Travail sur la détection des anomalies (5 points)

Mise en situation

Vous êtes responsable de la mise en service d'un nouveau réseau de transport en commun structurant pour votre municipalité. Dans la phase de déploiement, vous devez mettre au point une technique qui permettra de détecter des anomalies dans votre système. Ainsi, vous pourrez détecter des signaux permettant le déclenchement de l'arrêt du système avant que des bris importants se produisent.

Or, les données en votre possession ne sont pas étiquetées pour pouvoir résoudre un problème de classification. Ainsi, vous devez mettre au point une technique non supervisé d'apprentissage machine afin de faire une détection d'anomalie sur votre système avant sa mise en service.

Données

Le jeu de données se trouve sur le site internet Kaggle sous le nom de MetroPT-3 Train Dataset et consiste à des signaux provenant de plusieurs capteurs sur un compresseur à Air à l'intérieur du train. Vous enregistrez 15 signaux (8 digitales et 7 analogues) sous une fréquence de 1Hz entre le mois de février et août 2020. Voici la liste de 15 signaux enregistrés :

- TP2 (bar) – la mesure de la pression sur le compresseur.
- TP3 (bar) – la mesure de la pression générée au panneau pneumatique.
- H1 (bar) – la mesure de la pression générée en raison de la chute de pression lorsque la décharge du filtre séparateur cyclonique se produit.
- DV pressure (bar) – la mesure de la perte de pression générée lorsque les tours des sécheurs d'air se déchargent ; une lecture nulle indique que le compresseur fonctionne sous charge.
- Reservoirs (bar) – la mesure de la pression en aval des réservoirs, qui devrait être proche de la pression du panneau pneumatique (TP3).
- Motor Current (A) – la mesure du courant d'une phase du moteur triphasé ; il présente des valeurs proches de 0A - lorsqu'il est éteint, 4A - lorsqu'il fonctionne sans charge, 7A - lorsqu'il fonctionne sous charge, et 9A - lorsqu'il démarre.
- Oil Temperature (ºC) – la mesure de la température de l'huile sur le compresseur.
- COMP - le signal électrique de la vanne d'admission d'air sur le compresseur ; il est actif lorsqu'il n'y a pas d'admission d'air, indiquant que le compresseur est soit éteint, soit en fonctionnement sans charge.
- DV electric – le signal électrique qui contrôle la vanne de sortie du compresseur ; il est actif lorsque le compresseur fonctionne sous charge et inactif lorsque le compresseur est éteint ou fonctionne sans charge.
- TOWERS – le signal électrique qui définit la tour responsable du séchage de l'air et la tour responsable de l'évacuation de l'humidité extraite de l'air ; lorsqu'il n'est pas actif,

cela indique que la tour un est en fonctionnement ; lorsqu'il est actif, cela indique que la tour deux est en opération.

- MPG – le signal électrique responsable du démarrage du compresseur sous charge en activant la vanne d'admission lorsque la pression dans l'unité de production d'air (APU) tombe en dessous de 8.2 bar ; il active le capteur COMP, qui assume le même comportement que le capteur MPG.
- LPS – le signal électrique qui détecte et s'active lorsque la pression tombe en dessous de 7 bars.
- Pressure Switch - le signal électrique qui détecte la décharge dans les tours de séchage d'air.
- Oil Level – le signal électrique qui détecte le niveau d'huile sur le compresseur ; il est actif lorsque l'huile est en dessous des valeurs attendues.
- Caudal Impulse – le signal électrique qui compte les sorties d'impulsions générées par la quantité absolue d'air circulant de l'APU aux réservoirs.

*** Cette section est une traduction libre de l'anglais des informations sur kaggle. La liste des caractéristiques a été traduite à l'aide de ChatGPT.

Travail à faire

À l'aide des librairies d'apprentissage machine standard, vous devez implémenter un modèle d'apprentissage machine non supervisé pour la détection des anomalies en format « batch » et un modèle en format « streaming ». À la suite de cet entraînement, vous devez faire une analyse des résultats obtenus et déterminer si une méthode apporte de meilleurs résultats versus un autre.

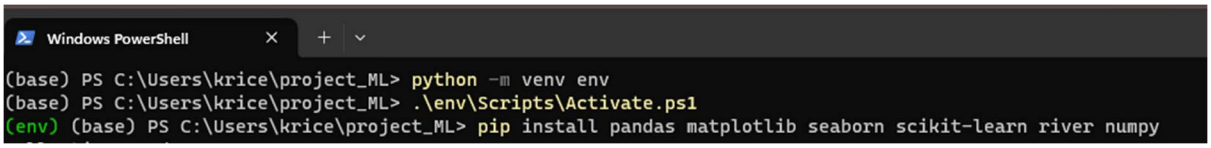
À fournir lors de la remise :

- Un script python pour l'entrainement des modèles ou un fichier Jupyter Notebook (1 point)
- Un rapport d'analyse contenant
    - Une brève explication des modèles utilisés. (1.5 points)
    - Une brève explication du concept de saisonnalité dans une série temporelle (1.5 points)
    - Une analyse comparative des résultats obtenues. (1 point)

Librairies python à utiliser (recommandation)

- Pandas pour la gestion des dataframes et des données
- Matplotlib/seaborn pour la création de visualisations
- Numpy pour le calcul matriciel en apprentissage machine
- Scikit-learn pour l'entrainement des modèles d'apprentissage machine en batch
- River pour l'entrainement des modèles d'apprentissage machine incrémentaux (stream)

Création d'un environnement virtuel python sous Windows



Étape de prétraitement des données pour ceux sans expérience en python

- Importation des données

```
import pandas as pd

df = pd.read_csv('./data/MetroPT3.csv')
df
```

- Transformation de l'index pandas avec la colonne « timestamp »

```
df['timestamp'] = pd.to_datetime(df['timestamp'])
df = df.drop(columns=['Unnamed: 0'], axis=1)
df.set_index('timestamp', inplace=True)
```

- Transformer le type des colonnes booléenne à bool

```
bool_column = ['COMP', 'DV_eletric', 'Towers', 'MPG', 'LPS', 'Pressure_switch', 'Oil_level', 'Caudal_impulses']


condition = df[bool_column] != 0.0
df[bool_column] = condition.astype(bool)
```

- Mettre à l'échelle les colonnes sous une échelle [0, 1]

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df_scaled = scaler.fit_transform(df)
df_scaled = pd.DataFrame(df_scaled, columns=df.columns)
df_scaled.index = df.index
df_scaled
```

- Entrainer un modèle pour la détection d'anomalie en batch et en streaming.

Site web avec la documentation technique pour River et Scikit-learn

- https://riverml.xyz/latest/
- https://scikit-learn.org/stable/

Suggestion de modèles

- IsolationForest (IF) -> Batch
- HalfSpaceTrees (HST) -> Streaming

Toutefois, vous pouvez utiliser d'autre modèle si vous le désirez.