**Refine your data understanding and prepare a DFD of solving the manufacturing quality problem**

The First Phase is the preprocessing steps to configure the SECOM dataset classified as predictive models. The SECOM dataset using the data cleaning, Feature selection. Pre-processing include data cleaning and feature selection method using SECOM dataset. Divide the SECOM dataset into training set (70%), testing set (30%). Oversampling uses as SMOTE. SMOTE is 1:2 balance and configured.

The second phase is to generate the prediction models and evaluation, Training set by the prediction model creation and utilization, LR, ANN, DT, RF to use.

In order to evaluate the prediction models used the confusion matrix.

The procedure for generate the fault prediction models including SMOTE based oversampling are as follows:

**Data cleaning:**

* Count in each attribute ‘not available data’ or missing values. If record set are missing more than 60%, then remove the attribute

**Feature selection:**

* Apply the PCA based future selection

**Oversampling**

* To balance the pass/fail used SMOTE based over sampling.

**Prediction model build:**

* Build a fault prediction model with LR, ANN, DT (C.4.5) and RF.
* Using the confusion matrix compares the precision, recall (sensitivity) and F-measure. Confusion matrix (TP: True Positive, FP: False Positive, FN: False Negative, TN: true Negative)

