Notebook

November 9, 2025

[1]:
```
import pandas as pd
import scipy.stats as stats
```

[2]:
```
df = pd.read_csv("data/original-data.csv")
df.rename(columns={'V1': 'X_Minimum'}, inplace=True)
df.rename(columns={'V2': 'X_Maximum'}, inplace=True)
df.rename(columns={'V3': 'Y_Minimum'}, inplace=True)
df.rename(columns={'V4': 'Y_Maximum'}, inplace=True)
df.rename(columns={'V5': 'Pixels_Areas'}, inplace=True)
df.rename(columns={'V6': 'X_Perimeter'}, inplace=True)
df.rename(columns={'V7': 'Y_Perimeter'}, inplace=True)
df.rename(columns={'V8': 'Sum_of_Luminosity'}, inplace=True)
df.rename(columns={'V9': 'Minimum_of_Luminosity'}, inplace=True)
df.rename(columns={'V10': 'Maximum_of_Luminosity'}, inplace=True)
df.rename(columns={'V11': 'Length_of_Conveyer'}, inplace=True)
df.rename(columns={'V12': 'TypesOfSteel_A300'}, inplace=True)
df.rename(columns={'V13': 'TypesOfSteel_A400'}, inplace=True)
df.rename(columns={'V14': 'Steel_Plate_Thickness'}, inplace=True)
df.rename(columns={'V15': 'Edges_Index'}, inplace=True)
df.rename(columns={'V16': 'Empty_Index'}, inplace=True)
df.rename(columns={'V17': 'Square_Index'}, inplace=True)
df.rename(columns={'V18': 'Outside_X_Index'}, inplace=True)
df.rename(columns={'V19': 'Edges_X_Index'}, inplace=True)
df.rename(columns={'V20': 'Edges_Y_Index'}, inplace=True)
df.rename(columns={'V21': 'Outside_Global_Index'}, inplace=True)
df.rename(columns={'V22': 'LogOfAreas'}, inplace=True)
df.rename(columns={'V23': 'Log_X_Index'}, inplace=True)
df.rename(columns={'V24': 'Log_Y_Index'}, inplace=True)
df.rename(columns={'V25': 'Orientation_Index'}, inplace=True)
df.rename(columns={'V26': 'Luminosity_Index'}, inplace=True)
df.rename(columns={'V27': 'SigmoidOfAreas'}, inplace=True)
df.rename(columns={'V28': 'Pastry'}, inplace=True)
df.rename(columns={'V29': 'Z_Scratch'}, inplace=True)
df.rename(columns={'V30': 'K_Scratch'}, inplace=True)
df.rename(columns={'V31': 'Stains'}, inplace=True)
df.rename(columns={'V32': 'Dirtiness'}, inplace=True)
df.rename(columns={'V33': 'Bumps'}, inplace=True)
df.rename(columns={'Class': 'Class'}, inplace=True) # Other_Faults
```

```
df
```

[2]:

```
      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_Scratch  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

      Class
0         1
1         1
2         1
3         1
4         1
…     …
1936      2
1937      2
1938      2
1939      2
1940      2

[1941 rows x 34 columns]
```

[3]:
```
df.duplicated().any()
```

[3]:
```
np.False_
```

No duplicates

1 Z-Score outlier detection

[4]:
```
# All Columns except the ones we're predicting ('Pastry', 'Z_Scratch',␣
↪'K_Scratch', 'Stains', 'Dirtiness', 'Bumps', 'Other_Faults_(Class)')
columns_to_check_outliers = [
    '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_Conveyer', 'TypesOfSteel_A300', 'TypesOfSteel_A400',␣
↪'Steel_Plate_Thickness',
    'Edges_Index', 'Empty_Index', 'Square_Index', 'Outside_X_Index',␣
↪'Edges_X_Index',
    'Edges_Y_Index', 'Outside_Global_Index', 'LogOfAreas', 'Log_X_Index',␣
↪'Log_Y_Index',
    'Orientation_Index', 'Luminosity_Index', 'SigmoidOfAreas'
]
# Ignoring certain columns without outliers - to use a more aggressive outlier␣
↪detection on the others
```

```
columns_to_ignore = [
    'X_Minimum', 'X_Maximum', 'Length_of_Conveyer', 'TypesOfSteel_A300',␣
 ↪'TypesOfSteel_A400',
    'Edges_Index','Square_Index','Edges_X_Index', 'Edges_Y_Index',␣
 ↪'Outside_Global_Index',
    'Orientation_Index', 'SigmoidOfAreas'
]
columns_to_check_outliers = [col for col in columns_to_check_outliers if col␣
 ↪not in columns_to_ignore]
print(columns_to_check_outliers)
z_score = stats.zscore(df[columns_to_check_outliers])
df_clean = df.copy()
df_clean = df_clean[(abs(z_score) < 4).all(axis=1)]
```

```
['Y_Minimum', 'Y_Maximum', 'Pixels_Areas', 'X_Perimeter', 'Y_Perimeter',
'Sum_of_Luminosity', 'Minimum_of_Luminosity', 'Maximum_of_Luminosity',
'Steel_Plate_Thickness', 'Empty_Index', 'Outside_X_Index', 'LogOfAreas',
'Log_X_Index', 'Log_Y_Index', 'Luminosity_Index']
```

2 Modifikovaná 5% 95% Metóda

[5]:
```
# All Columns except the ones we're predicting ('Pastry', 'Z_Scratch',␣
 ↪'K_Scratch', 'Stains', 'Dirtiness', 'Bumps', 'Other_Faults_(Class)')
columns_to_check_outliers = [
    'X_Minimum', 'X_Maximum', 'Y_Minimum', 'Y_Maximum', 'Pixels_Areas',␣
 ↪'X_Perimeter',
    'Y_Perimeter', 'Sum_of_Luminosity', 'Minimum_of_Luminosity',␣
 ↪'Maximum_of_Luminosity',
    'TypesOfSteel_A300', 'TypesOfSteel_A400', 'Steel_Plate_Thickness',␣
 ↪'Edges_Index',
    'Empty_Index', 'Square_Index', 'Outside_X_Index', 'Edges_X_Index',␣
 ↪'Edges_Y_Index',
    'Outside_Global_Index', 'LogOfAreas', 'Log_X_Index', 'Log_Y_Index',␣
 ↪'Orientation_Index',
    'Luminosity_Index', 'SigmoidOfAreas'
]
# Ignoring certain columns without outliers - to use a more aggressive outlier␣
 ↪detection on the others
columns_to_ignore = [
    'X_Minimum', 'X_Maximum', 'Length_of_Conveyer', 'TypesOfSteel_A300',␣
 ↪'TypesOfSteel_A400',
    'Edges_Index','Square_Index','Edges_X_Index', 'Edges_Y_Index',␣
 ↪'Outside_Global_Index',
    'Orientation_Index', 'SigmoidOfAreas',
    # Ignoring columns that were well handled by Z-Score already or don't␣
 ↪require more cleaning
```

```
    'Y_Minimum', 'Y_Maximum', 'Minimum_of_Luminosity', 'Maximum_of_Luminosity',␣
 ↪'Steel_Plate_Thickness', 'Empty_Index',
    'Log_X_Index', 'Log_Y_Index', 'Luminosity_Index'
]
columns_to_check_outliers = [col for col in columns_to_check_outliers if col␣
 ↪not in columns_to_ignore]
print(columns_to_check_outliers)

shaving_ranges = [
    (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98)
]

for idx, col in enumerate(columns_to_check_outliers):
    lower = 0
    higher = 1
    df_CO_lower = df_clean[col].quantile(shaving_ranges[idx][lower])
    df_CO_upper = df_clean[col].quantile(shaving_ranges[idx][higher])

    df_clean[col] = df[col].where(
        (df_clean[col] >= df_CO_lower) & (df_clean[col] <= df_CO_upper)
    )
    df_clean.dropna(inplace = True)
```

```
['Pixels_Areas', 'X_Perimeter', 'Y_Perimeter', 'Sum_of_Luminosity',
'Outside_X_Index', 'LogOfAreas']
```

[6]:
```
df_clean = df_clean.drop(columns=["Y_Minimum", "Y_Maximum", "Edges_Index",␣
 ↪"Empty_Index"], errors="ignore")
df_clean.describe()
```

[6]:

	X_Minimum	X_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	\
count	1613.000000	1613.000000	1613.000000	1613.000000	1613.000000	
mean	629.401116	655.522009	632.513329	48.324241	39.787353	
std	517.427471	507.119546	1334.099392	64.058738	47.378852	
min	0.000000	4.000000	2.000000	2.000000	1.000000	
25%	114.000000	181.000000	77.000000	14.000000	12.000000	
50%	563.000000	583.000000	141.000000	22.000000	21.000000	
75%	1090.000000	1106.000000	349.000000	47.000000	42.000000	
max	1705.000000	1713.000000	6277.000000	405.000000	330.000000	

	Sum_of_Luminosity	Minimum_of_Luminosity	Maximum_of_Luminosity	\
count	1613.000000	1613.000000	1613.000000	
mean	66874.983881	90.650961	129.055797	
std	139509.821414	27.155015	16.596791	
min	250.000000	0.000000	70.000000	
25%	8602.000000	77.000000	124.000000	
50%	16182.000000	95.000000	127.000000	

```
75%        37460.000000          109.000000           135.000000
max       652005.000000          179.000000           199.000000

       Length_of_Conveyer  TypesOfSteel_A300  …  Orientation_Index  \
count         1613.000000        1613.000000  …        1613.000000
mean          1468.679479           0.456293  …           0.121622
std            150.098130           0.498240  …           0.495092
min           1227.000000           0.000000  …          -0.970600
25%           1358.000000           0.000000  …          -0.250000
50%           1364.000000           0.000000  …           0.153900
75%           1656.000000           1.000000  …           0.545400
max           1794.000000           1.000000  …           0.946700

       Luminosity_Index  SigmoidOfAreas       Pastry    Z_Scratch  \
count       1613.000000     1613.000000  1613.000000  1613.000000
mean          -0.131680        0.524665     0.091754     0.115313
std            0.140006        0.321606     0.288769     0.319498
min           -0.609600        0.119000     0.000000     0.000000
25%           -0.199500        0.230000     0.000000     0.000000
50%           -0.132600        0.402500     0.000000     0.000000
75%           -0.057500        0.897100     0.000000     0.000000
max            0.457300        1.000000     1.000000     1.000000

         K_Scratch       Stains    Dirtiness        Bumps        Class
count  1613.000000  1613.000000  1613.000000  1613.000000  1613.000000
mean      0.106634     0.044637     0.034098     0.243025     1.364538
std       0.308743     0.206570     0.181537     0.429043     0.481450
min       0.000000     0.000000     0.000000     0.000000     1.000000
25%       0.000000     0.000000     0.000000     0.000000     1.000000
50%       0.000000     0.000000     0.000000     0.000000     1.000000
75%       0.000000     0.000000     0.000000     0.000000     2.000000
max       1.000000     1.000000     1.000000     1.000000     2.000000

[8 rows x 30 columns]
```

[7]:
```python
feature_cols = [*df_clean.columns[:len(df_clean.columns)-7]]
feature_cols
```

[7]:
```
['X_Minimum',
 'X_Maximum',
 'Pixels_Areas',
 'X_Perimeter',
 'Y_Perimeter',
 'Sum_of_Luminosity',
 'Minimum_of_Luminosity',
 'Maximum_of_Luminosity',
 'Length_of_Conveyer',
```

```
 'TypesOfSteel_A300',
 'TypesOfSteel_A400',
 'Steel_Plate_Thickness',
 'Square_Index',
 'Outside_X_Index',
 'Edges_X_Index',
 'Edges_Y_Index',
 'Outside_Global_Index',
 'LogOfAreas',
 'Log_X_Index',
 'Log_Y_Index',
 'Orientation_Index',
 'Luminosity_Index',
 'SigmoidOfAreas']
```

[8]:
```
predict_cols = ["Pastry", "Z_Scratch", "K_Scratch", "K_Scratch", "Stains",␣
 ↪"Dirtiness", "Bumps", "Class"]
```

[9]:
```
df_clean["Class"].nunique()
```

[9]:
```
2
```

[10]:
```
df_clean
```

[10]:
```
      X_Minimum  X_Maximum  Pixels_Areas  X_Perimeter  Y_Perimeter  \
0            42         50         267.0         17.0         44.0
1           645        651         108.0         10.0         30.0
2           829        835          71.0          8.0         19.0
3           853        860         176.0         13.0         45.0
4          1289       1306        2409.0         60.0        260.0
…           …          …             …            …            …
1936        249        277         273.0         54.0         22.0
1937        144        175         287.0         44.0         24.0
1938        145        174         292.0         40.0         22.0
1939        137        170         419.0         97.0         47.0
1940       1261       1281         103.0         26.0         22.0

      Sum_of_Luminosity  Minimum_of_Luminosity  Maximum_of_Luminosity  \
0               24220.0                     76                    108
1               11397.0                     84                    123
2                7972.0                     99                    125
3               18996.0                     99                    126
4              246930.0                     37                    126
…                   …                       …                     …
1936            35033.0                    119                    141
1937            34599.0                    112                    133
1938            37572.0                    120                    140
```

```
1939            52715.0                         117                          140
1940            11682.0                         101                          133

      Length_of_Conveyer  TypesOfSteel_A300  …  Orientation_Index  \
0                   1687                  1  …             0.8182
1                   1687                  1  …             0.7931
2                   1623                  1  …             0.6667
3                   1353                  0  …             0.8444
4                   1353                  0  …             0.9338
…                    …                 …   …                 …
1936                1360                  0  …            -0.4286
1937                1360                  0  …            -0.4516
1938                1360                  0  …            -0.4828
1939                1360                  0  …            -0.0606
1940                1360                  1  …            -0.2000

      Luminosity_Index  SigmoidOfAreas  Pastry  Z_Scratch  K_Scratch  Stains  \
0              -0.2913          0.5822       1          0          0       0
1              -0.1756          0.2984       1          0          0       0
2              -0.1228          0.2150       1          0          0       0
3              -0.1568          0.5212       1          0          0       0
4              -0.1992          1.0000       1          0          0       0
…                   …              …     …        …        …      …
1936            0.0026          0.7254       0          0          0       0
1937           -0.0582          0.8173       0          0          0       0
1938            0.0052          0.7079       0          0          0       0
1939           -0.0171          0.9919       0          0          0       0
1940           -0.1139          0.5296       0          0          0       0

      Dirtiness  Bumps  Class
0             0      0      1
1             0      0      1
2             0      0      1
3             0      0      1
4             0      0      1
…           …    …    …
1936          0      0      2
1937          0      0      2
1938          0      0      2
1939          0      0      2
1940          0      0      2

[1613 rows x 30 columns]
```

[11]:

```
normalized_df = (df_clean[feature_cols] - df_clean[feature_cols].mean())/
  ↪df_clean[feature_cols].std()
```

[12]:
```
normalized_df
```

[12]:
```
      X_Minimum  X_Maximum  Pixels_Areas  X_Perimeter  Y_Perimeter  \
0     -1.135234  -1.194042     -0.273978    -0.488992     0.088914
1      0.030147  -0.008917     -0.393159    -0.598267    -0.206576
2      0.385752   0.353917     -0.420893    -0.629489    -0.438747
3      0.432136   0.403215     -0.342188    -0.551435     0.110021
4      1.274766   1.282692      1.331600     0.182266     4.647910
…           …          …             …            …          …
1936  -0.735178  -0.746416     -0.269480     0.088602    -0.375428
1937  -0.938105  -0.947552     -0.258986    -0.067504    -0.333215
1938  -0.936172  -0.949524     -0.255238    -0.129947    -0.375428
1939  -0.951633  -0.957411     -0.160043     0.759861     0.152233
1940   1.220652   1.233394     -0.396907    -0.348496    -0.375428

      Sum_of_Luminosity  Minimum_of_Luminosity  Maximum_of_Luminosity  \
0             -0.305749              -0.539531              -1.268667
1             -0.397664              -0.244926              -0.364878
2             -0.422214               0.307458              -0.244372
3             -0.343194               0.307458              -0.184120
4              1.290626              -1.975729              -0.184120
…                   …                    …                    …
1936          -0.228242               1.043971               0.719669
1937          -0.231353               0.786191               0.237649
1938          -0.210042               1.080796               0.659417
1939          -0.101498               0.970319               0.659417
1940          -0.395621               0.381110               0.237649

      Length_of_Conveyer  TypesOfSteel_A300  …  Outside_X_Index  \
0               1.454519           1.091255  …        -0.514948
1               1.454519           1.091255  …        -0.556364
2               1.028131           1.091255  …        -0.552599
3              -0.770692          -0.915808  …        -0.496123
4              -0.770692          -0.915808  …        -0.217504
…                    …                  …  …              …
1936           -0.724056          -0.915808  …         0.083706
1937           -0.724056          -0.915808  …         0.166538
1938           -0.724056          -0.915808  …         0.110061
1939           -0.724056          -0.915808  …         0.223015
1940           -0.724056           1.091255  …        -0.138436

      Edges_X_Index  Edges_Y_Index  Outside_Global_Index  LogOfAreas  \
0         -0.718842       0.695123              0.836722    0.237173
1         -0.159904       0.526690              0.836722   -0.417695
2          0.488015       0.429070              0.836722   -0.721056
3         -0.425551       0.695123              0.836722   -0.064356
4         -1.527877       0.636955              0.836722    1.828612
```

```
…              …            …                     …           …
1936    -0.511940    -0.684202             -1.263874    0.253332
1937     0.291911    -0.780305             -1.263874    0.289482
1938     0.380028    -0.914342             -1.263874    0.301977
1939    -1.282100    -1.026631             -1.263874    0.563191
1940     0.570948    -0.684202             -1.263874   -0.452012

      Log_X_Index  Log_Y_Index  Orientation_Index  Luminosity_Index  \
0       -0.846876     0.865673           1.406965         -1.140097
1       -1.190733     0.399819           1.356268         -0.313702
2       -1.190733    -0.132917           1.100962          0.063426
3       -1.006553     0.890625           1.459885         -0.179421
4        0.054474     2.837130           1.640457         -0.482266
…             …            …                   …                …
1936     0.651061    -0.264622          -1.111353          0.959104
1937     0.772746    -0.196712          -1.157808          0.524836
1938     0.692907    -0.336648          -1.220827          0.977674
1939     0.847354     0.474417          -0.368057          0.818395
1940     0.248564    -0.264622          -0.649621          0.126995

      SigmoidOfAreas
0           0.178899
1          -0.703548
2          -0.962871
3          -0.010774
4           1.478005
…                …
1936        0.624165
1937        0.909918
1938        0.569750
1939        1.452819
1940        0.015345

[1613 rows x 23 columns]
```

[13]:
```python
df2 = normalized_df.join(df[predict_cols], how='inner', lsuffix='_caller',
    rsuffix='_other')
```

[14]:
```python
df2
```

[14]:
```
      X_Minimum  X_Maximum  Pixels_Areas  X_Perimeter  Y_Perimeter  \
0     -1.135234  -1.194042     -0.273978    -0.488992     0.088914
1      0.030147  -0.008917     -0.393159    -0.598267    -0.206576
2      0.385752   0.353917     -0.420893    -0.629489    -0.438747
3      0.432136   0.403215     -0.342188    -0.551435     0.110021
4      1.274766   1.282692      1.331600     0.182266     4.647910
…          …          …             …            …            …
```

```
1936  -0.735178  -0.746416     -0.269480     0.088602    -0.375428
1937  -0.938105  -0.947552     -0.258986    -0.067504    -0.333215
1938  -0.936172  -0.949524     -0.255238    -0.129947    -0.375428
1939  -0.951633  -0.957411     -0.160043     0.759861     0.152233
1940   1.220652   1.233394     -0.396907    -0.348496    -0.375428

      Sum_of_Luminosity  Minimum_of_Luminosity  Maximum_of_Luminosity  \
0             -0.305749              -0.539531              -1.268667
1             -0.397664              -0.244926              -0.364878
2             -0.422214               0.307458              -0.244372
3             -0.343194               0.307458              -0.184120
4              1.290626              -1.975729              -0.184120
…                     …                      …                      …
1936          -0.228242               1.043971               0.719669
1937          -0.231353               0.786191               0.237649
1938          -0.210042               1.080796               0.659417
1939          -0.101498               0.970319               0.659417
1940          -0.395621               0.381110               0.237649

      Length_of_Conveyer  TypesOfSteel_A300  …  Luminosity_Index  \
0               1.454519           1.091255  …         -1.140097
1               1.454519           1.091255  …         -0.313702
2               1.028131           1.091255  …          0.063426
3              -0.770692          -0.915808  …         -0.179421
4              -0.770692          -0.915808  …         -0.482266
…                      …                …  …                 …
1936           -0.724056          -0.915808  …          0.959104
1937           -0.724056          -0.915808  …          0.524836
1938           -0.724056          -0.915808  …          0.977674
1939           -0.724056          -0.915808  …          0.818395
1940           -0.724056           1.091255  …          0.126995

      SigmoidOfAreas  Pastry  Z_Scratch  K_Scratch  K_Scratch  Stains  \
0           0.178899       1          0          0          0       0
1          -0.703548       1          0          0          0       0
2          -0.962871       1          0          0          0       0
3          -0.010774       1          0          0          0       0
4           1.478005       1          0          0          0       0
…                …     …        …        …        …     …
1936        0.624165       0          0          0          0       0
1937        0.909918       0          0          0          0       0
1938        0.569750       0          0          0          0       0
1939        1.452819       0          0          0          0       0
1940        0.015345       0          0          0          0       0

      Dirtiness  Bumps  Class
0             0      0      1
```

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

[1613 rows x 31 columns]
```

[15]:
```
df2["Class"]-=1
df2_multiclass = df2.copy()
```

2.1 Feature Selection

2.1.1 Random Forest

[16]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel

columns_to_ignore = ["Pastry", "Z_Scratch", "K_Scratch", "Stains", "Dirtiness",␣
↪"Bumps", "Class"]
df_feature_sel = df2.drop(columns=columns_to_ignore)

# Selecting the target column
target = df2["Class"]

clf = RandomForestClassifier()
clf = clf.fit(df_feature_sel, target)
print(clf.feature_importances_)

model = SelectFromModel(clf, prefit=True)

mask = model.get_support()
selected_columns = df_feature_sel.columns[mask]

df_ranFor = df_feature_sel.loc[:, selected_columns]
```

```
[0.05692419 0.0593691  0.04443948 0.03865951 0.03056805 0.05036666
 0.05493708 0.04782427 0.06463809 0.01595412 0.01356195 0.06854246
 0.04278322 0.04671546 0.04287257 0.0451143  0.00442681 0.04734333
 0.03699062 0.03392335 0.05956042 0.05541509 0.03906988]
```

[17]:
```
df_feature_sel
```

[17]:

	X_Minimum	X_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	\
0	-1.135234	-1.194042	-0.273978	-0.488992	0.088914	
1	0.030147	-0.008917	-0.393159	-0.598267	-0.206576	
2	0.385752	0.353917	-0.420893	-0.629489	-0.438747	
3	0.432136	0.403215	-0.342188	-0.551435	0.110021	
4	1.274766	1.282692	1.331600	0.182266	4.647910	
…	…	…	…	…	…	
1936	-0.735178	-0.746416	-0.269480	0.088602	-0.375428	
1937	-0.938105	-0.947552	-0.258986	-0.067504	-0.333215	
1938	-0.936172	-0.949524	-0.255238	-0.129947	-0.375428	
1939	-0.951633	-0.957411	-0.160043	0.759861	0.152233	
1940	1.220652	1.233394	-0.396907	-0.348496	-0.375428	

	Sum_of_Luminosity	Minimum_of_Luminosity	Maximum_of_Luminosity	\
0	-0.305749	-0.539531	-1.268667	
1	-0.397664	-0.244926	-0.364878	
2	-0.422214	0.307458	-0.244372	
3	-0.343194	0.307458	-0.184120	
4	1.290626	-1.975729	-0.184120	
…	…	…	…	
1936	-0.228242	1.043971	0.719669	
1937	-0.231353	0.786191	0.237649	
1938	-0.210042	1.080796	0.659417	
1939	-0.101498	0.970319	0.659417	
1940	-0.395621	0.381110	0.237649	

	Length_of_Conveyer	TypesOfSteel_A300	…	Outside_X_Index	\
0	1.454519	1.091255	…	-0.514948	
1	1.454519	1.091255	…	-0.556364	
2	1.028131	1.091255	…	-0.552599	
3	-0.770692	-0.915808	…	-0.496123	
4	-0.770692	-0.915808	…	-0.217504	
…	…	…	…	…	
1936	-0.724056	-0.915808	…	0.083706	
1937	-0.724056	-0.915808	…	0.166538	
1938	-0.724056	-0.915808	…	0.110061	
1939	-0.724056	-0.915808	…	0.223015	
1940	-0.724056	1.091255	…	-0.138436	

	Edges_X_Index	Edges_Y_Index	Outside_Global_Index	LogOfAreas	\
0	-0.718842	0.695123	0.836722	0.237173	
1	-0.159904	0.526690	0.836722	-0.417695	
2	0.488015	0.429070	0.836722	-0.721056	
3	-0.425551	0.695123	0.836722	-0.064356	
4	-1.527877	0.636955	0.836722	1.828612	
…	…	…	…	…	
1936	-0.511940	-0.684202	-1.263874	0.253332	

```
1937      0.291911    -0.780305             -1.263874   0.289482
1938      0.380028    -0.914342             -1.263874   0.301977
1939     -1.282100    -1.026631             -1.263874   0.563191
1940      0.570948    -0.684202             -1.263874  -0.452012

      Log_X_Index  Log_Y_Index  Orientation_Index  Luminosity_Index  \
0       -0.846876     0.865673           1.406965         -1.140097
1       -1.190733     0.399819           1.356268         -0.313702
2       -1.190733    -0.132917           1.100962          0.063426
3       -1.006553     0.890625           1.459885         -0.179421
4        0.054474     2.837130           1.640457         -0.482266
…             …            …                  …                 …
1936     0.651061    -0.264622          -1.111353          0.959104
1937     0.772746    -0.196712          -1.157808          0.524836
1938     0.692907    -0.336648          -1.220827          0.977674
1939     0.847354     0.474417          -0.368057          0.818395
1940     0.248564    -0.264622          -0.649621          0.126995

      SigmoidOfAreas
0           0.178899
1          -0.703548
2          -0.962871
3          -0.010774
4           1.478005
…                …
1936        0.624165
1937        0.909918
1938        0.569750
1939        1.452819
1940        0.015345

[1613 rows x 23 columns]
```

[18]:
```
df_ranFor["Class"] = df2["Class"]
df_ranFor
```

[18]:
```
      X_Minimum  X_Maximum  Pixels_Areas  Sum_of_Luminosity  \
0     -1.135234  -1.194042     -0.273978          -0.305749
1      0.030147  -0.008917     -0.393159          -0.397664
2      0.385752   0.353917     -0.420893          -0.422214
3      0.432136   0.403215     -0.342188          -0.343194
4      1.274766   1.282692      1.331600           1.290626
…           …          …            …                  …
1936  -0.735178  -0.746416     -0.269480          -0.228242
1937  -0.938105  -0.947552     -0.258986          -0.231353
1938  -0.936172  -0.949524     -0.255238          -0.210042
1939  -0.951633  -0.957411     -0.160043          -0.101498
```

```
1940   1.220652   1.233394     -0.396907          -0.395621

      Minimum_of_Luminosity  Maximum_of_Luminosity  Length_of_Conveyer  \
0                 -0.539531              -1.268667            1.454519
1                 -0.244926              -0.364878            1.454519
2                  0.307458              -0.244372            1.028131
3                  0.307458              -0.184120           -0.770692
4                 -1.975729              -0.184120           -0.770692
…                     …                     …                  …
1936               1.043971               0.719669           -0.724056
1937               0.786191               0.237649           -0.724056
1938               1.080796               0.659417           -0.724056
1939               0.970319               0.659417           -0.724056
1940               0.381110               0.237649           -0.724056

      Steel_Plate_Thickness  Outside_X_Index  Edges_Y_Index  LogOfAreas  \
0                  0.035167        -0.514948       0.695123    0.237173
1                  0.035167        -0.556364       0.526690   -0.417695
2                  0.475065        -0.552599       0.429070   -0.721056
3                  4.654098        -0.496123       0.695123   -0.064356
4                  2.344632        -0.217504       0.636955    1.828612
…                     …                …              …          …
1936              -0.844629         0.083706      -0.684202    0.253332
1937              -0.844629         0.166538      -0.780305    0.289482
1938              -0.844629         0.110061      -0.914342    0.301977
1939              -0.844629         0.223015      -1.026631    0.563191
1940               0.035167        -0.138436      -0.684202   -0.452012

      Orientation_Index  Luminosity_Index  Class
0              1.406965         -1.140097      0
1              1.356268         -0.313702      0
2              1.100962          0.063426      0
3              1.459885         -0.179421      0
4              1.640457         -0.482266      0
…                 …                 …     …
1936          -1.111353          0.959104      1
1937          -1.157808          0.524836      1
1938          -1.220827          0.977674      1
1939          -0.368057          0.818395      1
1940          -0.649621          0.126995      1

[1613 rows x 14 columns]
```

2.2 Saving the Data

[19]:
```
df_ranFor
```

[19]:
```
      X_Minimum  X_Maximum  Pixels_Areas  Sum_of_Luminosity  \
0     -1.135234  -1.194042     -0.273978          -0.305749
1      0.030147  -0.008917     -0.393159          -0.397664
2      0.385752   0.353917     -0.420893          -0.422214
3      0.432136   0.403215     -0.342188          -0.343194
4      1.274766   1.282692      1.331600           1.290626
…           …          …            …                 …
1936  -0.735178  -0.746416     -0.269480          -0.228242
1937  -0.938105  -0.947552     -0.258986          -0.231353
1938  -0.936172  -0.949524     -0.255238          -0.210042
1939  -0.951633  -0.957411     -0.160043          -0.101498
1940   1.220652   1.233394     -0.396907          -0.395621

      Minimum_of_Luminosity  Maximum_of_Luminosity  Length_of_Conveyer  \
0                 -0.539531              -1.268667            1.454519
1                 -0.244926              -0.364878            1.454519
2                  0.307458              -0.244372            1.028131
3                  0.307458              -0.184120           -0.770692
4                 -1.975729              -0.184120           -0.770692
…                        …                      …                  …
1936               1.043971               0.719669           -0.724056
1937               0.786191               0.237649           -0.724056
1938               1.080796               0.659417           -0.724056
1939               0.970319               0.659417           -0.724056
1940               0.381110               0.237649           -0.724056

      Steel_Plate_Thickness  Outside_X_Index  Edges_Y_Index  LogOfAreas  \
0                  0.035167        -0.514948       0.695123    0.237173
1                  0.035167        -0.556364       0.526690   -0.417695
2                  0.475065        -0.552599       0.429070   -0.721056
3                  4.654098        -0.496123       0.695123   -0.064356
4                  2.344632        -0.217504       0.636955    1.828612
…                        …                …             …           …
1936              -0.844629         0.083706      -0.684202    0.253332
1937              -0.844629         0.166538      -0.780305    0.289482
1938              -0.844629         0.110061      -0.914342    0.301977
1939              -0.844629         0.223015      -1.026631    0.563191
1940               0.035167        -0.138436      -0.684202   -0.452012

      Orientation_Index  Luminosity_Index  Class
0              1.406965         -1.140097      0
1              1.356268         -0.313702      0
2              1.100962          0.063426      0
```

```
3          1.459885      -0.179421    0
4          1.640457      -0.482266    0
…               …             …    …
1936      -1.111353       0.959104    1
1937      -1.157808       0.524836    1
1938      -1.220827       0.977674    1
1939      -0.368057       0.818395    1
1940      -0.649621       0.126995    1

[1613 rows x 14 columns]
```

[20]:
```
df2_multiclass
```

[20]:
```
      X_Minimum  X_Maximum  Pixels_Areas  X_Perimeter  Y_Perimeter  \
0     -1.135234  -1.194042     -0.273978    -0.488992     0.088914
1      0.030147  -0.008917     -0.393159    -0.598267    -0.206576
2      0.385752   0.353917     -0.420893    -0.629489    -0.438747
3      0.432136   0.403215     -0.342188    -0.551435     0.110021
4      1.274766   1.282692      1.331600     0.182266     4.647910
…          …          …             …          …          …
1936  -0.735178  -0.746416     -0.269480     0.088602    -0.375428
1937  -0.938105  -0.947552     -0.258986    -0.067504    -0.333215
1938  -0.936172  -0.949524     -0.255238    -0.129947    -0.375428
1939  -0.951633  -0.957411     -0.160043     0.759861     0.152233
1940   1.220652   1.233394     -0.396907    -0.348496    -0.375428

      Sum_of_Luminosity  Minimum_of_Luminosity  Maximum_of_Luminosity  \
0             -0.305749              -0.539531              -1.268667
1             -0.397664              -0.244926              -0.364878
2             -0.422214               0.307458              -0.244372
3             -0.343194               0.307458              -0.184120
4              1.290626              -1.975729              -0.184120
…                   …                     …                     …
1936          -0.228242               1.043971               0.719669
1937          -0.231353               0.786191               0.237649
1938          -0.210042               1.080796               0.659417
1939          -0.101498               0.970319               0.659417
1940          -0.395621               0.381110               0.237649

      Length_of_Conveyer  TypesOfSteel_A300  …  Luminosity_Index  \
0               1.454519           1.091255  …         -1.140097
1               1.454519           1.091255  …         -0.313702
2               1.028131           1.091255  …          0.063426
3              -0.770692          -0.915808  …         -0.179421
4              -0.770692          -0.915808  …         -0.482266
…                    …                  …  …               …
1936           -0.724056          -0.915808  …          0.959104
```

```
1937          -0.724056          -0.915808  …          0.524836
1938          -0.724056          -0.915808  …          0.977674
1939          -0.724056          -0.915808  …          0.818395
1940          -0.724056           1.091255  …          0.126995

      SigmoidOfAreas  Pastry  Z_Scratch  K_Scratch  K_Scratch  Stains  \
0           0.178899       1          0          0          0       0
1          -0.703548       1          0          0          0       0
2          -0.962871       1          0          0          0       0
3          -0.010774       1          0          0          0       0
4           1.478005       1          0          0          0       0
…                …     …        …        …        …      …
1936        0.624165       0          0          0          0       0
1937        0.909918       0          0          0          0       0
1938        0.569750       0          0          0          0       0
1939        1.452819       0          0          0          0       0
1940        0.015345       0          0          0          0       0

      Dirtiness  Bumps  Class
0             0      0      0
1             0      0      0
2             0      0      0
3             0      0      0
4             0      0      0
…           …    …    …
1936          0      0      1
1937          0      0      1
1938          0      0      1
1939          0      0      1
1940          0      0      1

[1613 rows x 31 columns]
```

[21]:

```
df_ranFor.to_csv("data/norm_data.csv", index=False)
df2_multiclass.to_csv("data/norm_multiclass_data.csv", index=False)
```