0. Package Imports

Python packages import

In [ ]:

```
import os
import shutil
import time
import re
import numpy as np
import pandas as pd
import pickle as pk

from matplotlib import pyplot as plt
import seaborn as sns
from pprint import pprint

import random
from random import randint

import sklearn.datasets
import sklearn.metrics
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB, ComplementNB
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestClassifier

# import autosklearn.classification

# from imblearn.over_sampling import SMOTE
# from imblearn.over_sampling import RandomOverSampler

final_dir = 'Final/'
output_dir = 'Output/'

proc_nr_csv_out_fpath = output_dir+'Proc_Nr_2010a2020.csv'
proc_nr_csv_fpath = final_dir+'Proc_Nr_2010a2020.csv'
proc_data_csv_fpath = final_dir+'Proc_Julgados_Prot_2010a2020.csv'
docs_df_csv_fpath = final_dir+'Documentos_2010to2020.csv'
pchv_df_csv_fpath = final_dir+'Palavras_Chave.csv'
tab_he_csv_fpath = final_dir+'Tabela_HE.csv'

raw_df_pkl_fpath = final_dir+'raw_df.pkl'
docs_df_pkl_fpath = final_dir+'docs_df.pkl'
pchv_df_pkl_fpath = final_dir+'pchv_df.pkl'
res_df_pkl_fpath = final_dir+'res_df.pkl'
out_df_pkl_fpath = final_dir+'out_df.pkl'
high_df_pkl_fpath = final_dir+'high_df.pkl'
prev_df_pkl_fpath = final_dir+'prev_df.pkl'
admjud_df_pkl_fpath = final_dir+'admjud_df.pkl'
decl_df_pkl_fpath = final_dir+'decl_df.pkl'
conf_df_pkl_fpath = final_dir+'conf_df.pkl'
vol_df_pkl_fpath = final_dir+'vol_df.pkl'
prior_df_pkl_fpath = final_dir+'prior_df.pkl'
mcf_df_pkl_fpath = final_dir+'mcf_df.pkl'
jud_df_pkl_fpath = final_dir+'jud_df.pkl'
```

```
arr_df_pkl_fpath = final_dir+'arr_df.pkl'
rffp_df_pkl_fpath = final_dir+'rffp_df.pkl'
solid_df_pkl_fpath = final_dir+'solid_df.pkl'
geral_df_pkl_fpath = final_dir+'geral_df.pkl'
he_dics_pkl_fpath = final_dir+'he_dics.pkl'
ml_dics_pkl_fpath = final_dir+'ml_dics.pkl'
```

1. COLETA DE DADOS BRUTOS

1.1. Dados de Tramitação e HE - ReceitaData

Instrução: No HUE, rodar a query a seguir e baixar o resultado como arquivo CSV a ser renomeado para "Proc_Julgados_Prot_2010a2020.csv"

Filtragem --> protocolo entre 2010 e 2020 inclusive, atividades atual (sit=1) e anterior com nome 'Para Relatar' e equipe contém 'DRJ', 'he' registrado > 0

```
   sit_ant=1 AND ativ_pauta = 'Para Relatar' AND equipe_pauta LIKE '%DRJ%' AND sit_atual=1 AND
ativ_distr = 'Para Relatar' AND
    equipe_distr LIKE '%DRJ%' AND he_saido > 0 AND dt_protocolo > 20099999 AND dt_protocolo <
20210000
```

Dados --> número do processo, data do protocolo, data da distribuição ao relator, dia de início da sessão, dia de fim da sessão, horas estimadas

```
   proc_nr, dt_protocolo, dt_pauta_drj, sit_ant, ativ_pauta, equipe_pauta, dt_distr_drj, sit_atual,
ativ_distr, equipe_distr,
   nr_epro_reun_rnj_frr_dia, nr_epro_reun_rnj_irr_dia, he_float
```

QUERY -->

```
SELECT DISTINCT tab_proc.cd_num_epro_estq_processo AS proc_nr,
tab_proc.dt_dia_dtpr_eppr AS dt_protocolo
, tab_sjul_pauta.nr_epro_sjul_dia AS dt_pauta_drj,
tab_sjul_pauta.nr_epro_sjul_sit_anterior AS sit_ant
, tab_ativ_tipo_pauta.nm_epro_ativ_atividade AS ativ_pauta
, tab_eqip_equipe_pauta.nm_epro_eqip_equipe AS equipe_pauta
, tab_sjul_distr.nr_epro_sjul_dia AS dt_distr_drj, tab_sjul_distr.nr_epro_sjul_sit_atual
AS sit_atual
, tab_ativ_tipo_distr.nm_epro_ativ_atividade AS ativ_distr
, tab_eqip_equipe_distr.nm_epro_eqip_equipe AS equipe_distr
, tab_reun.nr_epro_reun_rnj_irr_dia AS sess_ini, tab_reun.nr_epro_reun_rnj_frr_dia
AS sess_fim
, tab_he.qt_epro_efra_hr_estim_proc AS he_float, tab_he.qt_epro_efra_proc_saido AS
he_saido
FROM wd_epro_estq_processo as tab_proc
LEFT JOIN wf_epro_sjul AS tab_sjul_pauta ON
tab_sjul_pauta.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo
LEFT JOIN wf_epro_ativ AS tab_ativ_pauta ON tab_ativ_pauta.nr_epro_ativ =
tab_sjul_pauta.nr_epro_sjul_ativ
LEFT JOIN wd_epro_ativ_atividade AS tab_ativ_tipo_pauta ON
tab_ativ_tipo_pauta.nr_epro_ativ_atividade = tab_ativ_pauta.nr_epro_ativ_atividade
LEFT JOIN wf_epro_eqip as tab_eqip_pauta ON tab_eqip_pauta.nr_epro_eqip =
tab_sjul_pauta.nr_epro_sjul_eqip
LEFT JOIN wd_epro_eqip_equipe AS tab_eqip_equipe_pauta ON
tab_eqip_equipe_pauta.nr_epro_eqip_equipe = tab_eqip_pauta.nr_epro_eqip_equipe
LEFT JOIN wf_epro_sjul AS tab_sjul_distr ON
tab_sjul_distr.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo
LEFT JOIN wf_epro_ativ AS tab_ativ_distr ON
tab_ativ_distr.nr_epro_ativ=tab_sjul_distr.nr_epro_sjul_ativ
LEFT JOIN wd_epro_ativ_atividade AS tab_ativ_tipo_distr ON
tab_ativ_tipo_distr.nr_epro_ativ_atividade=tab_ativ_distr.nr_epro_ativ_atividade
LEFT JOIN wf_epro_eqip as tab_eqip_distr ON
tab_eqip_distr.nr_epro_eqip=tab_sjul_distr.nr_epro_sjul_eqip
LEFT JOIN wd_epro_eqip_equipe AS tab_eqip_equipe_distr ON
tab_eqip_equipe_distr.nr_epro_eqip_equipe=tab_eqip_distr.nr_epro_eqip_equipe
LEFT JOIN wf_epro_reun AS tab_reun ON
tab_reun.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo AND
```

In [ ]:

```
# importar dados dos processos das principais tabelas do e-Processo no ReceitaData
trmt_dtypes = {
    'proc_nr':str, 'dt_protocolo':str, 'dt_pauta_drj':str, 'sit_ant':str, 'ativ_pauta':
str, 'equipe_pauta':str, 'dt_distr_drj':str,
    'sit_atual':str, 'ativ_distr':str, 'equipe_distr':str, 'sess_ini':str, 'sess_fim':s
tr, 'he_float':np.float64, 'he_saido':str,
}
raw_df = pd.read_csv(proc_data_csv_fpath, sep=',', usecols=list(trmt_dtypes.keys()), en
coding='utf-8', dtype=trmt_dtypes)
```

In [ ]:

```
# excluir duplicidades de proc_nr, mantendo sempre os com datas de distribuição, pauta
 e sessão mais atuais
proc_nr_list = list(set(raw_df['proc_nr']))
dupl_idx_list = list(raw_df[raw_df.duplicated(subset=['proc_nr'])].index)
unique_idx_list = list(set(raw_df.index) - set(dupl_idx_list))
new_trmt_df = pd.DataFrame(index=unique_idx_list, columns=raw_df.columns, data=None)
new_trmt_df['proc_nr'] = proc_nr_list
new_trmt_df.set_index('proc_nr', drop=True, inplace=True)
for idx in raw_df.index:
    row = raw_df.loc[idx]
    proc_nr = row['proc_nr']
    new_row = new_trmt_df.loc[proc_nr]
    if pd.isnull(new_row['dt_protocolo']):
        new_trmt_df.at[proc_nr] = row
    else:
        dt_distr_drj = row['dt_distr_drj']
        dt_pauta_drj = row['dt_pauta_drj']
        sess_ini = row['sess_ini']
        new_dt_distr_drj = new_row['dt_distr_drj']
        new_dt_pauta_drj = new_row['dt_pauta_drj']
        new_sess_ini = new_row['sess_ini']
        if not ( pd.isnull(dt_distr_drj) or pd.isnull(dt_pauta_drj) or pd.isnull(sess_i
ni) ):
            if (new_dt_distr_drj < dt_distr_drj) or (new_dt_pauta_drj < dt_pauta_drj) o
r (new_sess_ini < sess_ini):
                new_trmt_df.at[proc_nr] = row
raw_df = new_trmt_df.astype(trmt_dtypes.pop('proc_nr'))
del new_trmt_df

# salvar relação dos números dos processos objeto da análise em um arquivo texto (CSV)
 para input no ContÁgil na extração dos dados de documentos
proc_nr_list = [str(proc_nr) for proc_nr in raw_df.index]
proc_nr_str = '\n'.join(proc_nr_list)
with open(proc_nr_csv_fpath, 'w') as f:
    f.write(proc_nr_str)
with open(proc_nr_csv_out_fpath, 'w') as f:
    f.write(proc_nr_str)

# descrever dataframe dos processos
raw_df.describe(include='all')
```

Out[ ]:

	dt_protocolo	dt_pauta_drj	sit_ant	ativ_pauta	equipe_pauta	dt_distr_drj	sit_atual
count	50178	50178	50178	50178	50178	50178	50178
unique	2612	705	1	1	140	989	1
top	20121122.0	20190520.0	1.0	Para Relatar	(inativo) SP-DRJ-RPO / 03ª Turma de Julgamento	20190510.0	1.0
freq	425	3267	50178	50178	7475	3243	50178

1.2. Palavras-Chaves dos Documentos - ReceitaData

Instrução: No HUE, rodar a query a seguir e baixar o resultado como arquivo CSV a ser renomeado para "Documentos_2010to2020.csv"

Filtragem --> protocolo entre 2010 e 2020 inclusive, atividades atual (sit=1) e anterior com nome 'Para Relatar' e equipe contém 'DRJ', 'he' registrado > 0

sit_ant=1 AND ativ_pauta = 'Para Relatar' AND equipe_pauta LIKE '%DRJ%' AND sit_atual=1 AND ativ_distr = 'Para Relatar' AND equipe_distr LIKE '%DRJ%' AND
he_saido > 0 AND dt_protocolo > 20099999 AND dt_protocolo < 20210000

Dados --> número do processo, data da ciência, tipo do documento, data da atualização, data da inclusão, número das páginas inicial e final, situação e tamanho

proc_nr, tab_doc.nr_epro_docs_pch_ciencia_dia, tab_doc.nr_epro_docs_tp, tab_doc.dt_atualizacao, tab_doc.dt_inclusao,
tab_doc.dd_epro_docs_pch_num_pg_fim, tab_doc.dd_epro_docs_pch_num_pg_ini, tab_doc.nr_epro_docs_sit, tab_doc.vl_epro_docs_tam_doc

QUERY -->

```
SELECT DISTINCT tab_proc.cd_num_epro_estq_processo AS proc_nr,
tab_doc.nr_epro_docs_pch_ciencia_dia,tab_doc.nr_epro_docs_tp,
tab_doc.dt_atualizacao,
   tab_doc.dt_inclusao,tab_doc.dd_epro_docs_pch_num_pg_fim,
tab_doc.dd_epro_docs_pch_num_pg_ini, tab_doc.nr_epro_docs_sit,
tab_doc.vl_epro_docs_tam_doc
FROM wd_epro_estq_processo AS tab_proc
LEFT JOIN wf_epro_docs AS tab_doc ON
tab_doc.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo
LEFT JOIN wf_epro_sjul AS tab_sjul_pauta ON
tab_sjul_pauta.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo
LEFT JOIN wf_epro_ativ AS tab_ativ_pauta ON tab_ativ_pauta.nr_epro_ativ =
tab_sjul_pauta.nr_epro_sjul_ativ
LEFT JOIN wd_epro_ativ_atividade AS tab_ativ_tipo_pauta ON
tab_ativ_tipo_pauta.nr_epro_ativ_atividade = tab_ativ_pauta.nr_epro_ativ_atividade
LEFT JOIN wf_epro_eqip as tab_eqip_pauta ON tab_eqip_pauta.nr_epro_eqip =
tab_sjul_pauta.nr_epro_sjul_eqip
LEFT JOIN wd_epro_eqip_equipe AS tab_eqip_equipe_pauta ON
tab_eqip_equipe_pauta.nr_epro_eqip_equipe = tab_eqip_pauta.nr_epro_eqip_equipe
LEFT JOIN wf_epro_sjul AS tab_sjul_distr ON
tab_sjul_distr.dd_epro_estq_processo=tab_proc.cd_num_epro_estq_processo
LEFT JOIN wf_epro_ativ AS tab_ativ_distr ON tab_ativ_distr.nr_epro_ativ =
tab_sjul_distr.nr_epro_sjul_ativ
LEFT JOIN wd_epro_ativ_atividade AS tab_ativ_tipo_distr ON
tab_ativ_tipo_distr.nr_epro_ativ_atividade = tab_ativ_distr.nr_epro_ativ_atividade
LEFT JOIN wf_epro_eqip as tab_eqip_distr ON tab_eqip_distr.nr_epro_eqip =
tab_sjul_distr.nr_epro_sjul_eqip
LEFT JOIN wd_epro_eqip_equipe AS tab_eqip_equipe_distr ON
tab_eqip_equipe_distr.nr_epro_eqip_equipe = tab_eqip_distr.nr_epro_eqip_equipe
LEFT JOIN wf_epro_reun AS tab_reun ON tab_reun.dd_epro_estq_processo =
tab_proc.cd_num_epro_estq_processo AND
tab_reun.nr_epro_ativ=tab_sjul_pauta.nr_epro_sjul_ativ
LEFT JOIN wf_epro_frat as tab_he ON tab_he.dd_epro_estq_processo =
tab_proc.cd_num_epro_estq_processo AND
tab_he.nr_epro_ativ=tab_sjul_pauta.nr_epro_sjul_ativ
WHERE tab_sjul_pauta.nr_epro_sjul_sit_anterior = 1
AND tab_ativ_tipo_pauta.nm_epro_ativ_atividade = 'Para Relatar'
AND tab_eqip_equipe_pauta.nm_epro_eqip_equipe LIKE '%DRJ%'
AND tab_sjul_distr.nr_epro_sjul_sit_atual = 1
AND tab_ativ_tipo_distr.nm_epro_ativ_atividade = 'Para Relatar'
AND tab_eqip_equipe_distr.nm_epro_eqip_equipe LIKE '%DRJ%'
AND tab_he.qt_epro_efra_proc_saido > 0
AND tab_proc.dt_dia_dtpr_eppr > 20099999 AND tab_proc.dt_dia_dtpr_eppr <
20210000
LIMIT 10000
```

In [ ]:

```
# importar dados dos documentos, relacionados aos processos selecionados, da principal
 tabela de palavras-chave de documentos do e-Processo no ReceitaData
# docs_dtypes = {'proc_nr':str, 'nr_epro_docs_pch_ciencia_dia':str, 'nr_epro_docs_tp':s
tr, 'dt_atualizacao':str, 'dt_inclusao':str,\
#  'dd_epro_docs_pch_num_pg_fim':str, 'dd_epro_docs_pch_num_pg_ini':str, 'nr_epro_docs_
resp_anex_pf':str, 'nr_epro_docs_resp_anex_rh':str,\
#  'nr_epro_docs_sit':str, 'qt_epro_docs_doc':str, 'qt_epro_docs_pg_doc':str, 'vl_epro_
docs_tam_doc':str}
docs_dtypes = {'proc_nr':str, 'nr_epro_docs_pch_ciencia_dia':str, 'nr_epro_docs_tp':str
, 'dt_atualizacao':str, 'dt_inclusao':str,\
 'dd_epro_docs_pch_num_pg_fim':str, 'dd_epro_docs_pch_num_pg_ini':str,\
 'nr_epro_docs_sit':str, 'vl_epro_docs_tam_doc':str}
docs_df = pd.read_csv(docs_df_csv_fpath, sep=',', encoding='utf-8', usecols=list(docs_d
types.keys()), dtype=docs_dtypes)
docs_df.to_pickle(docs_df_pkl_fpath)

# descrever dataframe dos documentos
docs_df.describe(include='all')
```

Out[ ]:

	proc_nr	nr_epro_docs_pch_ciencia_dia	nr_epro_docs_tp	dt_atualizacao
count	1835447	1835447	1835447	1835447
unique	50178	3661	493	70329
top	16561720006201908	-8.0	226.0	2021-09-21 03:37:17.0
freq	6961	1685078	315892	4976

In [ ]:

```
# classifica docs como fisco ou contestacao conforme o código do tipo de documento
tipo_doc_dic = {
    'qtd_folhas_fisco': ['81.0', '156.0', '157.0', '178.0', '366.0', '367.0', '368.0',
'371.0', '372.0', '517.0', '584.0', '593.0', '595.0', '602.0', '609.0', '737.0', '1778
8.0'],\
    'qtd_folhas_contestacao': ['318.0', '319.0', '416.0', '549.0', '556.0', '3801.0',
'3811.0', '19791.0'],\
}
raw_df['qtd_folhas_fisco'] = 0.
raw_df['qtd_folhas_contestacao'] = 0.
raw_df['qtd_folhas_total'] = 0.
for doc_idx in docs_df.index:
    doc = docs_df.loc[doc_idx]
    proc_nr = doc['proc_nr']
    if doc['dd_epro_docs_pch_num_pg_fim']!='Não Informado' and doc['dd_epro_docs_pch_nu
m_pg_ini']!='Não Informado':
        doc_page_nr = int(doc['dd_epro_docs_pch_num_pg_fim']) - int(doc['dd_epro_docs_p
ch_num_pg_ini']) + 1
        if proc_nr in raw_df.index:
            accum_proc_page_nr = raw_df.at[proc_nr, 'qtd_folhas_total']
            raw_df.at[proc_nr, 'qtd_folhas_total'] = accum_proc_page_nr + doc_page_nr
            for k,v in tipo_doc_dic.items():
                for idx in v:
                    if doc['nr_epro_docs_tp']==idx:
                        accum_doc_page_nr = raw_df.at[proc_nr, k]
                        raw_df.at[proc_nr, k] = accum_doc_page_nr + doc_page_nr
raw_df.to_pickle(raw_df_pkl_fpath)

# descrever dataframe dos processos com informações dos documentos
raw_df.describe(include=np.number)
```

Out[ ]:

	qtd_folhas_fisco	qtd_folhas_contestacao	qtd_folhas_total
count	50178.000000	50178.000000	50178.000000
mean	12.366974	35.635318	527.628542
std	172.926976	163.134801	5535.888289
min	0.000000	0.000000	15.000000
25%	0.000000	3.000000	56.000000
50%	1.000000	12.000000	98.000000
75%	5.000000	30.000000	227.000000
max	15296.000000	14481.000000	391379.000000

In [ ]:

```
raw_df.describe(include=object)
```

Out[ ]:

	dt_protocolo	dt_pauta_drj	sit_ant	ativ_pauta	equipe_pauta	dt_distr_drj	sit_atual
count	50178	50178	50178	50178	50178	50178	50178
unique	2612	705	1	1	140	989	1
top	20121122.0	20190520.0	1.0	Para Relatar	(inativo) SP-DRJ-RPO / 03ª Turma de Julgamento	20190510.0	1.0
freq	425	3267	50178	50178	7475	3243	50178

1.3. Palavras-Chaves dos Processos - e-Processo

Instrução:
No ContÁgil, usando o plug-in do Farol, acessar o Extrator do e-Processo para consultar as palavras-chaves dos processos,
a partir de uma lista dos números dos processos e baixar o resultado como arquivo CSV a ser renomeado para "Documentos_2010to2020.csv"

Filtragem --> relação com os números dos processos

Dados --> todas as palavras-chaves dos processos existentes na consulta do Extrator (posteriormente restringidas via Pandas por código)

In [ ]:

```
# importar palavras-chaves dos processos selecionados a partir do arquivo extraído via
 ContÁgil/Farol/Extrator
pchv_incols_dtypes = {'Nº Processo':str,'Grupo Processo':str,'Tipo Processo':str,'Subti
po Processo':str,'Assunto COMPROT':str,'Controle Processual':str,\
       'Data do Protocolo':str,'Descrição Origem SIEF':str,'Indicador de Concessão de M
edida Cautelar Fiscal':str,'Indicador se Existe Nota de Processo':str,\
       'Indicador se Existe Processo de Acompanhamento Judicial':str,'Indicador se Exis
te Processo de Arrolamento':str,\
       'Indicador se Existe Processo de Representação para Fins Penais':str,'Infração':
str,'Número Lote Atual':str,'Processos Vinculados':str,\
       'Quantidade Volumes':str,'Tributo':str,'Prioridade do Processo':str,'Tipo Contri
buinte':str,'Porte Contribuinte':str,\
       'Idade Contribuinte':str,'Indicador Contribuinte Diferenciado':str,'Indicador Co
ntribuinte Especial':str,'Indicador de Optante pelo DTE':str,\
       'Indicador de Solicitante com Moléstia Grave':str,'Indicador se Existe Responsáv
el Solidário/Subsidiário':str,\
       'Indicador se Solicitada Prioridade Baseada no Estatuto do Idoso':str,'ACT - Áre
a de Concentração Temática':str,'ACT - Origem':str,'ACT - Tributo':str,\
       'ACT - Código':str,'ACT - Código do Tema':str,'ACT - Código Completo':str,'Data
 Sessão DRJ':str,'Equipe de Análise/Apreciação DRJ':str,\
       'Indicador de Julgamento em Lote':str,'Data Distribuição Última':str,'Valor do P
rocesso':str}
pchv_df = pd.read_csv(pchv_df_csv_fpath, usecols=list(pchv_incols_dtypes.keys()), sep=
';', encoding='iso-8859-1', dtype=pchv_incols_dtypes)
pchv_df.rename(columns={'Nº Processo': 'proc_nr'}, inplace=True)
pchv_df['Valor do Processo'] = pchv_df['Valor do Processo'].apply(lambda x: x.replace(
'.','').replace(',','.'))
pchv_df['Valor do Processo'] = pchv_df['Valor do Processo'].astype(float)
pchv_df['proc_nr'] = pchv_df['proc_nr'].apply(lambda x: x[:5]+x[6:12]+x[13:17]+x[18:20
])
pchv_df.set_index('proc_nr', drop=True, inplace=True)
pchv_df.to_pickle(pchv_df_pkl_fpath)
# descrever dataframe das palavras-chaves dos processos
pchv_df.describe(include='all')
```

Out[ ]:

	Grupo Processo	Tipo Processo	Subtipo Processo	Assunto COMPROT	Controle Processual	D Pro
count	50178	50178	50178	50178	16055	
unique	3	8	141	303	650	
top	PROCESSO TRIBUTÁRIO	LANÇAMENTO	AUTO DE INFRAÇÃO E/OU NOTIFICAÇÃO DE LANÇAMENT...	IMPUGNACAO (RECLAM/DEFESA) - RECURSO IRPF	02.13.001.0001 - CRETRI - CONTENCIOSO ADMINIST...	22/1
freq	49995	29466	15857	11049	1229	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	

11 rows × 38 columns

1.4. Importing Tabela de Apuração to dictionaries and cleaning them

Import Table from CSV file

In [ ]:

```
tab_ini_str = 'TABELAS DE QUESITOS E HORAS PARA APURAÇÃO DAS HORAS ESTIMADAS NECESSÁRIA
S AO\nJULGAMENTO DE PROCESSOS'
tab_end_str = 'DF GABINETE RFB'
with open(tab_he_csv_fpath, 'r', encoding='utf-8') as f:
    text_raw = f.read()
text_cut = text_raw[text_raw.find(tab_ini_str)+len(tab_ini_str)+1:text_raw.rfind(tab_en
d_str)-1]
pattern_list = [
    r'DF GABINETE RFB[\s\S]'\
    + r'Fl\. \d+[\s\S]'\
    + r'Verso em Branco - Documento nato-digital[\s\S]'\
    + r'Documento de \d+ página\(s\) assinado digitalmente\.[\s\S]{1,2}'\
    + r'Pode ser consultado no endereço https://cav\.receita\.fazenda\.gov\.br/eCAC/pub
lico/login\.aspx pelo[\s\S]'\
    + r'código de localização EP28\.0621\.22302\.UJ0E\.[\s\S]'\
    + r'Consulte a página de autenticação no final deste documento\.'\
    + r'\(Fl\. \d+ do Anexo IV da Portaria RFB nº 46, de 22 de junho de 2021\.\)[\s\S]'
,
    r'[\s\S]Aplica-se a Tabela \d\.\d, com a mesma estrutura \(quesitos/parâmetros/hora
s\), adaptando-se a[\s\S]numeração para \d\.\d onde consta \d\.\d\.'
]
text_str = text_cut
for pattern in pattern_list:
    text_str = re.sub(pattern, '', text_str, flags=re.DOTALL)
lines_raw = text_str.split('\n')

lines = []
for l_id in range(len(lines_raw)):
    l = lines_raw[l_id]
    if l[0] in ['T', 'Q', 'P']:
        lines.append(l)
    else:
        lines[-1] = lines[-1]+' '+l

he_dic, tabela, quesito, parametro = {}, '', '', ''
for l in lines:
    if l[0]=='T':
        tabela_pos = l.find('-')+2
        tabela = l[tabela_pos:]
        he_dic[tabela] = {}
    elif l[0]=='Q':
        if l[:94]!='Quantidade de folhas da resposta à diligência proposta pelas Turmas
das DRJ (incluindo provas)':
            first_b_pos = l.find(' ')
            second_b_rel_pos = l[first_b_pos+1:].find(' ')
            quesito = l[first_b_pos+second_b_rel_pos+2:]
            he_dic[tabela][quesito] = {}
    if l[0]=='P':
        first_b_pos = l.find(' ')
        second_b_rel_pos = l[first_b_pos+1:].find(' ')
        parametro_pos = first_b_pos+second_b_rel_pos+2
        he_pos = l.rfind(' ')+1
        parametro = l[parametro_pos:he_pos-1]
        he = l[he_pos:]
        he_dic[tabela][quesito][parametro] = he

drop_tab_list = ['IPI vinculados ao Comex', 'Imposto de Exportação (IE)', 'Contribuiçõe
s vinculadas ao Comex', 'Outros vinculados ao Comex', 'FINSOCIAL', 'COFINS', 'PASEP',
'CIDE']
```

```
for k in drop_tab_list:
    he_dic.pop(k, None)
rename_keys = {'IRPJ e tributos vinculados':'IRPJ','IPI, exceto vinculado ao Comex':'IP
I','Imposto de Importação (II)':'COMEX','PIS':'PIS/COFINS','CS - Contribuições Previden
ciárias':'CP'}
for k, v in rename_keys.items():
    he_dic[v] = he_dic.pop(k, None)
```

Tabela_x_Tributo Table

In [ ]:

```
tab_to_feature_dic = {
    '1_0': {'name': 'IRPF', 'tributos': ['IRPF'], 'patterns': ['IRPF'], 'tables_not':
[]},
    '2_0': {'name': 'IRPJ', 'tributos': ['IRPJ e tributos vinculados'], 'patterns': ['I
RPJ', 'MULDI'], 'tables_not':[]},
    '3_0': {'name': 'IRRF', 'tributos': ['IRRF'], 'patterns': ['IRRF'], 'tables_not':
[]},
    '4_0': {'name': 'IPI', 'tributos': ['IPI, exceto vinculado ao Comex'], 'patterns':
['IPI'], 'tables_not':['5_0']},
    '5_0': {'name': 'COMEX', 'tributos': ['Imposto de Importação (II)','IPI vinculados
 ao Comex', 'Imposto de Exportação (IE)', \
                    'Contribuições vinculadas ao Comex', 'Outros vinculados ao Comex'],
'patterns': ['II', 'IE'], 'tables_not':[]}, # ELABORAR patterns (no codigo???)!!!
    '6_1': {'name': 'PIS/COFINS', 'tributos': ['PIS', 'FINSOCIAL', 'COFINS', 'PASEP',
'CIDE'], 'patterns': ['PIS', 'FINSOCIAL', 'COFINS', 'PASEP', 'CIDE'], 'tables_not':[]},
    '6_4': {'name': 'CSLL', 'tributos': ['CSLL'], 'patterns': ['CSLL'], 'tables_not':
[]},
    '6_7': {'name': 'CP', 'tributos': ['CS - Contribuições Previdenciárias'], 'pattern
s': ['CP'], 'tables_not':[]},
    '7_0': {'name': 'SIMPLES', 'tributos': ['SIMPLES'], 'patterns': ['SIMPLES'], 'table
s_not':[]},
    '8_1': {'name': 'ITR', 'tributos': ['ITR'], 'patterns': ['ITR'], 'tables_not':[]},
    '8_2': {'name': 'IOF', 'tributos': ['IOF'], 'patterns': ['IOF'], 'tables_not':[]},
    '8_3': {'name': 'CPMF/IPMF', 'tributos': ['CPMF/IPMF'], 'patterns': ['CPMF', 'IPMF'
], 'tables_not':[]},
    '8_4': {'name': 'CPSS', 'tributos': ['CPSS'], 'patterns': ['CPSS'], 'tables_not':
[]},
    '9_0': {'name': 'OUTROS', 'tributos': ['OUTROS'], 'patterns': [], 'tables_not':['1_
0', '2_0', '3_0', '4_0', '5_0', '6_1', '6_4', '6_7', '7_0', '8_1', '8_2', '8_3', '8_4'
]},
}
```

Quesito_x_Feature Table

In [ ]:

```
attrib_dic = {
    'Aspectos gerais do processo horas':'aspectos_gerais',
    'Classificação fiscal - Quantidade de mercadorias a serem analisadas horas':'qtd_me
rcadorias',
    'Elementos de prova - DI/Adições/DDE/RE/DCR/Atos Concess./Laudos/Pareceres/Certific
ados de Origem/BL/Fatura (Soma) horas':'qtd_elementos_prova',
    'Quantidade de arquivos não-paginàveis horas':'qtd_arquivos_naopag',
    'Quantidade de autos de infração no e-processo horas':'qtd_ai',
    'Quantidade de autos de infração reflexos/ decorrentes horas':'qtd_ai_reflexos',
    'Quantidade de autos de infração reflexos/decorrentes horas':'qtd_ai_reflexos',
    'Quantidade de competências / Quantidade de períodos a serem apreciados - (em mese
s) (2) horas':'qtd_periodos',
    'Quantidade de estabelecimentos horas':'qtd_estabs',
    'Quantidade de folhas da impugnação / manifestação de inconformidade (até assinatur
a do responsável) horas':'qtd_folhas_contestacao',
    'Quantidade de folhas da impugnação/ manifestação de inconformidade (até assinatura
do responsável) (4) horas':'qtd_folhas_contestacao',
    'Quantidade de folhas da impugnação/ manifestação de inconformidade (até assinatura
do responsável) horas':'qtd_folhas_contestacao',
    'Quantidade de folhas da impugnação/manifestação de inconformidade (até assinatura
 do responsável) horas':'qtd_folhas_contestacao',
    'Quantidade de folhas das provas apresentadas com a impugnação/manifestação de inco
nformidade.':'qtd_folhas_prova_respdilig',
    'Quantidade de folhas do Despacho Decisório horas':'qtd_folhas_fisco',
    'Quantidade de folhas do Relatório fiscal horas':'qtd_folhas_fisco',
    'Quantidade de folhas do Termo de Verificação Fiscal/ Relatório Fiscal/ Informação
 Fiscal ou semelhantes (documento que deu suporte a AI/NL/DD) horas':'qtd_folhas_fisco'
,
    'Quantidade de folhas do e-processo horas':'qtd_folhas_total',
    'Quantidade de folhas do processo (3) horas':'qtd_folhas_total',
    'Quantidade de levantamentos (AIOP / NFLD) / Quantidade de temas a serem apreciados
- (matérias impugnadas) horas':'qtd_temas',
    'Quantidade de períodos a serem apreciados (em meses) horas':'qtd_periodos',
    'Quantidade de sujeitos passivos no caso de responsabilidade tributária que apresen
taram impugnação horas':'qtd_sujeitos',
    'Quantidade de temas a serem apreciados horas':'qtd_temas',
    'Quantidade de temas a serem apreciados ou tipos de produtos objeto de classificaçã
o fiscal / "Ex" tarifário horas':'qtd_temas',
    'Redução por apreciação em Lote/JAP? horas':'reducao_lote',
    'Redução por apreciação em Lote/JAP? pontos':'reducao_lote',
    'Temas/ Situações específicas horas':'tema_específico',
    'Temas/Matérias específicas 1 horas':'tema_específico',
    'Temas/Matérias específicas 2 horas':'tema_específico',
    'Temas/Matérias específicas 3 horas':'tema_específico',
    'Temas/Matérias específicas 4 horas':'tema_específico',
    'Temas/Situações específicas horas':'tema_específico',
    'Tipo de processo a ser apreciado Horas':'tipo_processo',
    'Tipo de processo a ser apreciado horas':'tipo_processo',
    'Tipos de processos horas':'tipo_processo',
    'Valor total a ser apreciado (Palavra-chave “Valor Processo” em “Sobre os Valores”,
do e-processo) horas':'valor_processo',
    'Valor total a ser apreciado (Palavra-chave “valor processo” em “sobre os valores”,
do e-processo) horas':'valor_processo',
    'folhas da resposta à diligência proposta pelas Turmas das DRJ (incluindo provas) h
oras':'qtd_folhas_diligencia',
    'folhas da resposta à diligência proposta pelas Turmas das DRJ (incluindo provas).
 horas':'qtd_folhas_diligencia',
}
old_he_dic = he_dic
```

```
he_dic = {}
for k,v in old_he_dic.items():
    he_dic[k] = {}
    for old_attrib in v.keys():
        new_attrib = attrib_dic[old_attrib]
        he_dic[k][new_attrib] = old_he_dic[k][old_attrib]

he_dic['SIMPLES']['qtd_folhas_contestacao']['de 51 a 100'] = '9:42'
he_dic['SIMPLES']['qtd_folhas_contestacao']['de 101 a 200'] = '9:42'
he_dic['SIMPLES']['qtd_folhas_contestacao']['de 201 a 400'] = '9:42'
he_dic['SIMPLES']['qtd_folhas_contestacao']['de 401 em diante'] = '9:42'
he_dic['ITR']['qtd_folhas_contestacao']['de 51 a 100'] = '8:00'
he_dic['ITR']['qtd_folhas_contestacao']['de 101 a 200'] = '8:00'
he_dic['ITR']['qtd_folhas_contestacao']['de 201 a 400'] = '8:00'
he_dic['ITR']['qtd_folhas_contestacao']['de 401 em diante'] = '8:00'
he_dic['IOF']['qtd_folhas_contestacao']['de 51 a 100'] = '4:42'
he_dic['IOF']['qtd_folhas_contestacao']['de 101 a 200'] = '4:42'
he_dic['IOF']['qtd_folhas_contestacao']['de 201 a 400'] = '4:42'
he_dic['IOF']['qtd_folhas_contestacao']['de 401 em diante'] = '4:42'
he_dic['CPMF/IPMF']['qtd_folhas_contestacao']['de 51 a 100'] = '5:12'
he_dic['CPMF/IPMF']['qtd_folhas_contestacao']['de 101 a 200'] = '5:12'
he_dic['CPMF/IPMF']['qtd_folhas_contestacao']['de 201 a 400'] = '5:12'
he_dic['CPMF/IPMF']['qtd_folhas_contestacao']['de 401 em diante'] = '5:12'
he_dic['CPSS']['qtd_folhas_contestacao']['de 51 a 100'] = '5:12'
he_dic['CPSS']['qtd_folhas_contestacao']['de 101 a 200'] = '5:12'
he_dic['CPSS']['qtd_folhas_contestacao']['de 201 a 400'] = '5:12'
he_dic['CPSS']['qtd_folhas_contestacao']['de 401 em diante'] = '5:12'
he_dic['OUTROS']['qtd_folhas_contestacao']['de 51 a 100'] = '4:00'
he_dic['OUTROS']['qtd_folhas_contestacao']['de 101 a 200'] = '4:00'
he_dic['OUTROS']['qtd_folhas_contestacao']['de 201 a 400'] = '4:00'
he_dic['OUTROS']['qtd_folhas_contestacao']['de 401 em diante'] = '4:00'
he_dic['CP']['qtd_folhas_contestacao']['2'] = '0:24'
he_dic['CP']['qtd_folhas_contestacao']['3'] = '0:24'
he_dic['CP']['qtd_folhas_contestacao']['4'] = '0:48'
he_dic['CP']['qtd_folhas_contestacao']['5'] = '0:48'
he_dic['CP']['qtd_folhas_contestacao']['6'] = '0:48'
he_dic['CP']['qtd_folhas_contestacao']['de 7 a 9'] = '0:48'
he_dic['CP']['qtd_folhas_contestacao']['de 10 a 12'] = '1:30'
he_dic['CP']['qtd_folhas_contestacao']['de 13 a 15'] = '1:30'
he_dic['CP']['qtd_folhas_contestacao']['de 16 a 19'] = '3:00'
he_dic['CP']['qtd_folhas_contestacao']['de 20 a 23'] = '3:00'
he_dic['CP']['qtd_folhas_contestacao']['de 24 a 28'] = '4:06'
he_dic['CP']['qtd_folhas_contestacao']['de 29 a 34'] = '4:06'
he_dic['CP']['qtd_folhas_contestacao']['de 35 a 41'] = '6:00'
he_dic['CP']['qtd_folhas_contestacao']['de 42 a 50'] = '7:30'
he_dic['CP']['qtd_folhas_contestacao']['de 51 a 100'] = '7:30'

he_dic['IRPF']['qtd_folhas_total']['de 1.001 a 3.000'] = '8:42'
he_dic['IRPF']['qtd_folhas_total']['de 3.001 a 5.000'] = '9:00'
he_dic['IRPF']['qtd_folhas_total']['de 5.001 a 6.000'] = '9:00'
he_dic['IRPF']['qtd_folhas_total']['de 6.001 a 10.000'] = '11:30'
he_dic['IRPF']['qtd_folhas_total']['de 10.001 em diante'] = '14:00'
he_dic['IRPJ']['qtd_folhas_total']['de 1.001 a 3.000'] = '9:00'
he_dic['IRPJ']['qtd_folhas_total']['de 3.001 a 5.000'] = '11:30'
he_dic['IRPJ']['qtd_folhas_total']['de 5.001 a 6.000'] = '11:30'
he_dic['IRPJ']['qtd_folhas_total']['de 6.001 a 10.000'] = '14:30'
he_dic['IRPJ']['qtd_folhas_total']['de 10.001 em diante'] = '18:00'
he_dic['IRRF']['qtd_folhas_total']['de 1.001 a 3.000'] = '8:42'
he_dic['IRRF']['qtd_folhas_total']['de 3.001 a 5.000'] = '10:54'
he_dic['IRRF']['qtd_folhas_total']['de 5.001 a 6.000'] = '10:54'
he_dic['IRRF']['qtd_folhas_total']['de 6.001 a 10.000'] = '13:06'
```

```
he_dic['IRRF']['qtd_folhas_total']['de 10.001 em diante'] = '17:24'
he_dic['IPI']['qtd_folhas_total']['de 1.001 a 3.000'] = '8:42'
he_dic['IPI']['qtd_folhas_total']['de 3.001 a 5.000'] = '10:54'
he_dic['IPI']['qtd_folhas_total']['de 5.001 a 6.000'] = '10:54'
he_dic['IPI']['qtd_folhas_total']['de 6.001 a 10.000'] = '13:06'
he_dic['IPI']['qtd_folhas_total']['de 10.001 em diante'] = '17:24'
he_dic['COMEX']['qtd_folhas_total']['de 1.001 a 3.000'] = '8:00'
he_dic['COMEX']['qtd_folhas_total']['de 3.001 a 5.000'] = '9:24'
he_dic['COMEX']['qtd_folhas_total']['de 5.001 a 6.000'] = '9:24'
he_dic['COMEX']['qtd_folhas_total']['de 6.001 a 10.000'] = '10:18'
he_dic['COMEX']['qtd_folhas_total']['de 10.001 em diante'] = '11:12'
he_dic['PIS/COFINS']['qtd_folhas_total']['de 1.001 a 3.000'] = '6:42'
he_dic['PIS/COFINS']['qtd_folhas_total']['de 3.001 a 5.000'] = '10:24'
he_dic['PIS/COFINS']['qtd_folhas_total']['de 5.001 a 6.000'] = '10:24'
he_dic['PIS/COFINS']['qtd_folhas_total']['de 6.001 a 10.000'] = '11:18'
he_dic['PIS/COFINS']['qtd_folhas_total']['de 10.001 em diante'] = '13:18'
he_dic['CSLL']['qtd_folhas_total']['de 1.001 a 3.000'] = '8:24'
he_dic['CSLL']['qtd_folhas_total']['de 3.001 a 5.000'] = '9:06'
he_dic['CSLL']['qtd_folhas_total']['de 5.001 a 6.000'] = '9:06'
he_dic['CSLL']['qtd_folhas_total']['de 6.001 a 10.000'] = '10:00'
he_dic['CSLL']['qtd_folhas_total']['de 10.001 em diante'] = '15:12'
he_dic['CP']['qtd_folhas_total']['de 1.001 a 3.000'] = '12:00'
he_dic['CP']['qtd_folhas_total']['de 3.001 a 5.000'] = '12:00'
he_dic['CP']['qtd_folhas_total']['de 5.001 a 6.000'] = '14:00'
he_dic['CP']['qtd_folhas_total']['de 6.001 a 10.000'] = '14:00'
he_dic['CP']['qtd_folhas_total']['de 10.001 em diante'] = '16:00'
he_dic['SIMPLES']['qtd_folhas_total']['de 1.001 a 3.000'] = '12:00'
he_dic['SIMPLES']['qtd_folhas_total']['de 3.001 a 5.000'] = '12:00'
he_dic['SIMPLES']['qtd_folhas_total']['de 5.001 a 6.000'] = '12:00'
he_dic['SIMPLES']['qtd_folhas_total']['de 6.001 a 10.000'] = '12:00'
he_dic['SIMPLES']['qtd_folhas_total']['de 10.001 em diante'] = '12:00'
he_dic['ITR']['qtd_folhas_total']['de 1.001 a 3.000'] = '6:18'
he_dic['ITR']['qtd_folhas_total']['de 3.001 a 5.000'] = '6:18'
he_dic['ITR']['qtd_folhas_total']['de 5.001 a 6.000'] = '6:18'
he_dic['ITR']['qtd_folhas_total']['de 6.001 a 10.000'] = '6:18'
he_dic['ITR']['qtd_folhas_total']['de 10.001 em diante'] = '6:18'
he_dic['IOF']['qtd_folhas_total']['de 1.001 a 3.000'] = '5:00'
he_dic['IOF']['qtd_folhas_total']['de 3.001 a 5.000'] = '5:00'
he_dic['IOF']['qtd_folhas_total']['de 5.001 a 6.000'] = '5:00'
he_dic['IOF']['qtd_folhas_total']['de 6.001 a 10.000'] = '5:00'
he_dic['IOF']['qtd_folhas_total']['de 10.001 em diante'] = '5:00'
he_dic['CPMF/IPMF']['qtd_folhas_total']['de 1.001 a 3.000'] = '4:24'
he_dic['CPMF/IPMF']['qtd_folhas_total']['de 3.001 a 5.000'] = '4:24'
he_dic['CPMF/IPMF']['qtd_folhas_total']['de 5.001 a 6.000'] = '4:24'
he_dic['CPMF/IPMF']['qtd_folhas_total']['de 6.001 a 10.000'] = '4:24'
he_dic['CPMF/IPMF']['qtd_folhas_total']['de 10.001 em diante'] = '4:24'
he_dic['CPSS']['qtd_folhas_total']['de 1.001 a 3.000'] = '4:24'
he_dic['CPSS']['qtd_folhas_total']['de 3.001 a 5.000'] = '4:24'
he_dic['CPSS']['qtd_folhas_total']['de 5.001 a 6.000'] = '4:24'
he_dic['CPSS']['qtd_folhas_total']['de 6.001 a 10.000'] = '4:24'
he_dic['CPSS']['qtd_folhas_total']['de 10.001 em diante'] = '4:24'
he_dic['OUTROS']['qtd_folhas_total']['de 1.001 a 3.000'] = '10:00'
he_dic['OUTROS']['qtd_folhas_total']['de 3.001 a 5.000'] = '10:00'
he_dic['OUTROS']['qtd_folhas_total']['de 5.001 a 6.000'] = '10:00'
he_dic['OUTROS']['qtd_folhas_total']['de 6.001 a 10.000'] = '10:00'
he_dic['OUTROS']['qtd_folhas_total']['de 10.001 em diante'] = '10:00'
```

Quesito_Parametro_x_Feature_Category Table

In [ ]:

```
attrib_to_features_dic = {
    'tributo': {
            'attrib_categ1': {'feature':'feat1', 'categ':list(range(1,6))},
    },
    'aspectos_gerais': {
            'Processos com lançamento de ano-calendário até 1996': {'feature':'feat1',
'categ':'categ1'},
            'Contestação de responsabilidade tributária pelo sujeito passivo': {'featur
e':'Solid', 'categ':'S'},
            'Processo com ação judicial não concomitante com a matéria objeto do lançam
ento': {'feature':'Jud', 'categ':'S'},
            'Processo com representação fiscal para fins penais': {'feature':'RFFP', 'c
ateg':'S'},
            'Contribuinte diferenciado ou de grande porte': {'feature':'CD', 'categ':
'S'},
            'Processo oriundo de delegacias especiais': {'feature':'feat1', 'categ':'ca
teg1'},
    },
    'qtd_ai': {
            '0': {'feature':'feat1', 'categ':'categ1'},
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            'de 3 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_ai_reflexos': {
            '0': {'feature':'feat1', 'categ':'categ1'},
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            '3': {'feature':'feat1', 'categ':'categ1'},
            '4': {'feature':'feat1', 'categ':'categ1'},
            '5': {'feature':'feat1', 'categ':'categ1'},
            'de 6 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_arquivos_naopag': {
            'de 1 a 5': {'feature':'feat1', 'categ':'categ1'},
            'de 6 a 10': {'feature':'feat1', 'categ':'categ1'},
            'de 11 a 15': {'feature':'feat1', 'categ':'categ1'},
            'de 16 a 20': {'feature':'feat1', 'categ':'categ1'},
            'de 21 a 25': {'feature':'feat1', 'categ':'categ1'},
            'de 26 a 30': {'feature':'feat1', 'categ':'categ1'},
            'de 31 a 35': {'feature':'feat1', 'categ':'categ1'},
            'de 36 a 40': {'feature':'feat1', 'categ':'categ1'},
            'de 41 a 45': {'feature':'feat1', 'categ':'categ1'},
            'de 46 a 50': {'feature':'feat1', 'categ':'categ1'},
            'de 51 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_elementos_prova': {
            'de 1 a 2': {'feature':'feat1', 'categ':'categ1'},
            'de 3 a 5': {'feature':'feat1', 'categ':'categ1'},
            'de 6 a 10': {'feature':'feat1', 'categ':'categ1'},
            'de 11 a 15': {'feature':'feat1', 'categ':'categ1'},
            'de 16 a 50': {'feature':'feat1', 'categ':'categ1'},
            'de 51 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_estabs': {
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            '3': {'feature':'feat1', 'categ':'categ1'},
            '4': {'feature':'feat1', 'categ':'categ1'},
```

```
            '5': {'feature':'feat1', 'categ':'categ1'},
            '6': {'feature':'feat1', 'categ':'categ1'},
            'de 7 a 16': {'feature':'feat1', 'categ':'categ1'},
            'de 17 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_folhas_contestacao': {
            '1': {'feature':'qtd_folhas_contestacao', 'categ':list(range(1,2))},
            '2': {'feature':'feat1', 'categ':list(range(2,3))},
            '3': {'feature':'feat1', 'categ':list(range(3,4))},
            '4': {'feature':'feat1', 'categ':list(range(4,5))},
            '5': {'feature':'feat1', 'categ':list(range(5,6))},
            '6': {'feature':'feat1', 'categ':list(range(6,7))},
            'de 7 a 9': {'feature':'feat1', 'categ':list(range(7,10))},
            'de 10 a 12': {'feature':'feat1', 'categ':list(range(10,13))},
            'de 13 a 15': {'feature':'feat1', 'categ':list(range(13,16))},
            'de 16 a 19': {'feature':'feat1', 'categ':list(range(16,20))},
            'de 20 a 23': {'feature':'feat1', 'categ':list(range(20,24))},
            'de 24 a 28': {'feature':'feat1', 'categ':list(range(24,29))},
            'de 29 a 34': {'feature':'feat1', 'categ':list(range(29,35))},
            'de 35 a 41': {'feature':'feat1', 'categ':list(range(35,42))},
            'de 42 a 50': {'feature':'feat1', 'categ':list(range(42,51))},
            'de 51 a 100': {'feature':'feat1', 'categ':list(range(51,101))},
            'de 101 a 200': {'feature':'feat1', 'categ':list(range(101,201))},
            'de 201 a 400': {'feature':'feat1', 'categ':list(range(201,401))},
            'de 401 em diante': {'feature':'feat1', 'categ':list(range(401,100000))},
    },
    'qtd_folhas_diligencia': {
            'de 1 a 5': {'feature':'feat1', 'categ':'categ1'},
            'de 6 a 10': {'feature':'feat1', 'categ':'categ1'},
            'de 11 a 25': {'feature':'feat1', 'categ':'categ1'},
            'de 26 a 50': {'feature':'feat1', 'categ':'categ1'},
            'de 51 a 100': {'feature':'feat1', 'categ':'categ1'},
            'de 101 a 150': {'feature':'feat1', 'categ':'categ1'},
            'de 151 a 200': {'feature':'feat1', 'categ':'categ1'},
            'de 201 a 250': {'feature':'feat1', 'categ':'categ1'},
            'de 251 a 300': {'feature':'feat1', 'categ':'categ1'},
            'de 301 a 350': {'feature':'feat1', 'categ':'categ1'},
            'de 351 a 400': {'feature':'feat1', 'categ':'categ1'},
            'de 401 a 450': {'feature':'feat1', 'categ':'categ1'},
            'de 451 a 500': {'feature':'feat1', 'categ':'categ1'},
            'de 501 a 1.000': {'feature':'feat1', 'categ':'categ1'},
            'de 1.001 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_folhas_fisco': {
            'de 1 a 5': {'feature':'qtd_folhas_fisco', 'categ':list(range(1,6))},
            'de 6 a 10': {'feature':'qtd_folhas_fisco', 'categ':list(range(6,11))},
            'de 11 a 15': {'feature':'qtd_folhas_fisco', 'categ':list(range(11,16))},
            'de 16 a 20': {'feature':'qtd_folhas_fisco', 'categ':list(range(16,21))},
            'de 21 a 30': {'feature':'qtd_folhas_fisco', 'categ':list(range(21,31))},
            'de 31 a 40': {'feature':'qtd_folhas_fisco', 'categ':list(range(31,41))},
            'de 41 a 50': {'feature':'qtd_folhas_fisco', 'categ':list(range(41,51))},
            'de 51 a 60': {'feature':'qtd_folhas_fisco', 'categ':list(range(51,61))},
            'de 61 em diante': {'feature':'qtd_folhas_fisco', 'categ':list(range(61,100
000))},
    },
    'qtd_folhas_prova_respdilig': {
            'de 1 a 5': {'feature':'feat1', 'categ':'categ1'},
            'de 6 a 10': {'feature':'feat1', 'categ':'categ1'},
            'de 11 a 25': {'feature':'feat1', 'categ':'categ1'},
            'de 26 a 50': {'feature':'feat1', 'categ':'categ1'},
            'de 51 a 100': {'feature':'feat1', 'categ':'categ1'},
```

```
            'de 101 a 150': {'feature':'feat1', 'categ':'categ1'},
            'de 151 a 200': {'feature':'feat1', 'categ':'categ1'},
            'de 201 a 250': {'feature':'feat1', 'categ':'categ1'},
            'de 251 a 300': {'feature':'feat1', 'categ':'categ1'},
            'de 301 a 350': {'feature':'feat1', 'categ':'categ1'},
            'de 351 a 400': {'feature':'feat1', 'categ':'categ1'},
            'de 401 a 450': {'feature':'feat1', 'categ':'categ1'},
            'de 451 a 500': {'feature':'feat1', 'categ':'categ1'},
            'de 501 a 1.000': {'feature':'feat1', 'categ':'categ1'},
            'de 1.001 em diante': {'feature':'feat1', 'categ':'categ1'},
            'de 1.001 a 5.000': {'feature':'feat1', 'categ':'categ1'},
            'de 5.001 a 10.000': {'feature':'feat1', 'categ':'categ1'},
            'de 10.001 a 20.000': {'feature':'feat1', 'categ':'categ1'},
            'de 20.001 em diante': {'feature':'feat1', 'categ':'categ1'},

    },
    'qtd_folhas_total': {
            'de 1 a 10': {'feature':'feat1', 'categ':list(range(1,11))},
            'de 11 a 20': {'feature':'feat1', 'categ':list(range(11,21))},
            'de 21 a 50': {'feature':'feat1', 'categ':list(range(21,51))},
            'de 51 a 100': {'feature':'feat1', 'categ':list(range(51,101))},
            'de 101 a 200': {'feature':'feat1', 'categ':list(range(101,201))},
            'de 201 a 300': {'feature':'feat1', 'categ':list(range(201,301))},
            'de 301 a 400': {'feature':'feat1', 'categ':list(range(301,401))},
            'de 401 a 500': {'feature':'feat1', 'categ':list(range(401,501))},
            'de 501 a 600': {'feature':'feat1', 'categ':list(range(501,601))},
            'de 601 a 700': {'feature':'feat1', 'categ':list(range(601,701))},
            'de 701 a 800': {'feature':'feat1', 'categ':list(range(701,801))},
            'de 801 a 900': {'feature':'feat1', 'categ':list(range(801,901))},
            'de 901 a 1.000': {'feature':'feat1', 'categ':list(range(901,1001))},
            'de 1.001 a 3.000': {'feature':'feat1', 'categ':list(range(1001,3001))},
            'de 3.001 a 5.000': {'feature':'feat1', 'categ':list(range(3001,5001))},
            'de 5.001 a 6.000': {'feature':'feat1', 'categ':list(range(5001,6001))},
            'de 6.001 a 10.000': {'feature':'feat1', 'categ':list(range(6001,10001))},
            'de 10.001 em diante': {'feature':'feat1', 'categ':list(range(10001,1000000
))},
    },
    'qtd_mercadorias': {
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            '3': {'feature':'feat1', 'categ':'categ1'},
            '4': {'feature':'feat1', 'categ':'categ1'},
            '5 ou mais': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_periodos': {
            'de 1 a 2': {'feature':'feat1', 'categ':'categ1'},
            'de 3 a 12': {'feature':'feat1', 'categ':'categ1'},
            'de 13 a 24': {'feature':'feat1', 'categ':'categ1'},
            'de 25 a 36': {'feature':'feat1', 'categ':'categ1'},
            'de 37 a 48': {'feature':'feat1', 'categ':'categ1'},
            'de 49 a 60': {'feature':'feat1', 'categ':'categ1'},
            'de 61 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_sujeitos': {
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            '3': {'feature':'feat1', 'categ':'categ1'},
            '4': {'feature':'feat1', 'categ':'categ1'},
            '5': {'feature':'feat1', 'categ':'categ1'},
            '6': {'feature':'feat1', 'categ':'categ1'},
            '7': {'feature':'feat1', 'categ':'categ1'},
```

```
            '8': {'feature':'feat1', 'categ':'categ1'},
            '9': {'feature':'feat1', 'categ':'categ1'},
            'de 10 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'qtd_temas': {
            '1': {'feature':'feat1', 'categ':'categ1'},
            '2': {'feature':'feat1', 'categ':'categ1'},
            '3': {'feature':'feat1', 'categ':'categ1'},
            '4': {'feature':'feat1', 'categ':'categ1'},
            '5': {'feature':'feat1', 'categ':'categ1'},
            '6': {'feature':'feat1', 'categ':'categ1'},
            '7': {'feature':'feat1', 'categ':'categ1'},
            '8': {'feature':'feat1', 'categ':'categ1'},
            '9': {'feature':'feat1', 'categ':'categ1'},
            'de 10 em diante': {'feature':'feat1', 'categ':'categ1'},
    },
    'reducao_lote': {
            'Sim - Lote/JAP nível 1': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 2': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 3': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 4': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 5': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 6': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 7': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 8': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 9': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 10': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 11': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 12': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 13': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 14': {'feature':'feat1', 'categ':'categ1'},
            'Sim - Lote/JAP nível 15': {'feature':'feat1', 'categ':'categ1'},
            'Não (retirado do Lote/JAP)': {'feature':'Lote', 'categ':'N'},
    },
    'tema_específico': {
            'Acréscimo patrimonial a descoberto': {'feature':'feat1', 'categ':'categ1'
},
            'AI decorrente de apuração não cumulativa': {'feature':'feat1', 'categ':'ca
teg1'},
            'Aplicação de multa qualificada/agravada': {'feature':'feat1', 'categ':'cat
eg1'},
            'Arbitramento / aferição indireta / cessão de mão de obra': {'feature':'fea
t1', 'categ':'categ1'},
            'Atividade rural': {'feature':'feat1', 'categ':'categ1'},
            'Cancelamento de isenção': {'feature':'feat1', 'categ':'categ1'},
            'Caracterização de segurado': {'feature':'feat1', 'categ':'categ1'},
            'Classificação fiscal (1 produto/insumo)': {'feature':'feat1', 'categ':'cat
eg1'},
            'Classificação fiscal (2 a 4 produtos)': {'feature':'feat1', 'categ':'categ
1'},
            'Classificação fiscal (5 ou mais produtos)': {'feature':'feat1', 'categ':'c
ateg1'},
            'Compensação de prejuízos fiscais ou bases negativas da CSLL (exceto trava
 dos 30%)': {'feature':'feat1', 'categ':'categ1'},
            'Construção civil - pessoa física': {'feature':'feat1', 'categ':'categ1'},
            'Construção civil - pessoa jurídica': {'feature':'feat1', 'categ':'categ1'
},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 1.0.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 2.0.12.02)': {'feature':'Solid', 'categ':'S'},
```

```
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 3.0.12.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 4.0.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 6.1.12.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 6.7.12.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 7.0.12.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 8.1.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 8.2.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 8.3.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 8.4.11.01)': {'feature':'Solid', 'categ':'S'},
            'Contestação de responsabilidade tributária pelo sujeito passivo (marcar pa
râmetro 9.0.11.01)': {'feature':'Solid', 'categ':'S'},
            'Conversão da pena de perdimento em multa, decorrente de outras hipóteses d
e dano ao Erário (incisos I a IV do art. 23 do Decreto-Lei nº 1.455, de 1976)': {'featu
re':'feat1', 'categ':'categ1'},
            'Conversão da pena de perdimento em multa, decorrente de ocultação ou inter
posição fraudulenta de terceiros (inciso V do art. 23 do Decreto-Lei nº 1.455, de 197
6)':{'feature':'feat1', 'categ':'categ1'},
            'Correção monetária de balanço': {'feature':'feat1', 'categ':'categ1'},
            'Dedução indevida de despesas de livro caixa': {'feature':'feat1', 'categ':
'categ1'},
            'Depósitos bancários de origem não comprovada': {'feature':'feat1', 'categ'
:'categ1'},
            'Depósito bancário de origem não comprovada': {'feature':'feat1', 'categ':
'categ1'},
            'Desconsideração de ato ou negócio jurídico': {'feature':'feat1', 'categ':
'categ1'},
            'Descaracterização de negócio jurídico': {'feature':'feat1', 'categ':'categ
1'},
            'Escrituração reconstituída ou Existência de reconstituição da escrita fisc
al no auto de infração': {'feature':'feat1', 'categ':'categ1'},
            'Ganho de capital': {'feature':'feat1', 'categ':'categ1'},
            'Ganhos líquidos em renda variável': {'feature':'feat1', 'categ':'categ1'},
            'Glosa de despesas/custos': {'feature':'feat1', 'categ':'categ1'},
            'Grau de utilização': {'feature':'feat1', 'categ':'categ1'},
            'Grupo econômico / solidariedade': {'feature':'feat1', 'categ':'categ1'},
            'Instituição financeira': {'feature':'feat1', 'categ':'categ1'},
            'Instituições financeiras': {'feature':'feat1', 'categ':'categ1'},
            'Juros sobre capital próprio': {'feature':'feat1', 'categ':'categ1'},
            'Lucro arbitrado': {'feature':'feat1', 'categ':'categ1'},
            'Lucro da exploração': {'feature':'feat1', 'categ':'categ1'},
            'Lucro inflacionário': {'feature':'feat1', 'categ':'categ1'},
            'Lucros e dividendos distribuídos aos sócios': {'feature':'feat1', 'categ':
'categ1'},
            'Lucros e rendimentos provenientes do exterior ou decorrentes de reavaliaçã
o de operações no exterior': {'feature':'feat1', 'categ':'categ1'},
            'Multa aplicada na hipótese de consumo de mercadoria de procedência estrang
eira introduzida clandestina ou irregularmente no país': {'feature':'feat1', 'categ':'c
ateg1'},
            'Multa de ofício com cobertura de crédito': {'feature':'feat1', 'categ':'ca
teg1'},
            'Multa por "cessão de nome" (art. 33 da Lei nº 11.488, de 2007)': {'featur
e':'feat1', 'categ':'categ1'},
```

```
            'Multa qualificada / agravada': {'feature':'feat1', 'categ':'categ1'},
            'Multa qualificada/agravada': {'feature':'feat1', 'categ':'categ1'},
            'Multas do controle administrativo das importações': {'feature':'feat1', 'c
ateg':'categ1'},
            'Multas exigidas de forma isolada (inciso II do caput do art. 44 da Lei nº
 9.430, de 1996) e/ou por simples descumprimento de prazos': {'feature':'feat1', 'cate
g':'categ1'},
            'Multas exigidas na importação, exportação, internação (ZFM), trânsito, bag
agem, embaraço, no controle aduaneiro e outras': {'feature':'feat1', 'categ':'categ1'},
            'Novas regras da contabilidade': {'feature':'feat1', 'categ':'categ1'},
            'Omissão de rendimentos recebidos no exterior e/ou compensação indevida de
 imposto pago no exterior': {'feature':'feat1', 'categ':'categ1'},
            'Operações no mercado financeiro e de capital (swap, hedge, futuro, fundo d
e investimento, etc.)': {'feature':'feat1', 'categ':'categ1'},
            'Perdas no recebimento de créditos – Seção III (arts. 9º a 14) do Capítulo
 I da Lei nº 9.430, de 1996': {'feature':'feat1', 'categ':'categ1'},
            'PER/Dcomp com saldo credor de IPI': {'feature':'feat1', 'categ':'categ1'},
            'PER/Dcomp de Saldo Negativo': {'feature':'feat1', 'categ':'categ1'},
            'PER/Dcomp decorrentes da apuração não cumulativa': {'feature':'feat1', 'ca
teg':'categ1'},
            'Planejamento tributário (incluindo ágio em suas diversas expressões)': {'f
eature':'feat1', 'categ':'categ1'},
            'Preços de transferências e demais transações com partes relacionadas que i
nfluenciem a apuração do lucro real': {'feature':'feat1', 'categ':'categ1'},
            'Procedimentos para desconsideração de atos ou negócios jurídicos, para fin
s tributários': {'feature':'feat1', 'categ':'categ1'},
            'Processos com lançamento de ano-calendário até 1996 (marcar parâmetro 2.0.
12.01)': {'feature':'feat1', 'categ':'categ1'},
            'Processos com lançamento de ano-calendário até 1996 (marcar parâmetro 3.0.
12.02)': {'feature':'feat1', 'categ':'categ1'},
            'Processos com lançamento de ano-calendário até 1996 (marcar parâmetro 6.4.
12.01)': {'feature':'feat1', 'categ':'categ1'},
            'Que contenha manifestação de inconformidade em Pedidos de Compensação / Re
stituição / Reembolso': {'feature':'feat1', 'categ':'categ1'},
            'Regime Tributário de Transição (RTT)': {'feature':'feat1', 'categ':'categ
1'},
            'Rendimento recebido em ação judicial e/ ou compensação de fonte em ação ju
dicial': {'feature':'feat1', 'categ':'categ1'},
            'Retorno de diligência': {'feature':'feat1', 'categ':'categ1'},
            'Risco ocupacional': {'feature':'feat1', 'categ':'categ1'},
            'RRA': {'feature':'feat1', 'categ':'categ1'},
            'Rural / agroindústria / cooperativa': {'feature':'feat1', 'categ':'categ1'
},
            'Subvenções para custeio ou para investimento': {'feature':'feat1', 'categ'
:'categ1'},
            'Suspensão da isenção ou imunidade': {'feature':'feat1', 'categ':'categ1'},
            'Valor da terra nua (com laudo)': {'feature':'feat1', 'categ':'categ1'},
            '1.40.nnn.1166 - crédito presumido indevido': {'feature':'feat1', 'categ':
'categ1'},
            '1.40.nnn.1707- venda sem emissão de nota fiscal apurada em decorrência de
 auditoria de produção': {'feature':'feat1', 'categ':'categ1'},
    },
    'valor_processo': {
            'de R$ 0,01 a R$ 1.000,00': {'feature':'proc_val', 'categ':'categ1'},
            'de R$ 1.000,01 a R$ 5.000,00': {'feature':'proc_val', 'categ':'categ1'},
            'de R$ 5.000,01 a R$ 10.000,00': {'feature':'proc_val', 'categ':'categ1'},
            'de R$ 10.000,01 a R$ 50.000,00': {'feature':'proc_val', 'categ':'categ1'},
            'de R$ 50.000,01 a R$ 100.000,00': {'feature':'proc_val', 'categ':'categ1'
},
            'de R$ 100.000,01 a R$ 500.000,00': {'feature':'proc_val', 'categ':'categ1'
},
```

```
            'de R$ 500.000,01 a R$ 1.000.000,00': {'feature':'proc_val', 'categ':'categ
1'},
            'de R$ 1.000.000,01 a R$ 5.000.000,00': {'feature':'proc_val', 'categ':'cat
eg1'},
            'de R$ 5.000.000,01 a R$ 10.000.000,00': {'feature':'proc_val', 'categ':'ca
teg1'},
            'de R$ 10.000.000,01 a R$ 50.000.000,00': {'feature':'proc_val', 'categ':'c
ateg1'},
            'de R$ 50.000.000,01 a R$ 100.000.000,00': {'feature':'proc_val', 'categ':
'categ1'},
            'de R$ 100.000.000,01 a R$ 500.000.000,00': {'feature':'proc_val', 'categ':
'categ1'},
            'de R$ 500.000.000,01 a R$ 1.000.000.000,00': {'feature':'proc_val', 'cate
g':'categ1'},
            'de R$ 1.000.000.000,01 a R$ 10.000.000.000,00': {'feature':'proc_val', 'ca
teg':'categ1'},
            'de R$ 10.000.000.000,01 em diante': {'feature':'proc_val', 'categ':'categ
1'},
    },
    'tipo_processo': {
            'AIOA - auto de infração para aplicação de penalidade por descumprimento de
obrigação acessória': {'feature':'desc_sief', 'categ':'Ação Fiscal'},
            'AIOP - auto de infração lavrado em ação fiscal externa para exigência de o
brigação principal': {'feature':'feat1', 'categ':'categ1'},
            'AIEOA - auto de infração eletrônico p/ aplicação de penalidade por descump
rimento de obrig. acessória': {'feature':'feat1', 'categ':'categ1'},
            'AIEOP- auto de infração eletrônico para exigência de obrigação principal':
{'feature':'feat1', 'categ':'categ1'},
            'NEOP - notificação eletrônica de obrigação principal': {'feature':'feat1',
'categ':'categ1'},
            'NEOA - notificação eletrônica de obrigação acessória': {'feature':'feat1',
'categ':'categ1'},
            'DCOMP - declaração de compensação decorrente de tratamento eletrônico': {
'feature':'feat1', 'categ':'categ1'},
            'DCOMP - declaração de compensação decorrente de tratamento manual': {'feat
ure':'feat1', 'categ':'categ1'},
            'PER - pedido de restituição/ressarcimento decorrente de tratamento eletrôn
ico': {'feature':'feat1', 'categ':'categ1'},
            'PER - pedido de restituição/ressarcimento decorrente de tratamento manual'
: {'feature':'feat1', 'categ':'categ1'},
            'Pedido de reconhecimento de benefícios fiscais, isenção e imunidade': {'fe
ature':'feat1', 'categ':'categ1'},
            'OUTROS - outros tipos de processo': {'feature':'feat1', 'categ':'categ1'},
    },
}
```

Saves Dics as pickle files

In [ ]:

```
pk_dic = {
    'attrib_dic': attrib_dic,
    'he_dic': he_dic,
    'tab_to_feature_dic': tab_to_feature_dic,
    'attrib_to_features_dic': attrib_to_features_dic,
}
with open(he_dics_pkl_fpath, 'wb') as f:
    pk.dump(pk_dic, f)
```

2. DATABASES JOINING AND DUPLICATED PROCESS TREATMENT

Joining Dataframes based on common index 'proc_nr'

In [ ]:

```
# Concatenação dos dataframes com dados dos processos e palavras-chaves
raw_df = pd.concat([raw_df, pchv_df], axis=1, join="inner")
raw_df.to_pickle(raw_df_pkl_fpath)
raw_df.describe(include=object)
```

Out[ ]:

	dt_protocolo	dt_pauta_drj	sit_ant	ativ_pauta	equipe_pauta	dt_distr_drj	sit_atual
count	50178	50178	50178	50178	50178	50178	50178
unique	2612	705	1	1	140	989	1
top	20121122.0	20190520.0	1.0	Para Relatar	(inativo) SP-DRJ-RPO / 03ª Turma de Julgamento	20190510.0	1.0
freq	425	3267	50178	50178	7475	3243	50178

4 rows × 50 columns

In [ ]:

```
raw_df.describe(include=np.number)
```

Out[ ]:

	qtd_folhas_fisco	qtd_folhas_contestacao	qtd_folhas_total	Valor do Processo
count	50178.000000	50178.000000	50178.000000	5.017800e+04
mean	12.366974	35.635318	527.628542	2.689093e+06
std	172.926976	163.134801	5535.888289	7.535788e+07
min	0.000000	0.000000	15.000000	0.000000e+00
25%	0.000000	3.000000	56.000000	0.000000e+00
50%	1.000000	12.000000	98.000000	2.611915e+03
75%	5.000000	30.000000	227.000000	2.618988e+04
max	15296.000000	14481.000000	391379.000000	9.339505e+09