2024/11/19

1. 修改 split data shuffle=True 因為資料很少，所以切完可能有些欄位會只有部份值存在。(例如在test set只有sometimes沒有其他no, always, frequently)
2. 修改 encoding，只使用 label encoder，因為有些欄位unique values不會出現在另外一個dataset，會導致欄位不一致。（例如 CALC\_Frequently, CALC\_Always沒有出現在test set)
3. 修改 drop columns 的程式，保留原有欄位順序，在train/predict的時候欄位順序必須一致
4. train, test, processed的檔案都有改變

Standardization, Normalization

Missing values (by mean, median, most\_frequent)

Feature Selection

* Removing low variance
* univariate feature selection
* select from model (random forest)

