Revision of Dataframe

Dataframe from the Previous model

In [1]:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
df = pd.read_csv('df.csv')
print("df.shape",df.shape)
item_count = df["target"].value_counts()
print("Number of each species in dataframe:\n",item_count)
```

```
df.shape (3607, 1505)
Number of each species in dataframe:
 A_luchuensis              318
A_trinidadensis           235
A_indologenus             232
A_welwitschiae            223
A_sclerotiicarbonarius    220
A_homomorphus             192
A_ibericus                160
A_japonicus               141
A_saccharolyticus         140
A_niger                   140
A_vadensis                137
A_costaricaensis          130
A_heteromorphus           130
A_carbonarius             120
A_brasiliensis            120
A_aculeatinus             120
A_sclerotioniger          119
A_ellipticus              110
A_eucalypticola           110
A_aculeatus               100
A_floridensis              90
A_tubingensis              90
A_neoniger                 80
A_brunneoviolaceus         80
A_uvarum                   70
Name: target, dtype: int64
```

Create New Dataframe

In [2]:

```
"""
Creating individual dataframe of 'A_costaricaensis' , 'A_neoniger' , 'A_tubingensis',
"""
df_class6 = df[df['target'] == 'A_costaricaensis']
df_class16 = df[df['target'] == 'A_neoniger']
df_class22 = df[df['target'] == 'A_tubingensis']
df_class17 = df[df['target'] == 'A_niger']
df_class25 = df[df['target'] == 'A_welwitschiae']
```

```
"""
Select 80 data randomly from each dataframe of 'A_costaricaensis' , 'A_neoniger' and '
Select 140 data randomly from each dataframe of 'A_niger' and 'A_welwitschiae'
(the number 80,140 based on the minimum of data in each group)
"""
df_class6_rd = df_class6.sample(n = 80)
df_class16_rd = df_class16.sample(n = 80)
df_class22_rd = df_class22.sample(n = 80)
df_class17_rd = df_class17.sample(n = 140)
df_class25_rd = df_class25.sample(n = 140)


"""
Selecting all data from the dataframe in which ‘target’ is not
'A_costaricaensis','A_neoniger','A_tubingensis', 'A_niger' and 'A_welwitschiae'
"""
target_cut = ['A_costaricaensis','A_neoniger','A_tubingensis','A_niger','A_welwitschia
df_cut = df.loc[~df['target'].isin(target_cut)]
print("df after cut 5 species out.shape",df_cut.shape)


"""
Concatenate dataframe (samples randomly) of
'A_costaricaensis','A_neoniger','A_tubingensis','A_niger','A_welwitschiae'
"""
list_5 = [df_class6_rd,df_class16_rd,df_class22_rd,df_class17_rd,df_class25_rd]
df_com5 = pd.concat(list_5, axis=0, ignore_index=True)
print("\nShape of combined 5 species dataframe :",df_com5.shape)
item_counts_5 = df_com5["target"].value_counts()
print("Number of each species :\n",item_counts_5)


"""
Renamed A_costaricaensis and A_neoniger in combined dataframe as A_tubingensis
and renamed A_welwitschiae as A_niger

since A_costaricaensis and A_neoniger are synnonyms of A_tubingensis
and A_welwitschiae is a synnonyms of A_niger
(Bian et al. 2022)
"""
df_com5_rename = df_com5.replace({'A_costaricaensis':'A_tubingensis',
                                  'A_neoniger':'A_tubingensis',
                                  'A_welwitschiae':'A_niger'})
print("\nShape of renamed dataframe :",df_com5_rename.shape)
print

item_counts = df_com5_rename["target"].value_counts()
print("Number of each species :\n",item_counts)


"""
Combine dataframe of cut dataframe and renamed dataframe
"""
list_2 = [df_cut,df_com5_rename]
df_all = pd.concat(list_2, axis=0, ignore_index=True)
df_all.to_csv('dataframe.csv', index=False)  # Save to csv file
print("\nSave file as: dataframe.csv")
print("Shape of dataframe :",df_all.shape)

print("Number of each species in dataframe:\n",df_all["target"].value_counts())
```

```
df after cut 5 species out.shape (2944, 1505)

Shape of combined 5 species dataframe : (520, 1505)
Number of each species :
 A_niger             140
A_welwitschiae      140
A_costaricaensis     80
A_neoniger           80
A_tubingensis        80
Name: target, dtype: int64

Shape of renamed dataframe : (520, 1505)
Number of each species :
 A_niger          280
A_tubingensis    240
Name: target, dtype: int64

Save file as: dataframe.csv
Shape of dataframe : (3464, 1505)
Number of each species in dataframe:
 A_luchuensis              318
A_niger                   280
A_tubingensis             240
A_trinidadensis           235
A_indologenus             232
A_sclerotiicarbonarius    220
A_homomorphus             192
A_ibericus                160
A_japonicus               141
A_saccharolyticus         140
A_vadensis                137
A_heteromorphus           130
A_aculeatinus             120
A_brasiliensis            120
A_carbonarius             120
A_sclerotioniger          119
A_ellipticus              110
A_eucalypticola           110
A_aculeatus               100
A_floridensis              90
A_brunneoviolaceus         80
A_uvarum                   70
Name: target, dtype: int64
```

Count no. of member in each set & Print class mapping encoder

In [3]:

```
cols =[x for x in df.columns if x not in ['target']]
rowused = []
for i in range (len(df)):
        if i % 10 == 0:
            rowused.append('test')

        elif i % 10 == 1:
            rowused.append('validate')

        else:
            rowused.append('train')

df['rowused'] = rowused
dd=df['rowused'].sample(len(df))
test_set=df[df['rowused']=='test']
validate_set=df[df['rowused']=='validate']
train_set=df[df['rowused']=='train']
print('Count test_set:\n',test_set['target'].value_counts())
print('\nCount validate_set:\n',validate_set['target'].value_counts())
print('\nCount train_set:\n',train_set['target'].value_counts())
print('------------------------------------------')
label_encoder = LabelEncoder()
data_y = df.loc[:, 'target']
encoded_y = label_encoder.fit_transform(data_y.values.ravel())
label_encoder_name_mapping = dict(zip(label_encoder.classes_,label_encoder.transform(
print('Mapping of Label Encoded Classes:', label_encoder_name_mapping, sep="\n")
```

```
Count test_set:
 A_luchuensis              32
A_trinidadensis           24
A_indologenus             23
A_sclerotiicarbonarius    22
A_welwitschiae            22
A_homomorphus             20
A_ibericus                16
A_saccharolyticus         14
A_japonicus               14
A_niger                   14
A_costaricaensis          13
A_vadensis                13
A_heteromorphus           13
A_carbonarius             12
A_brasiliensis            12
A_sclerotioniger          12
A_aculeatinus             12
A_ellipticus              11
A_eucalypticola           11
A_aculeatus               10
A_floridensis              9
A_tubingensis              9
A_neoniger                 8
A_brunneoviolaceus         8
A_uvarum                   7
Name: target, dtype: int64

Count validate_set:
 A_luchuensis              32
A_trinidadensis           24
A_indologenus             23
A_sclerotiicarbonarius    22
A_welwitschiae            22
A_homomorphus             19
A_ibericus                16
A_saccharolyticus         14
A_vadensis                14
A_japonicus               14
A_niger                   14
A_costaricaensis          13
A_heteromorphus           13
A_carbonarius             12
A_brasiliensis            12
A_sclerotioniger          12
A_aculeatinus             12
A_ellipticus              11
A_eucalypticola           11
A_aculeatus               10
A_floridensis              9
A_tubingensis              9
A_neoniger                 8
A_brunneoviolaceus         8
A_uvarum                   7
Name: target, dtype: int64

Count train_set:
 A_luchuensis              254
A_trinidadensis           187
A_indologenus             186
```

```
A_welwitschiae           179
A_sclerotiicarbonarius    176
A_homomorphus            153
A_ibericus               128
A_japonicus              113
A_saccharolyticus        112
A_niger                  112
A_vadensis               110
A_costaricaensis         104
A_heteromorphus          104
A_carbonarius             96
A_brasiliensis            96
A_aculeatinus             96
A_sclerotioniger          95
A_ellipticus              88
A_eucalypticola           88
A_aculeatus               80
A_floridensis             72
A_tubingensis             72
A_neoniger                64
A_brunneoviolaceus        64
A_uvarum                  56
Name: target, dtype: int64
------------------------------------------
Mapping of Label Encoded Classes:
{'A_aculeatinus': 0, 'A_aculeatus': 1, 'A_brasiliensis': 2, 'A_brunneoviolaceus': 3,
'A_carbonarius': 4, 'A_costaricaensis': 5, 'A_ellipticus': 6, 'A_eucalypticola': 7,
'A_floridensis': 8, 'A_heteromorphus': 9, 'A_homomorphus': 10, 'A_ibericus': 11, 'A_i
ndologenus': 12, 'A_japonicus': 13, 'A_luchuensis': 14, 'A_neoniger': 15, 'A_niger':
16, 'A_saccharolyticus': 17, 'A_sclerotiicarbonarius': 18, 'A_sclerotioniger': 19, 'A
_trinidadensis': 20, 'A_tubingensis': 21, 'A_uvarum': 22, 'A_vadensis': 23, 'A_welwit
schiae': 24}
```