Challenge

Conformity/quality asessement of products
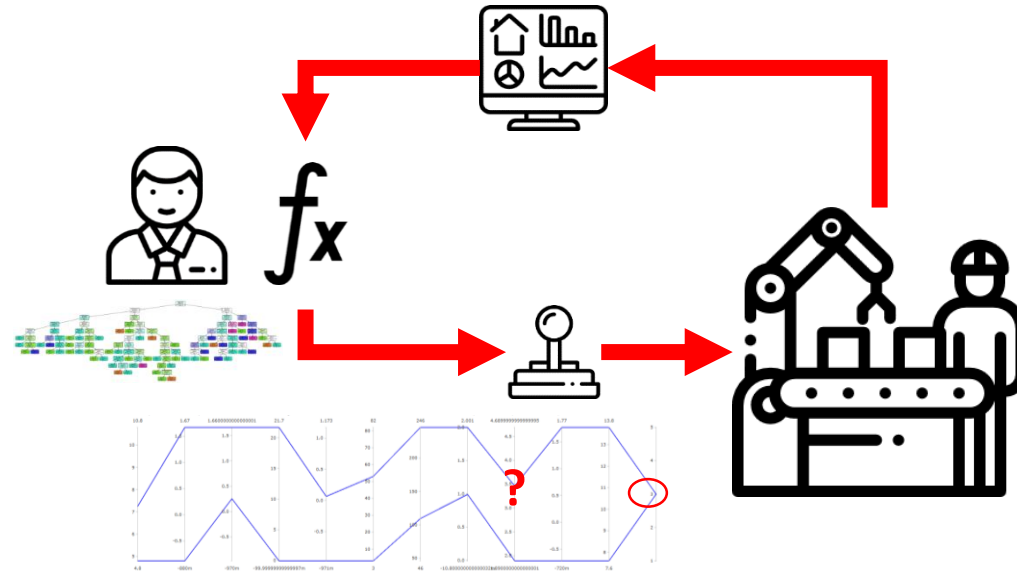
Key Process Parameter identification
Extraction of rules

Decision Tree ? Accuracy ?
Rules ?
Compared to PCA Results ?
Other clasifications? Why not?

Key Process Parameter identification
Definition of a regression model – assessment of the accuracy of the model

Rgression model , confusion matrix ?
Accuracy ?
Model reduction ?

Comments for PCA Use in R

```r
setwd("D:/Enseignement/Enseignement_ENSAM/Maths/Formation Data science/ACP")
rm(list=ls())
chooseCRANmirror(ind=29)

install.packages("FactoMineR")
install.packages("factoextra")
install.packages("corrplot")
library("FactoMineR")
library("factoextra")
library("corrplot")

donnee<-read.csv("factory_process_ACP5_KC4.csv", header=TRUE, sep=";")

res<-PCA(donnee,axes=c(1,5), graph=F)
vp<-fviz_eig(res)
var<-fviz_pca_var(res,axes=c(1,5))
ind<-fviz_pca_ind(res,axes=c(1,5),col.ind=donnee$KC4,label="none",gradient.cols="red")
bip<-fviz_pca_biplot(res,axes=c(1,5),col.ind=donnee$KC4,label="var",gradient.cols="red")
M <- cor(donnee)
corrplot(M, order = "hclust", addrect = 5)
print(ind)
print(vp)
print(bip)
print(var)
pdf("ACP15_KC4.pdf")
print(vp)
print(var)
print(ind)
print(bip)
corrplot(M, order = "hclust", addrect = 5)
dev.off()
```

Working directory address

Data file name (.csv)

PCA projection axis

Name of the considered KC

Number of clusters for the correlation analysis

Result file name (.pdf)

I have received some queries about the challenge. I have explained it as well during my last lecture and asked you for questions or comments and you have mentioned that you have understood the issues.

I will explain in this e-mail the tasks to work on . Be careful, the objective is not to apply what I have given and that's all. I think that you are in Master Degree and more tasks should be handled. An application of my codes is not an exam! The python and R codes are given to show you the outcomes and you can use them if necessary!

- You have two data (one with KCs values and one with conformity label (0 or 1) for the same KCs)
  The first data is not cleaned as well. One first task is to work on the pre-processing phase in order to let the data be used for AI techniques. The problems in the data are generally: missing values, defaults, … please take a look on the data before any application. For the pre-processing, you can see in Sklearn if you want or apply any other techniques like KNN,….

- Once the date (with Ks values) is cleaned, a second task is to determine most significant parameters to consider : we can apply PCA for example

- The third task is to identify most important rules for conformity to be used to build models after : we van focus on Decision Tree in this case; please identify the confusion matrix to assess the accuracy of the Decision Tree model. Which branch of the tree to select (with most significant informations, …)

- Then, models (basically regression) could be identified from extracted rules . the accuracy of the model should be studied (confusion matrix use for example) and then reduced to more precise and efficient (increase its accuracy)/ for the reduction you can focus on the technique used during the class or other , you can also used PCA results , you can use SHAP technique (you have the order of the significant parameters by SHAP), …

The length of the report is not too important for me also codes. The main aim of the project is to be able to interpret the results and to employ them as well
Please synthetize your work and identify conclusions of each part of the challenge.