1.Apply knn to the “Surface defects in stainless steel plates” and identify the differences

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import plot_confusion_matrix
```

In [5]:
```
#Dataframe df
df = pd.read_csv('faults.csv')
```

In [6]:
```
#To print data
print(df)
```

```
      X_Minimum  X_Maximum  Y_Minimum  Y_Maximum  Pixels_Areas  X_Perimeter
\
0            42         50     270900     270944           267           17
1           645        651    2538079    2538108           108           10
2           829        835    1553913    1553931            71            8
3           853        860     369370     369415           176           13
4          1289       1306     498078     498335          2409           60
...         ...        ...        ...        ...           ...          ...
1936        249        277     325780     325796           273           54
1937        144        175     340581     340598           287           44
1938        145        174     386779     386794           292           40
1939        137        170     422497     422528           419           97
1940       1261       1281      87951      87967           103           26

      Y_Perimeter  Sum_of_Luminosity  Minimum_of_Luminosity  \
0              44              24220                     76
1              30              11397                     84
2              19               7972                     99
3              45              18996                     99
4             260             246930                     37
...           ...                ...                    ...
1936           22              35033                    119
1937           24              34599                    112
1938           22              37572                    120
1939           47              52715                    117
1940           22              11682                    101

      Maximum_of_Luminosity  ...  Orientation_Index  Luminosity_Index  \
0                       108  ...             0.8182           -0.2913
1                       123  ...             0.7931           -0.1756
2                       125  ...             0.6667           -0.1228
3                       126  ...             0.8444           -0.1568
4                       126  ...             0.9338           -0.1992
...                     ...  ...                ...               ...
1936                    141  ...            -0.4286            0.0026
1937                    133  ...            -0.4516           -0.0582
1938                    140  ...            -0.4828            0.0052
1939                    140  ...            -0.0606           -0.0171
1940                    133  ...            -0.2000           -0.1139

      SigmoidOfAreas  Pastry  Z_Scratch  K_Scatch  Stains  Dirtiness  Bumps
\
0             0.5822       1          0         0       0          0      0
1             0.2984       1          0         0       0          0      0
2             0.2150       1          0         0       0          0      0
3             0.5212       1          0         0       0          0      0
4             1.0000       1          0         0       0          0      0
...              ...     ...        ...       ...     ...        ...    ...
1936          0.7254       0          0         0       0          0      0
1937          0.8173       0          0         0       0          0      0
1938          0.7079       0          0         0       0          0      0
1939          0.9919       0          0         0       0          0      0
1940          0.5296       0          0         0       0          0      0

      Other_Faults
0                0
1                0
```

```
2                0
3                0
4                0
...            ...
1936             1
1937             1
1938             1
1939             1
1940             1

[1941 rows x 34 columns]
```

In [7]:
```
#Size of data
print('n',df.shape)
```

```
n (1941, 34)
```

In [8]:
```
#To print column names of data
print('n',df.columns)
```

```
n Index(['X_Minimum', 'X_Maximum', 'Y_Minimum', 'Y_Maximum', 'Pixels_Areas',
       'X_Perimeter', 'Y_Perimeter', 'Sum_of_Luminosity',
       'Minimum_of_Luminosity', 'Maximum_of_Luminosity', 'Length_of_Conveye
r',
       'TypeOfSteel_A300', 'TypeOfSteel_A400', 'Steel_Plate_Thickness',
       'Edges_Index', 'Empty_Index', 'Square_Index', 'Outside_X_Index',
       'Edges_X_Index', 'Edges_Y_Index', 'Outside_Global_Index', 'LogOfArea
s',
       'Log_X_Index', 'Log_Y_Index', 'Orientation_Index', 'Luminosity_Index',
       'SigmoidOfAreas', 'Pastry', 'Z_Scratch', 'K_Scatch', 'Stains',
       'Dirtiness', 'Bumps', 'Other_Faults'],
      dtype='object')
```

In [9]:
```
#To check for datatype and presence of null value for each column
print('n',df.info())
print('n',df.isnull().values.any())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1941 entries, 0 to 1940
Data columns (total 34 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   X_Minimum              1941 non-null   int64  
 1   X_Maximum              1941 non-null   int64  
 2   Y_Minimum              1941 non-null   int64  
 3   Y_Maximum              1941 non-null   int64  
 4   Pixels_Areas           1941 non-null   int64  
 5   X_Perimeter            1941 non-null   int64  
 6   Y_Perimeter            1941 non-null   int64  
 7   Sum_of_Luminosity      1941 non-null   int64  
 8   Minimum_of_Luminosity  1941 non-null   int64  
 9   Maximum_of_Luminosity  1941 non-null   int64  
 10  Length_of_Conveyer     1941 non-null   int64  
 11  TypeOfSteel_A300       1941 non-null   int64  
 12  TypeOfSteel_A400       1941 non-null   int64  
 13  Steel_Plate_Thickness  1941 non-null   int64  
 14  Edges_Index            1941 non-null   float64
 15  Empty_Index            1941 non-null   float64
 16  Square_Index           1941 non-null   float64
 17  Outside_X_Index        1941 non-null   float64
 18  Edges_X_Index          1941 non-null   float64
 19  Edges_Y_Index          1941 non-null   float64
 20  Outside_Global_Index   1941 non-null   float64
 21  LogOfAreas             1941 non-null   float64
 22  Log_X_Index            1941 non-null   float64
 23  Log_Y_Index            1941 non-null   float64
 24  Orientation_Index      1941 non-null   float64
 25  Luminosity_Index       1941 non-null   float64
 26  SigmoidOfAreas         1941 non-null   float64
 27  Pastry                 1941 non-null   int64  
 28  Z_Scratch              1941 non-null   int64  
 29  K_Scatch               1941 non-null   int64  
 30  Stains                 1941 non-null   int64  
 31  Dirtiness              1941 non-null   int64  
 32  Bumps                  1941 non-null   int64  
 33  Other_Faults           1941 non-null   int64  
dtypes: float64(13), int64(21)
memory usage: 515.7 KB
n None
n False
```

In [25]:
```
#Last seven columns are type of error classification
label_columns = df.columns.values[-7:]
```

In [27]:
```
#Assigning the error classification values to variable targets
targets = (df.iloc[:,-7:]==1).idxmax(1)
print('n',label_columns)
print('n',targets)
```

```
n ['Pastry' 'Z_Scratch' 'K_Scatch' 'Stains' 'Dirtiness' 'Bumps'
 'Other_Faults']
n 0             Pastry
1              Pastry
2              Pastry
3              Pastry
4              Pastry
              ...
1936    Other_Faults
1937    Other_Faults
1938    Other_Faults
1939    Other_Faults
1940    Other_Faults
Length: 1941, dtype: object
```

In [12]:
```
#Dropping the 7 error classification columns and retaining only targets
dataset = df.drop(label_columns,axis=1)
```

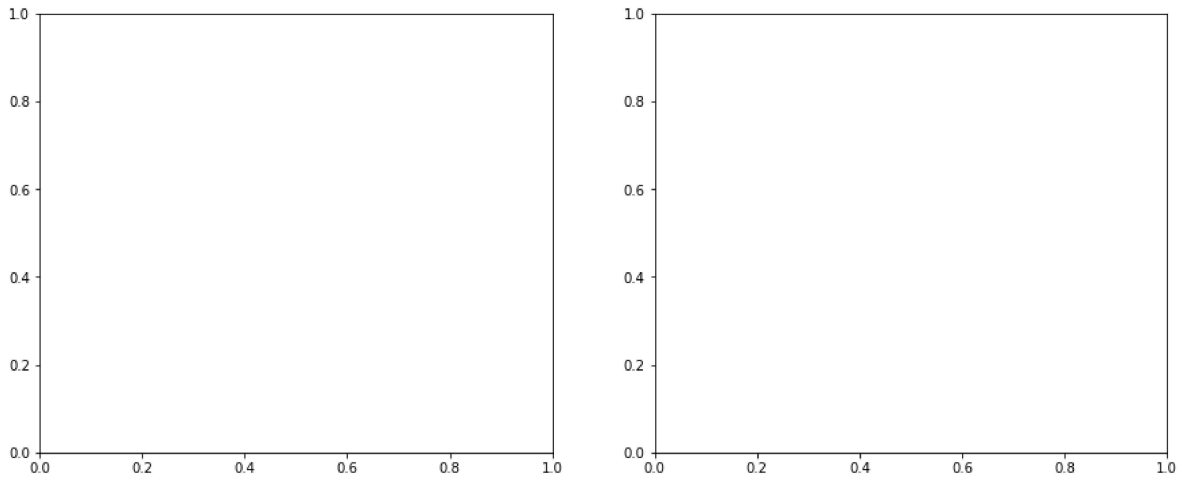
In [13]:
```
#creating a new column 'target' with all classification values
dataset['target']=targets
print('n',dataset)
```

```
n       X_Minimum  X_Maximum  Y_Minimum  Y_Maximum  Pixels_Areas  X_Perimeter
\
0            42         50     270900     270944           267           17
1           645        651    2538079    2538108           108           10
2           829        835    1553913    1553931            71            8
3           853        860     369370     369415           176           13
4          1289       1306     498078     498335          2409           60
...         ...        ...        ...        ...           ...          ...
1936        249        277     325780     325796           273           54
1937        144        175     340581     340598           287           44
1938        145        174     386779     386794           292           40
1939        137        170     422497     422528           419           97
1940       1261       1281      87951      87967           103           26

      Y_Perimeter  Sum_of_Luminosity  Minimum_of_Luminosity  \
0              44              24220                     76
1              30              11397                     84
2              19               7972                     99
3              45              18996                     99
4             260             246930                     37
...           ...                ...                    ...
1936           22              35033                    119
1937           24              34599                    112
1938           22              37572                    120
1939           47              52715                    117
1940           22              11682                    101

      Maximum_of_Luminosity  ...  Edges_X_Index  Edges_Y_Index  \
0                       108  ...         0.4706         1.0000
1                       123  ...         0.6000         0.9667
2                       125  ...         0.7500         0.9474
3                       126  ...         0.5385         1.0000
4                       126  ...         0.2833         0.9885
...                     ...  ...            ...            ...
1936                    141  ...         0.5185         0.7273
1937                    133  ...         0.7046         0.7083
1938                    140  ...         0.7250         0.6818
1939                    140  ...         0.3402         0.6596
1940                    133  ...         0.7692         0.7273

      Outside_Global_Index  LogOfAreas  Log_X_Index  Log_Y_Index  \
0                      1.0      2.4265       0.9031       1.6435
1                      1.0      2.0334       0.7782       1.4624
2                      1.0      1.8513       0.7782       1.2553
3                      1.0      2.2455       0.8451       1.6532
4                      1.0      3.3818       1.2305       2.4099
...                    ...         ...          ...          ...
1936                   0.0      2.4362       1.4472       1.2041
1937                   0.0      2.4579       1.4914       1.2305
1938                   0.0      2.4654       1.4624       1.1761
1939                   0.0      2.6222       1.5185       1.4914
1940                   0.0      2.0128       1.3010       1.2041

      Orientation_Index  Luminosity_Index  SigmoidOfAreas  target
0                0.8182           -0.2913          0.5822  Pastry
1                0.7931           -0.1756          0.2984  Pastry
2                0.6667           -0.1228          0.2150  Pastry
```

```
3                0.8444          -0.1568          0.5212        Pastry
4                0.9338          -0.1992          1.0000        Pastry
...                 ...              ...             ...           ...
1936            -0.4286           0.0026          0.7254  Other_Faults
1937            -0.4516          -0.0582          0.8173  Other_Faults
1938            -0.4828           0.0052          0.7079  Other_Faults
1939            -0.0606          -0.0171          0.9919  Other_Faults
1940            -0.2000          -0.1139          0.5296  Other_Faults

[1941 rows x 28 columns]
```

In [28]:
```
#Printing count of each type of error
print('n',dataset.target.value_counts())
fig, ax = plt.subplots(1,2,figsize=(15,6))
```

```
n Other_Faults    673
Bumps           402
K_Scatch        391
Z_Scratch       190
Pastry          158
Stains           72
Dirtiness        55
Name: target, dtype: int64
```

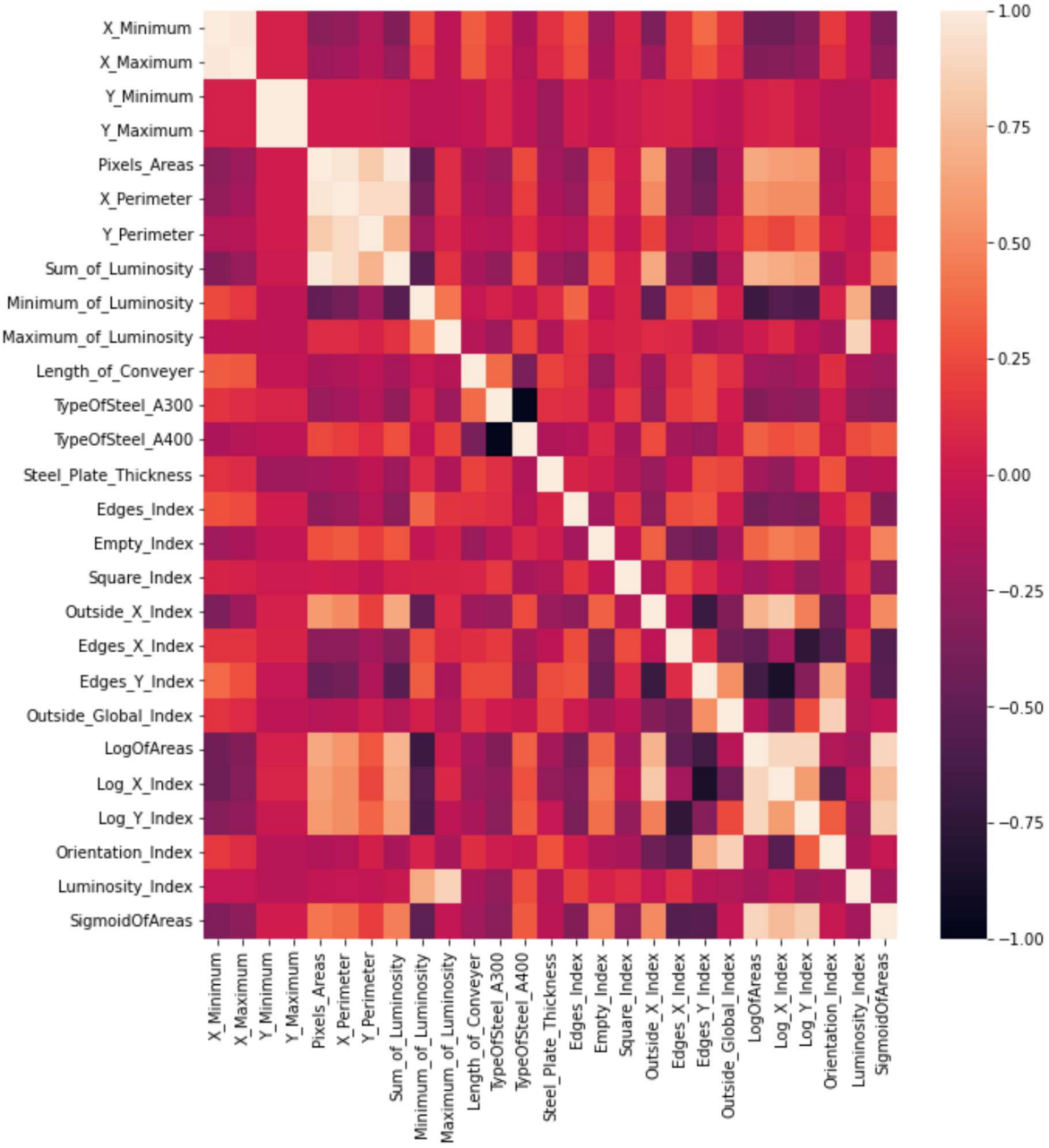


In [15]:
```
#Visualising the distribution of each error using histogram and pie chart
sns.countplot(x='target',data=dataset, ax=ax[0])
dataset['target'].value_counts().plot.pie(autopct = '%.1f', ax=ax[1])
```

Out[15]: <AxesSubplot:ylabel='target'>

In [17]:
```
#Visualising the correlation among each dataset feature
plt.figure(figsize=(10,11))
sns.heatmap(dataset.corr(),annot=False)
```

Out[17]: <AxesSubplot:>



In [18]:
```
#Dropping features with high correlation to others
dataset=dataset.drop('TypeOfSteel_A400',axis=1)
dataset=dataset.drop('X_Minimum',axis=1)
dataset=dataset.drop('Y_Minimum',axis=1)
```

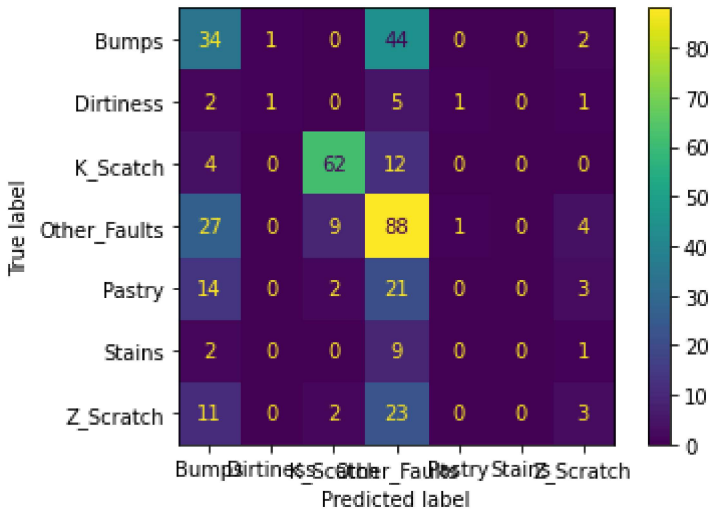
In [19]:
```
#Assigning feature column values to x
x = dataset.iloc[:,0:24]
```

In [20]:
```
#Assingning target column values to y
y = dataset.iloc[:,24]
```

In [21]:
```
#Splitting the dataset to training and testing sets
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2,random_st
```

In [22]:
```
#KNN model
knn = KNeighborsClassifier(n_neighbors= 19)
knn.fit(x_train,y_train)
classfier = knn.fit(x_train,y_train)
print("nKNN accuracy:",knn.score(x_test,y_test))
plot_confusion_matrix(classfier,x_test,y_test,labels=None, sample_weight=None,
plt.show()
```

```
nKNN accuracy: 0.4832904884318766

C:\Users\91830\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: F
utureWarning: Function plot_confusion_matrix is deprecated; Function `plot_co
nfusion_matrix` is deprecated in 1.0 and will be removed in 1.2. Use one of t
he class methods: ConfusionMatrixDisplay.from_predictions or ConfusionMatrixD
isplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: