Continuous Assessment - I
Introduction to Machine Learning ITC 2252
2022/04/29

1. Scenario

The file "molding_machine.csv" consists with machine operation data and quality results of productions on an injection molding machine (plastic manufacturing machine). Quality results are listed in the variable RESULT_QUALITY. Further, Good is good product, others (ColorUnevenness, Short, Mera) are defective products.

Client requirement

It is client's need to analyze which variable affects defective product generation for each type of defective product. Please perform analysis according to the following procedure.

1.1 Development environment

Python3.x
scikit-learn
Jupyter notebook

1.2 Submission

Jupyter notebook file

1.3 Analysis procedure

① Read the file " molding_machine.csv" and create the dataframe.

② Count the number of Good, ColorUnevenness, Short, Mera.

③ In the following, we will first analyze variables that have a large influence on the occurrence of ColorUnevenness.
Extract only the data of Good and ColorUnevenness

④ Calculate the number of missing values in each variable.

⑤ Calculate the basic statistics (average value, standard deviation, median value, maximum value, minimum value, etc.) of each variable with Good and ColorUnevenness as the aggregation axis

and output the result as csv file.

⑥ For each variable, investigate outliers, and if there are outliers, delete that line. Please survey outliers by visualizing the data.

⑦ Delete the variables which has more than 20% percentage of missing value. Others fill the missing values.

⑧ Identify the variables that can be used to distinguish Good from ColorUnevenness
Hint : Using two different colors for Good and ColorUnevenness and draw a histogram of each variable. Suppose there is a variable whose distributions of Good and ColorUnevenness are not overlap each other. In that case, we judge that the variable is a variable that distinguishes Good from ColorUnevenness.

Please submit on or before 15_05_2022