```
import pandas as pd
d=pd.read_csv("kdd_train.csv")
print(d)
```

```
       duration protocol_type   service flag  src_bytes  dst_bytes  land  \
0             0           tcp  ftp_data   SF        491          0     0
1             0           udp     other   SF        146          0     0
2             0           tcp   private   S0          0          0     0
3             0           tcp      http   SF        232       8153     0
4             0           tcp      http   SF        199        420     0
...         ...           ...       ...  ...        ...        ...   ...
36773         0           tcp  ftp_data   SF        334          0     0
36774         0           tcp      auth  REJ          0          0     0
36775         0           tcp      smtp   SF       1134        327     0
36776         0           tcp  ftp_data   SF       5336          0     0
36777         0           tcp    finger   S0          0          0     0

       wrong_fragment  urgent  hot  ...  dst_host_srv_count  \
0                   0       0    0  ...                  25
1                   0       0    0  ...                   1
2                   0       0    0  ...                  26
3                   0       0    0  ...                 255
4                   0       0    0  ...                 255
...               ...     ...  ...  ...                 ...
36773               0       0    0  ...                  32
36774               0       0    0  ...                  14
36775               0       0    0  ...                 162
36776               0       0    0  ...                  17
36777               0       0    0  ...                  24

       dst_host_same_srv_rate  dst_host_diff_srv_rate  \
0                        0.17                    0.03
1                        0.00                    0.60
2                        0.10                    0.05
3                        1.00                    0.00
4                        1.00                    0.00
...                       ...                     ...
36773                    1.00                    0.00
36774                    0.05                    0.07
36775                    0.82                    0.03
36776                    0.07                    0.07
36777                    0.09                    0.05

       dst_host_same_src_port_rate  dst_host_srv_diff_host_rate  \
0                             0.17                         0.00
1                             0.88                         0.00
2                             0.00                         0.00
3                             0.03                         0.04
4                             0.00                         0.00
...                            ...                          ...
36773                         1.00                         0.12
36774                         0.00                         0.00
```

✓ 0s completed at 10:28 AM

```
36777                         0.00                         0.00

       dst_host_serror_rate  dst_host_srv_serror_rate  dst_host_rerror_rate  \
0                      0.00                      0.00                  0.05
1                      0.00                      0.00                  0.00
2                      1.00                      1.00                  0.00
3                      0.03                      0.01                  0.00
4                      0.00                      0.00                  0.00
```

```
d.shape
```

```
(36778, 42)
```

```
d.head()
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment
0	0	tcp	ftp_data	SF	491	0	0	0
1	0	udp	other	SF	146	0	0	0
2	0	tcp	private	S0	0	0	0	0
3	0	tcp	http	SF	232	8153	0	0
4	0	tcp	http	SF	199	420	0	0

5 rows × 42 columns

```
d.tail()
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_frag
36773	0	tcp	ftp_data	SF	334	0	0	
36774	0	tcp	auth	REJ	0	0	0	
36775	0	tcp	smtp	SF	1134	327	0	
36776	0	tcp	ftp_data	SF	5336	0	0	
36777	0	tcp	finger	S0	0	0	0	

5 rows × 42 columns

```
d.columns
```

```
Index(['duration', 'protocol_type', 'service', 'flag', 'src_bytes',
       'dst_bytes', 'land', 'wrong_fragment', 'urgent', 'hot',
       'num_failed_logins', 'logged_in', 'num_compromised', 'root_shell',
       'su_attempted', 'num_root', 'num_file_creations', 'num_shells',
       'num_access_files', 'num_outbound_cmds', 'is_host_login',
       'is_guest_login', 'count', 'srv_count', 'serror_rate',
       'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
       'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count',
       'dst_host_srv_count', 'dst_host_same_srv_rate',
       'dst_host_diff_srv_rate', 'dst_host_same_src_port_rate',
       'dst_host_srv_diff_host_rate', 'dst_host_serror_rate',
       'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
       'dst_host_srv_rerror_rate', 'labels'],
      dtype='object')
```

```
d.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36778 entries, 0 to 36777
Data columns (total 42 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   duration                     36778 non-null  int64  
 1   protocol_type                36778 non-null  object 
 2   service                      36778 non-null  object 
 3   flag                         36778 non-null  object 
 4   src_bytes                    36778 non-null  int64  
 5   dst_bytes                    36778 non-null  int64  
 6   land                         36778 non-null  int64  
 7   wrong_fragment               36778 non-null  int64  
 8   urgent                       36778 non-null  int64  
 9   hot                          36778 non-null  int64  
 10  num_failed_logins            36778 non-null  int64  
 11  logged_in                    36778 non-null  int64  
 12  num_compromised              36778 non-null  int64  
 13  root_shell                   36778 non-null  int64  
 14  su_attempted                 36778 non-null  int64  
 15  num_root                     36778 non-null  int64  
 16  num_file_creations           36778 non-null  int64  
 17  num_shells                   36778 non-null  int64  
 18  num_access_files             36778 non-null  int64  
 19  num_outbound_cmds            36778 non-null  int64  
 20  is_host_login                36778 non-null  int64  
 21  is_guest_login               36778 non-null  int64  
 22  count                        36778 non-null  int64  
 23  srv_count                    36778 non-null  int64  
 24  serror_rate                  36778 non-null  float64
 25  srv_serror_rate              36778 non-null  float64
 26  rerror_rate                  36778 non-null  float64
```

```
 27  srv_rerror_rate              36778 non-null  float64
 28  same_srv_rate                36778 non-null  float64
 29  diff_srv_rate                36778 non-null  float64
 30  srv_diff_host_rate           36778 non-null  float64
 31  dst_host_count               36778 non-null  int64
 32  dst_host_srv_count           36778 non-null  int64
 33  dst_host_same_srv_rate       36778 non-null  float64
 34  dst_host_diff_srv_rate       36778 non-null  float64
 35  dst_host_same_src_port_rate  36778 non-null  float64
 36  dst_host_srv_diff_host_rate  36778 non-null  float64
 37  dst_host_serror_rate         36778 non-null  float64
 38  dst_host_srv_serror_rate     36778 non-null  float64
 39  dst_host_rerror_rate         36778 non-null  float64
 40  dst_host_srv_rerror_rate     36777 non-null  float64
 41  labels                       36777 non-null  object
dtypes: float64(15), int64(23), object(4)
memory usage: 11.8+ MB
```

```
d["labels"].value_counts()
```

```
normal             19621
neptune            12078
satan               1045
ipsweep             1043
portsweep            866
smurf                751
nmap                 439
back                 283
warezclient          281
teardrop             261
pod                   54
guess_passwd          15
buffer_overflow       11
warezmaster            9
imap                   5
multihop               4
rootkit                4
phf                    2
land                   2
ftp_write              1
loadmodule             1
spy                    1
Name: labels, dtype: int64
```

```
d["labels"].unique
```

```
<bound method Series.unique of 0             normal
1             normal
2            neptune
3             normal
4             normal
             ...
36773    warezclient
```

```
36774        neptune
36775         normal
36776         normal
36777            NaN
Name: labels, Length: 36778, dtype: object>
```

```
d["service"].describe()
```

```
count     36778
unique       67
top        http
freq      11717
Name: service, dtype: object
```

```
d.describe()
```

	duration	src_bytes	dst_bytes	land	wrong_fragment	u
count	36778.000000	3.677800e+04	3.677800e+04	36778.000000	36778.000000	36778.0
mean	302.837131	1.912880e+04	3.329609e+03	0.000082	0.022649	0.0
std	2679.159511	1.997686e+06	8.342668e+04	0.009031	0.254076	0.0
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
25%	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
50%	0.000000	4.400000e+01	0.000000e+00	0.000000	0.000000	0.0
75%	0.000000	2.780000e+02	5.297500e+02	0.000000	0.000000	0.0
max	42862.000000	3.817091e+08	5.153771e+06	1.000000	3.000000	3.0

8 rows × 38 columns

```
d["labels"] = d["labels"].replace(['neptune', 'warezclient', 'ipsweep', 'portsweep',
       'teardrop', 'nmap', 'satan', 'smurf', 'pod', 'back',
       'guess_passwd', 'ftp_write', 'multihop', 'rootkit',
       'buffer_overflow', 'imap', 'warezmaster', 'phf', 'land',
       'loadmodule', 'spy', 'perl'], 'attack')
```

```
d["labels"].unique()
```

```
array(['normal', 'attack', nan], dtype=object)
```

```
d1=d.drop(['protocol_type','service','flag','labels'],axis=1)
d1
```

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_
0	0	491	0	0	0	0	0	
1	0	146	0	0	0	0	0	
2	0	0	0	0	0	0	0	
3	0	232	8153	0	0	0	0	
4	0	199	420	0	0	0	0	
...	...	...	...	...	...	...	...	
36773	0	334	0	0	0	0	0	
36774	0	0	0	0	0	0	0	
36775	0	1134	327	0	0	0	0	
36776	0	5336	0	0	0	0	0	
36777	0	0	0	0	0	0	0	

36778 rows × 38 columns

```
x=d.iloc[:, :-1].values
d.describe()
```

	duration	src_bytes	dst_bytes	land	wrong_fragment	u
count	36778.000000	3.677800e+04	3.677800e+04	36778.000000	36778.000000	36778.0
mean	302.837131	1.912880e+04	3.329609e+03	0.000082	0.022649	0.0
std	2679.159511	1.997686e+06	8.342668e+04	0.009031	0.254076	0.0
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
25%	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
50%	0.000000	4.400000e+01	0.000000e+00	0.000000	0.000000	0.0
75%	0.000000	2.780000e+02	5.297500e+02	0.000000	0.000000	0.0
max	42862.000000	3.817091e+08	5.153771e+06	1.000000	3.000000	3.0

8 rows × 38 columns

```
y = d.iloc[:, 41].values
d.describe()
```

	duration	src_bytes	dst_bytes	land	wrong_fragment	u
count	36778.000000	3.677800e+04	3.677800e+04	36778.000000	36778.000000	36778.0
mean	302.837131	1.912880e+04	3.329609e+03	0.000082	0.022649	0.0
std	2679.159511	1.997686e+06	8.342668e+04	0.009031	0.254076	0.0
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
25%	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.0
50%	0.000000	4.400000e+01	0.000000e+00	0.000000	0.000000	0.0
75%	0.000000	2.780000e+02	5.297500e+02	0.000000	0.000000	0.0
max	42862.000000	3.817091e+08	5.153771e+06	1.000000	3.000000	3.0

8 rows × 38 columns

```
display(x)
```

```
array([[0, 'tcp', 'ftp_data', ..., 0.0, 0.05, 0.0],
       [0, 'udp', 'other', ..., 0.0, 0.0, 0.0],
       [0, 'tcp', 'private', ..., 1.0, 0.0, 0.0],
       ...,
       [0, 'tcp', 'smtp', ..., 0.0, 0.0, 0.0],
       [0, 'tcp', 'ftp_data', ..., 0.0, 0.0, 0.0],
       [0, 'tcp', 'finger', ..., 1.0, 0.0, nan]], dtype=object)
```

```
display(y)
```

```
array(['normal', 'normal', 'attack', ..., 'normal', 'normal', nan],
      dtype=object)
```

```
import numpy as np
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_x_1 = LabelEncoder()
labelencoder_x_2 = LabelEncoder()
labelencoder_x_3 = LabelEncoder()
```

```
x[:, 1] = labelencoder_x_1.fit_transform(x[:, 1])
x[:, 2] = labelencoder_x_2.fit_transform(x[:, 2])
x[:, 3] = labelencoder_x_3.fit_transform(x[:, 3])


Labelencoder_y= LabelEncoder()
y=Labelencoder_y.fit_transform(y)

df=pd.DataFrame(y)
display(df.sample(n=10))
```

	0
18331	1
35473	0
23510	0
15692	0
25871	0
17014	0
33866	0
18571	1
32787	1
13884	1

Colab paid products - Cancel contracts here