DONNEES RAFFINERIE 2020

Ce jeu de données est issu de données industrielle provenant de diverses raffineries dans le monde.

Les raffineries mettent en œuvre des procédés (de raffinage) qui ont pour objectif de transformer une entrée (une charge pétrolière) en sorties (ou effluents) plus aptes à être valorisées

On souhaite prédire certaines propriétés des sorties d'un procédé donné en fonction des entrées

Les données ont déjà été partiellement prétraitées.

En entrée, nous disposons des informations suivantes :

1. Densité de la charge "FEED_D154"
2. Soufre (de la charge) "FEED_SULFUR"
3. Azote de la charge "FEED_NITROGEN"
4. Mean Average Boiling Point de la charge "FEED_MeABP" : c'est une information employée par les praticiens.
5. Kwatson "FEED_KWATSON": c'est un index utilisé par les raffineurs.
6. Distillation de la charge "FEED_DS_005", "FEED_DS_010", "FEED_DS_030", "FEED_DS_050","FEED_DS_070", "FEED_DS_090", "FEED_DS_095" : c'est un ensemble de points qui relient le pourcentage (volumique ou poids suivant les méthodes) distillé à la température nécessaire pour obtenir le pourcentage en question.
7. Distillation de la coupe GO (Gas-Oil) "DIESEL_DS_005", "DIESEL_DS_010", "DIESEL_DS_030", "DIESEL_DS_050", "DIESEL_DS_070", "DIESEL_DS_090", "DIESEL_DS_095"
8. TMP : c'est une donnée sur la charge exploitée par les raffineurs.
9. Quelques données procédés : "Proc1",
10. "Proc2",
11. "Proc3"

La sortie est dénommée "actual.values".

Les données sont organisées par site et par cycle de maintenance. Le numéro du site est donné dans la dernière colonne des data. On a rajouté dans le fichier 'CycleNumber.txt' le numéro du cycle. Les dates des observations sont dans le fichier 'Date.txt'.

Les données ont été anonymisées de la façon suivante :

- Le minimum des valeurs des items 1,2,3,4,5 et 8,9,10 est forcé à 0, le maximum à 1
- Le minimum du minimum des courbes 6 et 7 est forcé à 0, le maximum du maximum à 1.
- La sortie a pour minimum 0 et maximum 1.

Les données contiennent beaucoup de valeurs manquantes. Elles contiennent aussi beaucoup de valeurs suspectes (ou aberrantes).

En sortie, nous souhaitons modéliser la sortie à l'aide d'un modèle linéaire. Ce modèle peut être construit site par site ou globalement (tous les sites confondus). L'écart de prédiction acceptable est de 0.005.

On propose la démarche suivante

- Lire les données. Les données sont stockées en format Matlab dans un fichier nommé 'Database.mat'.
- Les repésenter (histogrammes, ACP, représentation site par site, 'pairplots', 'parallel coordinate plots', …
- Préparer les données :
- Supprimer les colonnes inutiles. Certaines variables sont redondantes.
    - En effet, le TMP se déduit des valeurs de la distillation.
    - Les valeurs à 30, 50 et 70% des distillations se déduisent avec une très grande fiabilité des autres valeurs de la distillation
    - La MeABP est une donnée construite aussi à partir de la distillation.
- Supprimer les lignes inutiles. Pour qu'une observation soit prise en compte dans les régressions, il faut qu'elle ne contienne aucune valeur manquante.
- Supprimer les sites insuffisamment renseignés. Pour être pris en compte, un site doit contenir au moins 50 valeurs.
- Éliminer si c'est faisable les valeurs suspectes. On suggère ici de faire une moyenne mobile sur une période de 31 jours par site et cycle. Construire autour de cette moyenne mobile des intervalles de confiance en supposant les données gaussiennes avec un risque de 2% (suggéré) de se tromper.
- Construire une base de calibration et de validation. Plusieurs stratégies sont possibles : par Kennard and Stone, ou par 'stratified sampling'. On veillera à ce que les bases de calibration (et surtout de validation) contiennent des observations de tous les sites.
- Construire les modèles (site par site et global). Le critère de sélection du modèle est à choisr parmi plusieurs possibilités :
    - RMSEP ('Root Mean Square Error of Prediction').
    - MAE ('Mean Absolute Error')
    - …
- Construire les modèles (site par site et global). Là encore plusieurs stratégies sont possibles.
    - On peut utiliser lm si l'on estime que les hypothèses des modèles linéaires sont vérifiées.
    - Ou, on peut procéder de façon systématique en fabriquant tous les modèles possibles et ne retenant que ceux dont la RMSEP est la meilleure. Cette stratégie est possible car le nombre de variables est assez faible (15 au final), et cela fait en tout

2^15 modèles=32768 régressions, ce qui est largement faisable avec les puissances de calcul actuelles.

- D’autres possibilités existent :
  - la régression quantile, adaptée au critère MAE comme la régression aux moindres carrés est adaptée au critère RMSE.
  - PLS ou autre compte tenu qu’on a beaucoup de variables. L’avantage de la PLS est qu’on s’occupe peu des redondances entre variables.