Titre du rapport

SOUS-TITRE DU RAPPORT

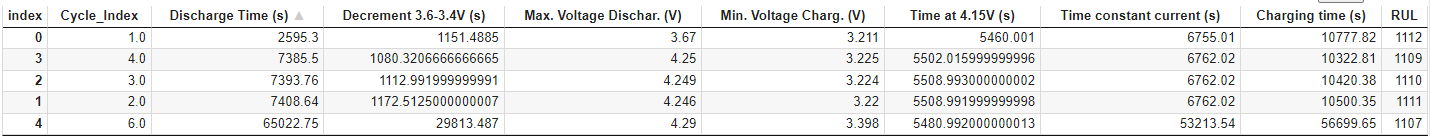
Nom | Intitulé du cours | Date

# Exploitation des données

Dans un premier temps, nous ouvrons la base de données avec google collab et la bibliothèque Pandas. Nous pouvons ainsi visualiser les données

Une image contenant texte, Police, capture d’écran, blanc

Description générée automatiquement



Nous pouvons ensuite a l’aide de la bibliothèque mathplotlib , afficher quelques courbes relatives aux données.

Cette premiere etape permettra de remarquer d’eventuels couplages entre les features, ce qui nous sera très utile pour réaliser le feature ingeniering

Une image contenant capture d’écran, texte, diagramme, ligne

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

Une image contenant texte, diagramme, carte, capture d’écran

Description générée automatiquement

Une image contenant ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant diagramme, ligne, Tracé

Description générée automatiquementUne image contenant diagramme, ligne, Tracé, texte

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

# Features Engineering

Pour le feature engineering, nous allons réaliser les étapes suivantes :

* Sélection des variables pertinentes
* Combinaison des variables redondantes
* Suppression des valeurs aberrantes

Dans un premier temps , on réalise la matrice de corrélation de data Set , présentée ci-dessous sous forme de heatmap.

Une image contenant texte, capture d’écran, diagramme, Parallèle

Description générée automatiquement

Cet outil nous permet de visualiser facilement les feature forement correléles.

On remarque notamment que charging time et discharging time sont très fortement coréles. Prendre en compte ces deux paramètres serait assez redondant. On purra alors , soit combiner les deux feature , soit siplement ne prendre en compte qu’un des deux.

La matrice de correlation nous permet également de sélectionner les attributs particulieremnt pertinents pour la prediction du parametre RUL. En effet , on remarque que RUL est très corellé a max volage discharge, ce qui en fait un feature a prendre en compte en, priorité dans notre model de prediction de RUL.

Nous réalisant dans un second temps l’analyse en composantes principales du data SET.

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

Cet outil permet de visualiser les relations entre les données. En particulier , on peut visualiser facilement quels attributs devront avoir un poids important pour la prediction de RUL en comparant les pourcenages de variance expliquées. On remarque notamment que les composantes 5 , 6 et 7 ne seront pas très utiles pour notre modele comparées aux features 1 , 2 et 3 qui sont le discharge time , le decrement et le max voltage discharge.

Cette premiere phase d’etude des données permet de tirer les conclusios suivantes :

* Correlation entre discharge time et time constant current
* Correlation entre charging tie et discharging time
* Correlation relativement forte entre RUL et max voltage discharg
* le discharge time , le decrement et le max voltage discharge sont les composantes principales de RUL

Choix d’un model d’IA

En premier lieu , rul étant une valeur continue , nous nous concentrerons sur les algorithmes de régression.

Disposant d’une base de données, nous choisirons un entrainement supervisé.

Parmi les algorithmes les plus connus qi s’offrent a nous, il y a :

* Regression SvM
* Arbre de regression
* GpR KRieging
* Processus gaussien

Regression SVM

Cet algorithme est l’adaptation continue de la machine a vecteur de support, permettant de séparer les classes par un hyperplan. Cette technique sera particuliereemnt utile dans le cas d’une prédiction predant en entree de très nombvreux paramètres. Nous n,avons que 8 parametres dont certains très fortement correles , la situation ne semble donc as exploiter au mieux les points forts du SVM.

Processus Gaussien

Cet algorithme repose sur l’estimation de la distribution de probabilité de la variable de sortie en fonction des paramètres d’entree. Elle pourra s’adapter a des relations complexes entre les données , sera relativeemnt interpretable et permettra de quantifier facilemtn l’incertitude de la prediction. Cependant , il risquera d’etre assez exigeant en mémoire et assez lent .

Arbre de regression vs GPR Kriging

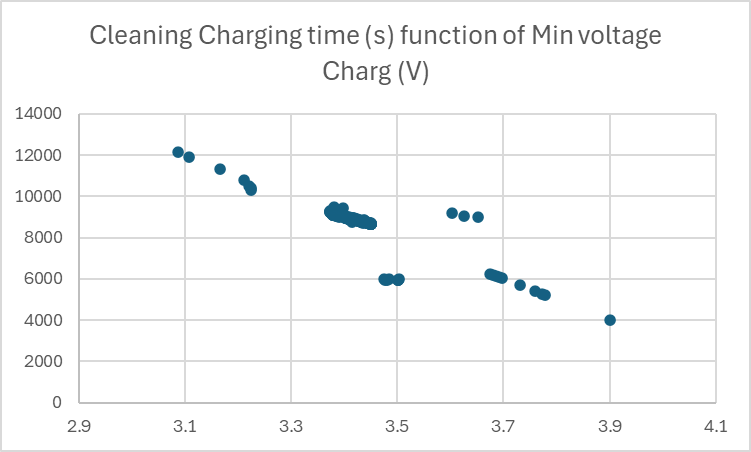
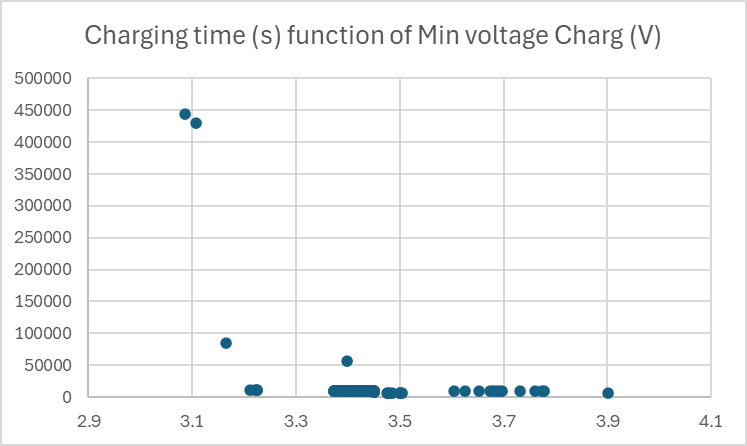
Ces deux algorithmes vont se distinguer sur 3 points : la complexitédes donnees , l’interpretabilité et la taille des données.

L’arbre de regression sra plus performant dans l cas de données simples, sera également plus rapide dans le cas de rands ensembles données , et aura l’avanage d’etre interpretable.

Le GPR krieging lui saura mieux exploiter des features reliées non linéairement , risquera d’etre plus lent dans loe cas de grands ensembles de données , mais sera probablement plus précis.

L’etude preliminaire de l’ensemble de données laisse supposer que les relation inter feature sont plutôt linéaires. L’arbre de decision devrait être efficace et rapide. Cependant , il sera intéressant d’implmenter un Gpr krieging afin de quantifier le rapport gain en precision sur perte en rapîdité afin de savoir quel model sera le plus pertinant.

En conclusion , nous implementerons le processus gaussion , le krieging et l’arbre de regression



Implementation et améliorations de la methode GPR