

```
In [1]: import pandas as pd
import numpy as np
train = pd.read_csv('UNSW_NB15_training-set.csv')
test = pd.read_csv('UNSW_NB15_testing-set.csv')
train.shape, test.shape
```

Out[1]: ((175341, 45), (82332, 45))

```
In [2]: #Combine into file - 'concat_data':
train['source']= 'train'
test['source'] = 'test'
concat_data=pd.concat([train, test],ignore_index=True)
concat_data.shape
```

Out[2]: (257673, 46)

```
In [3]: concat_data.head(257673)
```

Out[3]:

	id	dur	proto	service	state	spkts	dpkts	sbytes	dbytes	rate	...	ct_dst_src_ltm	is_ftp_login	ct_ftp_cmd
0	1	0.121478	tcp	-	FIN	6	4	258	172	74.087490	...	1	0	0
1	2	0.649902	tcp	-	FIN	14	38	734	42014	78.473372	...	2	0	0
2	3	1.623129	tcp	-	FIN	8	16	364	13186	14.170161	...	3	0	0
3	4	1.681642	tcp	ftp	FIN	12	12	628	770	13.677108	...	3	1	1
4	5	0.449454	tcp	-	FIN	10	6	534	268	33.373826	...	40	0	0
...
257668	82328	0.000005	udp	-	INT	2	0	104	0	200000.005100	...	2	0	0
257669	82329	1.106101	tcp	-	FIN	20	8	18062	354	24.410067	...	1	0	0
257670	82330	0.000000	arp	-	INT	1	0	46	0	0.000000	...	1	0	0
257671	82331	0.000000	arp	-	INT	1	0	46	0	0.000000	...	1	0	0
257672	82332	0.000009	udp	-	INT	2	0	104	0	111111.107200	...	1	0	0

257673 rows × 46 columns



```
In [4]: concat_data = pd.DataFrame(concat_data)
fe = concat_data.groupby('proto').size()/len(concat_data)
concat_data.loc[:, 'proto_freq_encode'] = concat_data['proto'].map(fe)
fe = concat_data.groupby('service').size()/len(concat_data)
concat_data.loc[:, 'service_freq_encode'] = concat_data['service'].map(fe)
fe = concat_data.groupby('state').size()/len(concat_data)
concat_data.loc[:, 'state_freq_encode'] = concat_data['state'].map(fe)
concat_data
```

Out[4]:

	id	dur	proto	service	state	spkts	dpkts	sbytes	dbytes	rate	...	ct_flw_http_mthd	ct_src_ltm	ct_srv_dsl
0	1	0.121478	tcp	-	FIN	6	4	258	172	74.087490	...	0	1	1
1	2	0.649902	tcp	-	FIN	14	38	734	42014	78.473372	...	0	1	€
2	3	1.623129	tcp	-	FIN	8	16	364	13186	14.170161	...	0	2	€
3	4	1.681642	tcp	ftp	FIN	12	12	628	770	13.677108	...	0	2	1
4	5	0.449454	tcp	-	FIN	10	6	534	268	33.373826	...	0	2	3€
...
257668	82328	0.000005	udp	-	INT	2	0	104	0	200000.005100	...	0	2	1
257669	82329	1.106101	tcp	-	FIN	20	8	18062	354	24.410067	...	0	3	2
257670	82330	0.000000	arp	-	INT	1	0	46	0	0.000000	...	0	1	1
257671	82331	0.000000	arp	-	INT	1	0	46	0	0.000000	...	0	1	1
257672	82332	0.000009	udp	-	INT	2	0	104	0	111111.107200	...	0	1	1

257673 rows × 49 columns



```
In [5]: var_to_encode = ['proto','service','state']
concat_data = pd.get_dummies(concat_data, columns=var_to_encode)
concat_data.columns
concat_data
```

Out[5]:

	id	dur	spkts	dpkts	sbytes	dbytes	rate	sttl	dttl	sload	...	state_CLO	state_CON	state_ECO	s
0	1	0.121478	6	4	258	172	74.087490	252	254	1.415894e+04	...	0	0	0	
1	2	0.649902	14	38	734	42014	78.473372	62	252	8.395112e+03	...	0	0	0	
2	3	1.623129	8	16	364	13186	14.170161	62	252	1.572272e+03	...	0	0	0	
3	4	1.681642	12	12	628	770	13.677108	62	252	2.740179e+03	...	0	0	0	
4	5	0.449454	10	6	534	268	33.373826	254	252	8.561499e+03	...	0	0	0	
...
257668	82328	0.000005	2	0	104	0	200000.005100	254	0	8.320000e+07	...	0	0	0	
257669	82329	1.106101	20	8	18062	354	24.410067	254	252	1.241044e+05	...	0	0	0	
257670	82330	0.000000	1	0	46	0	0.000000	0	0	0.000000e+00	...	0	0	0	
257671	82331	0.000000	1	0	46	0	0.000000	0	0	0.000000e+00	...	0	0	0	
257672	82332	0.000009	2	0	104	0	111111.107200	254	0	4.622222e+07	...	0	0	0	

257673 rows × 203 columns



```
In [6]: concat_data.drop('attack_cat',axis = 1,inplace = True)
```

```
In [7]: train = concat_data.loc[concat_data['source']=='train']
test = concat_data.loc[concat_data['source']=='test']
```

```
In [8]: train.drop('source',axis=1,inplace=True)
test.drop('source',axis=1,inplace=True)
```

C:\Users\admin\anaconda3\lib\site-packages\pandas\core\frame.py:4308: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
return super().drop(

```
In [9]: train.to_csv('UNSW_NB15_ohe_freq_enc_training_set.csv',index=False)
test.to_csv('UNSW_NB15_ohe_freq_enc_testing_set.csv',index=False)
```

```
In [ ]:
```