

# Sistema de Alerta Processual e Avaliação de Unidades Judiciais

## Tecnologias e ferramentas utilizadas

A linguagem utilizada para a parte do sistema que realiza o processamento dos dados, clusterização, avaliação e alertas foi o Python. Tudo que compõe esse sistema, das bibliotecas de processamento de dados ao modelo de cluster, estão identificadas abaixo:

```
import numpy as np
import pandas as pd
import seaborn as sns
import networkx as nx
import matplotlib.pyplot as plt
import json
import datetime
import pickle
import operator

from google.colab import drive
from unidecode import unidecode
from matplotlib.cm import ScalarMappable
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.image as mpimg
from sklearn.preprocessing import normalize
from sklearn.cluster import Birch
```

## Informações sobre os clusters

O principal ponto do desafio, e também alicerce do sistema desenvolvido, é a clusterização. Para isso, optamos por utilizar o modelo [Birch](#), na implementação do scikit-learn.

O motivo da escolha é a flexibilidade do modelo Birch. Sendo implementado a partir de uma estrutura de árvore, o modelo tem suporte ao *online training* (treino em etapas, utilizando a função `birch.fit(X)`) e é ideal para situações onde o

conjunto de dados é grande e há o risco de outliers. Tais informações partem do próprio [scikit-learn](#).

## Métricas do Modelo

O modelo BIRCH faz parte da classe de algoritmos não-supervisionados, de modo que não é uma tarefa trivial (e nem mesmo comum) estabelecer métricas de avaliação do modelo. O fluxo padrão em algoritmos de clustering é criar os clustering e realizar análises, o que foi feito.

Como dito, o BIRCH foi escolhido por ser extremamente eficiente ao lidar com grandes quantidades de dados, como dito, inclusive, no seu [artigo de origem](#). Essa eficiência é alcançada ao se construir um sumário das informações e distribuições dos dados e executando a clusterização sobre esses dados.

O funcionamento do Birch se torna ainda mais interessante ao fornecer a opção de especificar um valor de *threshold*, que determina o raio dos subclusters que o algoritmo monta, de modo que há uma certa flexibilidade na avaliação dos clusters.

## Áreas do Conhecimento e Técnicas Envolvidas

A solução envolve, principalmente, conhecimentos e técnicas relacionadas ao tratamento de dados. Como a base de dados do DATAJUD tinha bastante problemas de padronização, padronizar e tratar os dados foram tarefas que consumiram bastante tempo.

Além disso, envolveu-se conhecimentos em inteligência artificial, para o uso e aplicação do modelo de aprendizado não-supervisionado.

Por fim, há conhecimentos em webdevelopment, necessários para a criação e configuração da interface que exhibe os dados gerados pela solução.

## Formato dos dados

### Processos

Os processos prontos, cujos dataframes estão no repositório sob o formato .pkl, possuem as seguintes colunas:

- *millisInsercao*: milissegundo de inserção (data-hora em formato UNIX)
- *grau*
- *siglaTribunal*
- *valorCausa*
- *nivelSigilo*
- *numero*

- *codigoOrgaoJulgador*
- *procEl*: Indicador se o processo tramita em meio eletrônico ou físico
- *classeProcessual*
- *tamanhoProcesso*
- *dscSistema*: Descritor do sistema onde o processo tramita
- *competencia*: identificador da competência a que pertence o processo, ou da competência a que ele se destina caso se trate de processo inicial
- *codigoLocalidade*
- *totalAssuntos*
- *dataAjuizamento*
- *movimentoNacional*: Lista de movimentos Nacionais, cada elemento é composto por [codigoMovimento,dataHora]
- *movimentoLocal*: Lista de movimentos Locais, cada elemento é composto por [codigoMovimento,codigoPaiNacional,dataHora]
- *assuntoNacional*: Lista de assuntos Nacionais, cada elemento é composto por [codigo,dataHora]
- *assuntoLocal*: Lista de assuntos Locais, cada elemento é composto por [codigo,codigoPaiNacional,dataHora]
- *tempo\_proc\_dias*: tempo do processo, em dias, desde a dataAjuizamento ao último movimento
- *data\_ultimo\_mov*: data do último movimento do processo.
- *total\_movs\_loc*: total de movimentos locais
- *total\_movs\_nac*: total de movimentos nacionais
- *total\_movs*: total de movimentos (soma de locais e nacionais)
- *1\_mov\_nac\_freq*: movimento nacional mais frequente
- *2\_mov\_nac\_freq*: segundo movimento nacional mais frequente
- *desc\_mov1*: descrição do movimento nacional mais frequente
- *desc\_mov2*: descrição do segundo movimento nacional mais frequente
- *tempoMedioEntreMovimentos*: tempo médio entre os movimentos, calculado dividindo o tempo total do processo pela quantidade total de movimentos
- *cluster\_processo*: id do cluster do processo
- *pontuacao\_processo*: pontuação do processo, com cálculo explicado abaixo

Imagem do formato dos dados (algumas colunas foram retiradas para melhorar a visualização):

grau	siglaTribunal	numero	totalAssuntos	dataJuizamento	tempo_proc_dias	data_ultimo_mov	total_movs_loc	total_movs_nac	total_movs	1_mov_nac_freq	2_mov_nac_freq	desc_mov1	desc_mov2	tempoMedioEntreMovimentos
0	1	TRT17_00717005820055170003	0	18446681937950751616	10916	20190826145155	1.0	0.0	1.0	0	0	0	0	10916.000000
1	1	TRT17_00401001420085170003	0	18446681937950751616	10396	20190823153928	2.0	0.0	2.0	0	0	0	0	5198.000000
2	1	TRT17_00186008620085170003	0	18446681937950751616	10396	20190823175437	1.0	0.0	1.0	0	0	0	0	10396.000000
3	1	TRT17_02041017819925170002	0	18446742238930339616	7218	20130118163227	50.0	0.0	50.0	0	0	0	0	144.360000
4	1	TRT17_01317019519945170002	0	18446742238930339616	8314	20160330151926	2.0	0.0	2.0	0	0	0	0	4157.000000
5	1	TRT17_01863014219895170002	0	623646000000	5383	20090817130708	15.0	0.0	15.0	0	0	0	0	358.866667
6	1	TRT17_00352017219915170001	0	668314800000	6346	20130308182207	3.0	0.0	3.0	0	0	0	0	2115.333333
7	1	TRT17_00353012719915170001	0	668314800000	5782	20120815124131	2.0	0.0	2.0	0	0	0	0	2891.000000
8	1	TRT17_01102011019935170001	0	734842800000	4661	20110222094627	2.0	0.0	2.0	0	0	0	0	2330.500000
9	1	TRT17_01708019119935170002	0	740977200000	6958	20180502154156	3.0	0.0	3.0	0	0	0	0	2319.333333
10	1	TRT17_01533011319935170131	0	785383200000	-1238	19960425001700	17.0	0.0	17.0	0	0	0	0	-72.823529
11	1	TRT17_00059015419955170121	0	814413600000	5431	20101206135101	4.0	0.0	4.0	0	0	0	0	1357.750000
12	1	TRT17_00760013419955170121	0	845431200000	8652	20191010160650	7.0	0.0	7.0	0	0	0	0	1236.000000
13	1	TRT17_01073008419985170101	0	895633200000	7824	20170822145126	45.0	0.0	45.0	0	0	0	0	173.866667
14	1	TRT17_00611014619955170121	0	924058800000	2752	20040618114315	95.0	0.0	95.0	0	0	0	0	28.968421
15	1	TRT17_01778009420125170131	0	937278000000	8267	20191111154940	1.0	6.0	7.0	60	85	Expedição de documento	Expedição de documento	1181.000000
16	1	TRT17_03277003020015170005	0	1006740000000	7455	20170831131454	101.0	0.0	101.0	0	0	0	0	73.811881
17	1	TRT17_008470003200935170131	0	1052449200000	8233	20200114105546	2.0	24.0	26.0	11010	51	Mero expediente	Mero expediente	316.653846
18	1	TRT17_00834006920045170131	0	1085367600000	7945	20191127100600	376.0	0.0	376.0	0	0	0	0	21.130319
19	1	TRT17_00825004920055170132	0	1096340400000	7860	20191114144903	1.0	0.0	1.0	0	0	0	0	7860.000000

totalAssuntos	dataJuizamento	tempo_proc_dias	data_ultimo_mov	total_movs_loc	total_movs_nac	total_movs	1_mov_nac_freq	2_mov_nac_freq	desc_mov1	desc_mov2	tempoMedioEntreMovimentos	cluster_processo	pontuacao_processo
0	18446681937950751616	10916	20190826145155	1.0	0.0	1.0	0	0	0	0	10916.000000	3	368.050921
0	18446681937950751616	10396	20190823153928	2.0	0.0	2.0	0	0	0	0	5198.000000	3	257.317193
0	18446681937950751616	10396	20190823175437	1.0	0.0	1.0	0	0	0	0	10396.000000	3	349.589383
0	18446742238930339616	7218	20130118163227	50.0	0.0	50.0	0	0	0	0	144.360000	3	111.193406
0	18446742238930339616	8314	20160330151926	2.0	0.0	2.0	0	0	0	0	4157.000000	3	201.879324
0	623646000000	5383	20090817130708	15.0	0.0	15.0	0	0	0	0	358.866667	3	82.427253
0	668314800000	6346	20130308182207	3.0	0.0	3.0	0	0	0	0	2115.333333	3	130.701809
0	668314800000	5782	20120815124131	2.0	0.0	2.0	0	0	0	0	2891.000000	3	134.459205
0	734842800000	4661	20110222094627	2.0	0.0	2.0	0	0	0	0	2330.500000	3	104.610093
0	740977200000	6958	20180502154156	3.0	0.0	3.0	0	0	0	0	2319.333333	3	145.187016
0	785383200000	-1238	19960425001700	17.0	0.0	17.0	0	0	0	0	-72.823529	0	53.146057
0	814413600000	5431	20101206135101	4.0	0.0	4.0	0	0	0	0	1357.750000	3	101.010880
0	845431200000	8652	20191010160650	7.0	0.0	7.0	0	0	0	0	1236.000000	3	156.027253
0	895633200000	7824	20170822145126	45.0	0.0	45.0	0	0	0	0	173.866667	3	122.474590
0	924058800000	2752	20040618114315	95.0	0.0	95.0	0	0	0	0	28.968421	3	29.866929
0	937278000000	8267	20191111154940	1.0	6.0	7.0	60	85	Expedição de documento	Expedição de documento	1181.000000	3	148.216602
0	1006740000000	7455	20170831131454	101.0	0.0	101.0	0	0	0	0	73.811881	3	114.148174
0	1052449200000	8233	20200114105546	2.0	24.0	26.0	11010	51	Mero expediente	Mero expediente	316.653846	2	274.924323
0	1085367600000	7945	20191127100600	376.0	0.0	376.0	0	0	0	0	21.130319	3	121.911223
0	1096340400000	7860	20191114144903	1.0	0.0	1.0	0	0	0	0	7860.000000	3	259.553880

Este formato foi escolhido como padrão para lidar com as tabelas massivas de dados, devido a se tratar de uma [notação de objetos de Python](#), assim como o JSON é para o JavaScript. A vantagem é que, enquanto arquivos JSON são lidos em Python como strings ou dicionários, ocupando muito espaço tanto em disco quanto em memória RAM, o Pickle é uma tradução binária de objetos.

Desta forma, nosso fluxo de trabalho envolveu ler os arquivos JSON, converter as colunas para objetos de tamanhos menores (uint, em sua maioria), armazená-los em uma estrutura de dados de consulta bem mais rápida (Pandas DataFrames), e persistir essas estruturas em Pickles, que posteriormente foram ainda [compactados](#) (em .zip, individualmente) para possibilitar o upload no GitHub para a submissão.

Isso nos permitiu sair de 673 MB dos arquivos originais zipados, para 13.5 GB combinados dos arquivos JSON, para um tamanho final, após todas as análises e operações nos dados, de aproximadamente 1.3 GB.

Isso seria simplesmente impossível caso tivéssemos permanecido com os arquivos JSON, além de dificultar muito a análise, como pode ser visto nas células que tratam o TRT 5, cujos arquivos eram grandes demais para caber em um só DataFrame em memória, e ele permaneceu dividido em 2 por grande parte do desafio, sendo combinado novamente apenas no início da etapa de Clustering,

quando ele já havia sido bem reduzido através dos métodos e operações mencionados acima.

## Pontuação de Processos

Após realização de processamento nos dados (limpeza e formatação) e também engenharia de recursos (criação de novas colunas), a clusterização de processos foi feita. Para isso, levou-se em conta as seguintes colunas:

- *grau*
- *totalAssuntos*
- *tempo\_proc\_dias*
- *total\_movs\_loc*
- *total\_movs\_nac*
- *total\_movs*
- *1\_mov\_nac\_freq*
- *2\_mov\_nac\_freq*
- *tempoMedioEntreMovimentos*

A pontuação do processo leva em conta:

- Pontuação do cluster: os processos foram divididos em 4 clusters. Após essa divisão feita pelo modelo Birch, nós analisamos os clusters e montamos estatísticas sobre seu comportamento, que estão disponíveis no repositório.
- Desvio de tempo entre movimentos: compara o tempo entre movimentos do processo com o tempo médio entre movimentos daquele cluster
- Desvio de tempo total: compara o tempo total do processo com o tempo médio total de processos daquele cluster.

A imagem abaixo mostra esse arquivo, contendo estatísticas do TRT1, que são usadas para computar as pontuações. Duas observações:

- O BIRCH foi treinado com dados de **todos** os TRTs e o TST, de modo que os clusters são similares em todos os DataFrames, por mais que os dados variem um pouco;
- As médias negativas nos campos relacionados a dias são um problema que enfrentamos devido às datas em formato UNIX e ISO que se encontram nos dados. Pela lógica do nosso código, há uma inversão de sinal quando o formato é ISO. Isso **não** influencia o código, os clusters ou as conclusões, pois:
  - a) o BIRCH separou de modo perfeito todas as médias negativas e
  - b) o cálculo de média é feito *por cluster*.

	media_dias	media_movs	media_movs_loc	media_movs_nac	media_assuntos	media_dias_entre_movs	mov_mais_comum1	mov_mais_comum2	qtd_processos	pontuacao	id_cluster
0	-573.835146	118.715313	117.981146	0.734167	0.053904	-33.221806	0	0	16548	26	0
1	1065.318070	35.035400	0.965657	34.069743	3.814019	107.431683	51	11010	5678	14	1
2	353.923010	36.929066	0.927912	36.001153	3.929642	45.728723	1051	51	3468	19	2
3	3951.602104	39.797733	10.732043	29.065690	2.317428	320.307114	60	0	15969	11	3

## Unidades Judiciais

Já as unidades judiciais, também presentes no repositório sob o formato .csv, possuem as seguintes colunas:

- *ranking*: ranking da unidade com base na pontuação.
- *SEQ\_ORGAO*
- *NOMEDAVARA*
- *SEQ\_ORGAO\_PAI*
- *GRAU\_ORGAO*
- *SEQ\_CIDADE*
- *DSC\_CIDADE*
- *SIG\_UF*
- *COD\_IBGE*
- *DSC\_TIP\_ORGAO*
- *TIP\_ESFERA\_JUSTICA*
- *INT\_ORDEM\_ORGAO*
- *LATITUDE*
- *LONGITUDE*
- *CAPITAL*: informação se a unidade está localizada em uma capital ou não.
- *qtd\_processos*: quantidade de processos atualmente naquela unidade
- *tempo\_medio\_proc*: tempo, em dias, médio dos processos
- *tempo\_entre\_movs*: tempo médio, em dias, entre movimentações dos processos
- *cluster\_processual\_mais\_comum*: id do cluster mais comum dos processos daquela unidade (calculado por moda)
- *cluster\_unidade*: id do cluster da unidade
- *procs\_em\_alerta*: número de processos em alerta
- *pontuacao\_unidade*: pontuação da unidade

	SEQ_ORGAO	NOME/ABR	GRAU_ORGAO	DSC_CIDADE	SIG_UF	CAPITAL	qtd_processos	tempo_medio_proc	tempo_entre_movs	cluster_processual_mais_comum	cluster_unidade	procs_em_alerta	ranking
0	13196	Conselho Superior da Justiça do Trabalho	CONST	NaN	NaN	0	852636	917.650771	109.630173	0	1	0	0
1	13197	Tribunal Regional do Trabalho da 1ª Região	TRIBT	rio de janeiro	RJ	1	24603	981.677181	171.197779	0	1	0	0
2	13198	Tribunal Regional do Trabalho da 2ª Região	TRIBT	sao paulo	SP	1	45040	889.716040	263.043108	0	1	0	0
3	13199	Tribunal Regional do Trabalho da 3ª Região	TRIBT	belo horizonte	MG	1	51922	1814.812192	108.810920	2	1	0	0
4	13200	Tribunal Regional do Trabalho da 4ª Região	TRIBT	porto alegre	RS	1	46315	1032.923983	146.501466	0	1	0	0
5	13201	Tribunal Regional do Trabalho da 5ª Região	TRIBT	salvador	BA	1	41014	1528.124971	39.740457	3	1	0	0
6	13202	Tribunal Regional do Trabalho da 6ª Região	TRIBT	recife	PE	1	38020	1196.201450	79.852748	1	1	0	0
7	13203	Tribunal Regional do Trabalho da 7ª Região	TRIBT	fortaleza	CE	1	39596	983.067452	200.133463	0	1	0	0
8	13204	Tribunal Regional do Trabalho da 8ª Região	TRIBT	belem	PA	1	42440	1313.456089	98.423746	1	1	0	0
9	13205	Tribunal Regional do Trabalho da 9ª Região	TRIBT	curitiba	PR	1	40707	823.074717	135.886299	3	1	0	0
10	13206	Tribunal Regional do Trabalho da 10ª Região	TRIBT	brasilia	DF	1	41167	289.735016	26.480121	1	1	0	0
11	13207	Tribunal Regional do Trabalho da 11ª Região	TRIBT	manaus	AM	1	34424	381.597940	76.777191	1	1	0	0
12	13208	Tribunal Regional do Trabalho da 12ª Região	TRIBT	florianopolis	SC	1	42487	1205.527680	52.689954	3	1	0	0
13	13209	Tribunal Regional do Trabalho da 13ª Região	TRIBT	joao pessoa	PB	1	30338	1018.570038	292.550414	3	1	0	0
14	13210	Tribunal Regional do Trabalho da 14ª Região	TRIBT	porto velho	RO	1	27994	516.444009	40.829240	1	1	0	0

## Pontuação das Unidades

O cluster de unidades foi feito após o cluter de processos, sendo que o modelo foi alimentado com os seguintes dados:

- *CAPITAL*
- *qtd\_processos*
- *tempo\_entre\_movs*
- *cluster\_processual\_mais\_comum*

A pontuação das unidades:

- Leva em conta a pontuação do cluster mais comum dos processos;
- Leva em conta a média de dias da unidade em relação à média de dias do cluster;
- Leva em conta a média de dias entre movimentos da unidade em relação à média de dias do cluster;
- Leva em conta a quantidade de processos na unidade.

A imagem do arquivo contendo os detalhes dos clusters.

	media_dias	media_dias_entre_movs	qtd_media_procs	pontuacao_cluster_proc	pontuacao	id_cluster
0	-948.314154	-147.343819	79.875000	25	48	0
1	1387.773044	99.756838	1555.184802	16	27	1
2	6.382353	0.053317	1.176471	21	40	2
3	2284.378721	484.818272	159.532775	8	15	3

Assim, a metodologia de avaliação é a mesma: os aspectos avaliados são os mesmos para todas as unidades e todos os processos. Entretanto, tem-se a ponderação de cluster: **os processos e unidades só são comparados com**

**similares.** Desse modo, as informações e os rankings são mais valiosos e refletem de maneira mais fidedigna a realidade.

## Licenças

Utilizamos apenas bibliotecas de licença aberta. Todas, sem exceção, permitem uso livre e, inclusive, comercial.

Não foram utilizados programas e ferramentas que possuem licenças pagas, assim como não foram utilizadas soluções de *cloud*.