Les methodes Autoencoder, SVM, LOF et DBSCAN ont été testées pour différents Scaler et différents paramètres d’initialisation de chaque méthode.

Pour SVM, LOF et DBSCAN, j’ai appliqué les scalers suivant :

- No Scaler, les données ne sont pas preprocessées et envoyées telles quel dans les méthodes.

- Standard Scaler, it removes the mean and scales the data to unit variance.

- MinMaxScaler rescales the data set such that all feature values are in the range [0, 1]

- MaxAbsScaler is similar to MinMaxScaler except that the values are mapped in the range [0, 1].

- Robust Scaler, the centering and scaling statistics of RobustScaler is based on percentiles and are therefore not influenced by a few numbers of very large marginal outliers.

- PowerTransformer applies a power transformation to each feature to make the data more Gaussian-like in order to stabilize variance and minimize skewness. Currently the Yeo-Johnson and Box-Cox transforms are supported and the optimal scaling factor is determined via maximum likelihood estimation in both methods.

- QuantileTransformer applies a non-linear transformation such that the probability density function of each feature will be mapped to a uniform or Gaussian distribution.

- The Normalizer rescales the vector for each sample to have unit norm, independently of the distribution of the samples.

On compare chaque résultat pour chaque scaler de sorte à choisir le scaler qui nous donne le résultat pour la détection d’anomalie.

Dans le but d’obtenir les meilleurs résultats pour chaque méthode, j’ai joué avec les paramètres de chaque méthode.

For SVM, I have modified the parameter “nu” which is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. En diminuant ce parameter, nous obtenons de meilleurs résultats pour la détection de comportements normaux mais de mauvais résultats pour la détection d’anomalie. En augmentant ce paramètre, on peut obtenir de meilleurs résultats dépendants des scalers.

Pour LOF, on va jouer sur le nombre de voisins pour le calcul de la déviation de densité d’un échantillon. En modifiant ce paramètre, on observe des comportements différents suivant les scalers, ce qui peut donner de bons résultats pour une faible valeur du nombre de voisins pour certains scalers alors que pour d’autres scalers, il faut un grand nombre de voisins pour avoir une détection correcte d’anomalie.

Pour DBSCAN, je joue avec les paramètres qui représentent la distance maximum entre deux échantillons pour qu’ils soient considérés dans le même échantillonnage et le nombre maximum d’échantillon dans un voisinage pour que cela soit considéré comme un core point. Ces deux paramètres sont paramétrés de sorte à ce que DBSCAN ne détecte un seul cluster, celui regroupant le comportement normal du processus et tout les outliers sont les points détectant une anomalie.