Data Exploration & Preprocessing

In [1]:
```
%matplotlib inline

from sklearn.preprocessing import StandardScaler
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Data ¶

In [2]:
```
df = pd.read_csv("./data/HTRU2/HTRU_2.csv", names = ['Profile_mean'
, 'Profile_stdev', 'Profile_skewness',
                                                     'Profile_kurt
osis', 'DM_mean', 'DM_stdev', 'DM_skewness',
                                                     'DM_kurtosis'
, 'class'])

df
```

Out[2]:

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	DM_mean
0	140.562500	55.683782	-0.234571	-0.699648	3.199833
1	102.507812	58.882430	0.465318	-0.515088	1.677258
2	103.015625	39.341649	0.323328	1.051164	3.121237
3	136.750000	57.178449	-0.068415	-0.636238	3.642977
4	88.726562	40.672225	0.600866	1.123492	1.178930
...	...	...	...	...	...
17893	136.429688	59.847421	-0.187846	-0.738123	1.296823
17894	122.554688	49.485605	0.127978	0.323061	16.409699
17895	119.335938	59.935939	0.159363	-0.743025	21.430602
17896	114.507812	53.902400	0.201161	-0.024789	1.946488
17897	57.062500	85.797340	1.406391	0.089520	188.306020

17898 rows × 9 columns

Exploration

Lost values

```
In [3]: df.isnull().values.any() # Has no NaN/lost values
```

```
Out[3]: False
```

```
In [4]: df.isna().values.any()
```

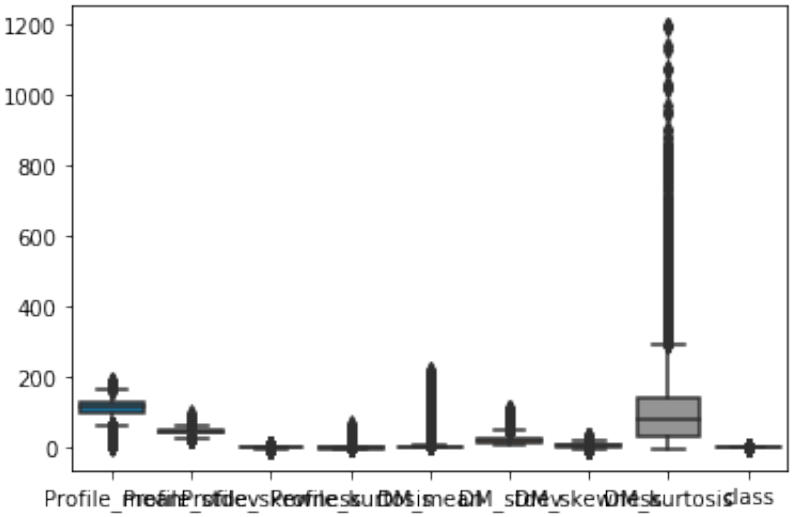
```
Out[4]: False
```

Statistic analysis

Data boxplot:

```
In [5]: sns.boxplot(data = df, palette="colorblind")
```

```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x108237d30>
```



Description of the features:

In [6]:
```
df.describe()
```

Out[6]:

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	DM_mea
count	17898.000000	17898.000000	17898.000000	17898.000000	17898.0000
mean	111.079968	46.549532	0.477857	1.770279	12.614400
std	25.652935	6.843189	1.064040	6.167913	29.472897
min	5.812500	24.772042	-1.876011	-1.791886	0.213211
25%	100.929688	42.376018	0.027098	-0.188572	1.923077
50%	115.078125	46.947479	0.223240	0.198710	2.801839
75%	127.085938	51.023202	0.473325	0.927783	5.464256
max	192.617188	98.778911	8.069522	68.101622	223.392140

In [7]:
```
pd.set_option('display.max_columns', 500)
df.groupby('class').describe()
#df.describe()
```

Out[7]:

	Profile_mean						
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**
class							
0	16259.0	116.562726	17.475932	17.210938	105.253906	117.257812	128.2851
1	1639.0	56.690608	30.007707	5.812500	31.777344	54.296875	79.27734

Plotting 2 to 2 and densities:

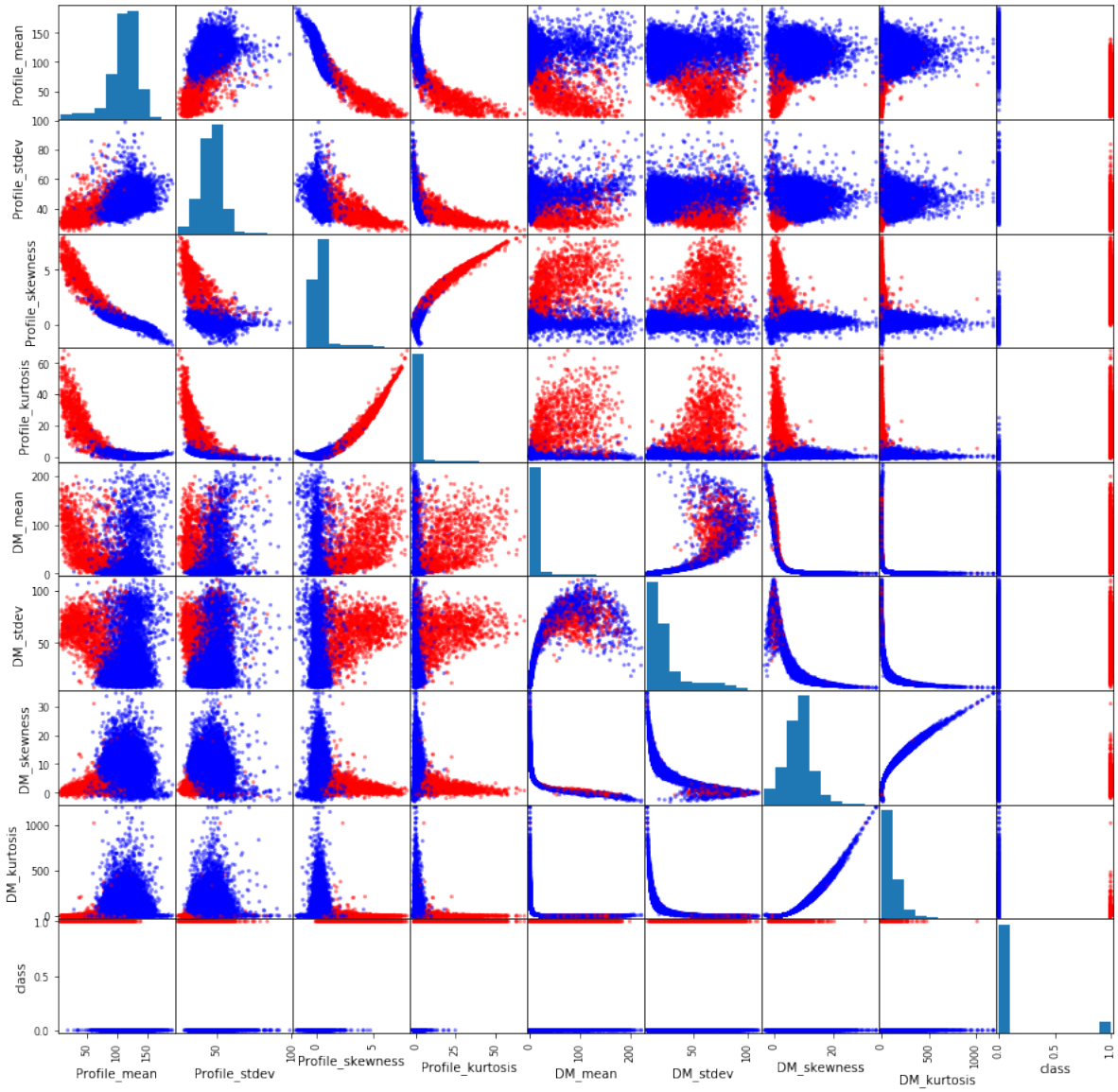
In [8]:
```
col = df['class'].map({1:'r', 0:'b'})
pd.plotting.scatter_matrix(df, c=col, figsize=(15,15))
```

Out[8]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x11b8687
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bc1b3
90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bc4a4
a8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bc7b8
60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bcade
10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bceb4
00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bd1c9
b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bd4ef
98>,
```

```
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bd4ef
d0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11bdbdb
00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bdfc0
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11be2c6
a0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11be5fc
50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11be9d2
40>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11becd7
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bf00d
a0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bf3d3
90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bf6d9
40>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11bfa2e
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11bfdf4
e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c00fa
90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c0500
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c07f6
30>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c0b1b
e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c0f11
d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c1207
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c152d
30>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11c18e3
20>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c1c18
d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c1f3e
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c2304
70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c261a
20>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11c296f
d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e64c5
c0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e67fb
70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e6bf1
60>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11e6ef7
```

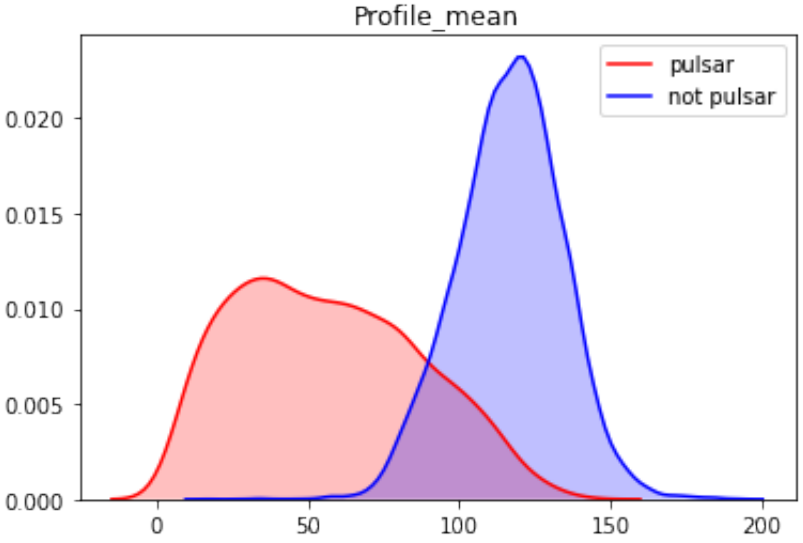
```
10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e720c
c0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e75f2
b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e7908
60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e7c2e
10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e7ff4
00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e8309
b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e862f
60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e89e5
50>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11e8d1b
00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e90f0
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e93e6
a0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e973c
50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e9b12
40>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11e9e17
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ea13d
a0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ea523
90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ea829
40>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11eab5e
f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eaf24
e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eb24a
90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eb630
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eb946
30>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ebc4b
e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ec031
d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ec337
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ec65d
30>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11eca43
20>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ecd48
d0>,
```

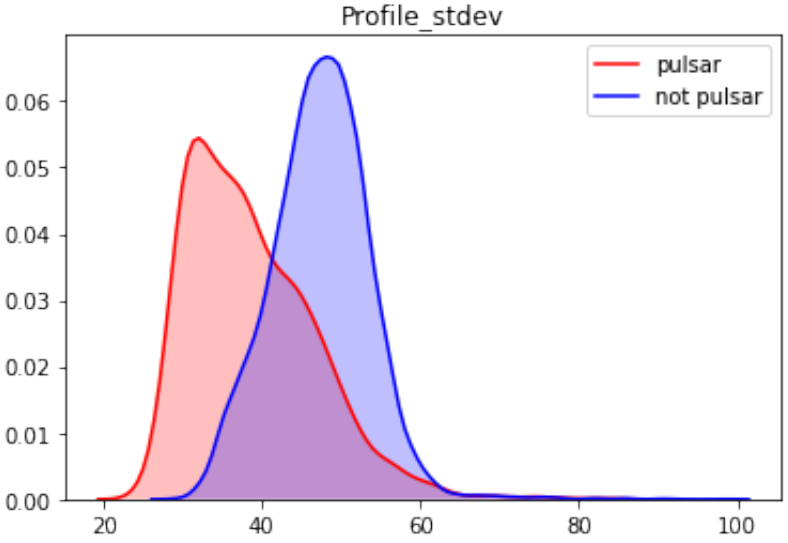
```
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ed06e
80>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ed454
70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ed74a
20>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eda9f
d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ede45
c0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ee17b
70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ee561
60>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x11ee877
10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eeb9c
c0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eef62
b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ef288
60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ef5ae
10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11ef994
00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11efc99
b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11eff9f
60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x11f0395
50>]],
      dtype=object)
```

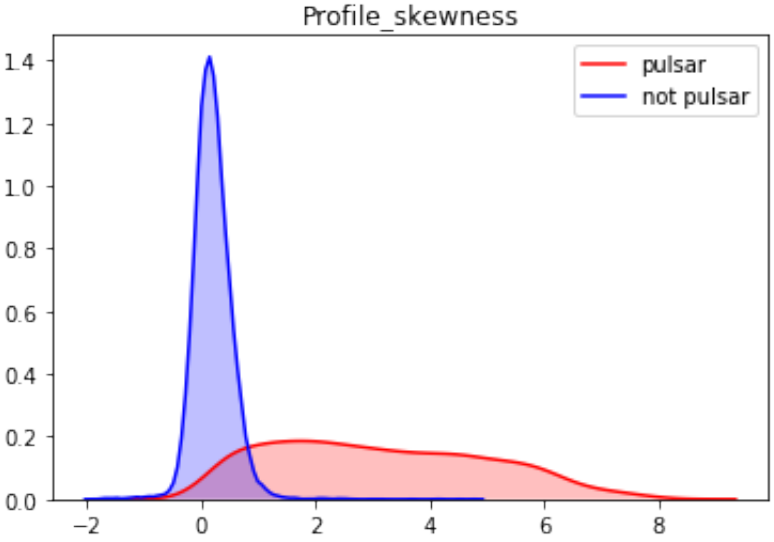


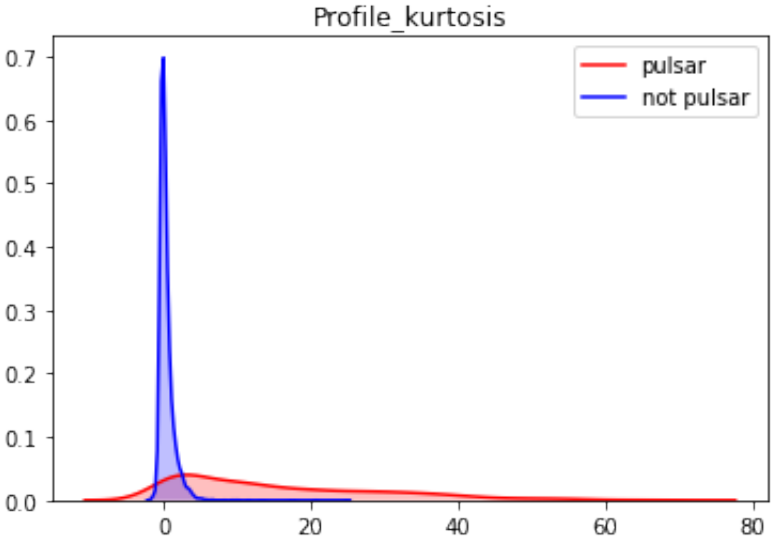
Densities per class and feature:

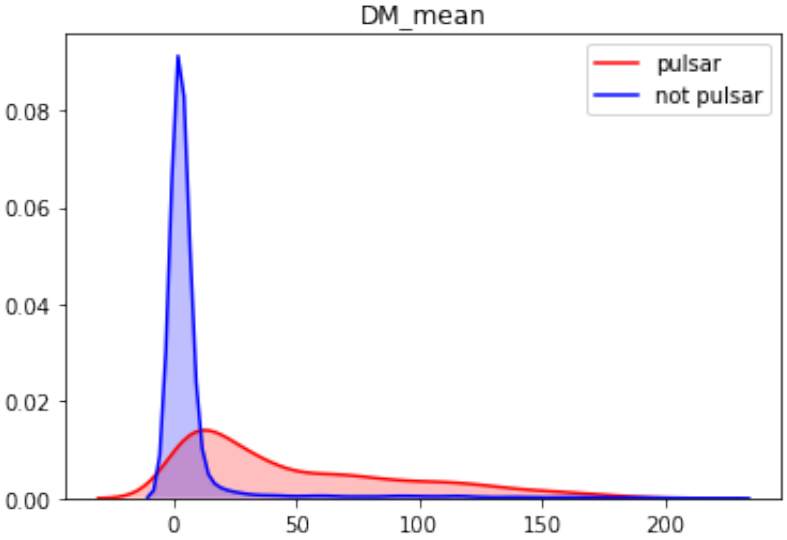
In [11]:
```
dfPulsar = df.loc[df['class'] == 1]
dfNotPulsar = df.loc[df['class'] == 0]

for column in df.columns[:-1]:
    pulsars = dfPulsar[column]
    notPulsars = dfNotPulsar[column]

    p1=sns.kdeplot(pulsars, shade=True, color="r", label="pulsar")
    p1=sns.kdeplot(notPulsars, shade=True, color="b", label="not pu
lsar")
    plt.title(column)
    plt.show()
```

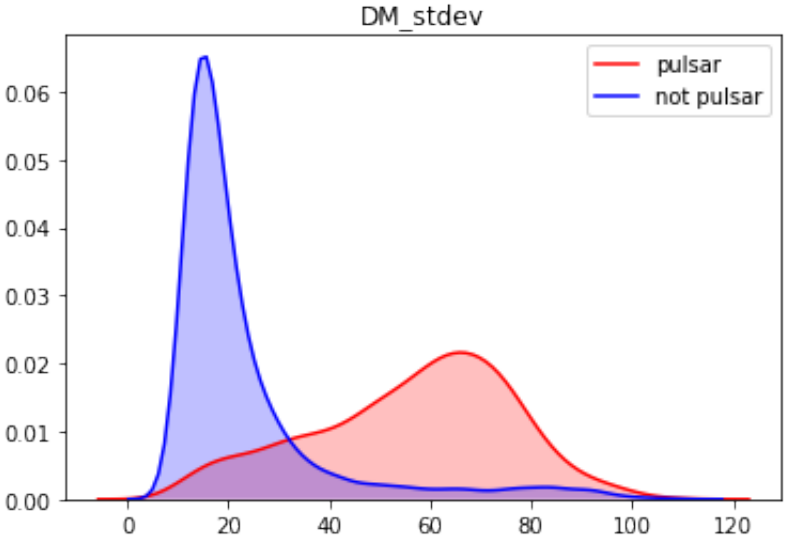

Profile_mean
pulsar
not pulsar
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200

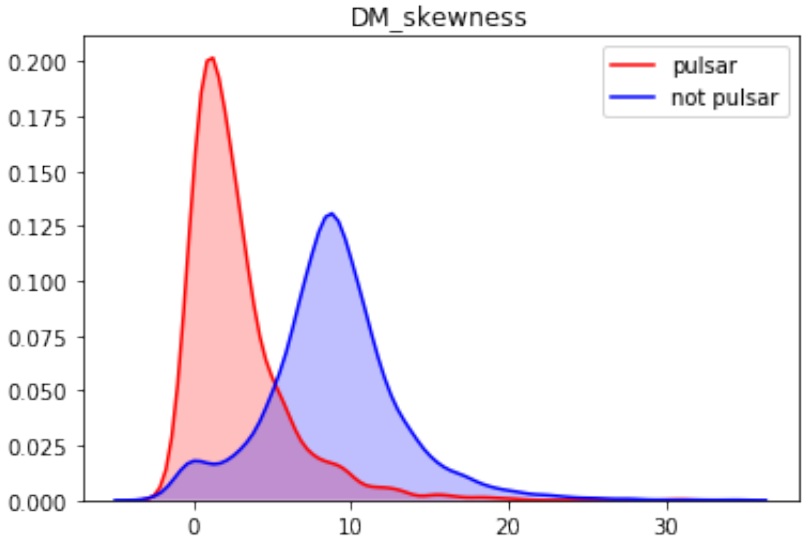

Profile_stdev
pulsar
not pulsar
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100

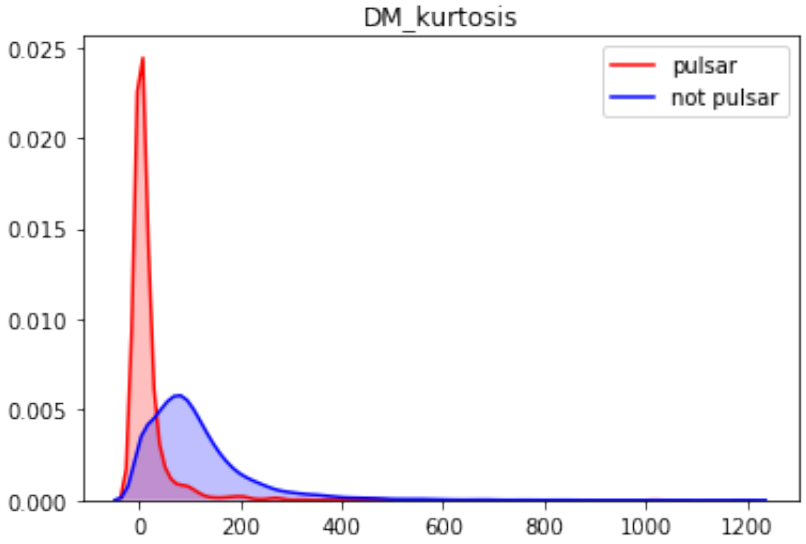

Profile_skewness
pulsar
not pulsar
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8


Profile_kurtosis
pulsar
not pulsar
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80


DM_mean
pulsar
not pulsar
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200


DM_stdev
pulsar
not pulsar
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
120





Preprocessing

Standarization

In [8]:
```
scaler = StandardScaler()

scaledData = scaler.fit_transform(df.drop(['class'], axis = 1))

stdDf = pd.DataFrame(scaledData, columns = df.columns[:-1])
stdDfWithClass = pd.concat([stdDf, df[['class']]], axis = 1)
```

In [9]:
```
stdDfWithClass.to_csv("./data/stdHTRU_2.csv", index = False)
```

Feature Extraction

Correlation Matrix of Data:

In [10]:
```
corrStd = stdDf.corr()
corrStd.style.background_gradient(cmap='coolwarm')
```

Out[10]:

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	I
Profile_mean	1	0.547137	-0.873898	-0.738775	-
Profile_stdev	0.547137	1	-0.521435	-0.539793	(
Profile_skewness	-0.873898	-0.521435	1	0.945729	(
Profile_kurtosis	-0.738775	-0.539793	0.945729	1	(
DM_mean	-0.298841	0.00686873	0.414368	0.412056	
DM_stdev	-0.307016	-0.0476316	0.43288	0.41514	(
DM_skewness	0.234331	0.0294294	-0.341209	-0.328843	-
DM_kurtosis	0.144033	0.0276915	-0.214491	-0.204782	-

In order to improve the performance of ML models that will be affected by the correlation of features and irrelevant variables, we will remove the correlated features (with correlation higher than 0.9).

In [11]:
```
features = np.full((corrStd.shape[0],), True, dtype=bool)
for i in range(corrStd.shape[0]):
    for j in range(i+1, corrStd.shape[0]):
        if corrStd.iloc[i,j] >= 0.9:
            if features[j]:
                features[j] = False

selectedFeatures = stdDf.columns[features]

noCorrStdData = stdDf[selectedFeatures]
```

In [12]:
```
noCorrStdDfWithClassData = pd.concat([noCorrStdData, df[['class']]]
, axis = 1)
```

In [13]:
```
corrNoCorrStd = noCorrStdDfWithClassData.corr()
corrNoCorrStd.style.background_gradient(cmap='coolwarm')
```

Out[13]:

	Profile_mean	Profile_stdev	Profile_skewness	DM_mean	DM_s
Profile_mean	1	0.547137	-0.873898	-0.298841	-0.30
Profile_stdev	0.547137	1	-0.521435	0.00686873	-0.04
Profile_skewness	-0.873898	-0.521435	1	0.414368	0.432
DM_mean	-0.298841	0.00686873	0.414368	1	0.796
DM_stdev	-0.307016	-0.0476316	0.43288	0.796555	1
DM_skewness	0.234331	0.0294294	-0.341209	-0.615971	-0.80
class	-0.673181	-0.363708	0.791591	0.400876	0.491

In [14]:
```
noCorrStdDfWithClassData.to_csv("./data/noCorrStdHTRU_2.csv", index
= False)
```