1. sometime problem with how datasets been saved to load before processing the data, so i need to make it a standard before analyzing
    - 3 types of data saving container extension normally used to save (will used in future)
        - .pkl
            - for .pkl file, we need to analyze first what kind of data exist inside
            - asked which keyname is wavenumber and which keyname is spectrum intensity data (give selection by checking if the value is list, df, np)
            - can group data based on keyname with value string
        - .csv
            - for csv also need to declare which data is wavenumber, as other is spectrum intensity data
        - .txt
            - asked user to give same input format file
            - asked user punctuation/symbol used to seperate data
            - how many column data
2. processing data will be save in dataframe/dict

```
{ 'benign': {'benign': 50, 'cancer': 29},
'cancer': {'benign': 16, 'cancer': 202}}
(' precision recall f1-score support\n'
'\n'
' benign 0.76 0.63 0.69 79\n'
' cancer 0.87 0.93 0.90 218\n'
'\n'
' accuracy 0.85 297\n'
' macro avg 0.82 0.78 0.79 297\n'
'weighted avg 0.84 0.85 0.84 297\n')
CV Accuracy: 0.845 ± 0.025
Decision Function Score: 0.357 ± 0.642
```

PROF

```
Predicting 1450 test samples with 27947 features.
{ 'label_percentages': { np.str_('benign'): 0.16206896551724137,
np.str_('cancer'): 0.8379310344827586},
'most_common_label': np.str_('cancer')}
```