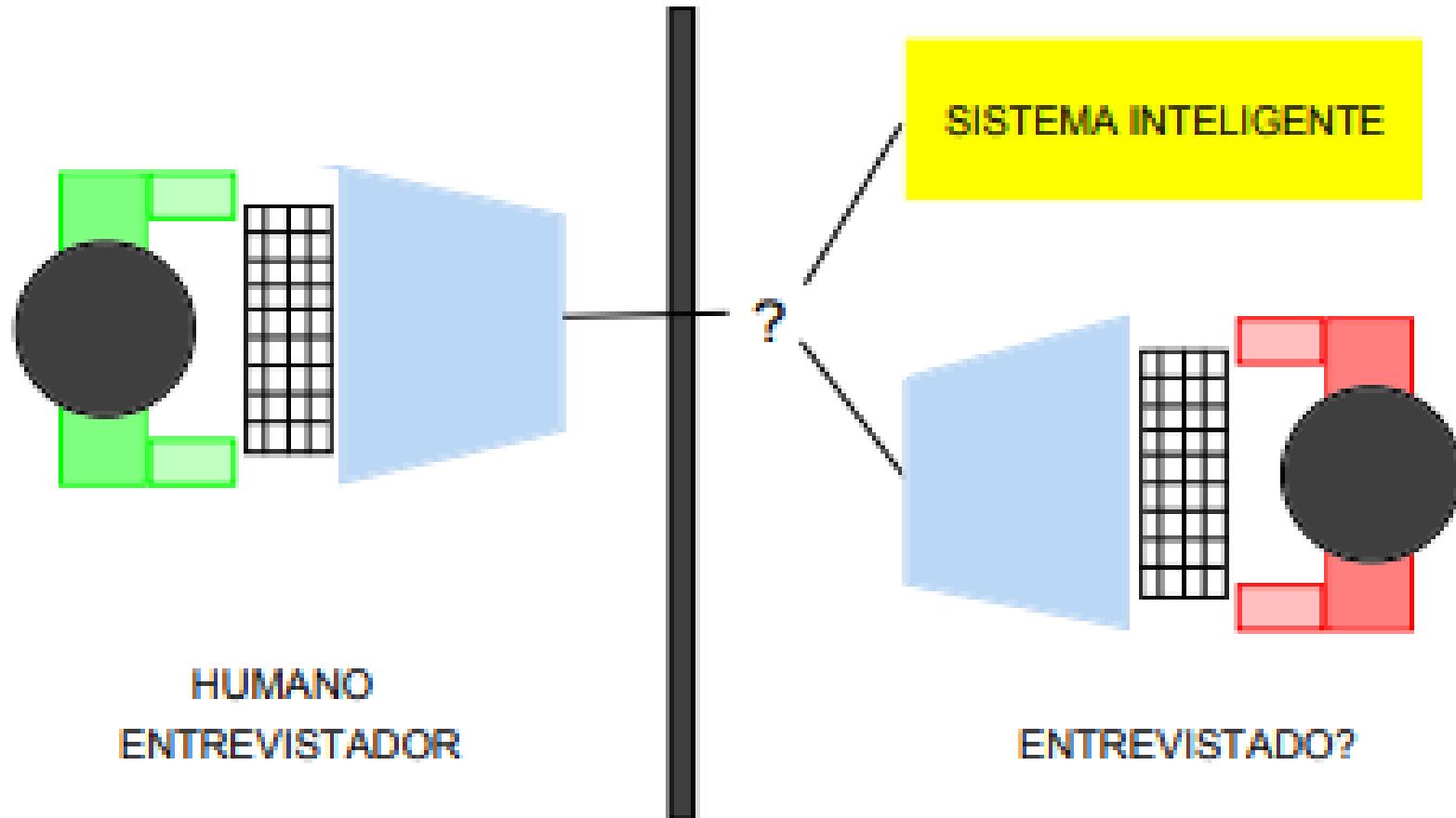
- **IA FRACA**

Limitada a atividades específicas

- **IA FORTE**

É Capaz de realizar qualquer tarefa intelectual que um ser humano possa resolver.

# O Teste de Turing



# Áreas de aplicação da Inteligência Artificial

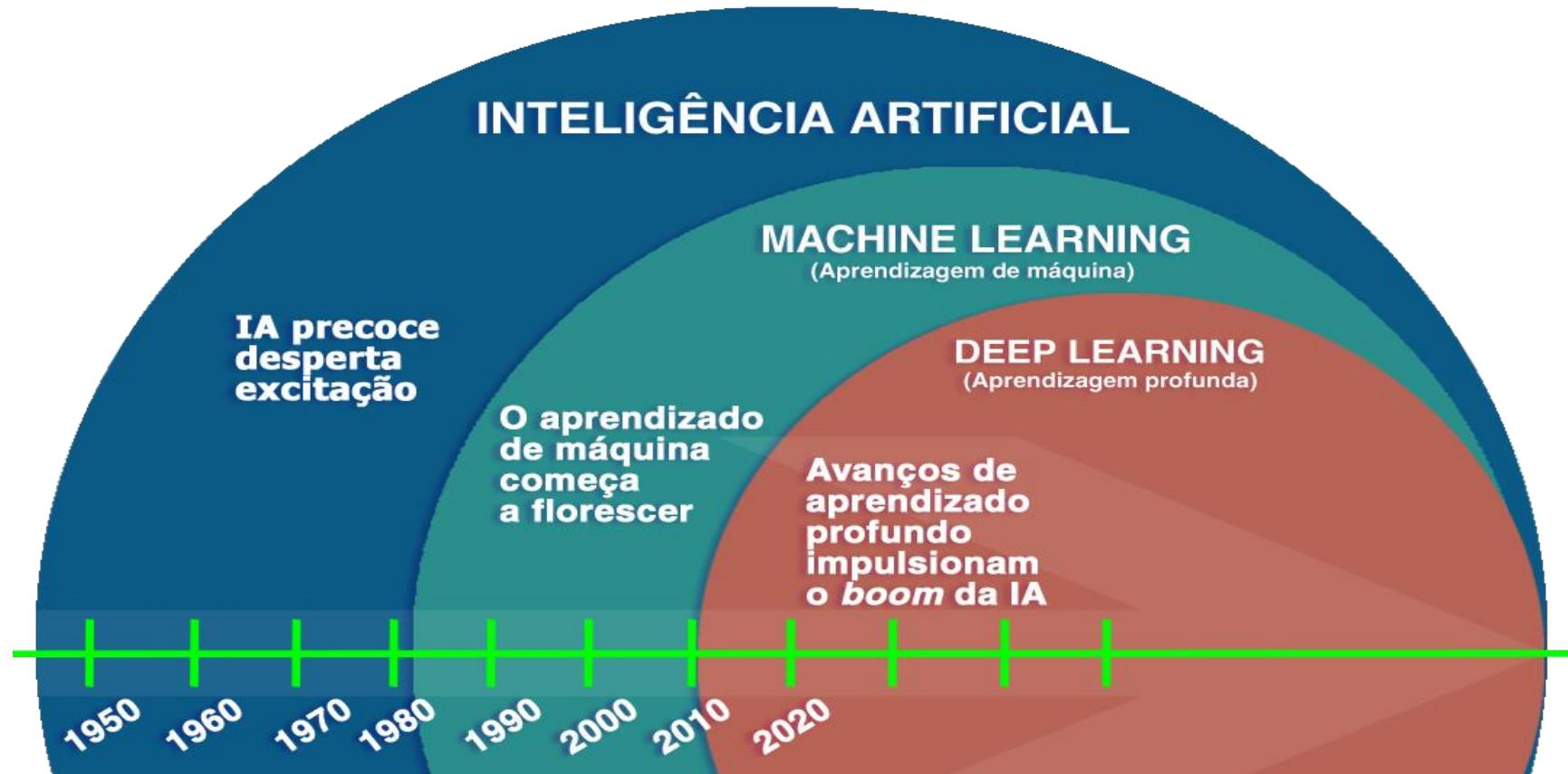
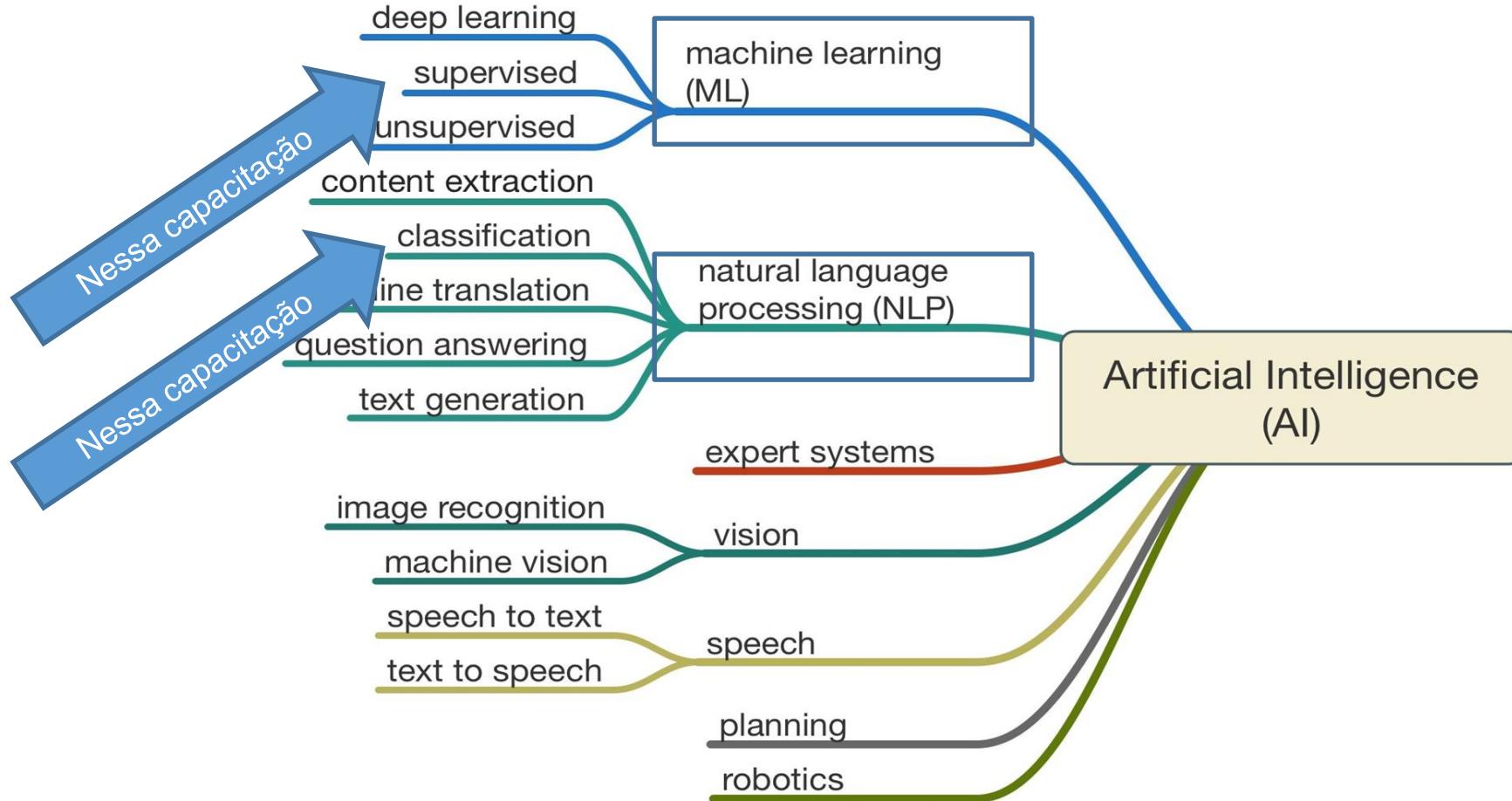


Figura: Buzzrobot/Tradução: Thiago Zina Crepaldi

# Áreas de aplicação da Inteligência Artificial



# Algumas vantagens da IA

- **Redução de erros:** Uma vez que são máquinas tem reduzidas as chances de falharem, tendo maior grau de precisão.
- **Exploração:** Máquinas podem realizar um trabalho mais laborioso e duro, superando as limitações humanas.
- **Aplicações diárias:** A sua utilização está presente em vários mecanismos do nosso cotidiano.
- **Sem pausas:** As máquinas, ao contrário dos seres humanos, não precisam de intervalos frequentes.
- **Velocidade:** Apresentam soluções muito mais rapidamente que outros sistemas.
- **Adaptabilidade:** São capazes de se adaptar as mudanças de condições de operação.

# Algumas desvantagens da IA

- **Alto custo:** devido a sua complexidade o seu custo de produção é alto.
- **Falta de criatividade:** A inteligência artificial não é desenvolvida ao ponto de atuar como o cérebro humano, de forma criativa.
- **Causa o desemprego:** Como são capazes de executar tarefas antes exclusivas aos humanos de maneira mais otimizada e eficiente, tendem a substituir a atividade humana em larga escala.
- **Representação do conhecimento:** para criar sistemas de inteligência artificial é necessário desenvolver um sistema de representação do conhecimento, o que geralmente é dispendioso.

# O que é IA responsável?

- A **IA responsável** é um conceito que representa uma **mudança de paradigma** na forma de enxergar a IA. Em vez de apenas pensar no desempenho de um modelo, os profissionais devem pensar também no **compromisso ético** e nas implicações desses sistemas em um contexto real, prático.
- Ou seja, é entender a IA e seus **impactos para as pessoas** e para as empresas, impactos reais e poderosos para gerar **benefícios** e **malefícios**. O ideal é amplificar os benefícios e estar ciente dos possíveis riscos.

# Como funciona a IA responsável

IA responsável é um padrão que garante **segurança, confiança e imparcialidade nos resultados, garantindo modelos mais éticos, eficientes e explicáveis**. O objetivo final é garantir que sua organização compreenda e cumpra os quatro pilares básicos para uma governança cooperativa:

- **Transparência**
- **Inclusão**
- **Sustentabilidade**
- **Segurança**

# Exemplos de empresas que passaram por problemas éticos - Tiveram que rever seus modelos de IA



**OLHAR DIGITAL**

NOTÍCIAS VÍDEOS EDITORIAS SUPORTE OD SEGURANÇA OFERTAS 🔍

Nós do **Olhar Digital** e nossos parceiros utilizamos *cookies*, *localStorage* e outras tecnologias semelhantes para personalizar conteúdo, anúncios, recursos de mídia social, análise de tráfego e melhorar sua experiência neste site, de acordo com nossos [Termos de Uso](#) e [Privacidade](#). Ao continuar navegando, você concorda com estas condições.

[Continuar](#)

**NOTÍCIAS**

## Inteligência artificial da Amazon exercitava preconceito

Amazon descarta algoritmo que se tornou sexista e descartava candidatos a emprego



**veja** RADAR RADAR ECONÔMICO POLÍTICA ECONOMIA SAÚDE MUNDO CULTURA ESPORTE AGENDA VERDE

Tecnologia

## Exposto à internet, robô da Microsoft vira racista em 1 dia

Projeto de inteligência artificial da gigante da tecnologia foi tirado do ar em menos de 24 horas depois que passou a reproduzir ofensas escabrosas ao interagir com trolls nas redes

Por Da Redação 24 mar 2016, 19h48



ECONOMIA

## Facebook pede desculpas após rotular vídeo de homens negros como “primatas”

Rede social chamou episódio de “erro inaceitável” e desabilitou software de inteligência artificial que disparou a mensagem

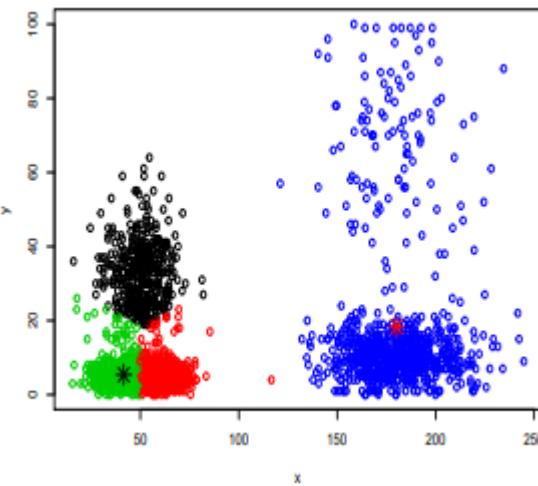
New York Times

05/09/2021 - 18:02

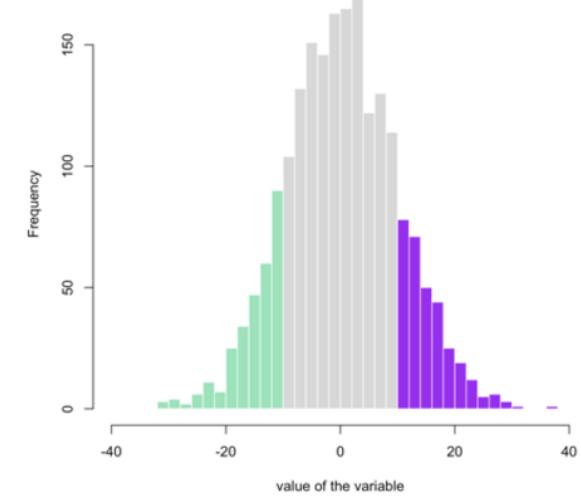
# PARTE II **CIÊNCIA DE DADOS**

# Introdução à Ciência de Dados

Ciência de dados é um ramo multidisciplinar da ciência que envolve **técnicas de computação, matemática aplicada, inteligência artificial, estatística e otimização** com o intuito de resolver problemas analiticamente complexos, utilizando grandes conjuntos de dados como núcleo de operação.



id	idade	nmal	parasit	ngest	idgest	sexrn	pesorn	estrn	pcefal
1	25	0	0	3	38	2	3665	46	36
2	30	0	0	9	37	1	2880	44	33
3	40	0	0	1	41	1	2960	52	35
4	26	0	0	2	40	1	2740	47	34
5	.	0	0	1	38	1	2975	50	33
6	18	0	0	.	38	2	2770	48	33
7	20	0	0	1	41	1	2755	48	34
8	15	0	0	1	39	1	2860	49	32
9	.	0	0	.	42	2	3000	50	35
10	18	0	0	1	40	1	3515	51	34
11	17	0	0	2	40	1	3645	54	35
12	18	1	1	3	40	2	2665	48	35
13	30	0	0	6	40	2	2995	49	33
14	19	0	0	1	40	1	2972	46	34
15	32	0	0	5	41	2	3045	50	35
16	32	0	0	8	38	2	3150	44	35
17	18	0	0	2	40	1	2650	48	33.5
18	18	0	0	1	41	1	3200	50	37
19	19	0	0	1	39	1	3140	48	32
20	18	0	0	2	40	1	3150	47	35



## Inteligência Artificial

*Qualquer técnica que permita aos computadores imitar a inteligência humana, usando lógica, árvores de decisão e aprendizado de máquina (incluindo aprendizado profundo)*

- Visão Computacional
- Processamento de Linguagem Natural
- Reconhecimento de Fala
- Robótica

Está inserido em todos os processos

## Aprendizado de Máquina

*Um contribuinte da IA que inclui técnicas de estatística e algoritmos que capacitam máquinas a melhorarem tarefas com experiência.*

- Supervisionado
- Não-supervisionado
- Aprendizado Reforçado

## Aprendizagem Profunda

*Subconjunto do aprendizado de máquina composto de algoritmos que permitem ao software treinar a si mesmo para realizar tarefas, como reconhecimento de voz e imagem, expondo uma grande quantidade de dados à multicamadas de redes neurais.*

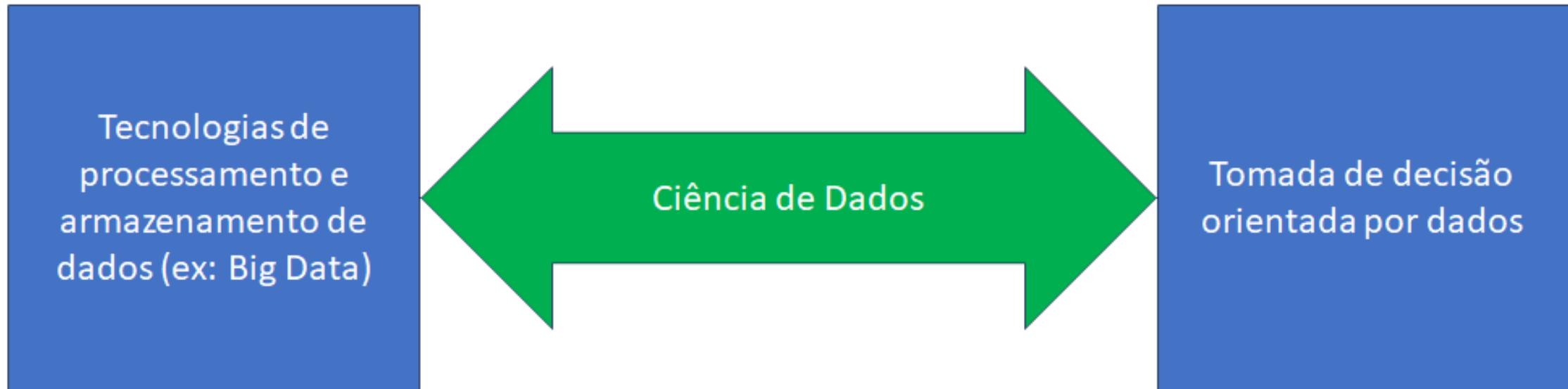
- Redes Neurais Artificiais
- Redes Neurais Recorrentes
- Redes Neurais Convolucionais

## Ciência de Dados

*Uma área interdisciplinar, que localiza-se em uma interface entre a estatística e a ciência da computação, que utiliza o método científico; processos, algoritmos e sistemas. É focada na descoberta de insights acionáveis a partir de grandes conjuntos de dados.*

*BI, Big Data, Data Warehouse, Data Lake, Analytics, ...*

# PAPEL DA CIÊNCIA DE DADOS



# PAPEL DA CIÊNCIA DE DADOS – GERAR CONHECIMENTO



# Etapas da Ciência de Dados



# Entender o problema

- Ao menos o suficiente para se comunicar com quem tem o problema!
- Que dados existem?
- Que dados deveriam existir?
- **ALERTA: não devemos fazer Data Science sem entender o problema!**



# Achar Dados

- Achar = localizar, identificar, etc
- Que dados existem relacionados ao problema em questão?
- Que dados estão disponíveis?
  - É preciso coletar mais/outros?
  - Como acessar os dados?
  - Existem formas prontas?
  - Preciso replicar/amostrar?
  - **Qual é o volume destes dados e no que isto impacta a coleta?**

A	B	C	D	E	F
Num_Serviço	Data	Nível	Prioridade	Situação	G
0001	01/05/10	1	Baixa	Pendente	Falhas no sistema
0002	23/06/10	2	Normal	Aberta	Falhas no sistema
0003	05/05/10	3	Alta	Pendente	Falhas no sistema
0004	18/05/10	1	Baixa	Aberta	Falhas no sistema
0005	15/06/10	2	Normal	Aberta	Falhas no sistema
0006	04/04/10	3	Alta	Aberta	Falhas no sistema
0007	01/05/10	1	Baixa	Pendente	Falhas no sistema
0008	09/06/10	2	Normal	Aberta	Falhas no sistema
0009	08/05/10	3	Alta	Pendente	Falhas no sistema
0010	15/05/10	1	Baixa	Aberta	Falhas no sistema
0011	01/06/10	2	Normal	Aberta	Falhas no sistema
0012	08/04/10	3	Alta	Aberta	Falhas no sistema
0013	06/05/10	1	Baixa	Pendente	Falhas no sistema
0014	06/06/10	2	Normal	Aberta	Falhas no sistema
0015	05/05/10	3	Alta	Pendente	Falhas no sistema
0016	05/05/10	1	Baixa	Aberta	Falhas no sistema
0017	05/06/10	2	Normal	Aberta	Falhas no sistema
0018	05/04/10	3	Alta	Aberta	Falhas no sistema
0019	04/05/10	1	Baixa	Pendente	Falhas no sistema
0020	04/06/10	2	Normal	Aberta	Falhas no sistema
0021	04/05/10	3	Alta	Pendente	Falhas no sistema

# Entender a organização dos Dados

- **Antes do processamento:**
  - Como os dados são representados?
  - Tabelas, documentos, imagens, relações, mistura?
  - Os dados estão em um formato útil para resolver nosso problema?
    - Como transformar?
    - Qual é o tamanho desta tarefa?



Agora que conhecemos um pouco da teoria,  
vamos a primeira parte prática, primeiro  
visualizando algumas informações no ORANGE



# Fazendo classificações em bases de dados Jurídicas

O principal objetivo desta aula, é o de demonstrar como podemos utilizar a ferramenta Orange Canvas com modelos de aprendizado de máquina para realizarmos a classificação de textos jurídicos.

Neste exemplo iremos partir de uma planilha feita em excel na qual iremos ter uma pequena informação sobre um processo e qual é a sua classificação, podendo ser do tipo: Civil, Criminal ou Trabalhista.

# Planilha no Excel

	A	B	C	D
1	ID	Texto do Processo	Tipo de Processo	
2	1	Funcionário alega salários não pagos após demissão injusta.	Trabalhista	
3	2	Acusado de vandalismo em propriedade pública durante protesto.	Criminal	
4	3	Disputa sobre quebra de contrato de prestação de serviços de consultoria.	Cível	
5	4	Empregado busca indenização por assédio moral no ambiente de trabalho.	Trabalhista	
6	5	Julgamento de fraude em transações financeiras envolvendo esquema de pirâmide.	Criminal	
7	6	Pedido de guarda de menor após divórcio e disputa pela custódia.	Cível	
8	7	Reclamação de discriminação racial no emprego e ambiente hostil.	Trabalhista	
9	8	Julgamento de agressão física resultando em lesões graves.	Criminal	
10	9	Disputa sobre propriedade de imóvel herdado entre herdeiros.	Cível	
11	10	Reivindicação de pagamento de horas extras não remuneradas e falta de intervalos.	Trabalhista	
12	11	Acusação de fraude acadêmica em exame universitário.	Criminal	
13	12	Disputa de patente entre empresas de tecnologia.	Cível	
14	13	Investigação de fraude em licitação pública.	Criminal	
15	14	Funcionário alega retaliação após denunciar má conduta.	Trabalhista	
16	15	Processo de divórcio com questões de pensão alimentícia.	Cível	
17	16	Caso de difamação em redes sociais.	Cível	
18	17	Acusação de plágio acadêmico em tese de doutorado.	Criminal	
19	18	Empregador enfrenta processo por não pagamento de férias.	Trabalhista	
20	19	Julgamento de violência doméstica e ordem de restrição	Criminal	

IREMOS APLICAR ALGUMAS TÉCNICAS DE  
CIÊNCIAS DE DADOS E APRENDIZADO DE  
MÁQUINA NESSA BASE DE DADOS JURIDICA.

*Aprendizagem é uma propriedade essencialmente humana*

### Aprender

Significa mudar para fazer melhor quando uma situação similar acontecer.

Aprender não é memorizar e sim ter a **capacidade de generalizar um dado comportamento para uma nova situação**

- Computadores memorizam facilmente.
- Dificuldade em generalizar.

# Aprendizado de Máquina (Machine Learning)

## Funções de mapeamento.

- Aplicação de modelagem matemática.
- Correlação entre a entrada e a saída.
- Presença da estatística e probabilidade.
- Resposta inexata.
- Acurácia e sua dependência do caso de estudo.
- A escolha do algoritmo mais adequado irá depender do banco de dados e da categoria de resposta esperada.

# Tipos de Aprendizagem

- **Supervisionada**
- **Não Supervisionada**
- **Semi-supervisionada**
- **Por Reforço**



Vamos detalhar duas nessa capacitação: **Supervisionado** e **Não Supervisionado**.

# Tipos de Aprendizagem

- **Não Supervisionada**

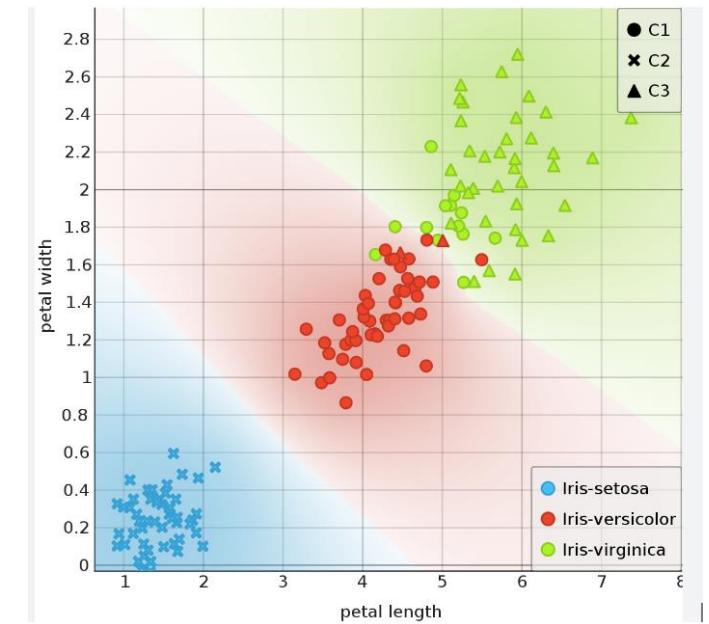
Não há uma saída desejada associada a cada padrão, de modo que os dados são não-rotulados. Neste cenário, desejamos que o modelo seja capaz de capturar, representar ou expressar propriedades existentes no conjunto de dados.

- Classes não são conhecidas
- Algoritmo deve definir quais são as classes em função de um determinado critério
  - Descoberta de Conhecimento

# EXEMPLO DE UMA BASE SEM RÓTULO



SepalLength	SepalWidth	PetalLength	PetalWidth
7.2	3.2	6.	1.8
5.	3.6	1.4	0.2
6.4	2.8	5.6	2.2
4.9	2.4	3.3	1.
6.6	2.9	4.6	1.3
5.	3.5	1.6	0.6
5.	3.3	1.4	0.2
4.9	3.1	1.5	0.1
6.3	3.3	6.	2.5
6.	2.7	5.1	1.6
5.2	2.7	3.9	1.4
6.7	2.5	5.8	1.8
5.7	2.8	4.1	1.3
5.6	2.8	4.9	2.
4.5	2.3	1.3	0.3
6.6	3.	4.4	1.4



# Voltando a nossa base – Porque ela é supervisionada?

F11	A	B	C	D
	ID	Texto do Processo	Tipo de Processo	
1	1	Funcionário alega salários não pagos após demissão injusta.	Trabalhista	
2	2	Acusado de vandalismo em propriedade pública durante protesto.	Criminal	
3	3	Disputa sobre quebra de contrato de prestação de serviços de consultoria.	Cível	
4	4	Empregado busca indenização por assédio moral no ambiente de trabalho.	Trabalhista	
5	5	Julgamento de fraude em transações financeiras envolvendo esquema de pirâmide.	Criminal	
6	6	Pedido de guarda de menor após divórcio e disputa pela custódia.	Cível	
7	7	Reclamação de discriminação racial no emprego e ambiente hostil.	Trabalhista	
8	8	Julgamento de agressão física resultando em lesões graves.	Criminal	
9	9	Disputa sobre propriedade de imóvel herdado entre herdeiros.	Cível	
10	10	Reivindicação de pagamento de horas extras não remuneradas e falta de intervalos.	Trabalhista	
11	11	Acusação de fraude acadêmica em exame universitário.	Criminal	
12	12	Disputa de patente entre empresas de tecnologia.	Cível	
13	13	Investigação de fraude em licitação pública.	Criminal	
14	14	Funcionário alega retaliação após denunciar má conduta.	Trabalhista	
15	15	Processo de divórcio com questões de pensão alimentícia.	Cível	
16	16	Caso de difamação em redes sociais.	Cível	
17	17	Acusação de plágio acadêmico em tese de doutorado.	Criminal	
18	18	Empregador enfrenta processo por não pagamento de férias.	Trabalhista	
19	19	Julgamento de violência doméstica e ordem de restrição	Criminal	

# Tipos de Aprendizagem

- **Supervisionada**

Para cada dado (ou padrão) de treinamento disponível, existe uma resposta desejada que é conhecida. Neste caso, dizemos que os dados são rotulados.

- Fornecemos a “resposta correta” durante o treinamento.
- Classes são conhecidas a priori
- Ajustamos os pesos em função das respostas corretas que conhecemos

# Transformando um arquivo xlsx em csv

Análises e Modelos de Machine Learning, trabalham muito bem com arquivos em formato CSV.

Seguindo essa premissa, nosso primeiro passo será o de transformar o arquivo que está em formato XLSX (Excel) para um arquivo no formato CSV (Separado por tabulação).

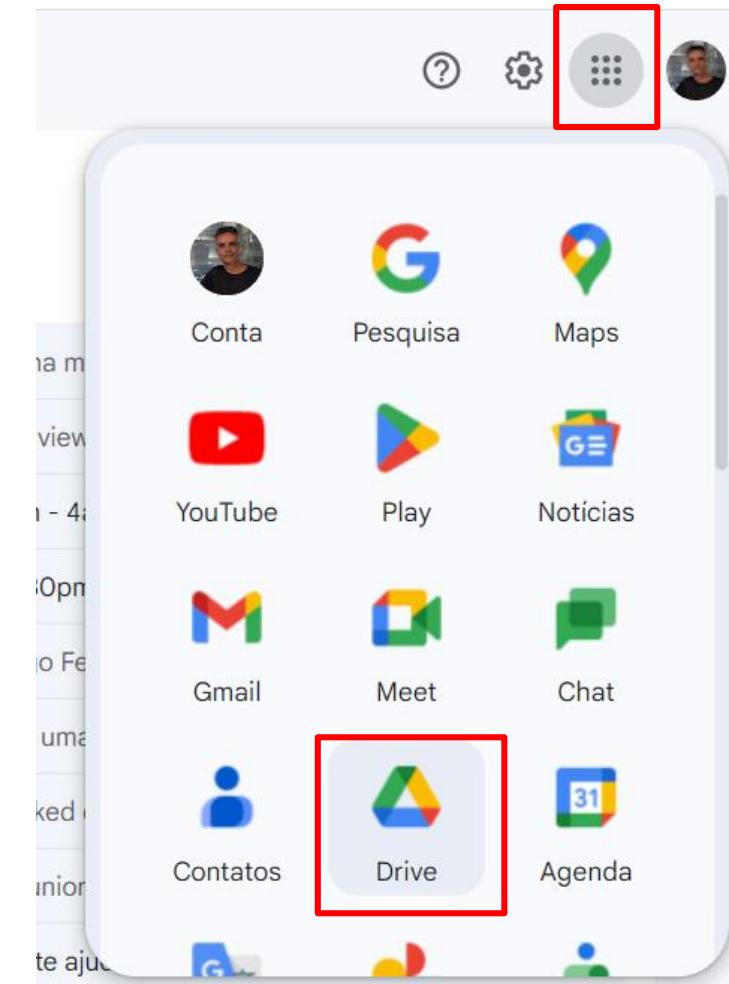
O arquivo pode ser obtido [aqui](#).

# Transformando um arquivo xlsx em csv

Uma maneira rápida para transformar essa planilha em um arquivo CSV, é entrando com um email do google, criar uma pasta em seu drive, enviar o arquivo XLSX original para esta pasta e abrindo-o utilizando o planilhas do google.

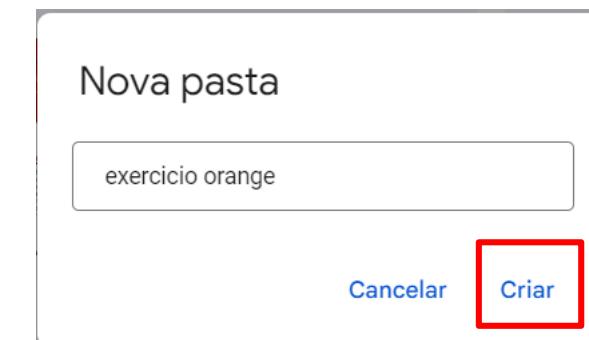
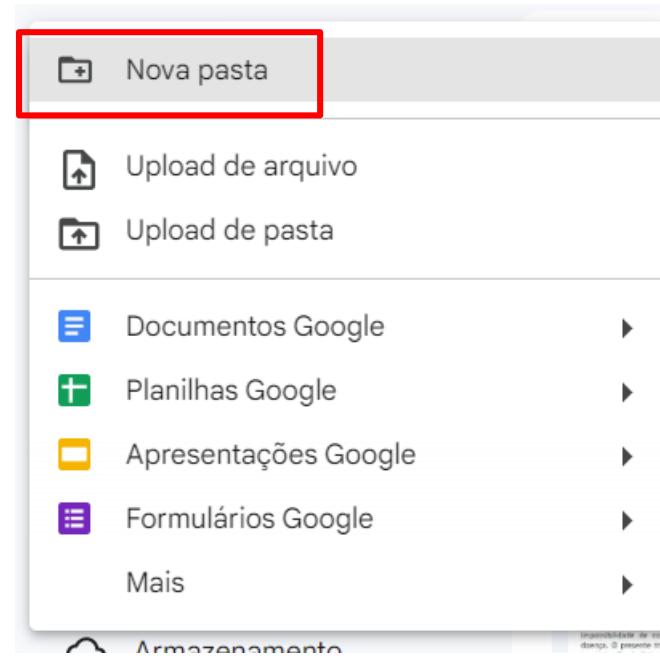
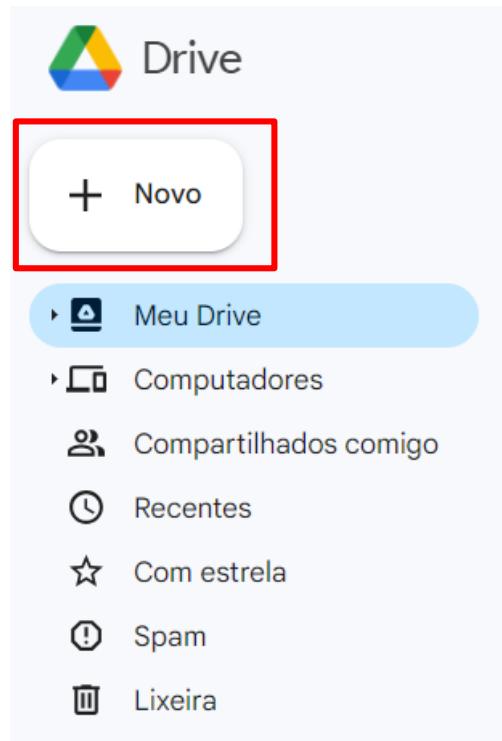
# Criando uma pasta em seu drive no google

Abra seu Email do google e  
selecione a ferramenta Drive



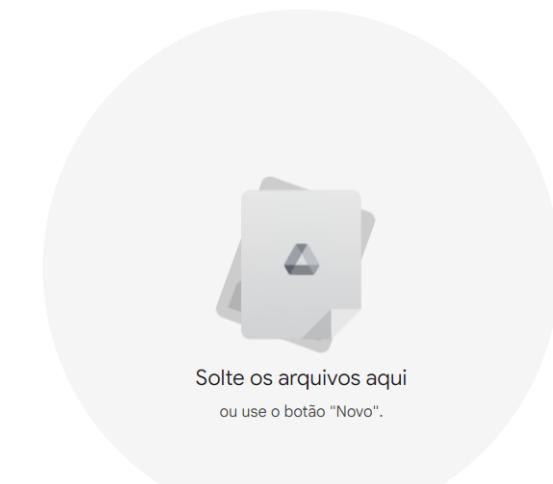
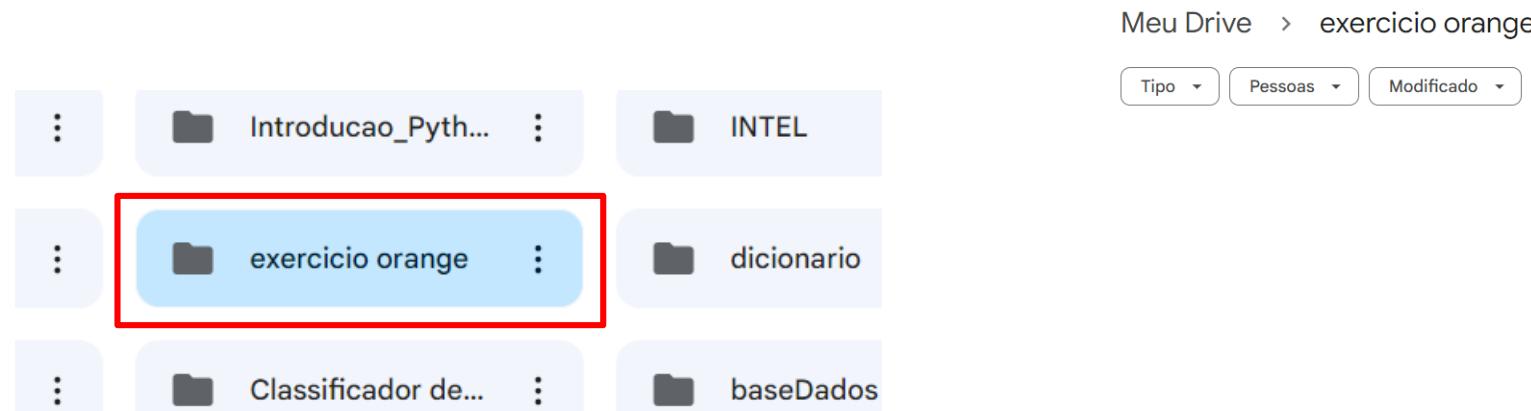
# Criando uma pasta em seu drive no google

Na janela do Drive, clique em Novo e selecione para criar uma nova pasta em seu drive. Salve-a como exercício Orange.



# Criando uma pasta em seu drive no google

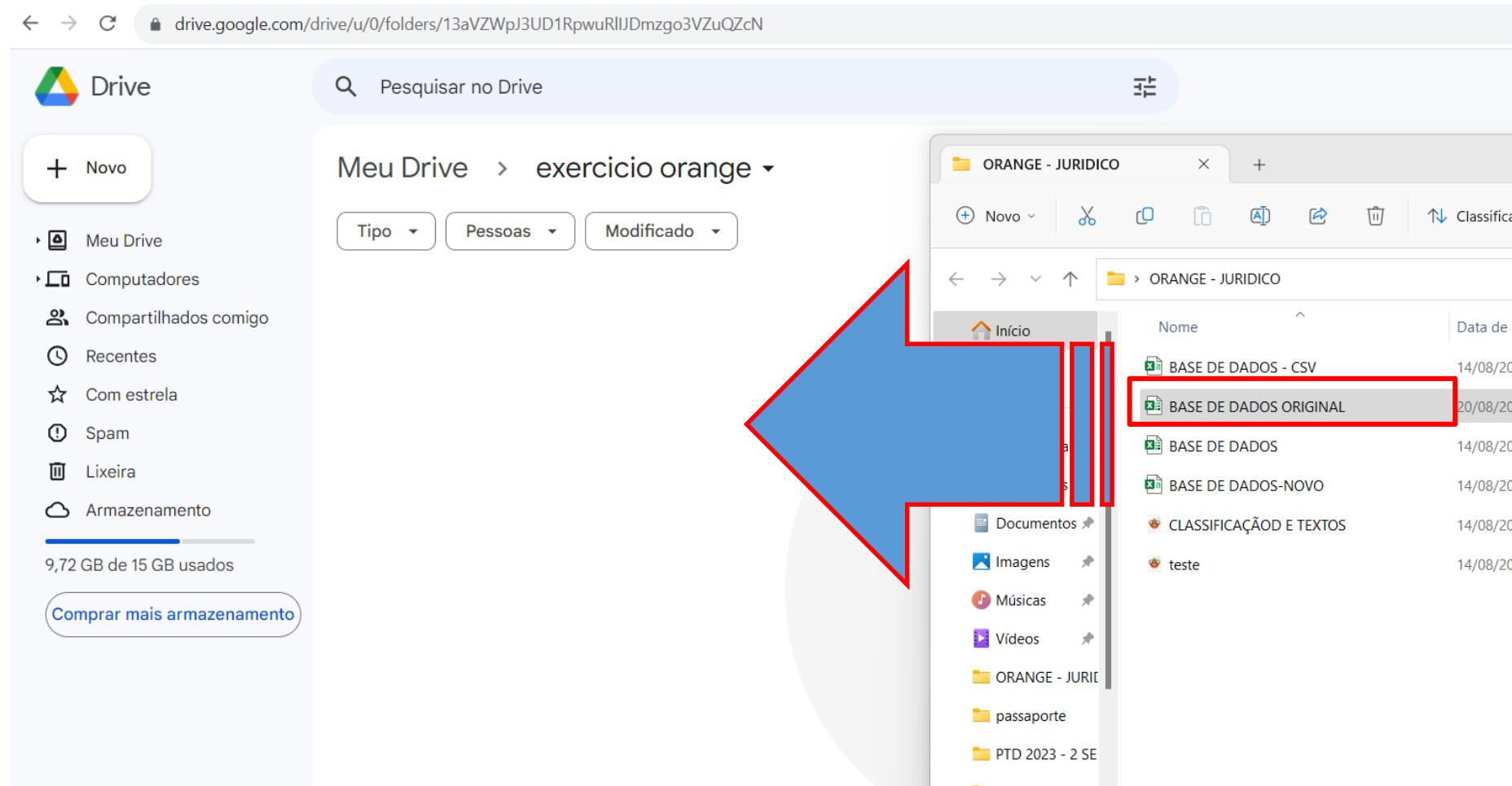
Abra a pasta criada. Clique duplo sobre a pasta.



# Enviando um arquivo para pasta em seu drive no google

Para enviar um arquivo, basta arrastá-lo para a área em branco da pasta ou clicar sobre o botão novo e selecionar upload de arquivo.

# Enviando um arquivo para pasta em seu drive no google



# Enviando um arquivo para pasta em seu drive no google

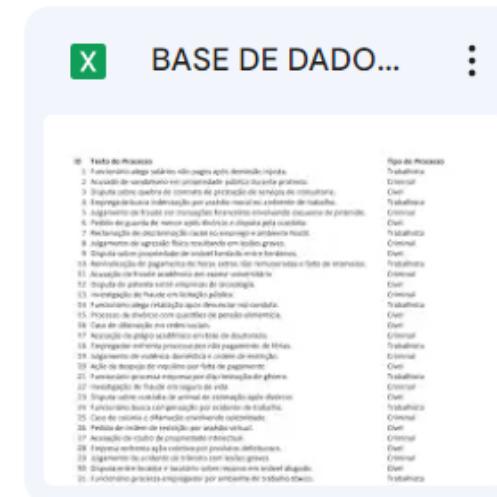
Meu Drive > exercício orange ▾

Tipo ▾

Pessoas ▾

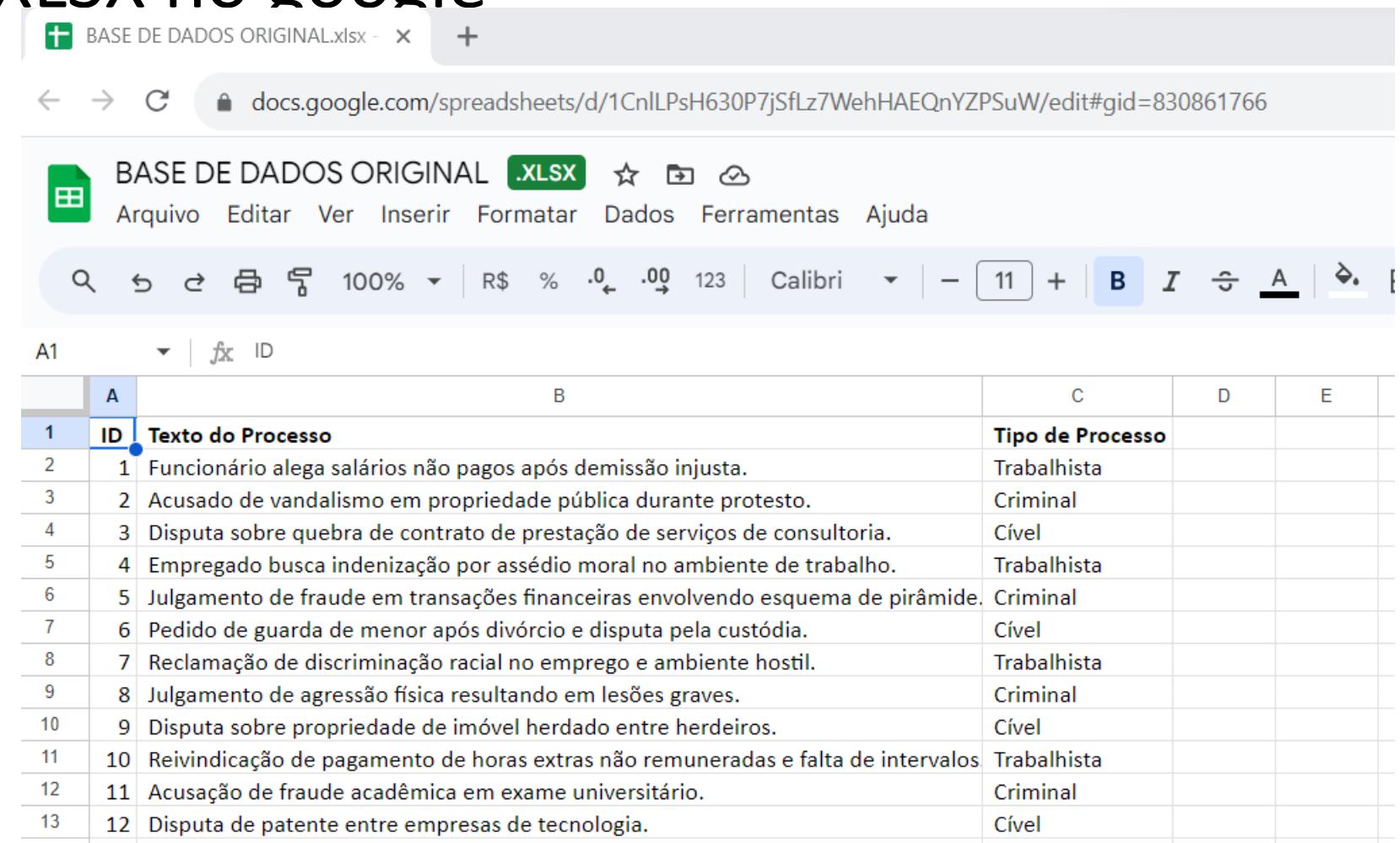
Modificado ▾

Arquivos



# Abra o arquivo XLSX no google

Basta dar um duplo clique sobre o arquivo e ele será aberto em seu navegador.



A	B	C	D	E
1	<b>ID</b>			
1	<b>Texto do Processo</b>			
2	1 Funcionário alega salários não pagos após demissão injusta.	Trabalhista		
3	2 Acusado de vandalismo em propriedade pública durante protesto.	Criminal		
4	3 Disputa sobre quebra de contrato de prestação de serviços de consultoria.	Cível		
5	4 Empregado busca indenização por assédio moral no ambiente de trabalho.	Trabalhista		
6	5 Julgamento de fraude em transações financeiras envolvendo esquema de pirâmide.	Criminal		
7	6 Pedido de guarda de menor após divórcio e disputa pela custódia.	Cível		
8	7 Reclamação de discriminação racial no emprego e ambiente hostil.	Trabalhista		
9	8 Julgamento de agressão física resultando em lesões graves.	Criminal		
10	9 Disputa sobre propriedade de imóvel herdado entre herdeiros.	Cível		
11	10 Reivindicação de pagamento de horas extras não remuneradas e falta de intervalos	Trabalhista		
12	11 Acusação de fraude acadêmica em exame universitário.	Criminal		
13	12 Disputa de patente entre empresas de tecnologia.	Cível		

# Salvar o arquivo em formato CSV

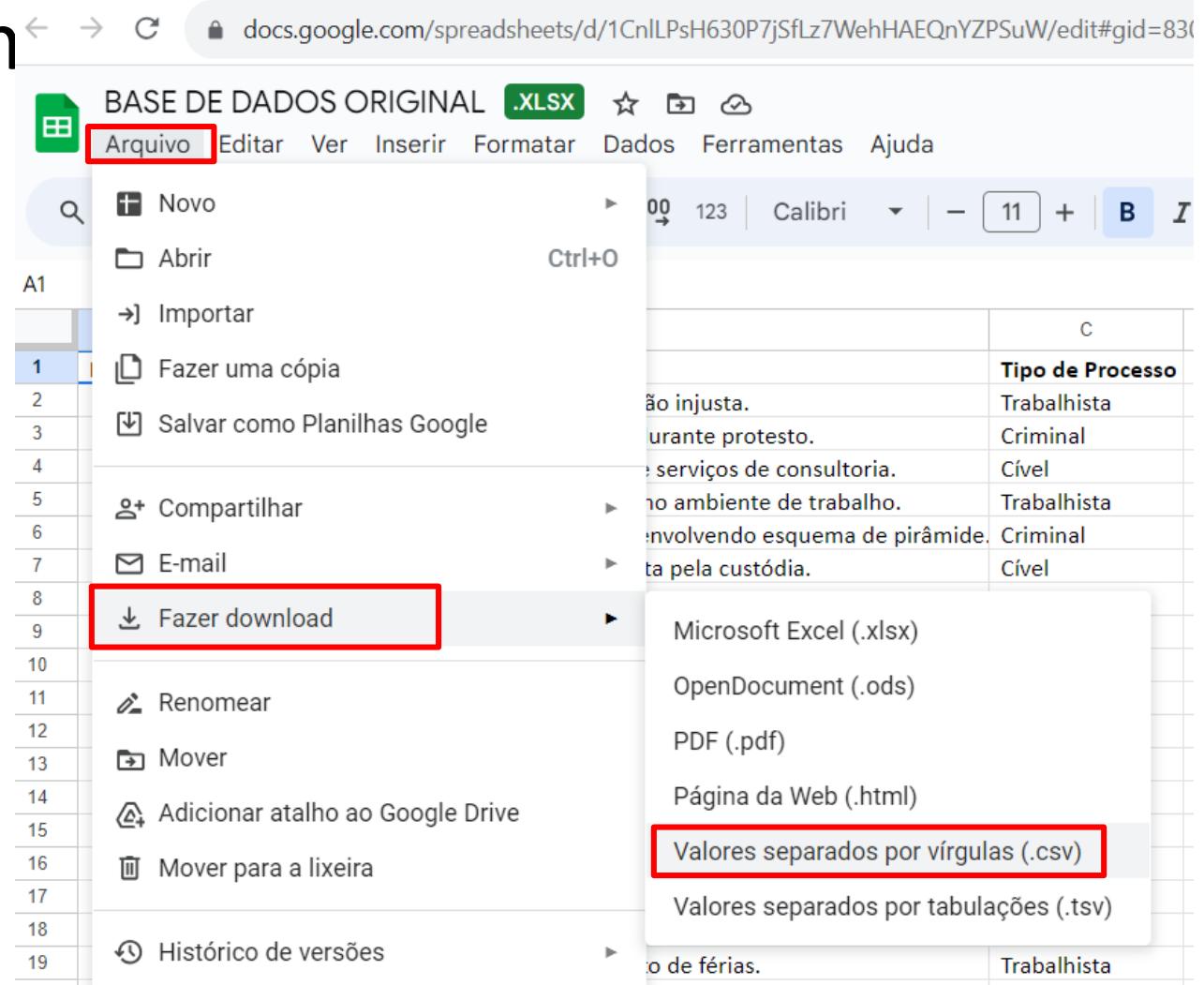
Para salvar clique em

ARQUIVO →

FAZER DOWNLOAD →

VALORES SEPARADOS POR VÍRGULAS (.CSV)

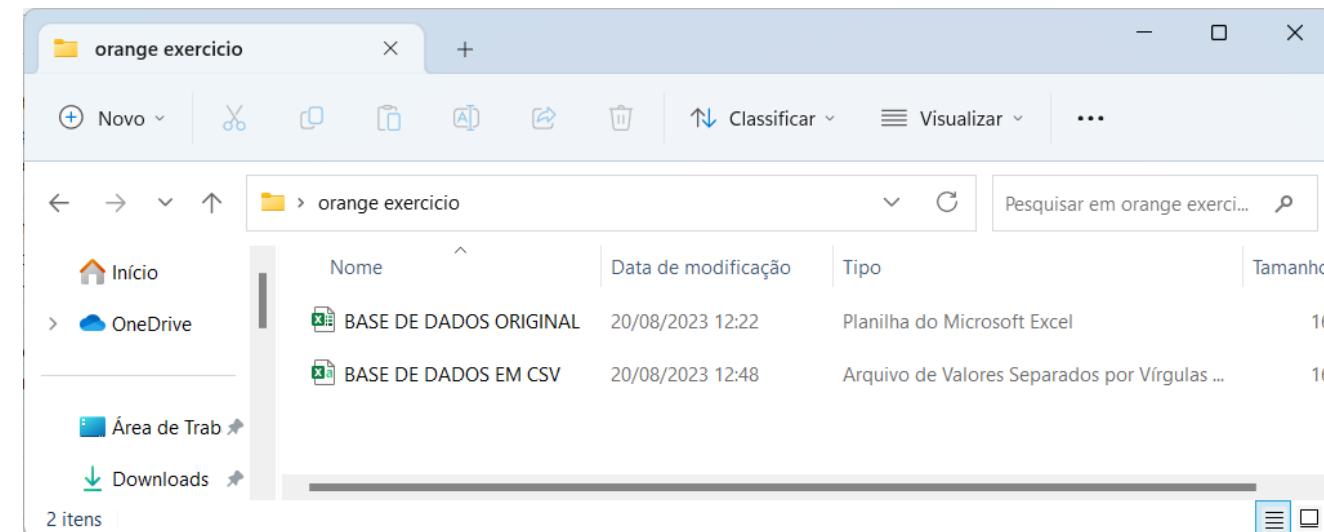
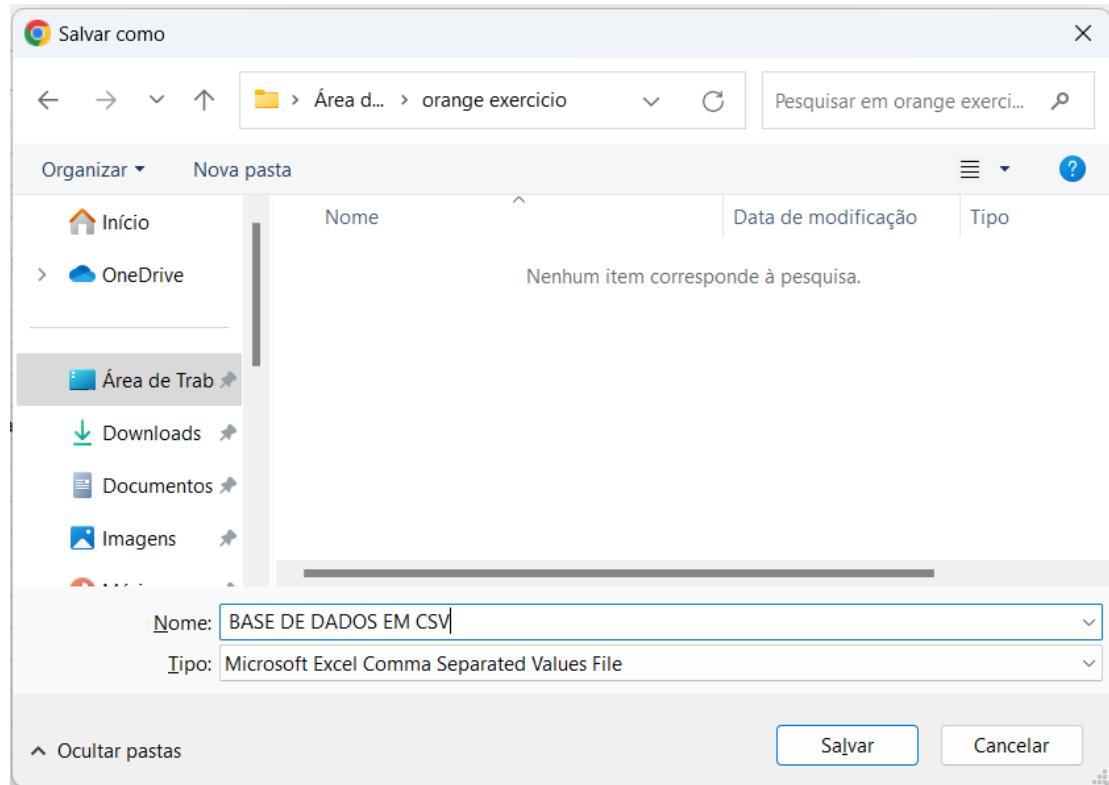
Escolha uma pasta em seu computador.



The screenshot shows a Google Sheets interface with a table titled "BASE DE DADOS ORIGINAL". The "Arquivo" menu is open, and the "Fazer download" option is highlighted with a red box. A dropdown menu is displayed, also with a red box around the "Valores separados por vírgulas (.csv)" option.

Tipo de Processo
ão injusta.
urante protesto.
e serviços de consultoria.
no ambiente de trabalho.
envolvendo esquema de pirâmide.
ta pela custódia.

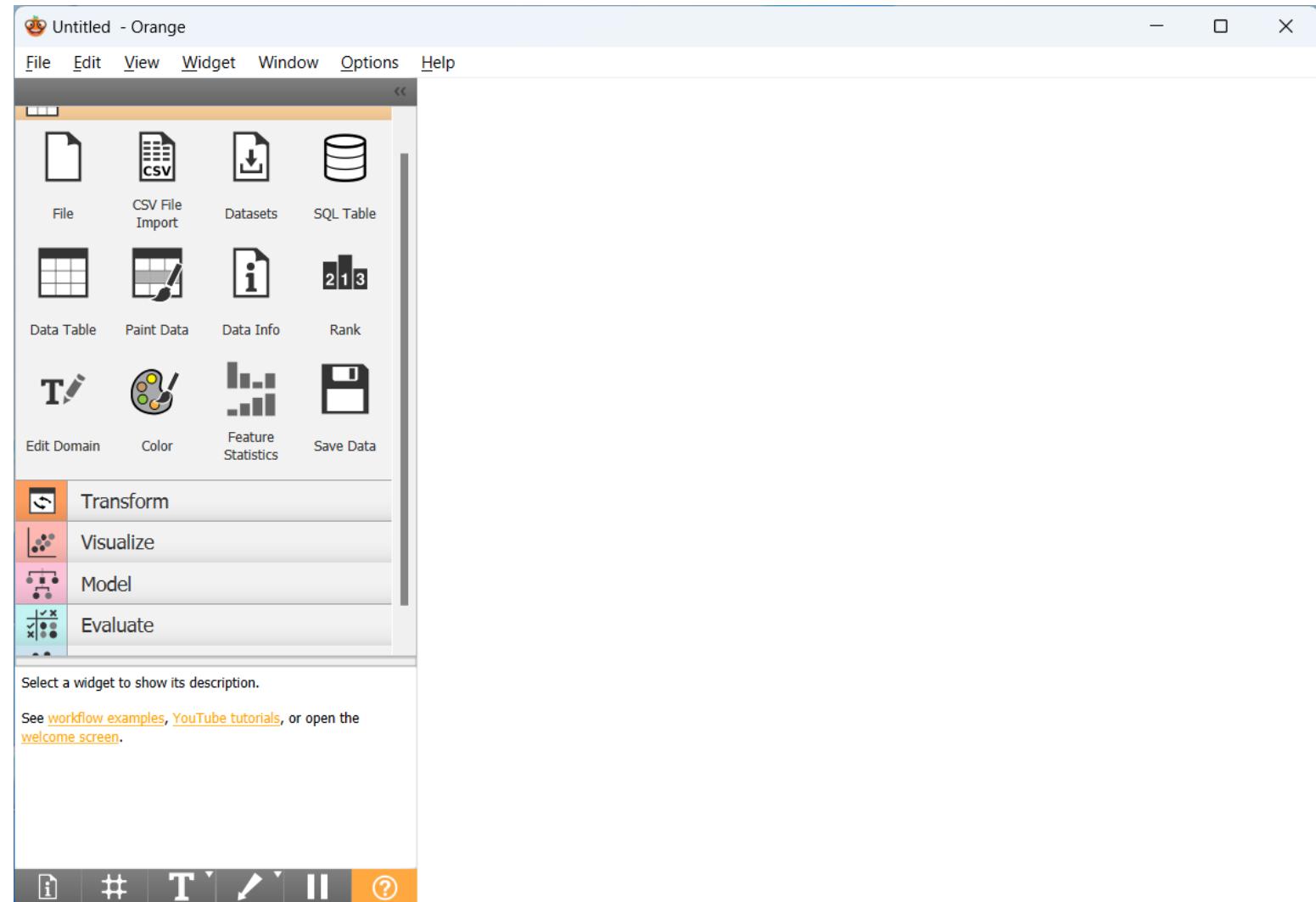
# Salvar o arquivo em formato CSV em sua máquina.



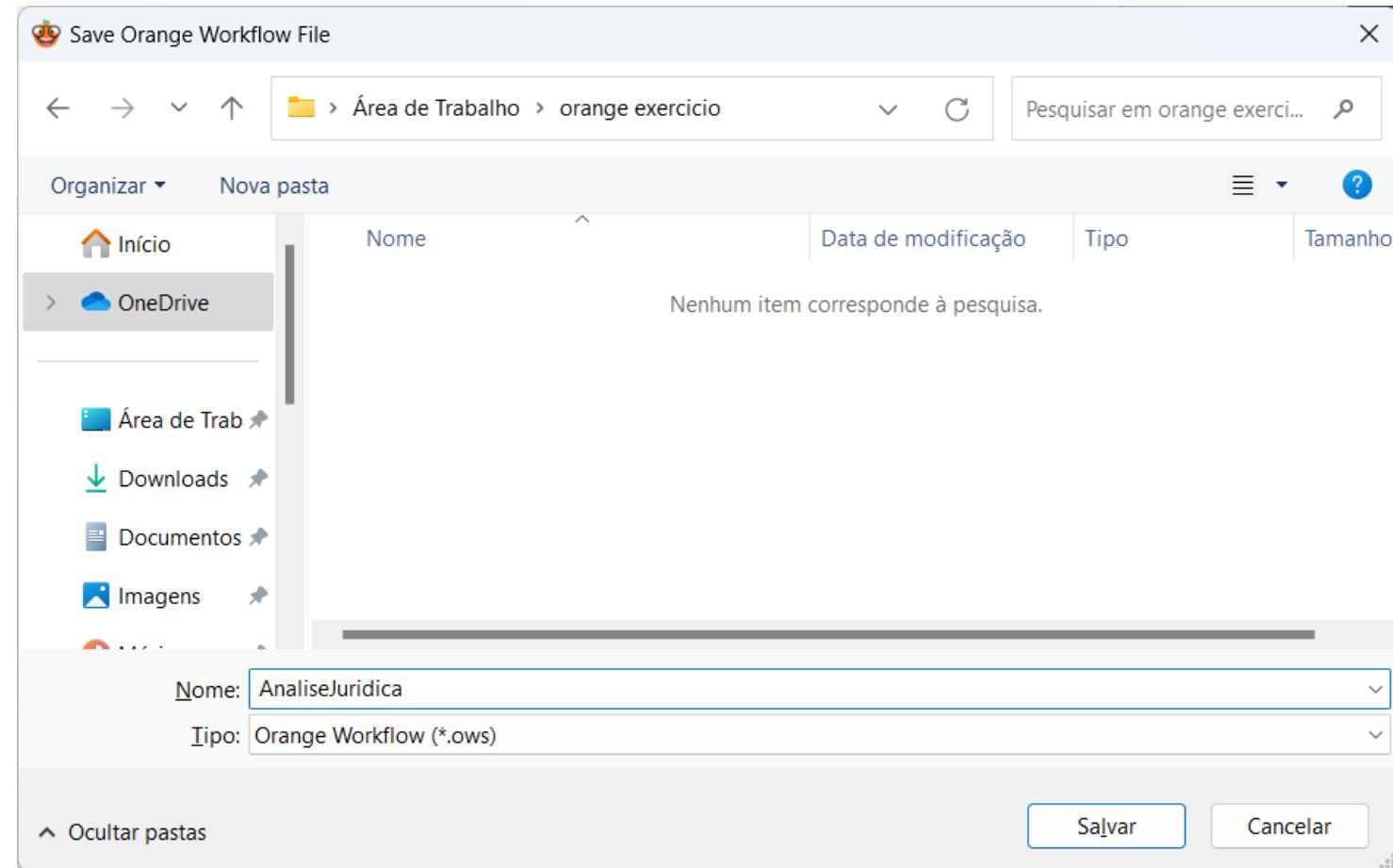
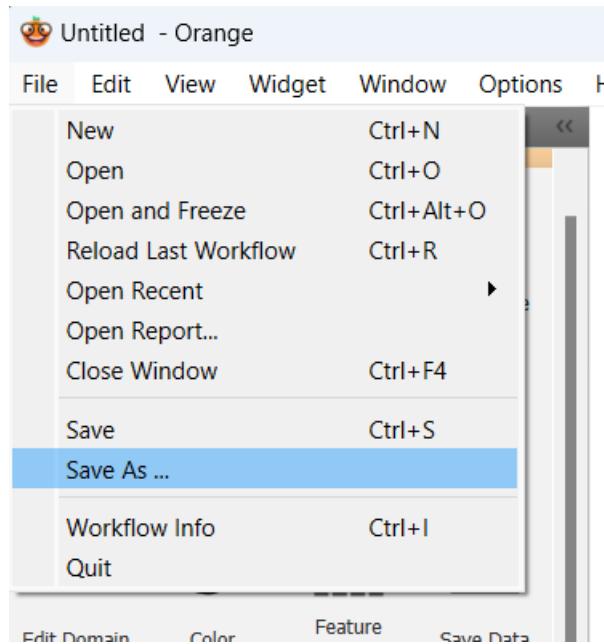
Agora que já temos a nossa base de dados convertida para um arquivo do tipo CSV, vamos começar a analisa-la por meio do Orange Canvas.

# Orange

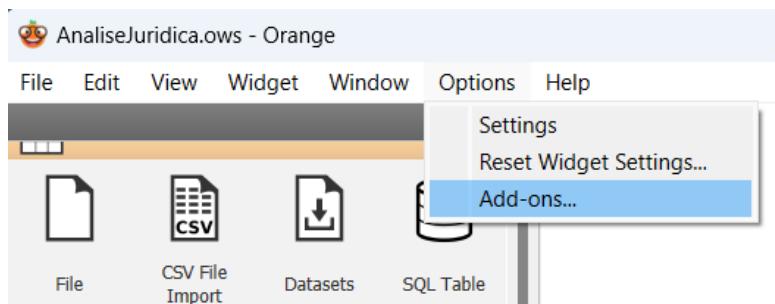
Antes de iniciar, vamos salvar nosso projeto como AnaliseJuridica e depois instalar um AdOn para análises em documentos do tipo texto.



# Orange



# Orange – Adicionando o plugin Text



**Installer - Orange**

	Name	Version	Action
<input type="checkbox"/>	Geo	0.4.0	
<input type="checkbox"/>	Image Analytics	0.11.0	
<input type="checkbox"/>	Network	1.8.0	
<input type="checkbox"/>	Prototypes	0.19.0	
<input type="checkbox"/>	Single Cell	1.5.0	
<input type="checkbox"/>	Spectroscopy	0.6.10	
<input checked="" type="checkbox"/>	Text	1.13.1	Install
<input type="checkbox"/>	Textable	3.1.11	
<input type="checkbox"/>	Timeseries	0.5.3	
<input type="checkbox"/>	Survival Analysis	0.5.1	
<input type="checkbox"/>	World Happiness	0.1.9	

**Orange 3**

Orange is a component-based data mining software. It includes a range of data visualization, exploration, preprocessing and modeling techniques. It can be used through a nice and intuitive user interface or, for more advanced users, as a module for the Python programming language.

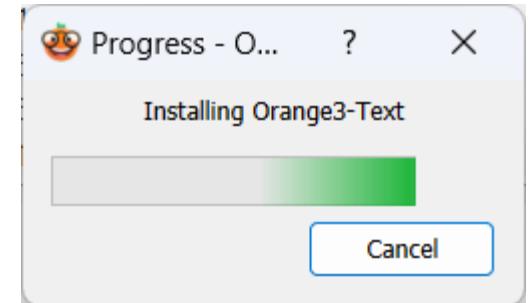
This is the latest version of Orange (for Python 3). The deprecated version of Orange 2.7 (for Python 2.7) is still available ([binaries](#) and [sources](#)).

**Installing with pip**

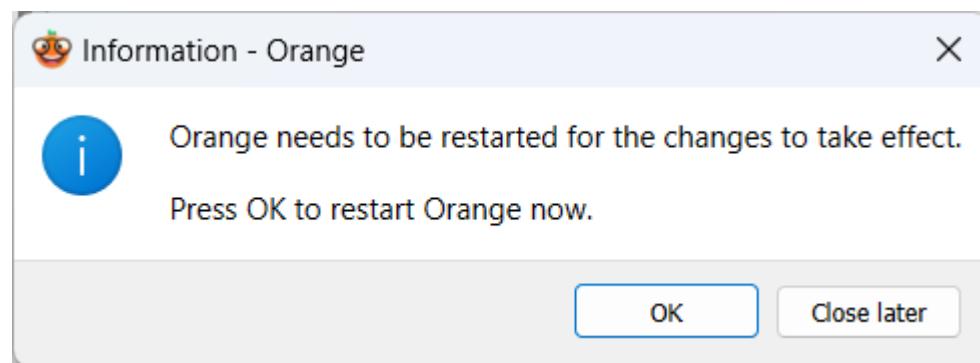
To install Orange with pip, run the following.

```
# Install some build requirements via your system's package manager
sudo apt install virtualenv build-essential python3-dev
```

**OK**   **Cancel**

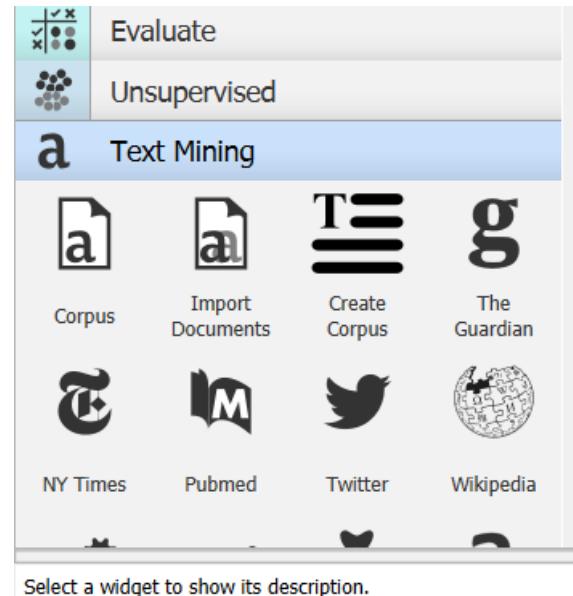


Sempre que precisar instalar alguma biblioteca adicional ao Orange, será necessário o restart da aplicação. Clique para reiniciar o Orange.

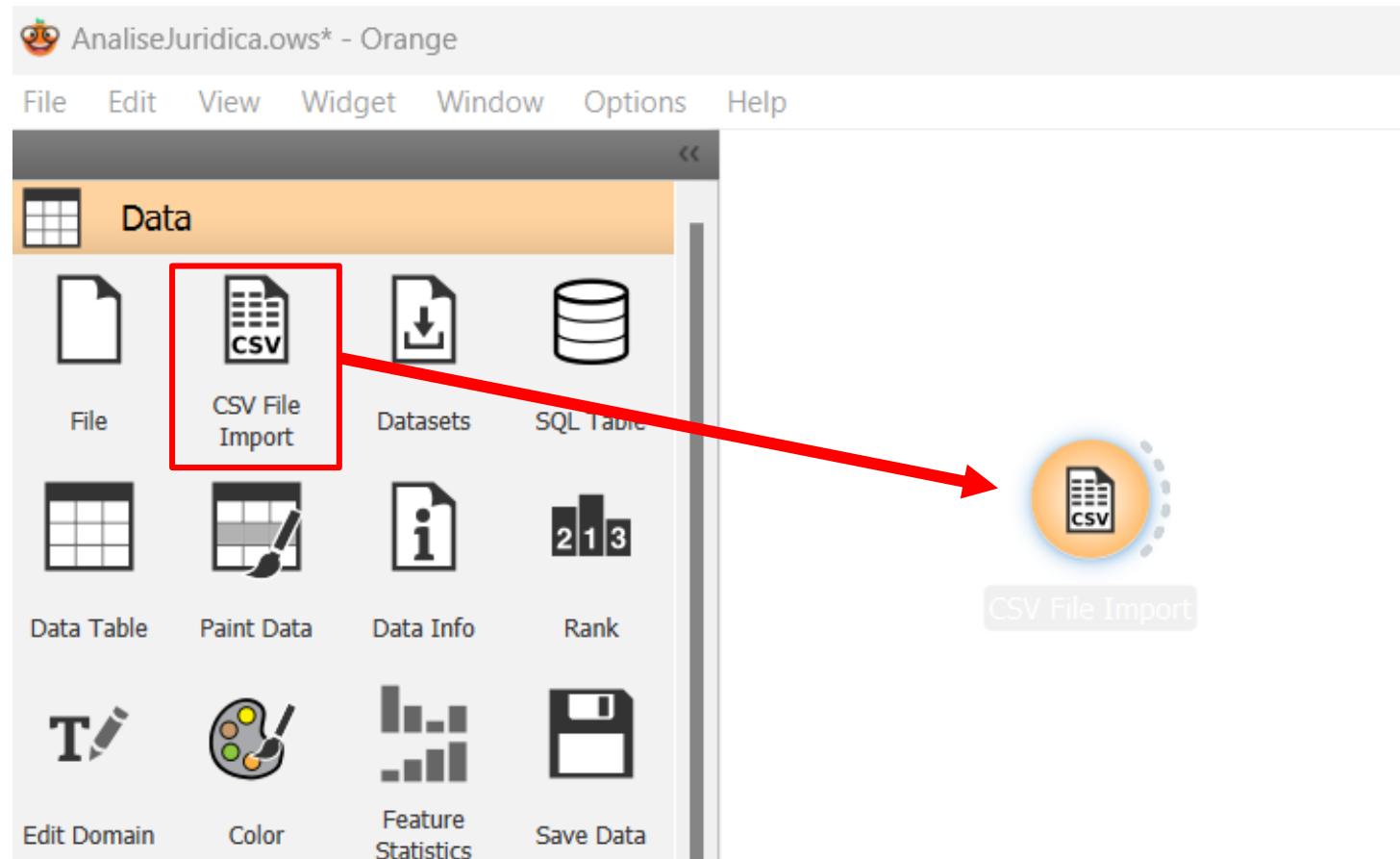


# Orange

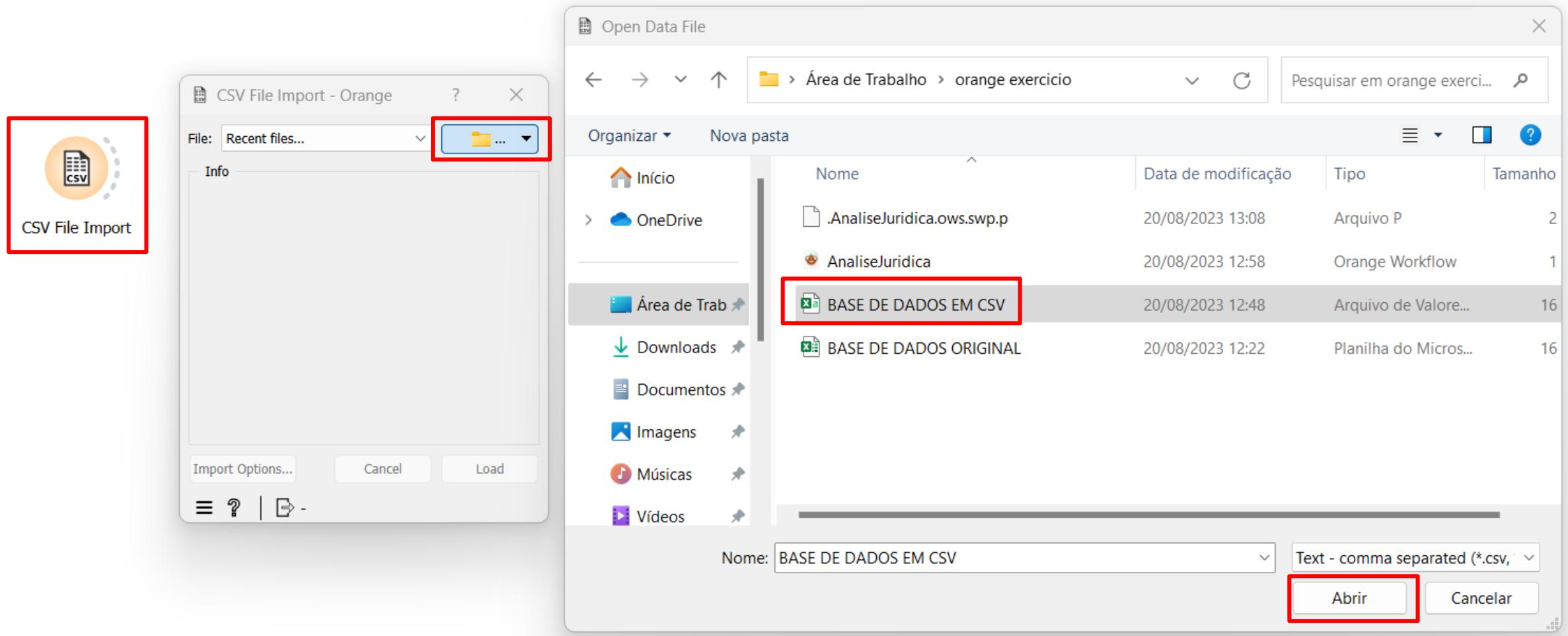
Após reiniciar, reabra seu arquivo (analisjuridica.ows) e verifique que agora temos um grupo de ferramentas para Text Mining



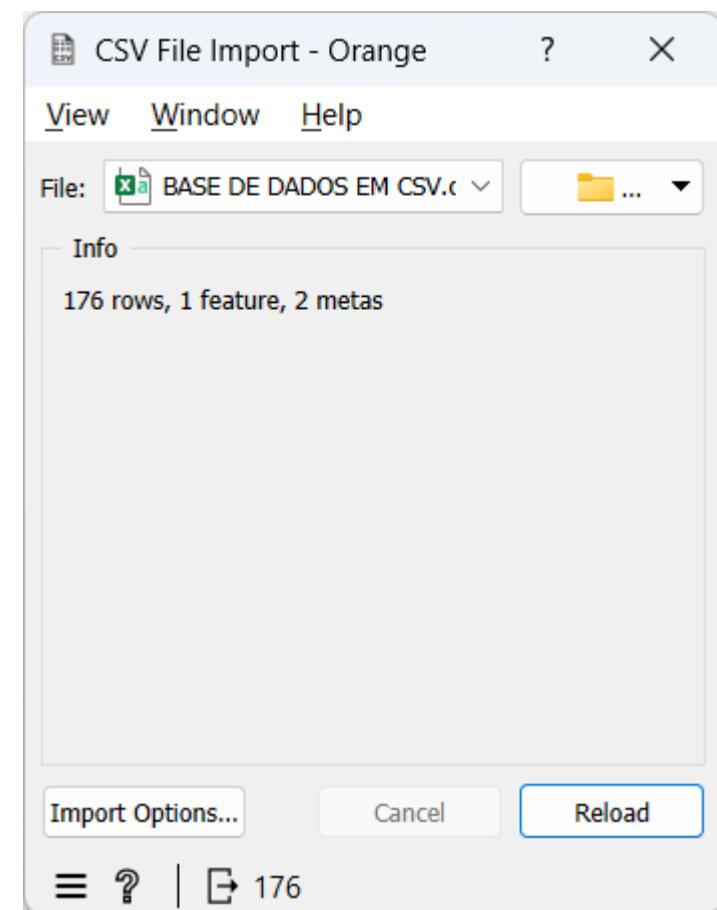
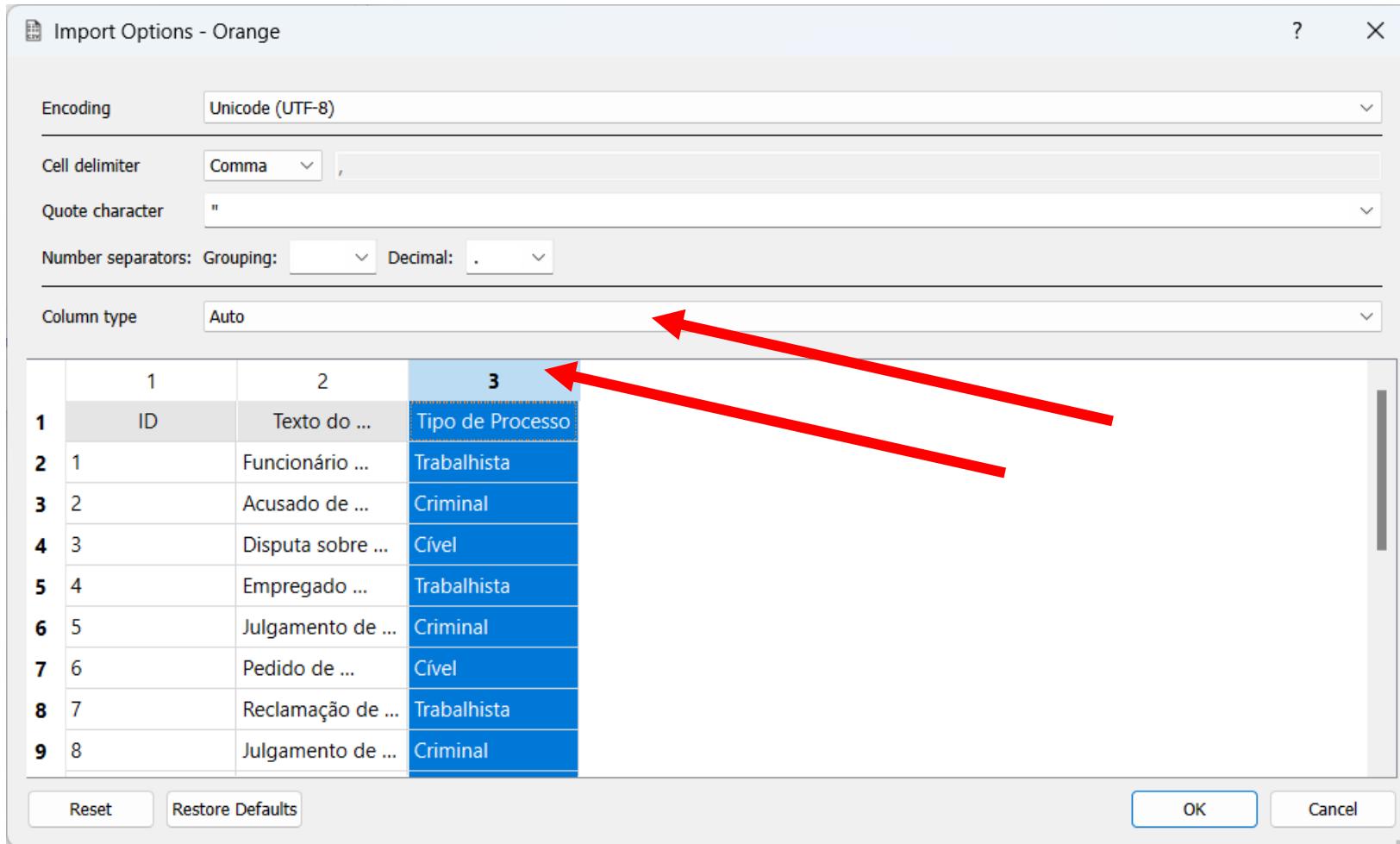
# Orange – Carregando a Base de Dados em CSV



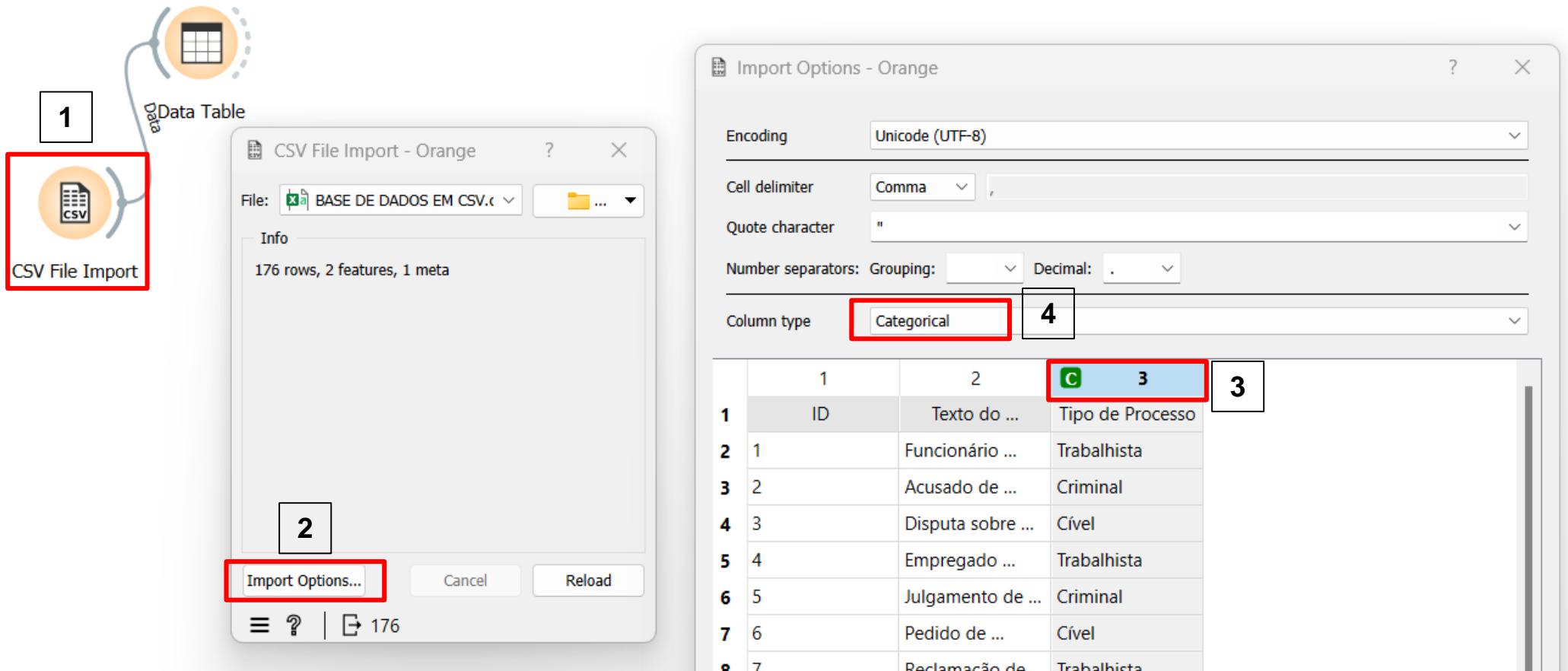
# Orange – Carregando a Base de Dados em CSV



# Orange – Carregando a Base de Dados em CSV



# Orange – Modificando a coluna Tipo para Categorical



The screenshot illustrates the process of modifying a column type in the Orange data mining software.

**CSV File Import Dialog (Left):**

- Step 1:** A **Data Table** node is connected to a **CSV File Import** node (highlighted with a red box).
- Step 2:** The **Import Options...** button is highlighted with a red box.

**Import Options - Orange Dialog (Right):**

- Step 3:** The **Column type** dropdown is set to **Categorical** (highlighted with a red box).
- Step 4:** The **Column type** dropdown is shown again, indicating the change has been applied (highlighted with a red box).

**Table View (Bottom):**

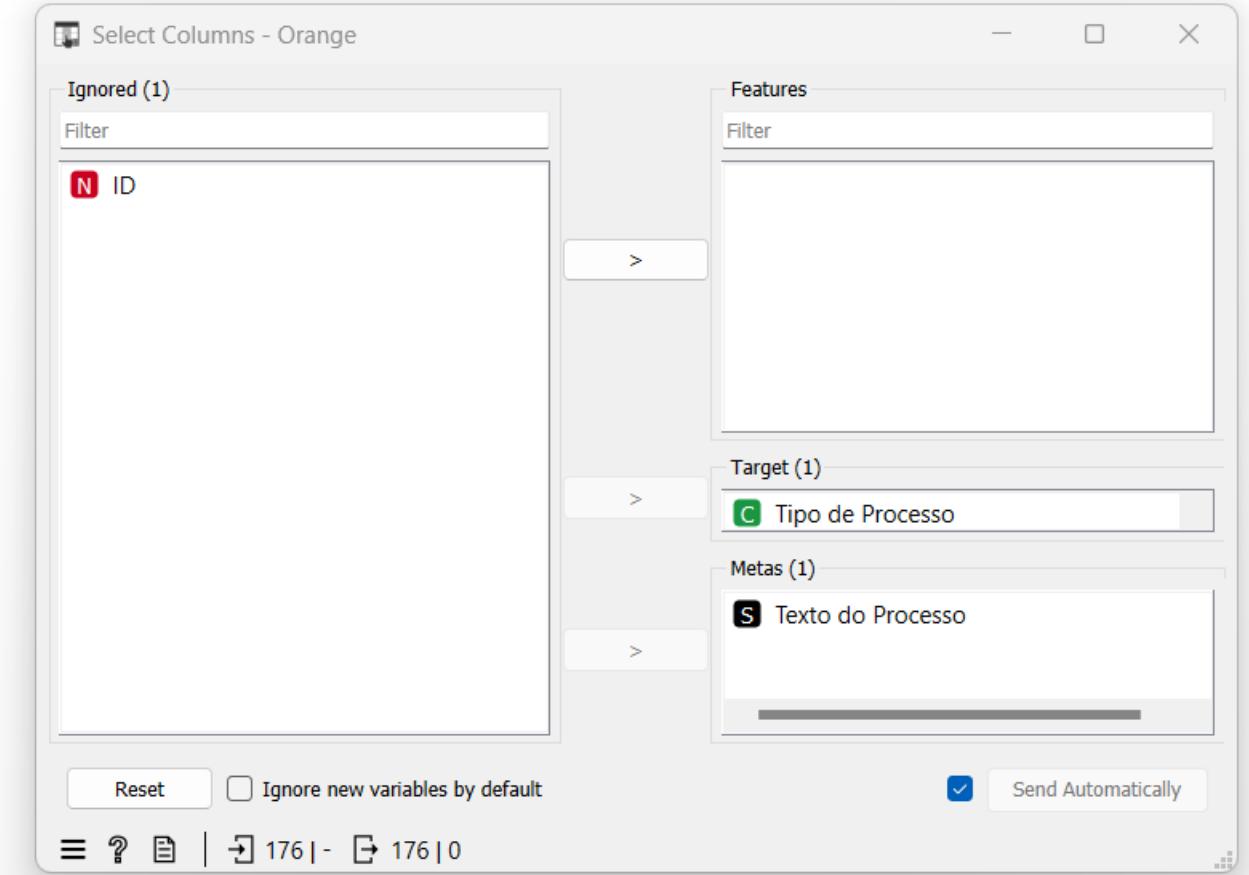
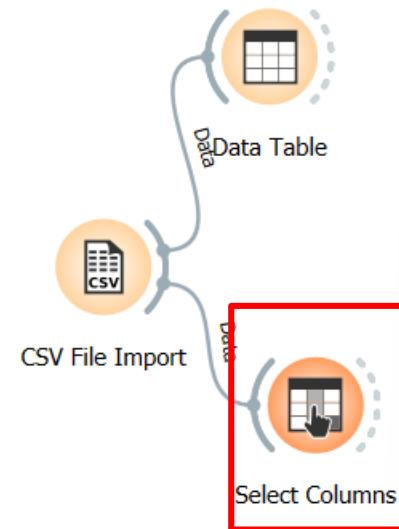
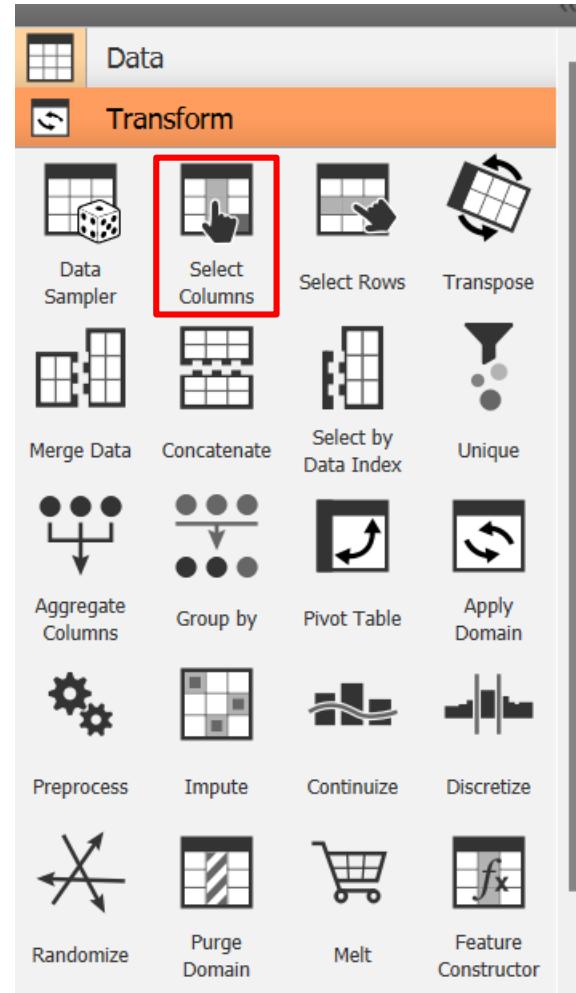
	1	2	C	3
1	ID	Texto do ...	Tipo de Processo	
2	1	Funcionário ...	Trabalhista	
3	2	Acusado de ...	Criminal	
4	3	Disputa sobre ...	Cível	
5	4	Empregado ...	Trabalhista	
6	5	Julgamento de ...	Criminal	
7	6	Pedido de ...	Cível	
8	7	Reclamação de ...	Trabalhista	

# Orange – Visualizando os dados em uma tabela

The screenshot shows the Orange data mining software interface. The top menu bar includes File, Edit, View, Widget, Window, Options, and Help. The main window has a toolbar on the left with various icons for Data, Transform, Visualize, Model, Evaluate, and Unsupervised tasks. A red box highlights the 'Data Table' icon in the Data section. The central workspace shows a flow diagram where a 'CSV File Import' node is connected to a 'Data Table' node, also highlighted with a red box. To the right is a detailed view of the 'Data Table - Orange' window, which displays 176 instances with no missing data. The table has columns for 'Texto do Processo', 'Tipo de Processo', and 'ID'. The data is as follows:

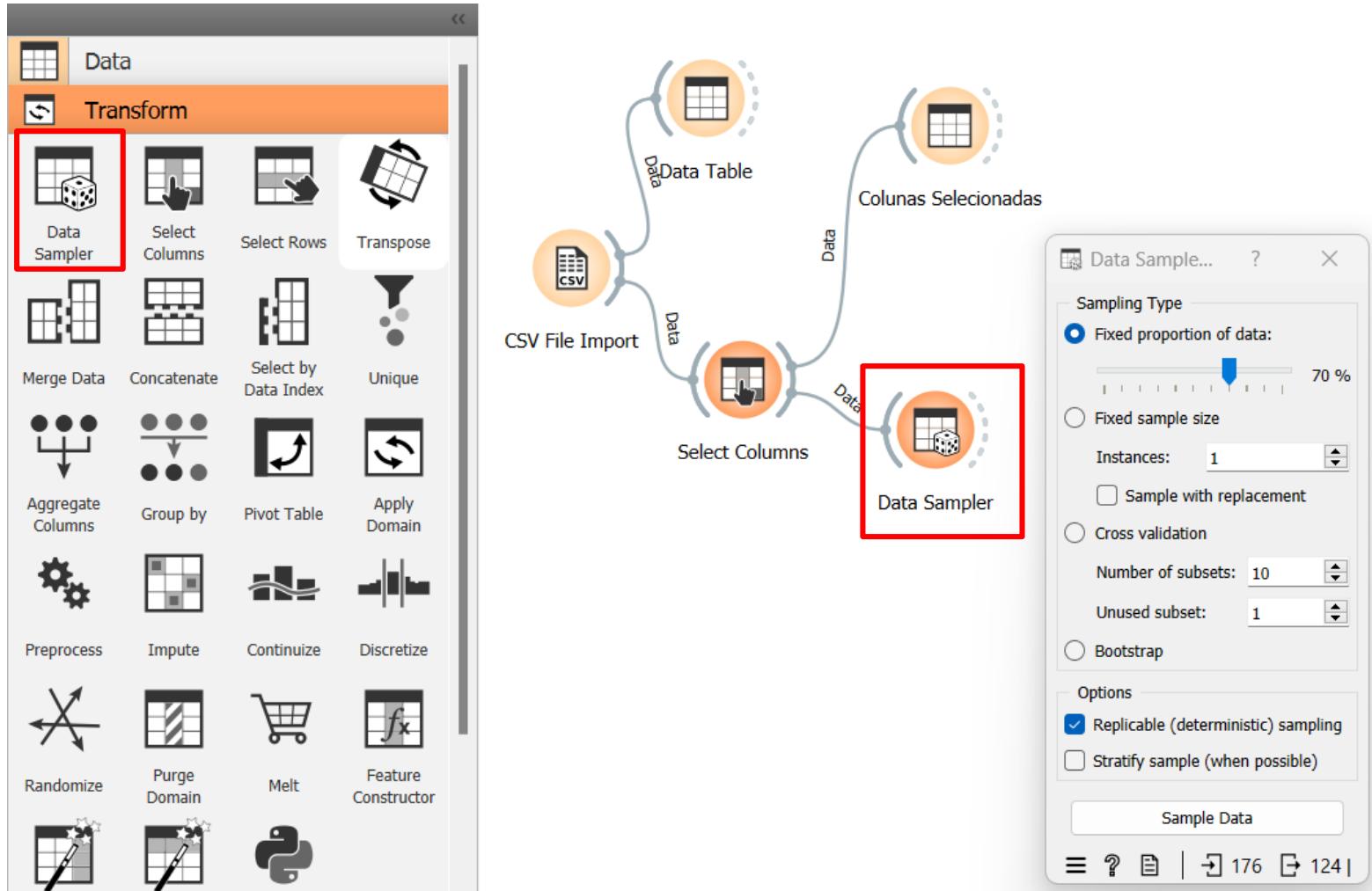
ID	Texto do Processo	Tipo de Processo
1	Funcionário alega salários não pagos após demissão injusta.	Trabalhista
2	Acusado de vandalismo em propriedade pública durante protesto.	Criminal
3	Disputa sobre quebra de contrato de prestação de serviços de consultoria.	Cível
4	Empregado busca indenização por assédio moral no ambiente de trabalho.	Trabalhista
5	Julgamento de fraude em transações financeiras envolvendo esquema de pirâmide.	Criminal
6	Pedido de guarda de menor após divórcio e disputa pela custódia.	Cível
7	Reclamação de discriminação racial no emprego e ambiente hostil.	Trabalhista
8	Julgamento de agressão física resultando em lesões graves.	Criminal
9	Disputa sobre propriedade de imóvel herdado entre herdeiros.	Cível
10	Reivindicação de pagamento de horas extras não remuneradas e falta de intervalos.	Trabalhista
11	Acusação de fraude acadêmica em exame universitário.	Criminal
12	Disputa de patente entre empresas de tecnologia.	Cível
13	Investigação de fraude em licitação pública.	Criminal
14	Funcionário alega retaliação após denunciar má conduta.	Trabalhista
15	Processo de divórcio com questões de pensão alimentícia.	Cível
16	Caso de difamação em redes sociais.	Cível

# Orange – Selecionando coluna Target

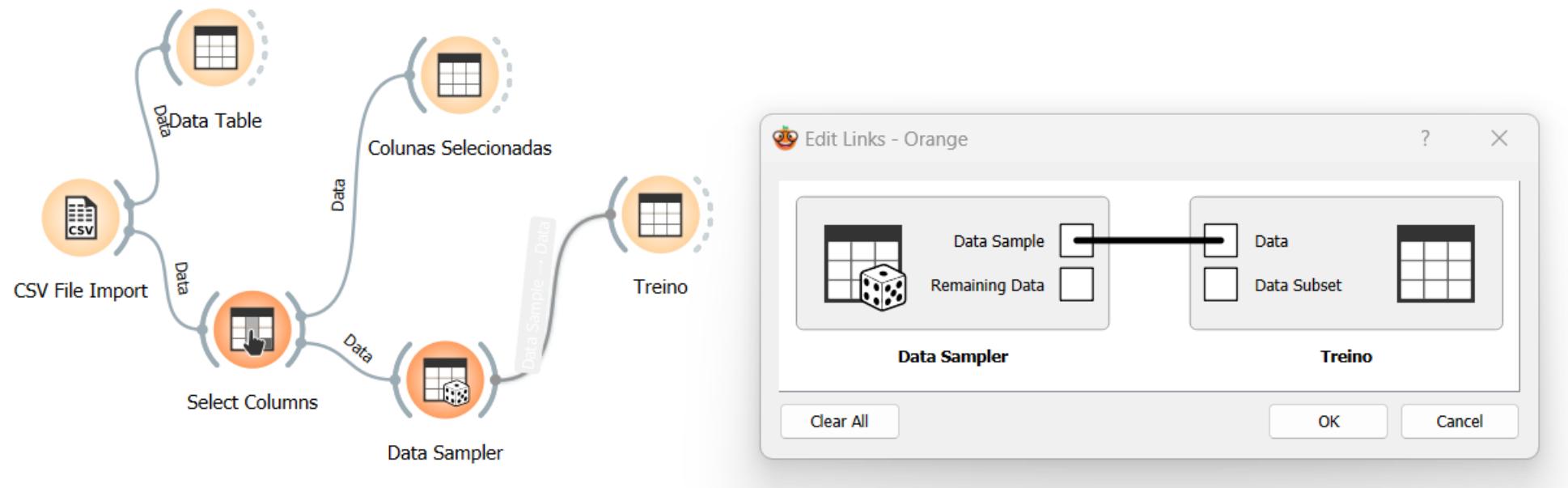


Após os dados estarem corretamente dispostos e sem falhas (dados faltantes ou incorretos) podemos separar nosso conjunto de dados em Dados de Treino e Dados de Teste (85/15)

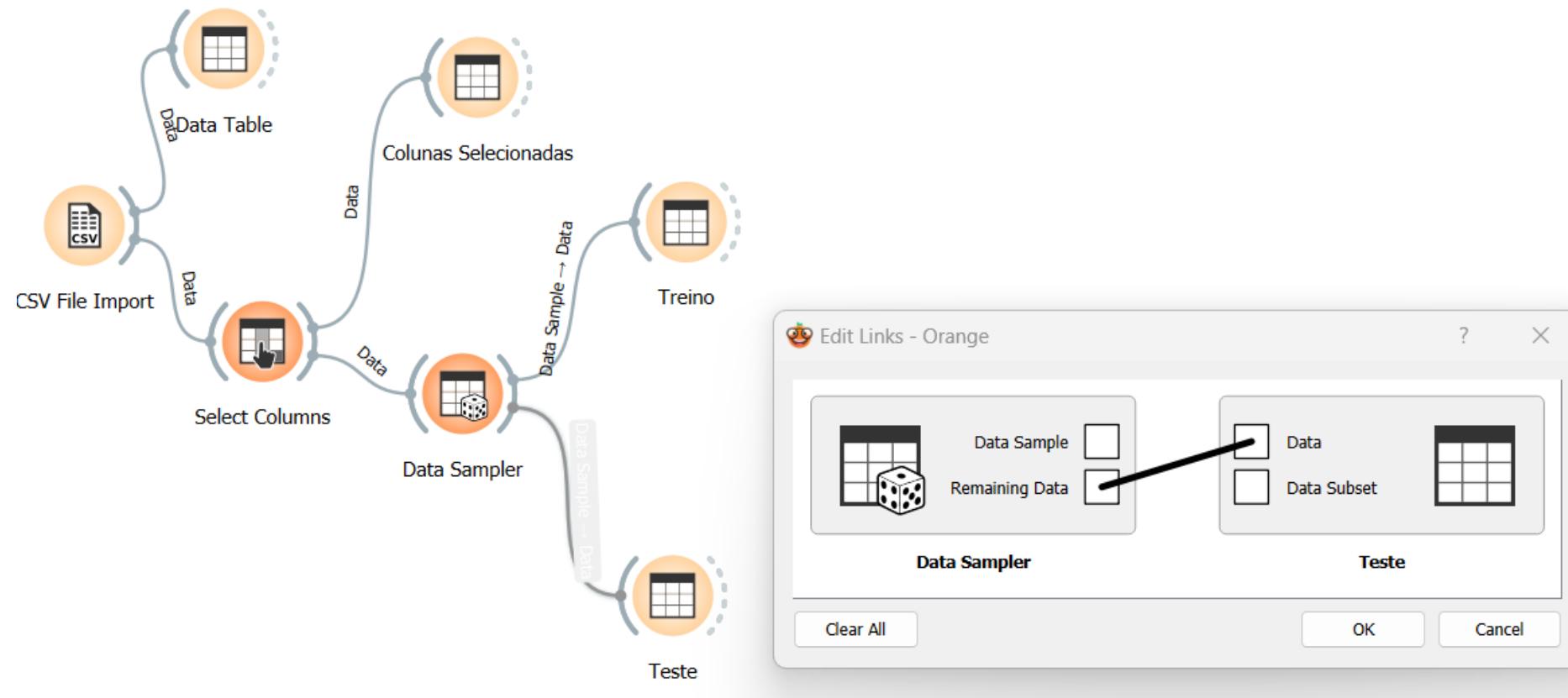
# Orange – Separando em Treino e Teste



# Orange – Visualizando Treino



# Orange – Visualizando Teste



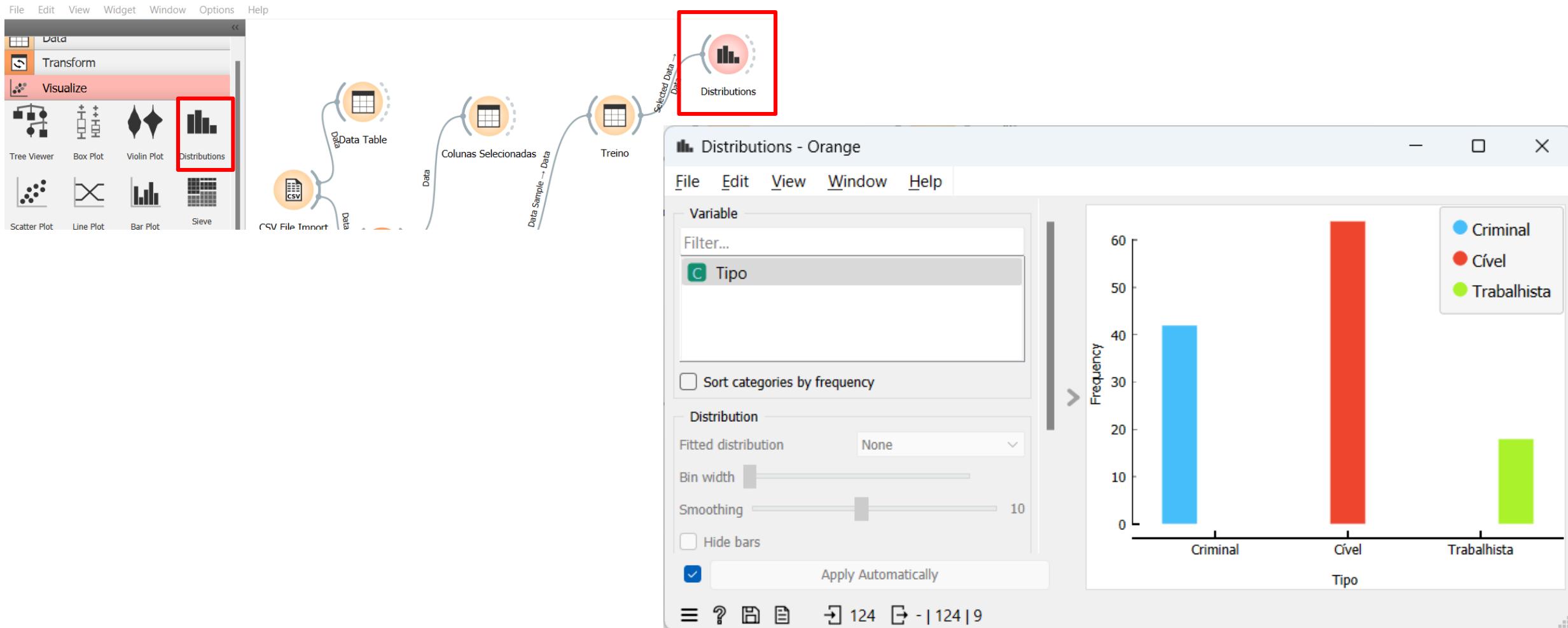
Dividimos nosso conjunto de dados em Dados de Treino e Dados de Teste (70/30).

Tínhamos inicialmente no total 176 registros, fazendo a divisão agora temos:

- **Treino: 124 (70%)**
- **Teste: 52 (30%)**

Iremos utilizar mais tarde os dados de Teste para validar os modelos de Machine Learning.

# Orange – Histograma dos dados de Treino



Agora vamos transformar os dados em Corpus. A principal particularidade do Corpus widget é que ele define os recursos de texto para aplicar a mineração de texto.

# Orange – Aplicando Corpus

File Edit View Widget Window Options Help

VISUALIZE Model Evaluate Unsupervised Text Mining Corpus Import Documents Create Corpus The Guardian NY Times Pubmed Twitter Wikipedia Preprocess Text Corpus to Network Bag of Words Document Embedding

CSV File Import Data Table Colunas Selecionadas Treino Distributions Corpus

Select Columns Data Sampler Remaining Data

Corpus - Orange

Corpus file: book-excerpts.tab

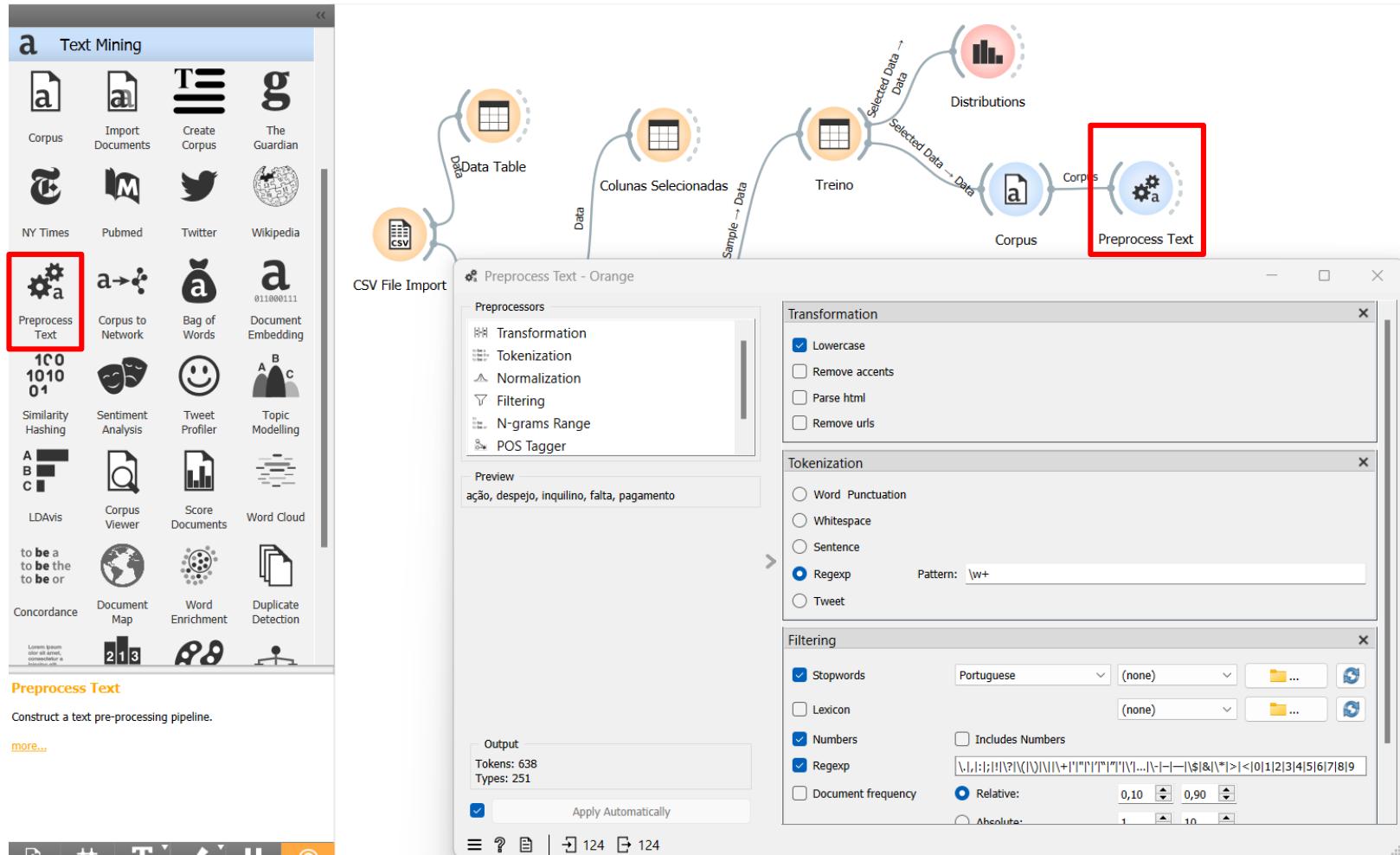
Corpus settings:

- Title variable: Processo
- Language: Portuguese
- Used text features: Processo
- Ignored text features:

Browse documentation corpora

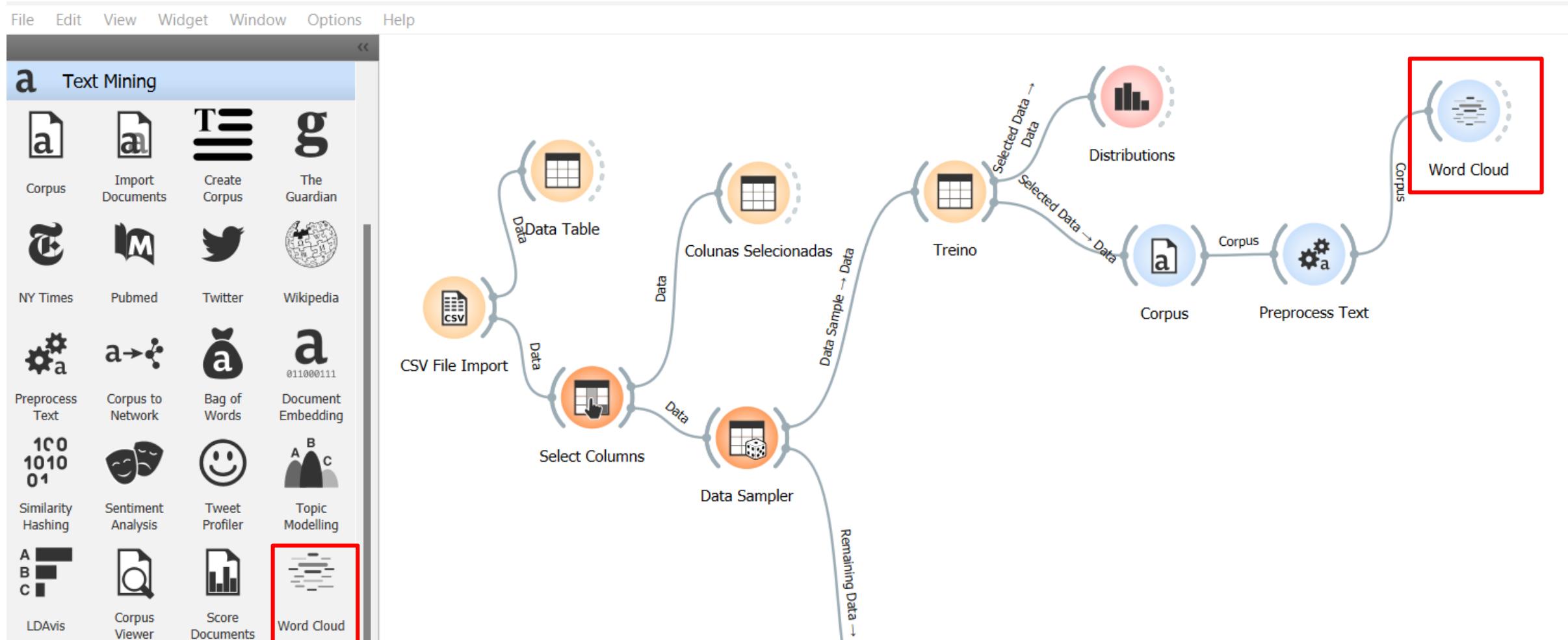
Depois de definido o Corpus de nossa base, vamos fazer um pré-processamento eliminando caracteres especiais, artigos, preposições, pronomes, etc (Stop Words).

# Orange – Pré-processamento Stop Words

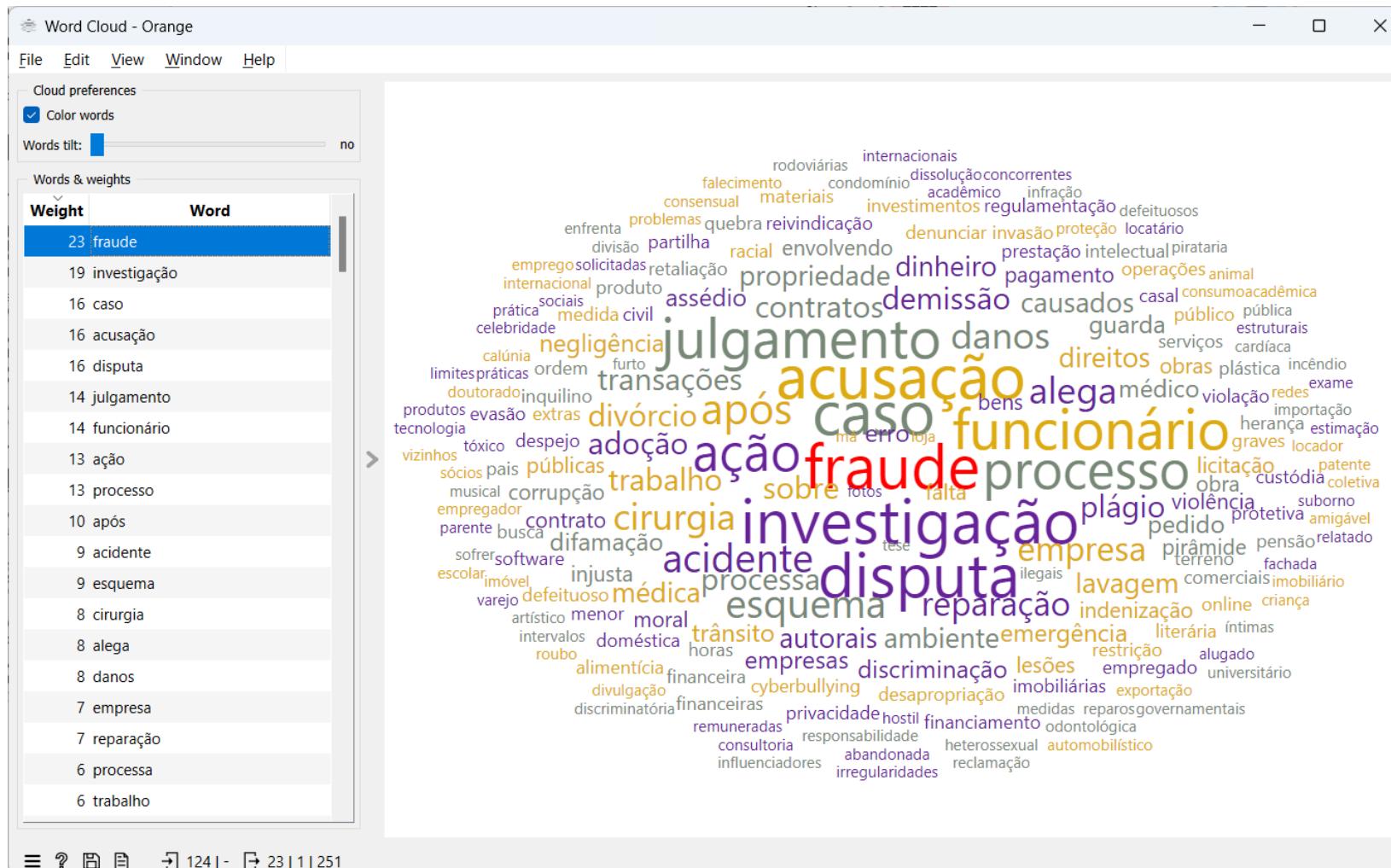


Um recurso bastante utilizado em análises de texto é a Nuvem de Palavras. Na qual podemos ver quais são as palavras mais utilizadas dentro de uma base de dados. Nela também é possível ver qual a frequência de utilização delas.

# Orange – Nuvem de Palavras

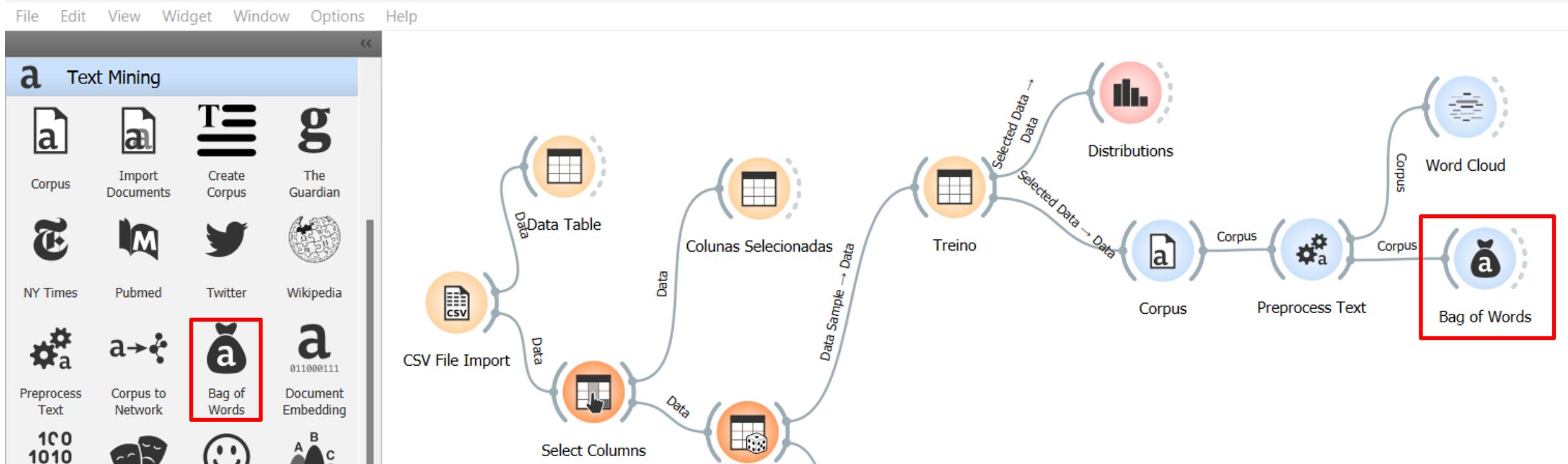


# Orange – Nuvem de Palavras



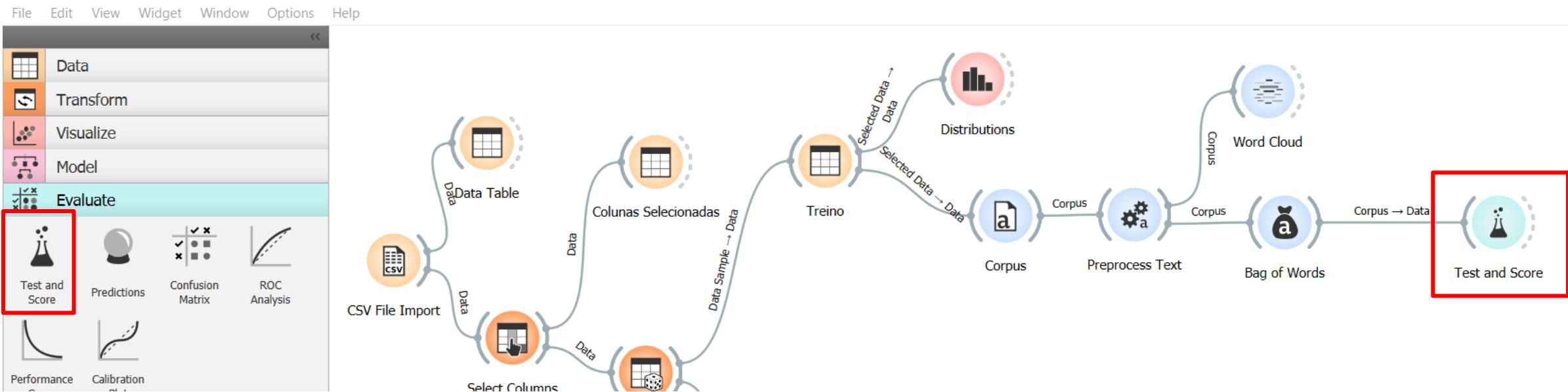
Outro recurso é Bag of Words que cria um corpus com contagens de palavras para cada instância de dados (documento). A contagem pode ser absoluta, binária (contém ou não contém) ou sublinear (logaritmo do termo frequência). O modelo Bag of words pode ser usado para modelagem preditiva.

# Orange – Bag of Words



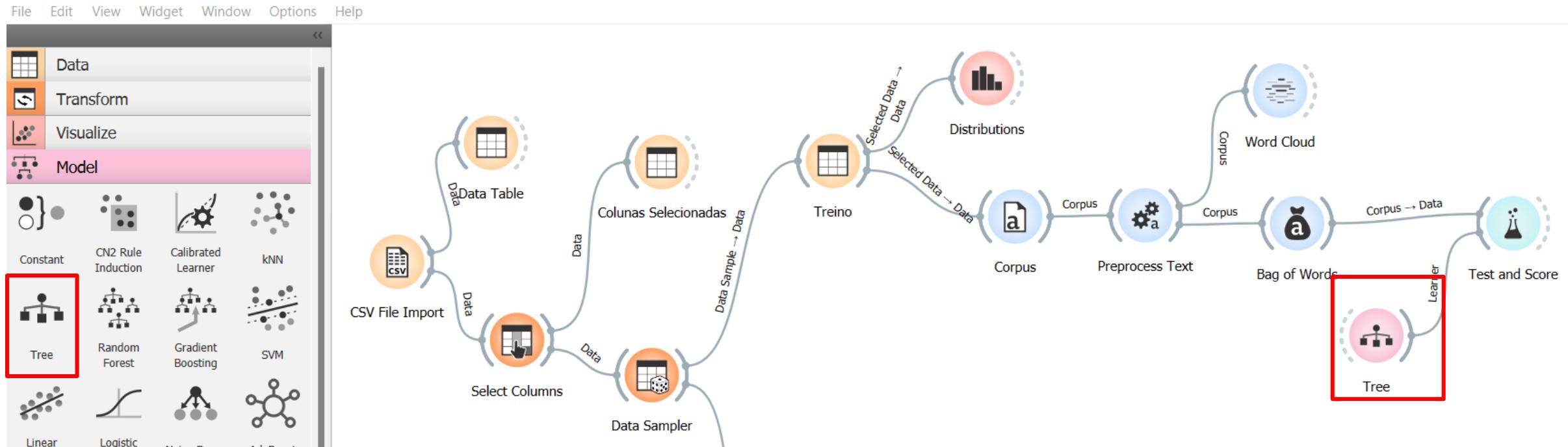
Na próxima etapa já iniciaremos a utilização de algoritmos de machine learning para a realização da classificação dos documentos de nossa base de dados. Para isso utilizaremos um widget chamado Test and Score que em conjunto com os widgets dos modelos de machine learning, trará os resultados de Acurácia e Precisão dos modelos possibilitando a comparação de vários destes.

# Orange – Test and Score

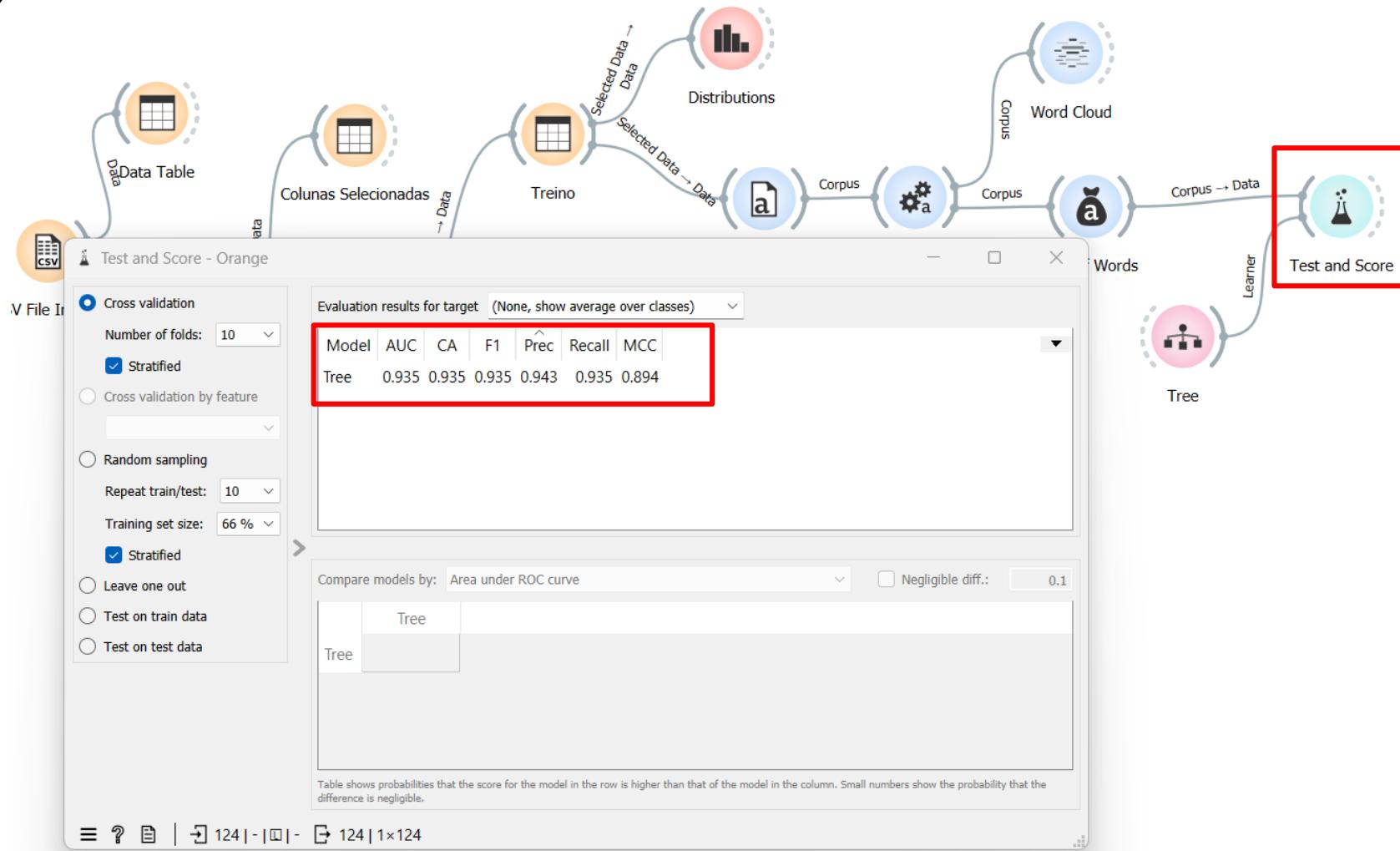


Para visualizar estes resultados devemos selecionar um ou mais modelos de algoritmos que realizam classificação. Vamos começar selecionando um algoritmo chamado Árvore de Decisão.

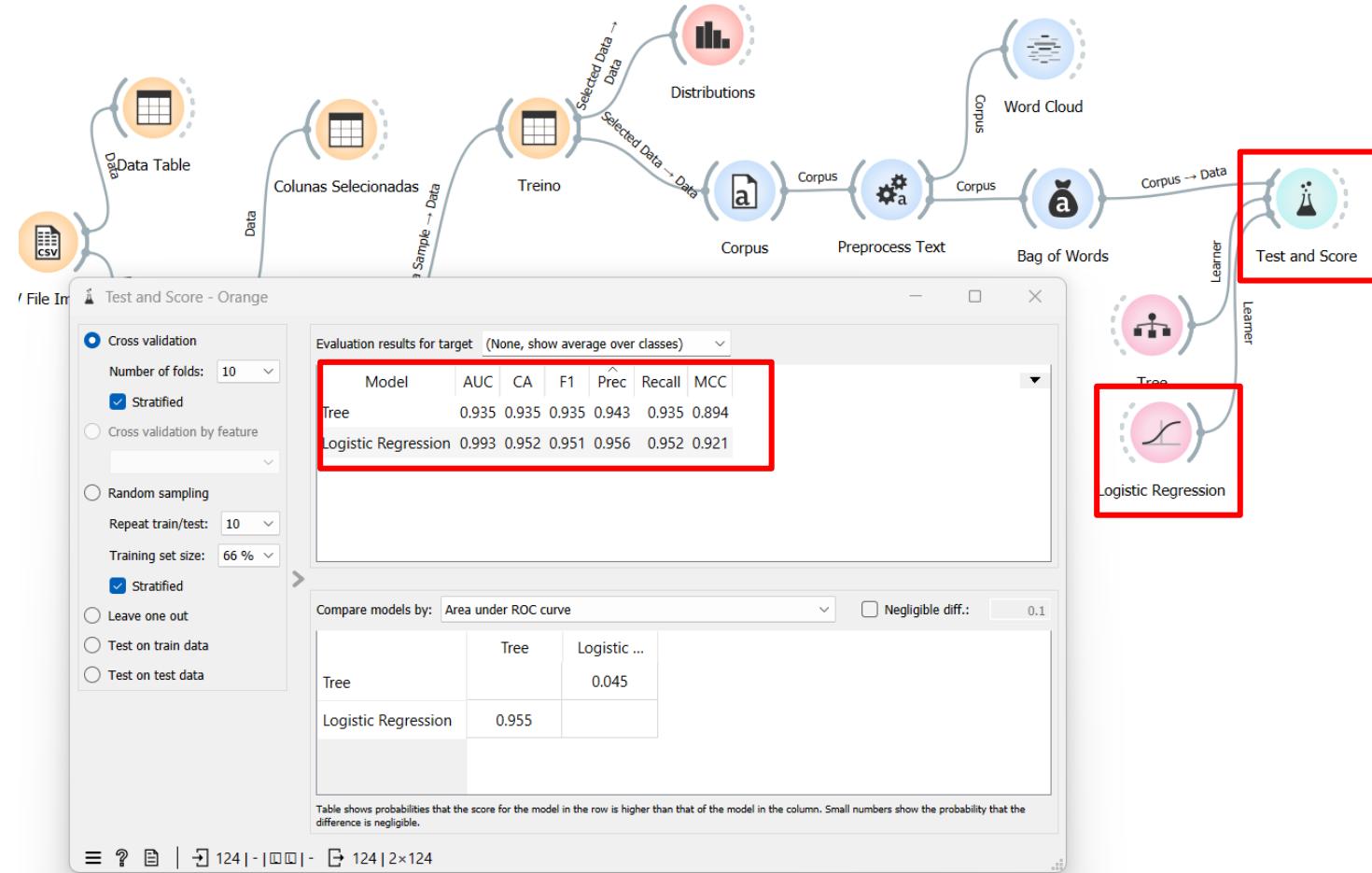
# Orange – Árvore de Decisão



# Orange – Árvore de Decisão



# Orange – Regressão Logística

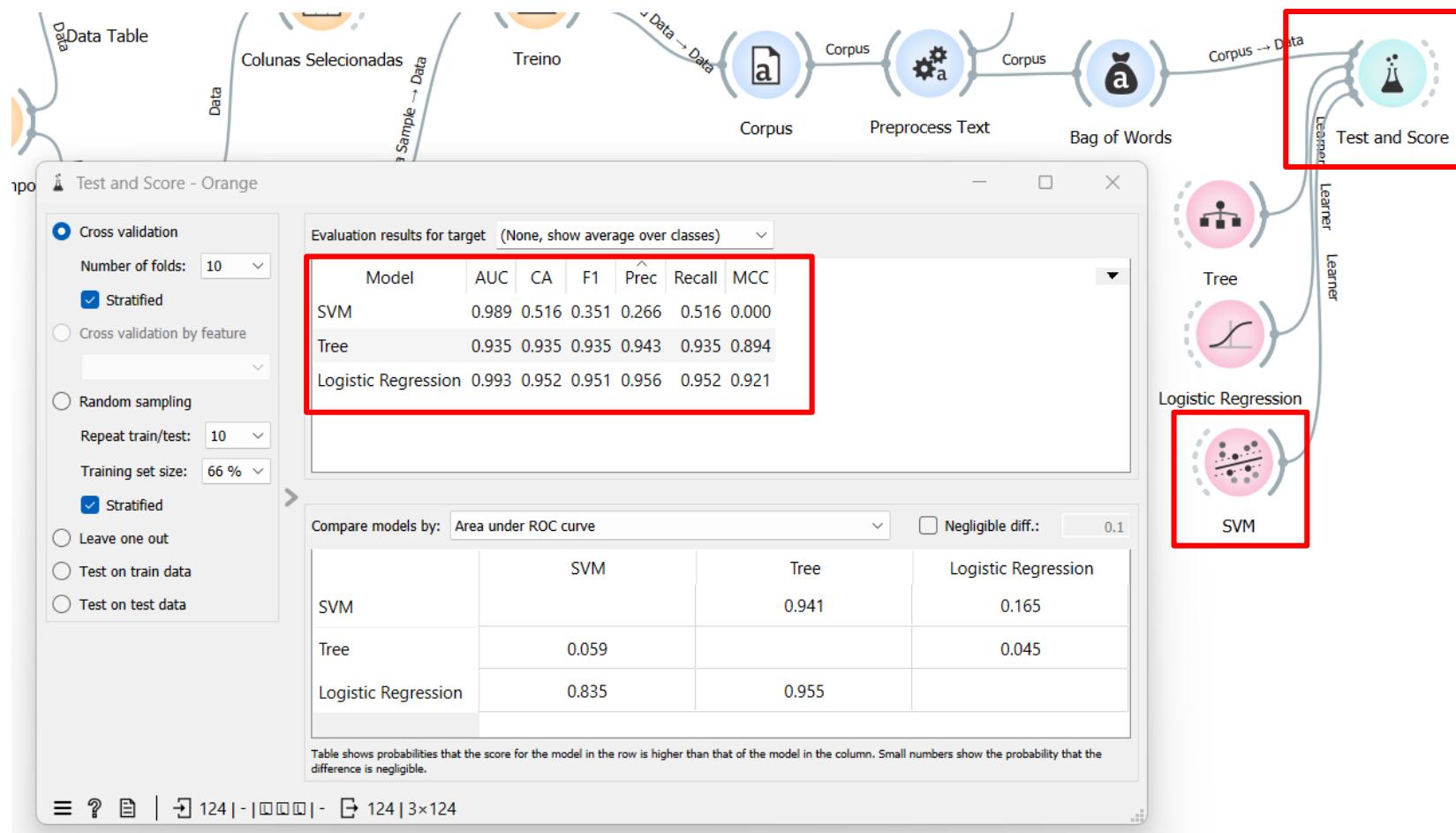


Até aqui, aplicamos 2 modelos de Machine Learning e já é possível verificar alguns pontos interessantes. Vamos ver como foi a Acurácia de cada um deles:

Model	AUC	CA	F1	$\hat{Prec}$	Recall	MCC
Tree	0.935	0.935	0.935	0.943	0.935	0.894
Logistic Regression	0.993	0.952	0.951	0.956	0.952	0.921

Vamos adicionar mais um modelo chamado SVM (Suport Vector Machine) e avaliar o desempenho.

# Orange – SVM



Observe atentamente aos resultados.

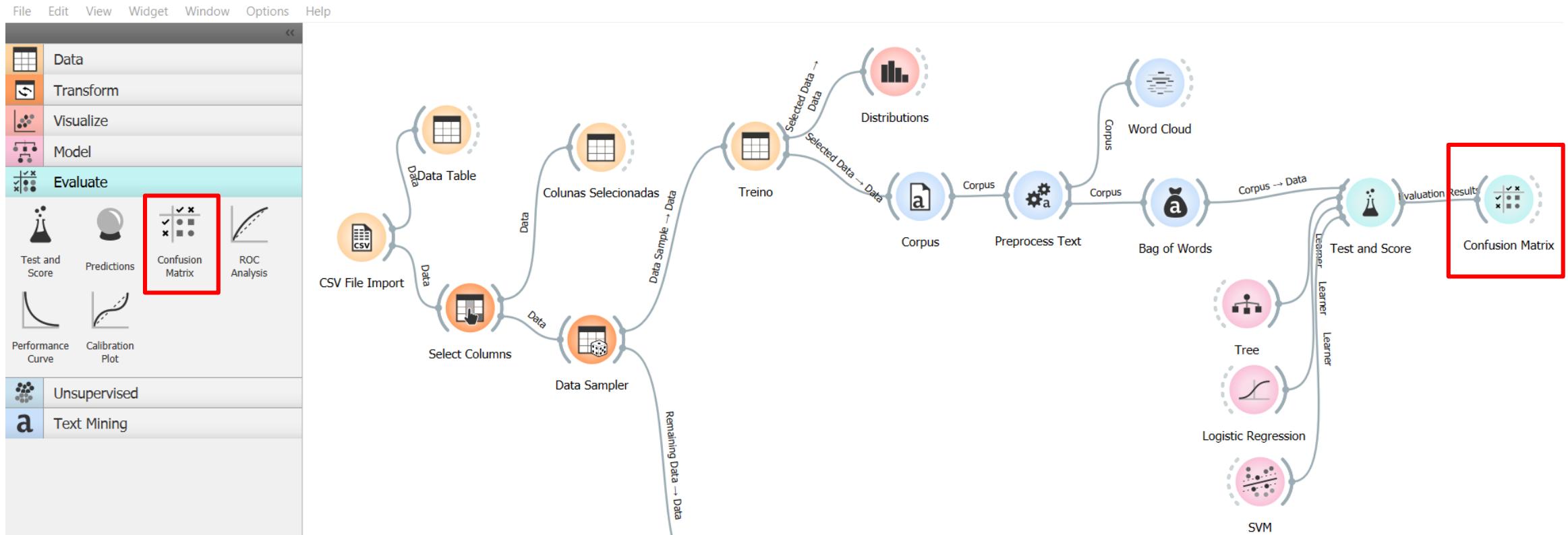


Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.989	0.516	0.351	0.266	0.516	0.000
Tree	0.935	0.935	0.935	0.943	0.935	0.894
Logistic Regression	0.993	0.952	0.951	0.956	0.952	0.921

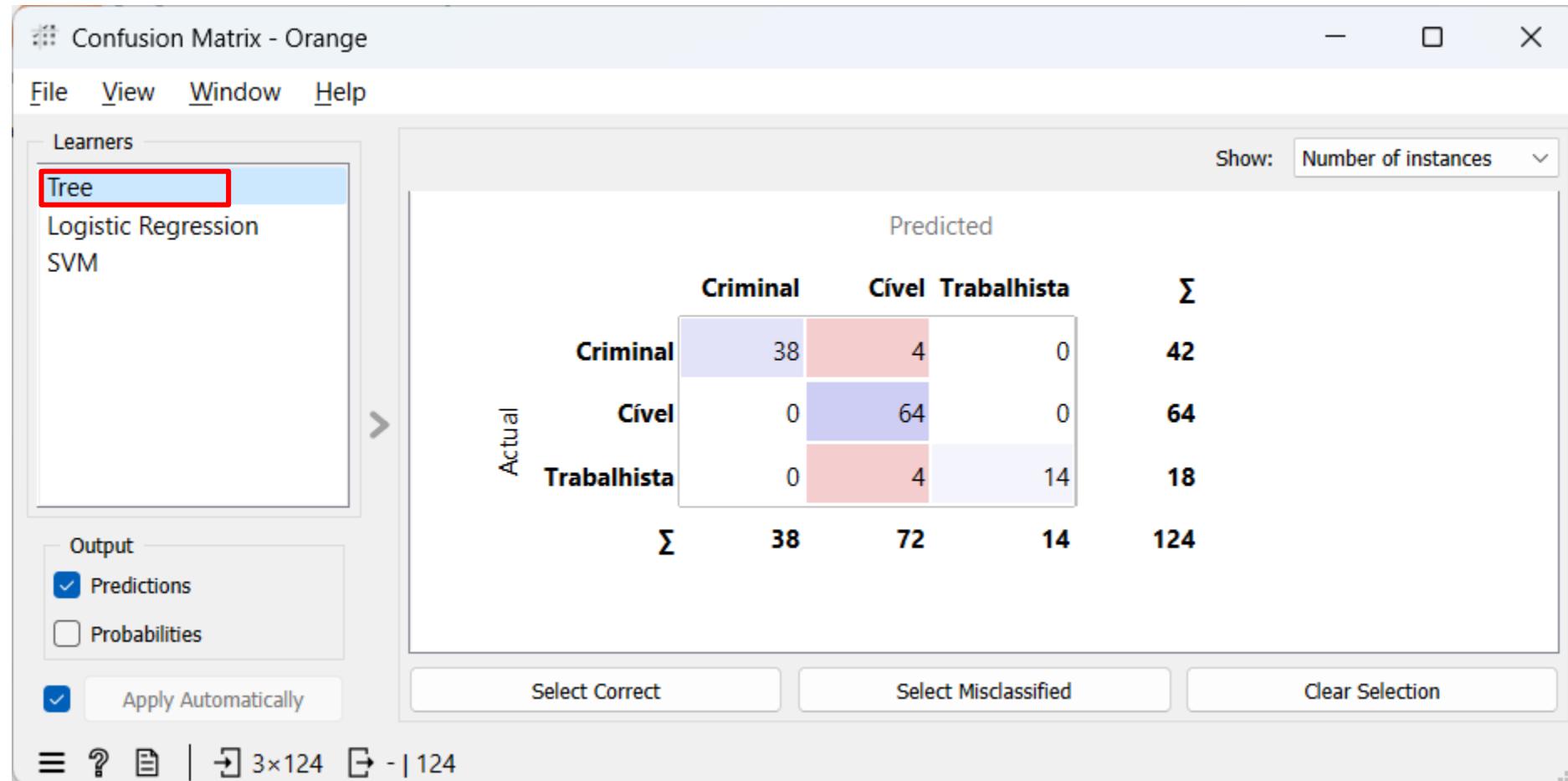
Apesar da Acurácia ter ficado próximo aos modelos anteriormente testados, observe que a coluna F1 ficou muito diferente dos demais. Valores ideais são aqueles próximos de 1.

Para entender um pouco mais o que aconteceu com nossos modelos, vamos utilizar um outro recurso chamado Matriz de Confusão.

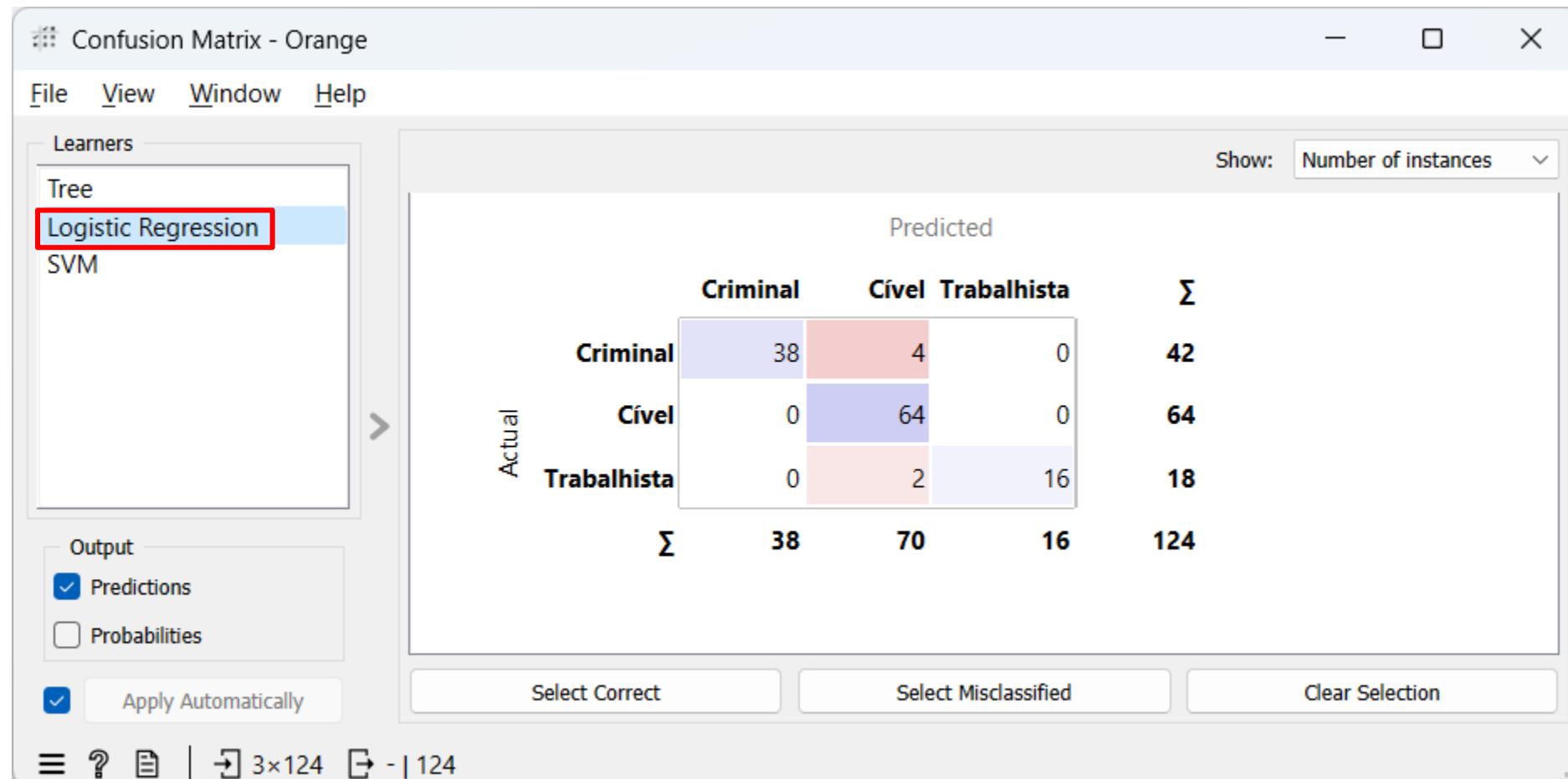
# Orange – Matriz de Confusão



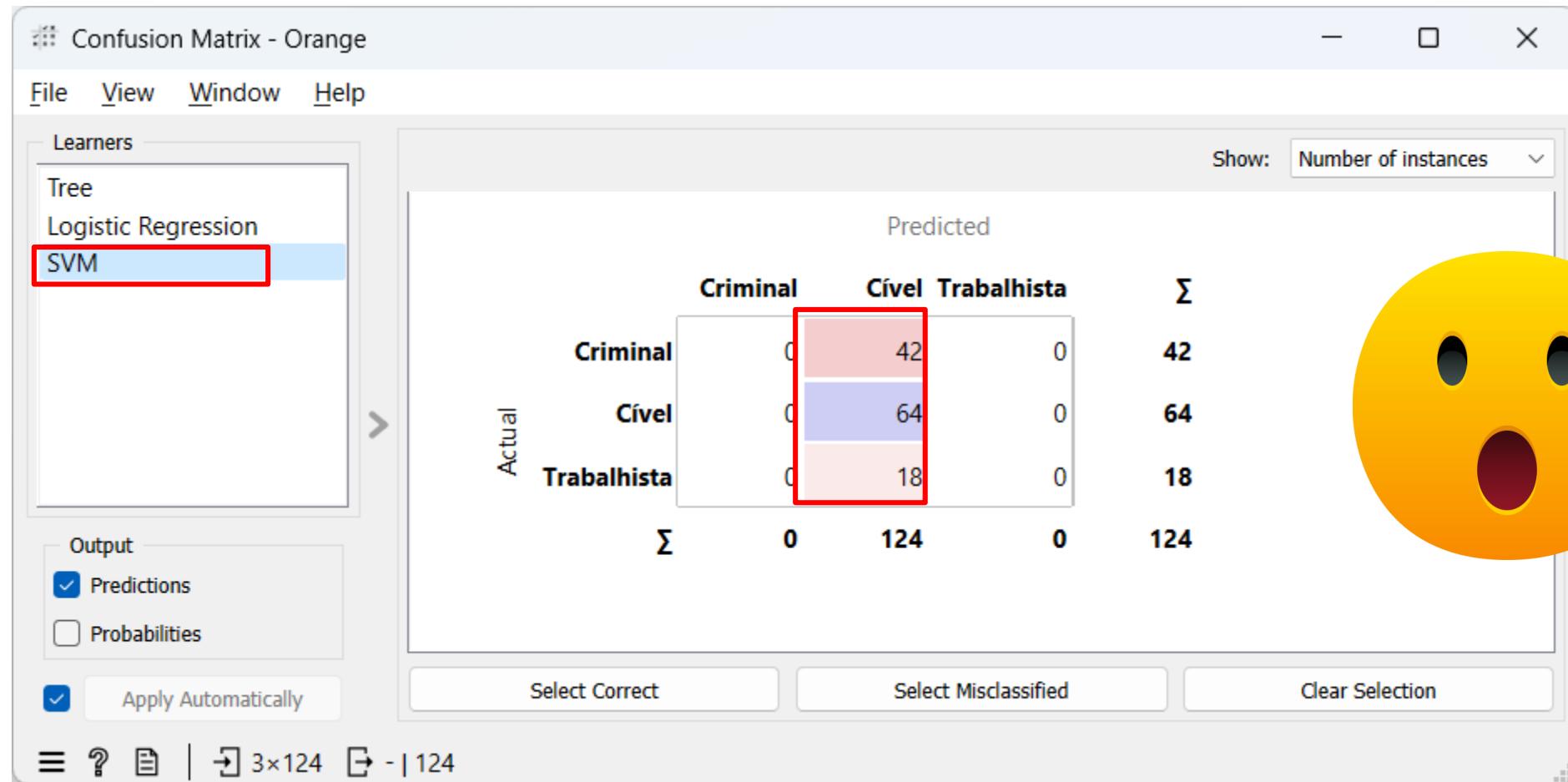
# Orange – Matriz de Confusão - Árvore



# Orange – Matriz de Confusão – Regressão Logistica



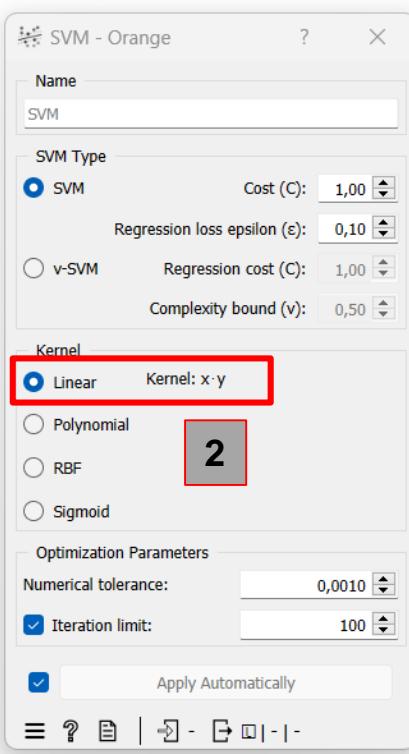
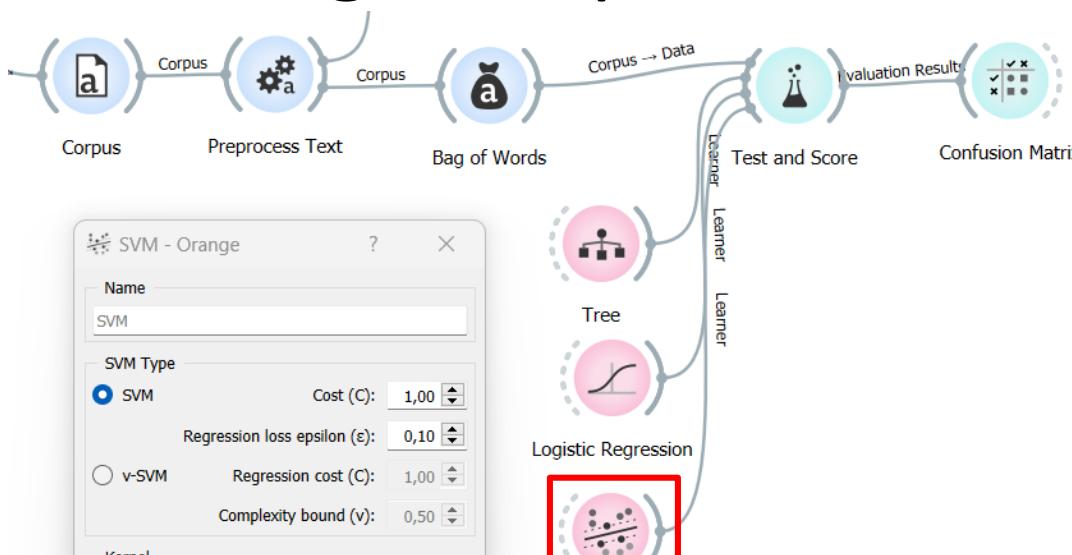
# Orange – Matriz de Confusão – SVM



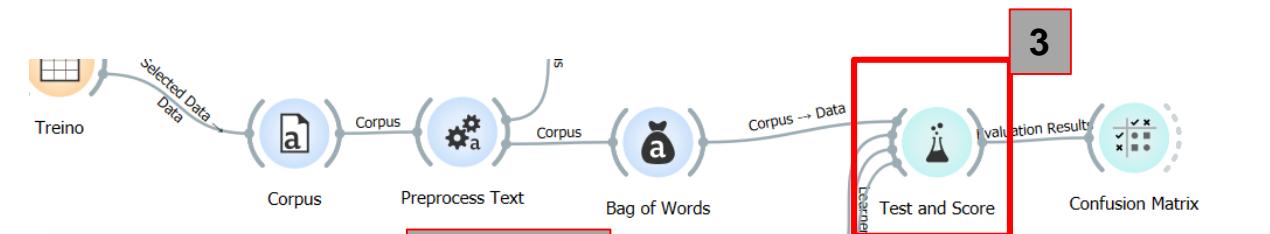
Alguns modelos de algoritmos performam muito bem com suas configurações básicas de parâmetros, porém outros, como acabamos de ver, o SVM, foi um dos que tiveram uma performance baixa. Uma das alternativas pode ser o ajuste de seus hiper-parâmetros.

Vamos alterar um de seus parâmetros para visualizar sua nova performance.

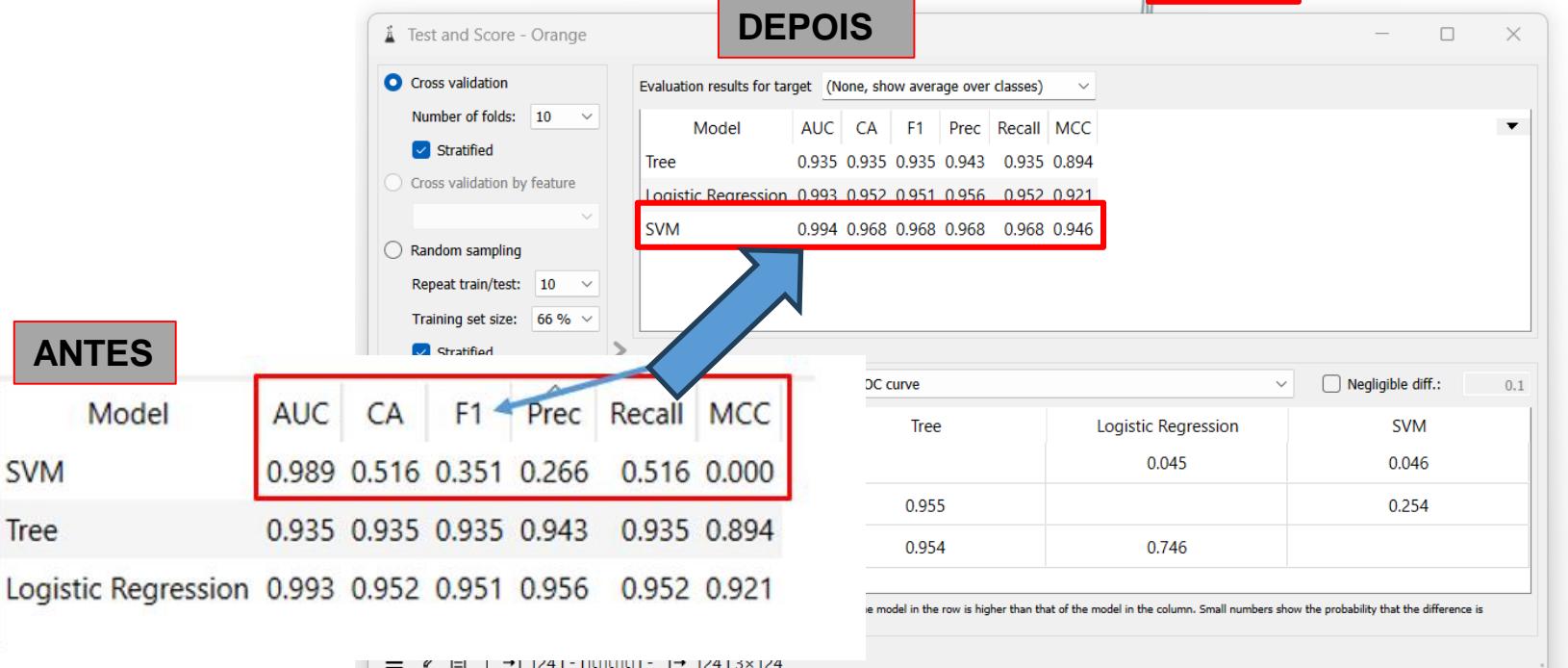
# Orange – Hiper Parâmetros – SVM



1  
2



**DEPOIS**



Screenshot showing the results of the 'Test and Score - Orange' dialog and the resulting table:

**Test and Score - Orange**

- Cross validation selected, Number of folds: 10, Stratified checked
- Evaluation results for target (None, show average over classes) table:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.935	0.935	0.935	0.943	0.935	0.894
Logistic Regression	0.993	0.952	0.951	0.956	0.952	0.921
<b>SVM</b>	<b>0.994</b>	<b>0.968</b>	<b>0.968</b>	<b>0.968</b>	<b>0.968</b>	<b>0.946</b>

**ANTES**

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.989	0.516	0.351	0.266	0.516	0.000
Tree	0.935	0.935	0.935	0.943	0.935	0.894
Logistic Regression	0.993	0.952	0.951	0.956	0.952	0.921

**DC curve**

Tree	Logistic Regression	SVM
0.955	0.045	0.254
0.954	0.746	

ie model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is

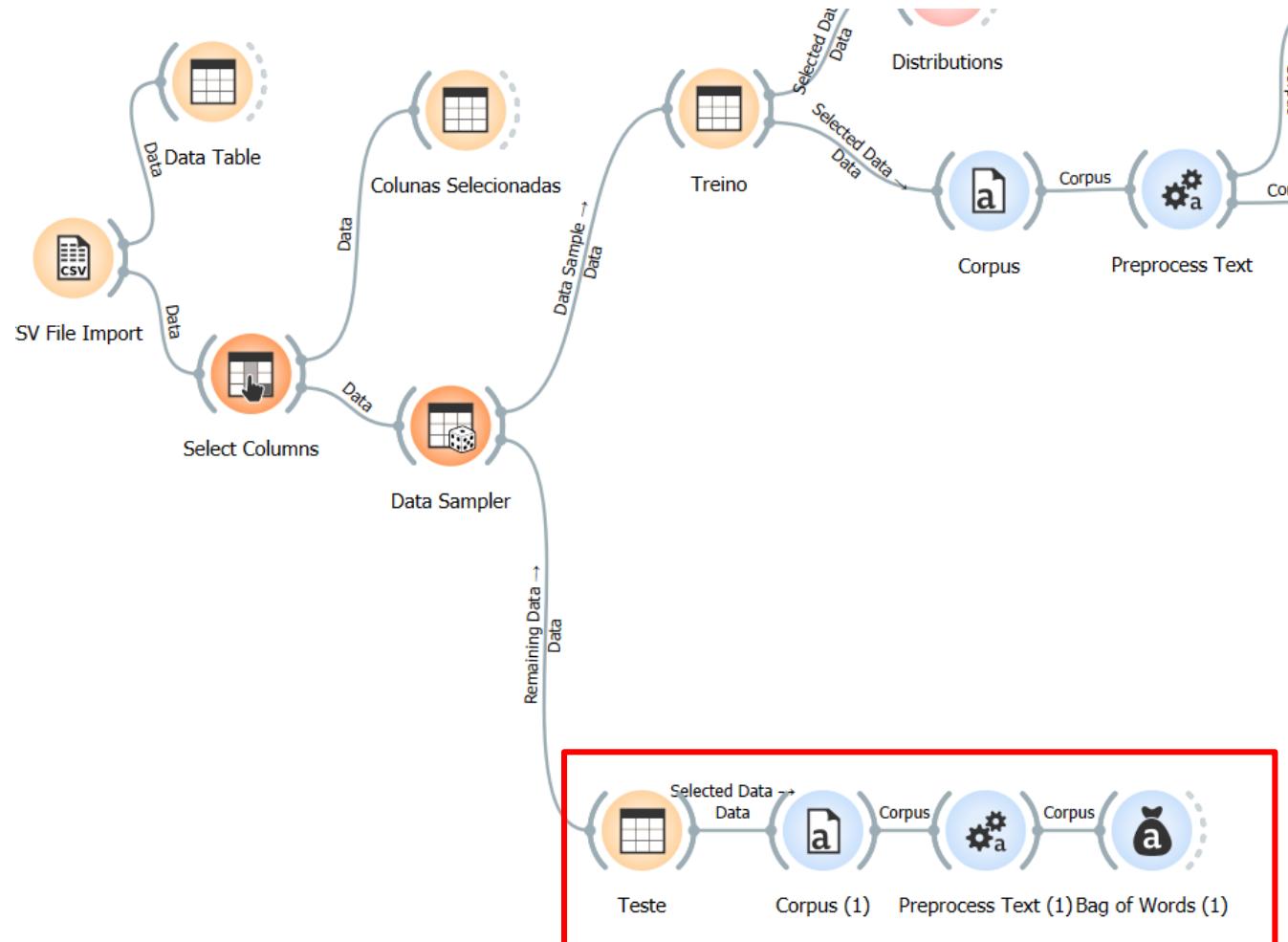
Ficou claro que uma pequena mudança afetou bastante os resultados. Compare:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.989	0.516	0.351	0.266	0.516	0.000
Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.994	0.968	0.968	0.968	0.968	0.946

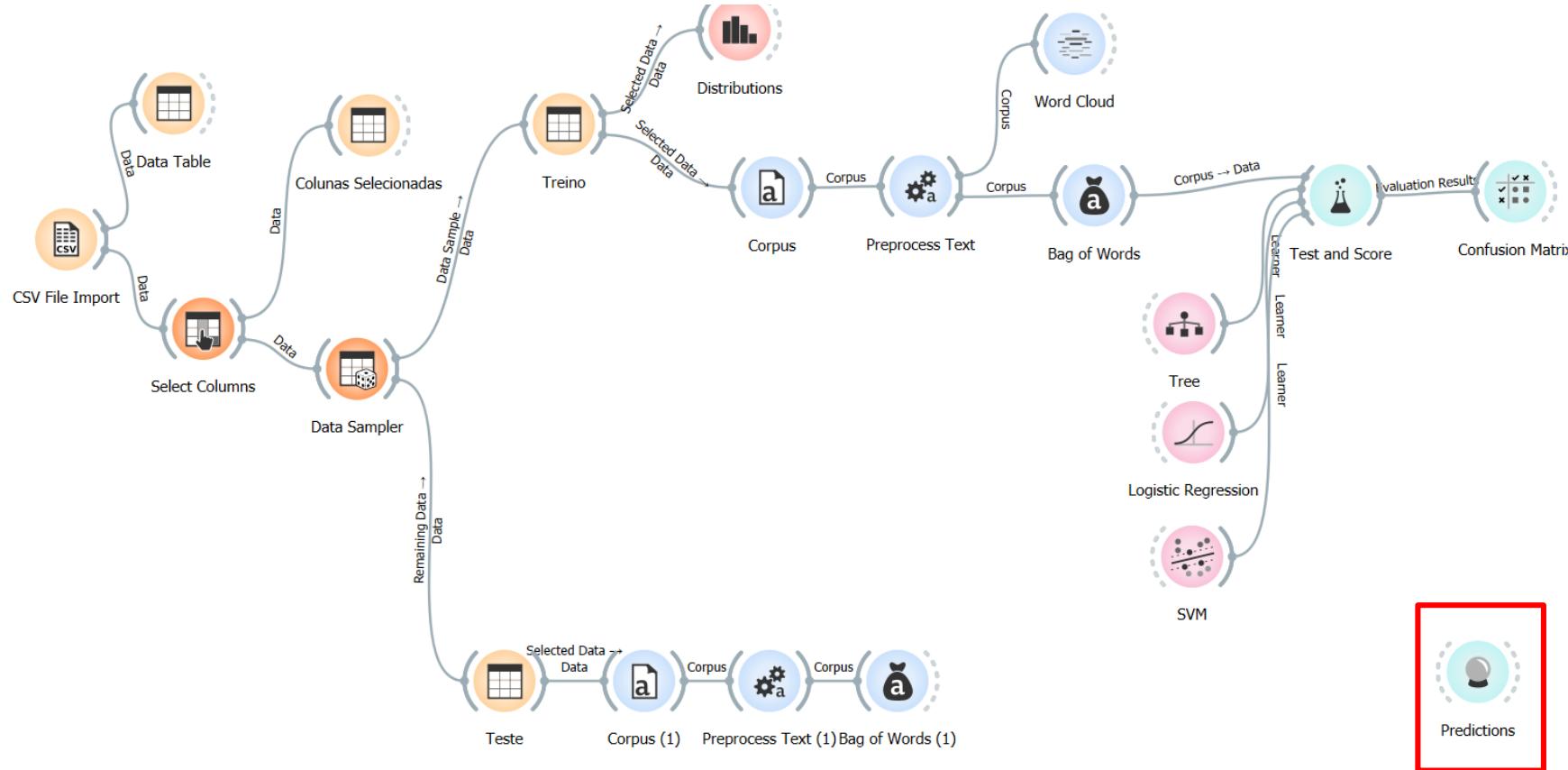
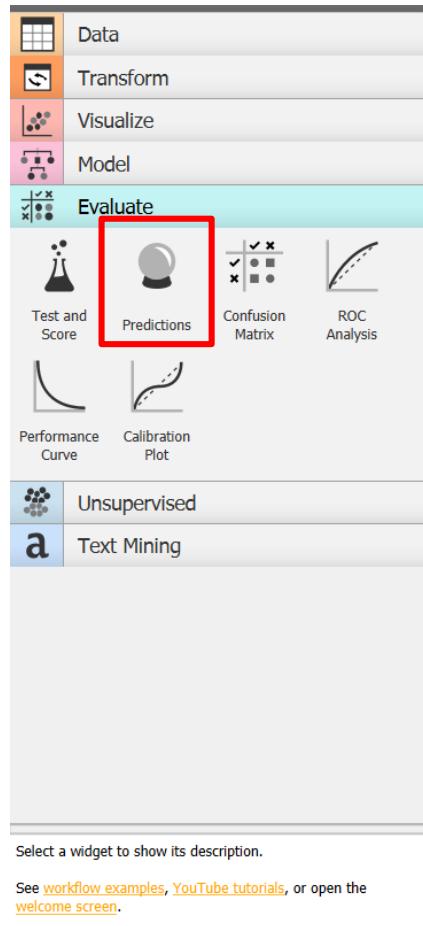
Agora que temos pelo menos 3 modelos treinados e que boa precisão, vamos voltar nossa atenção aos 30% de dados que no início desta aula deixamos separados e que serão utilizados agora para efetuarmos teste de previsão com nossos modelos. Lembre-se que que estes são dados nunca vistos pelos nossos modelos, então será uma boa prova para testar o quanto nossos modelos estão acertando de fato!

Nos dados restantes (30%) aplique as etapas de Corpus, Preprocess Text e Bag of Words.

# Orange – Dados de Teste (30%)

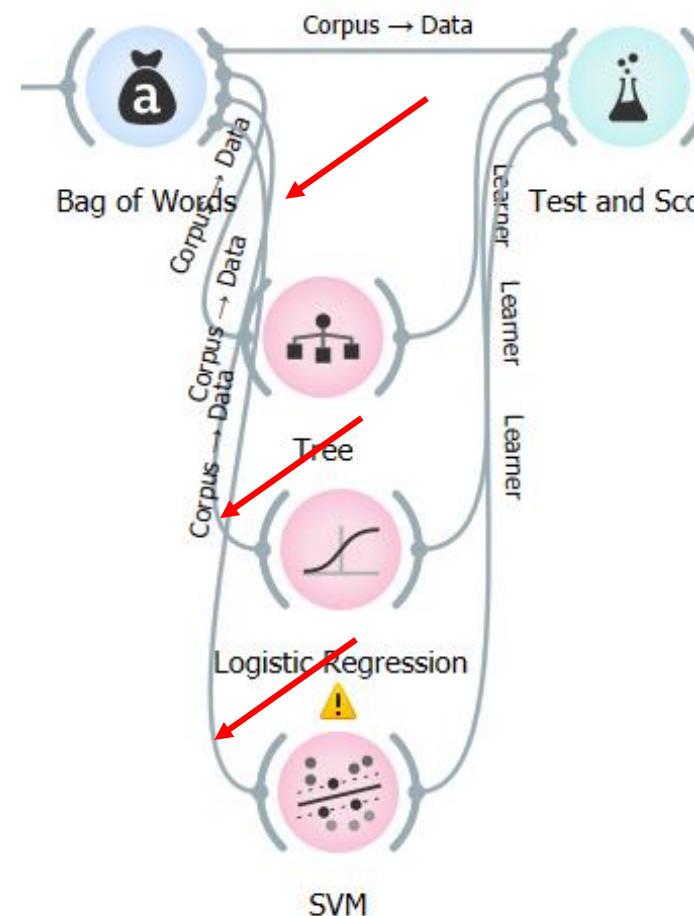


# Orange – Adicionar o widget Predictions

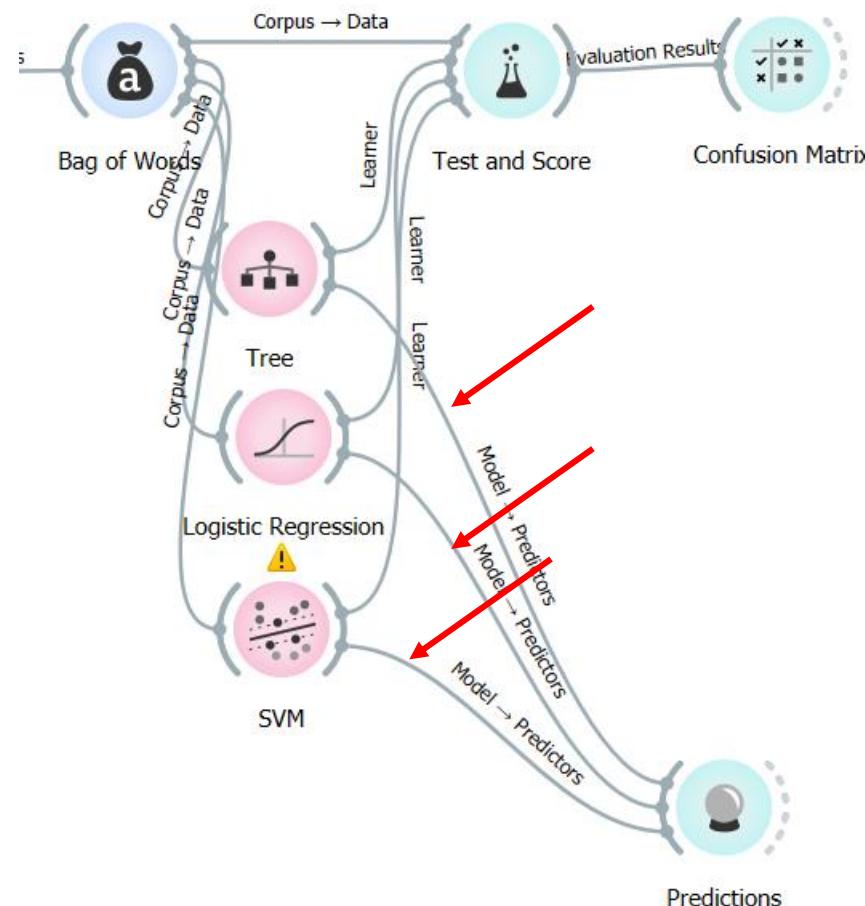


Vamos dividir esta última tarefa em 3 partes (devido as ligações visuais dos widgets).

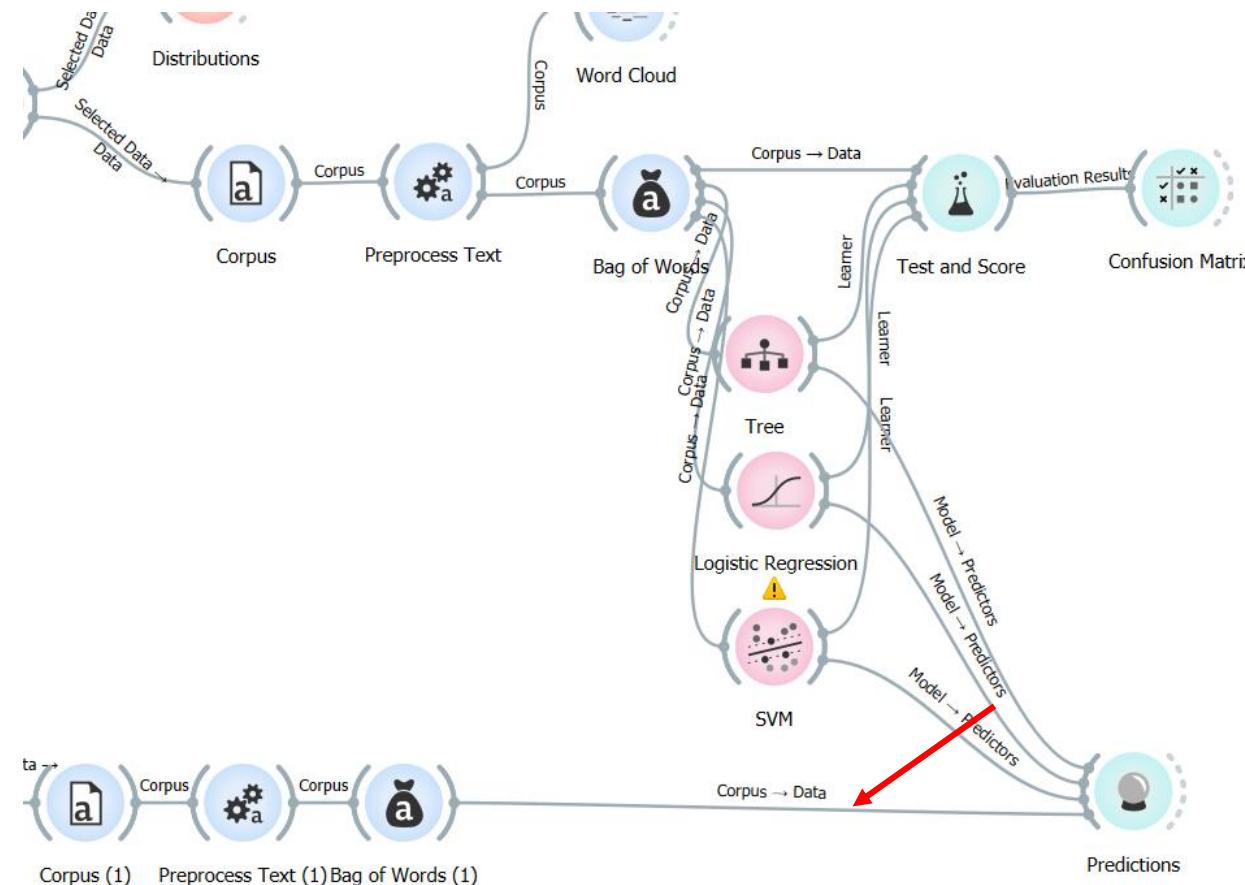
# Orange – Final Predictions – 1<sup>a</sup> Parte



# Orange – Final Predictions – 2<sup>a</sup> Parte



# Orange – Final Predictions – 3<sup>a</sup> Parte



# Orange – Visualizando Predictions

Predictions - Orange

Show probabilities for (None)  Show classification errors

	Tree	error	Logistic Regression	error	SVM	error	Tipo	Processo	(...)
3	Cível	0.045	Cível	0.128	Cível	0.046	Cível	Ação de indeni...	ação, causadas,...
4	Trabalhi...	0.000	Trabalhista	0.109	Trabalhi...	0.014	Trabalhista	Funcionário pro...	demissão, empr...
5	Cível	0.045	Cível	0.057	Cível	0.051	Cível	Disputa por dir...	autorais, cinem...
6	Trabalhi...	0.000	Trabalhista	0.098	Trabalhi...	0.014	Trabalhista	Funcionário ale...	alega, após, ass...
7	Cível	0.045	Cível	0.064	Cível	0.039	Cível	Julgamento de ...	cirurgia, emerg...
8	Cível	0.045	Cível	0.143	Cível	0.121	Cível	Ação de indeni...	acidente, ação, ...
9	Trabalhi...	0.000	Trabalhista	0.127	Trabalhi...	0.015	Trabalhista	Funcionário pro...	discriminação, ...
10	Cível	0.045	Cível	0.073	Cível	0.073	Cível	Disputa sobre p...	disputa, herdad...
11	Trabalhi...	0.000	Trabalhista	0.139	Trabalhi...	0.020	Trabalhista	Funcionário ale...	alega, após, co...
12	Trabalhi...	0.000	Trabalhista	0.159	Trabalhi...	0.030	Trabalhista	Funcionário pro...	assédio, funcio...
13	Cível	0.045	Cível	0.040	Cível	0.044	Cível	Ação de repar...	ação, barragem...
14	Cível	0.045	Cível	0.082	Cível	0.052	Cível	Ação de despej...	aluguel, ação, d...
15	Cível	0.970	Cível	0.946	Cível	0.950	Trabalhista	Empregador en...	empregador, e...
16	Criminal	0.000	Criminal	0.074	Criminal	0.029	Criminal	Investigação de...	dinheiro, imobil...
17	Cível	0.045	Cível	0.038	Cível	0.044	Cível	Ação de repar...	ação, causados,...
18	Criminal	0.000	Criminal	0.272	Criminal	0.108	Criminal	Investigação de...	empresa, impos...
19	Cível	0.985	Cível	0.859	Cível	0.893	Criminal	Julgamento de ...	ambiente, caso,...
20	Criminal	0.000	Criminal	0.051	Criminal	0.023	Criminal	Acusação de fra...	acusação, fraud...
21	Trabalhi...	0.000	Trabalhista	0.180	Trabalhi...	0.022	Trabalhista	Empregado ale...	alega, após, de...
22	Cível	0.045	Cível	0.133	Cível	0.118	Cível	Processo de ali...	alienação, cônj...

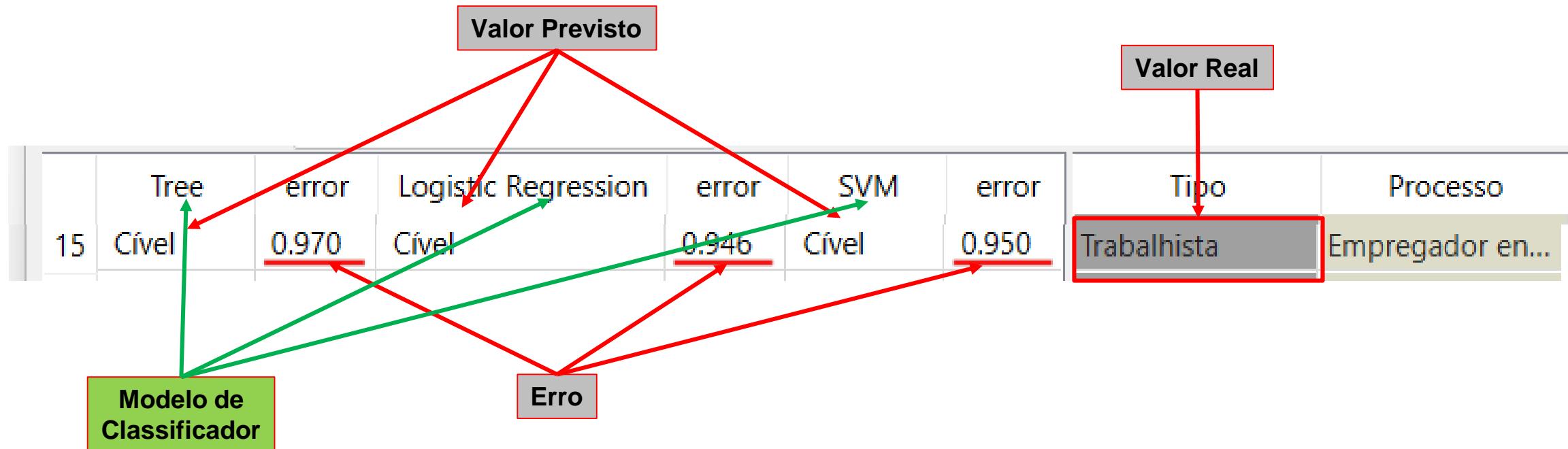
Show performance scores  Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.947	0.942	0.942	0.948	0.942	0.906
Logistic Regression	0.996	0.942	0.942	0.948	0.942	0.906
SVM	0.996	0.942	0.942	0.948	0.942	0.906

☰ ? 📄 | ↗ 52 | 🖼 52 | 3x52



# Orange – Visualizando Predictions





Obrigado!

