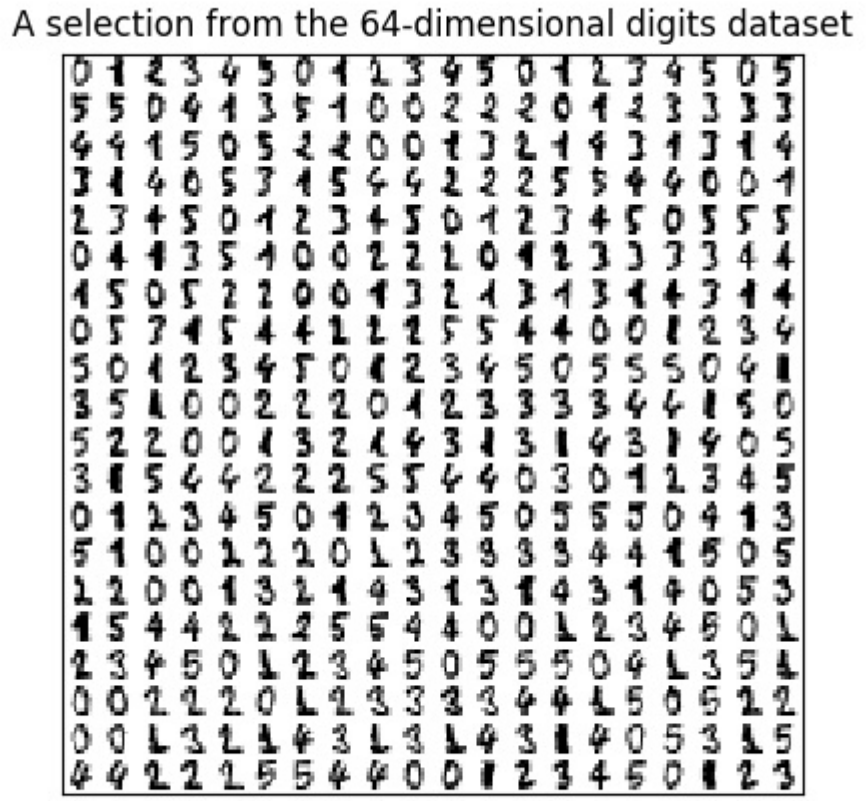
TP 2

Exercice 1

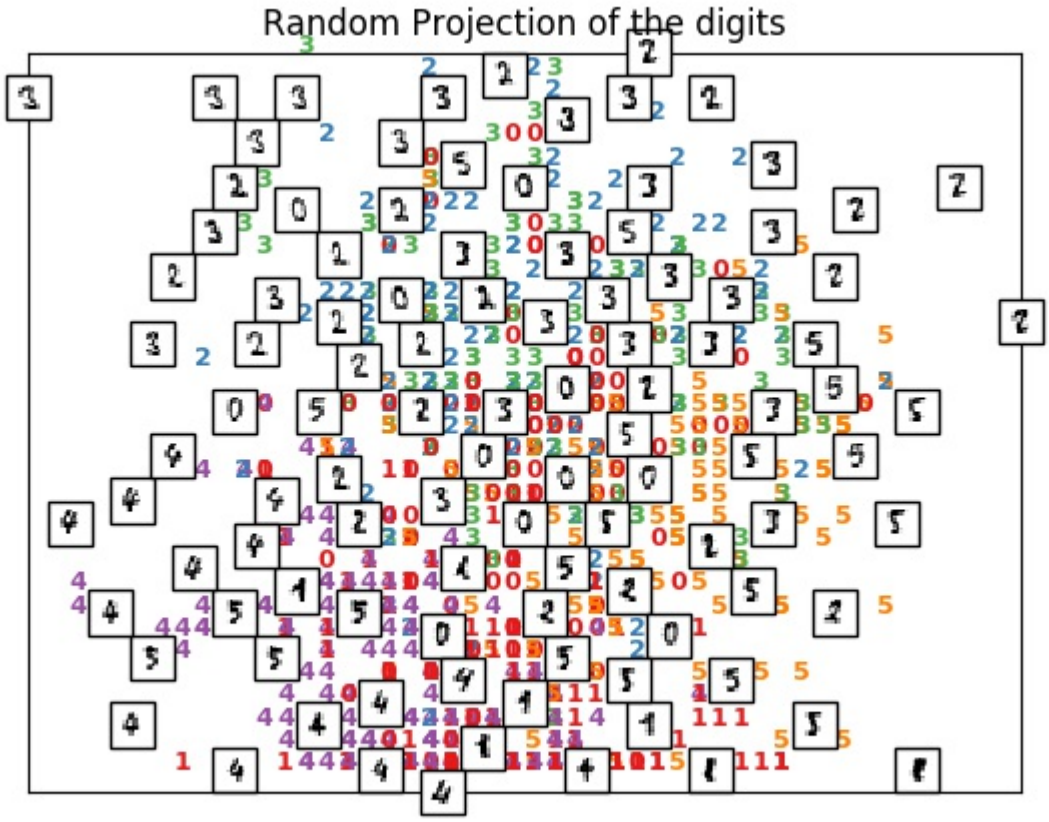
```
from sklearn.datasets.mldata import fetch_mldata

mnist = fetch_mldata('MNIST Original', data_home = custom_data_home)
```

- figure 1



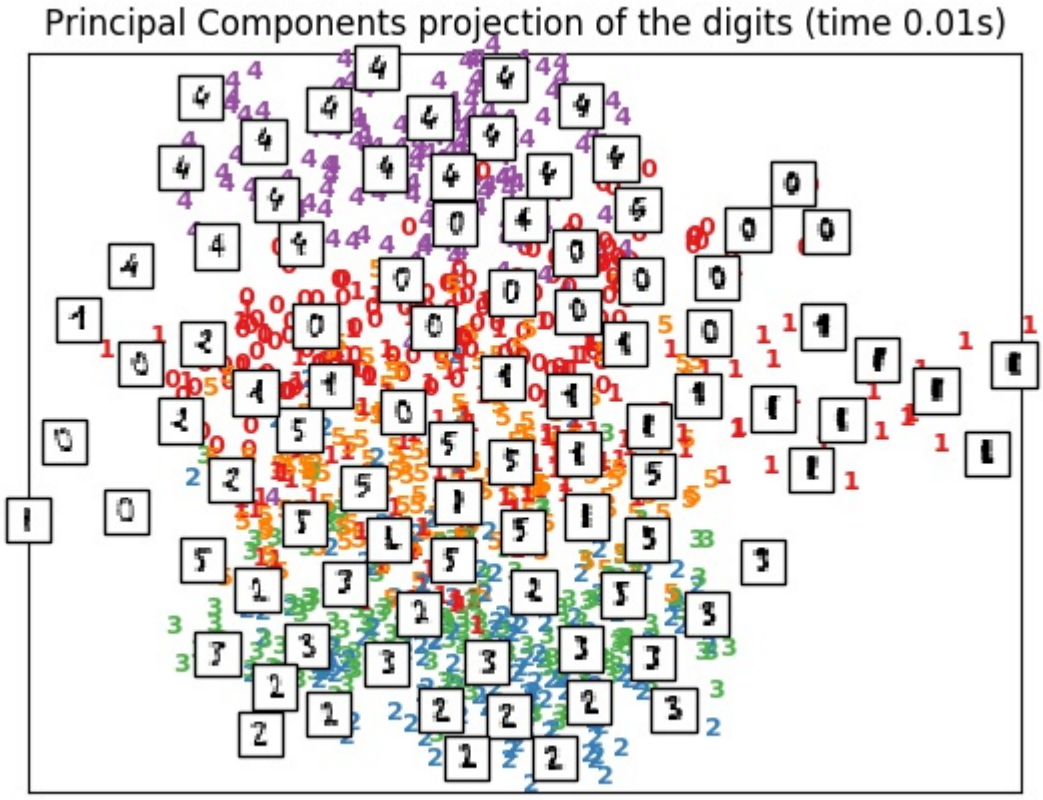
On a ici une représentation sous forme de matrice 20x20 des chiffres choisie aléatoirement dans une base de dimension 64.

- figure 2



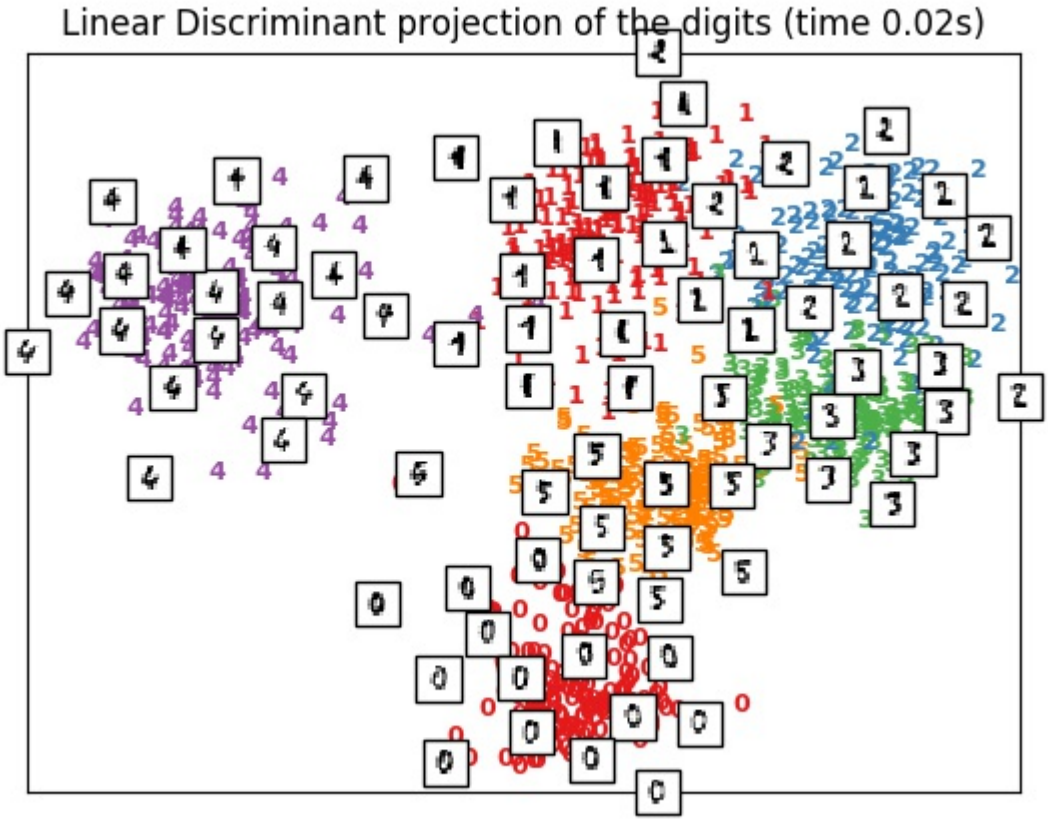
Nous avons ici une projection des points aléatoires des nombres.

- Figure 3



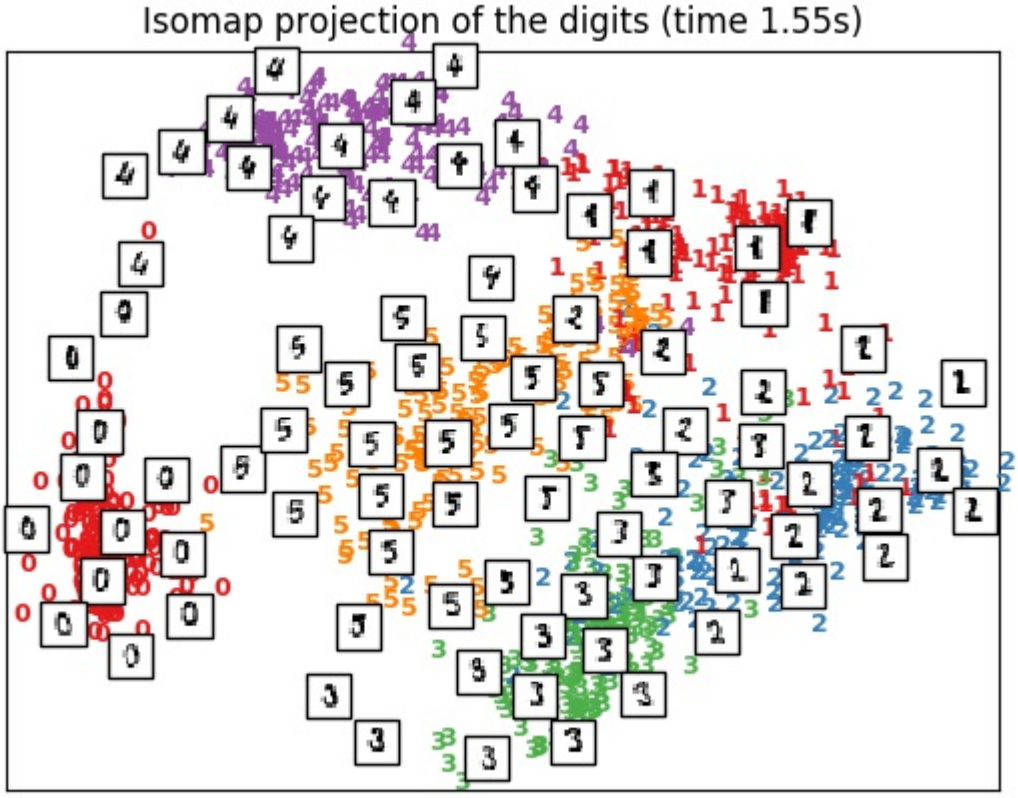
Nous avons ici un ensemble principal des projections des nombres. On remarque que les points sont un peu près réunit par bloc mais pas totalement.

- Figure 4



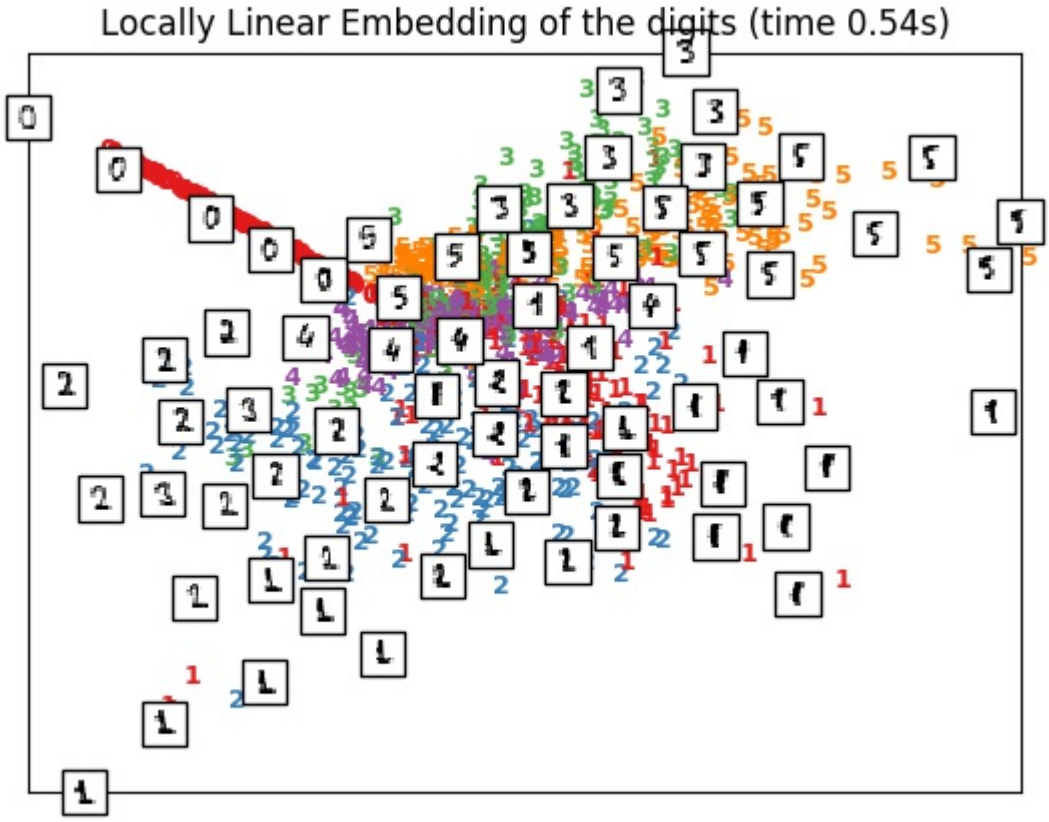
On voit déjà qu'il y a une meilleure représentation des points des nombres, en effet, il sont regroupé par une projection linéaire discriminatoire.

- Figure 5

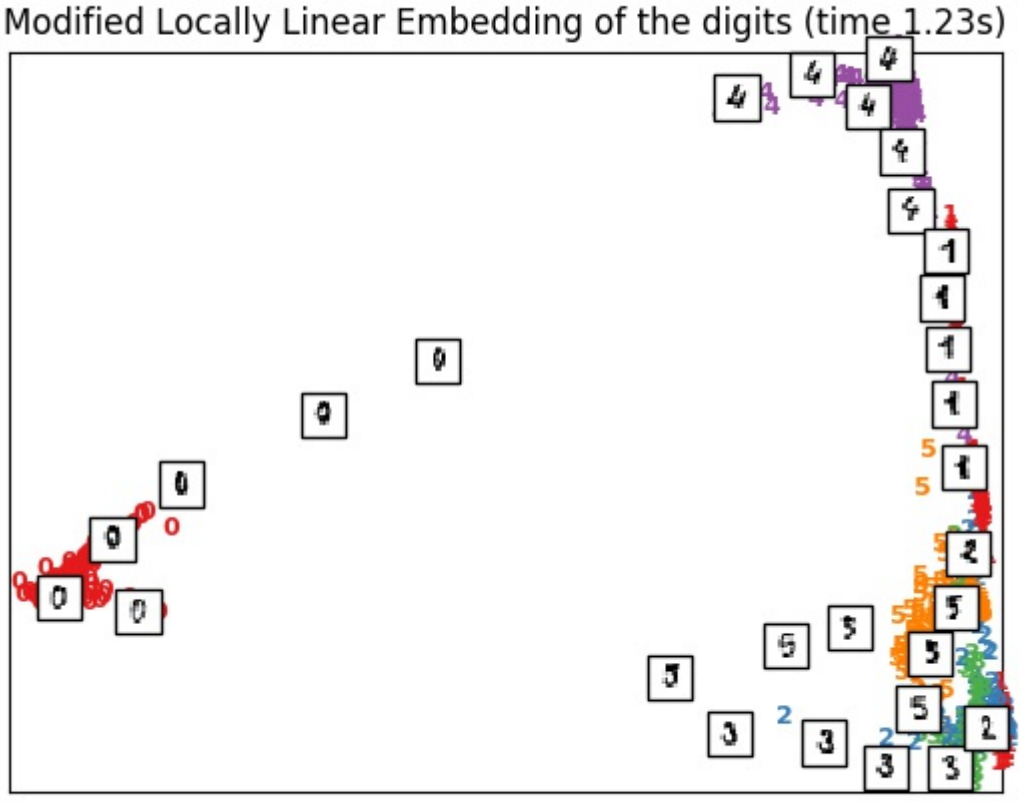


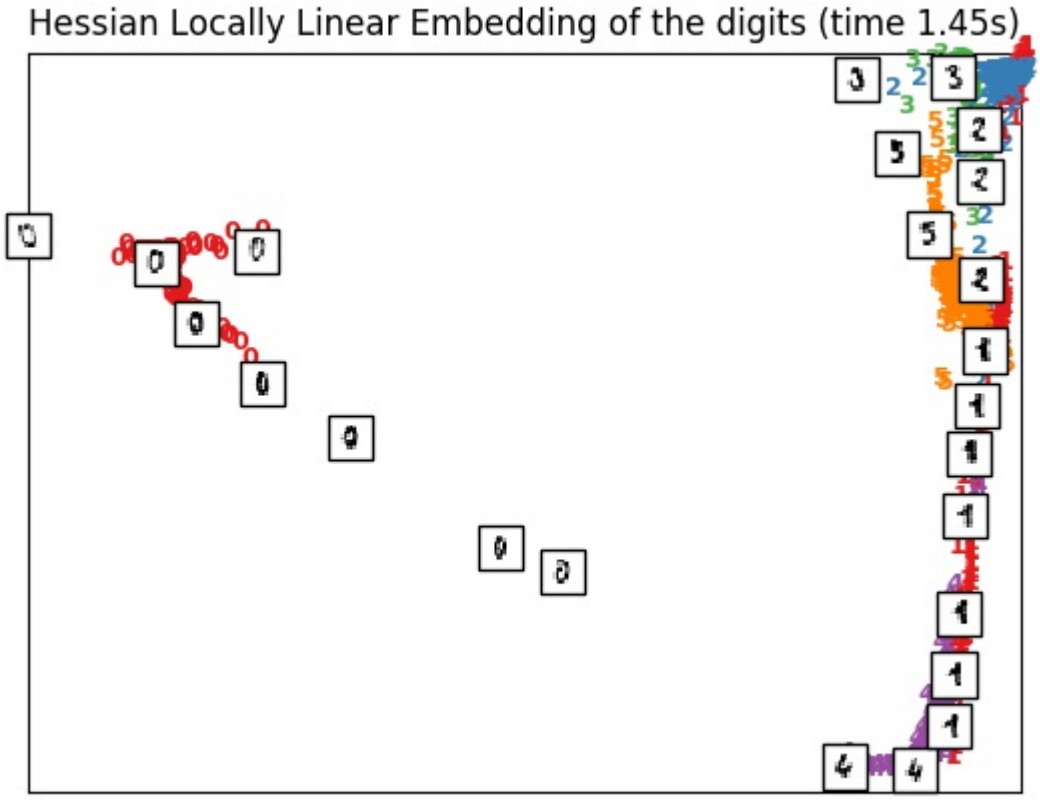
La projection isomap à réussi à regouper les points de nombres mais ce n'est pas la meilleur projection puisqu'il y a des petites quantités des nombres qui sont mélangées avec d'autres.

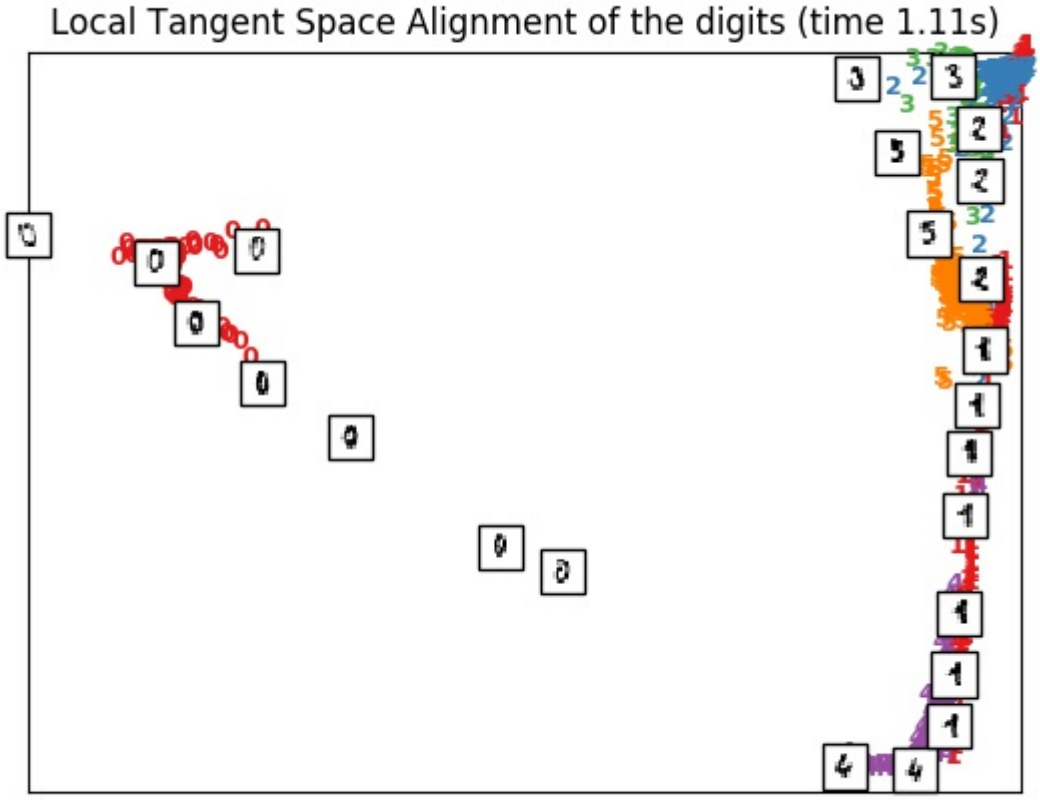
- Figure 6 (Intégration linéaire des nombres locaux)

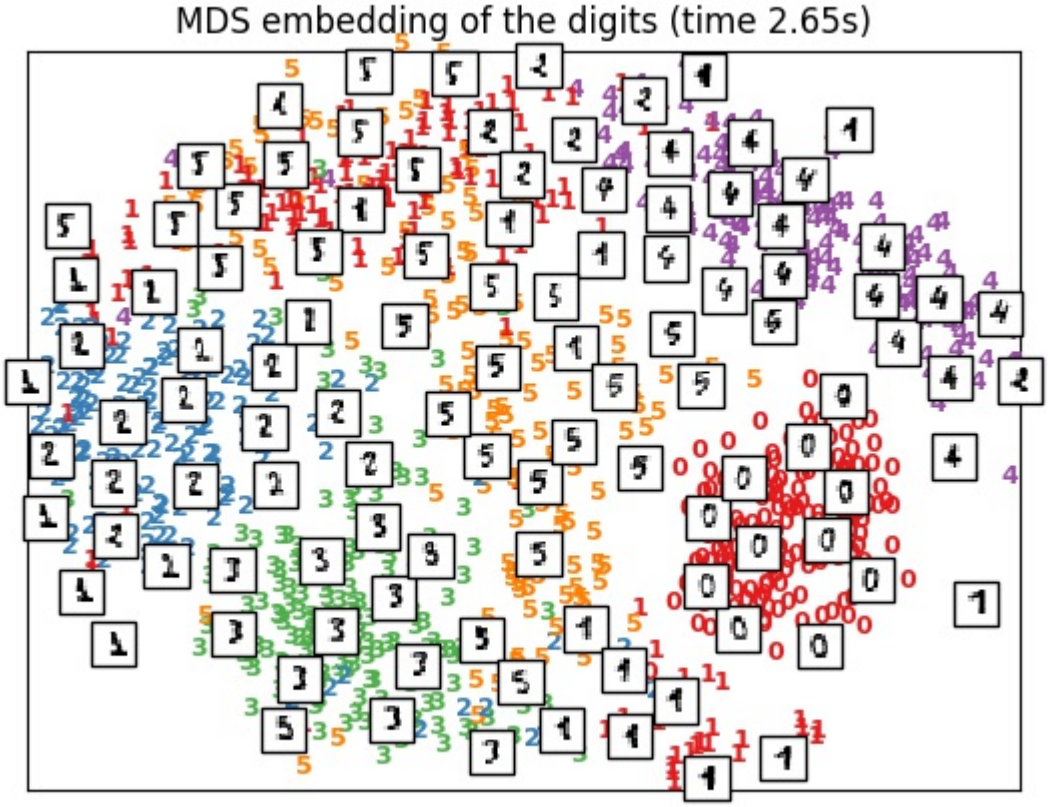


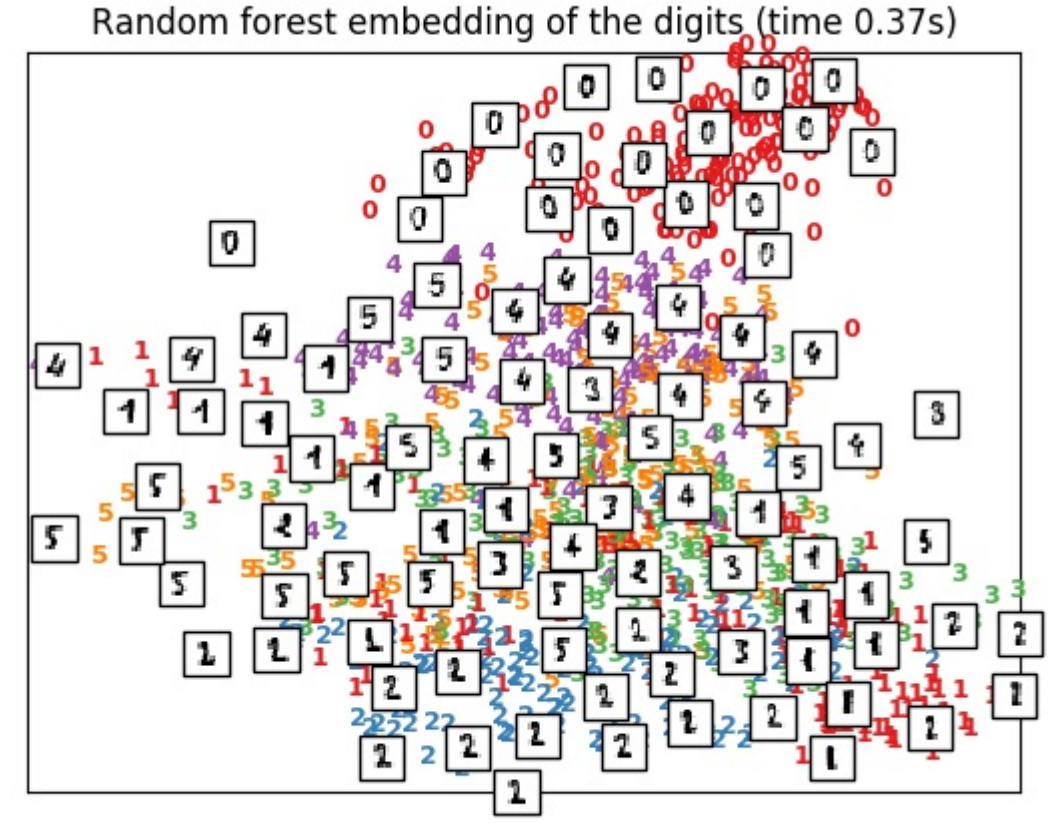
On a ici une intégration linéaire, mais

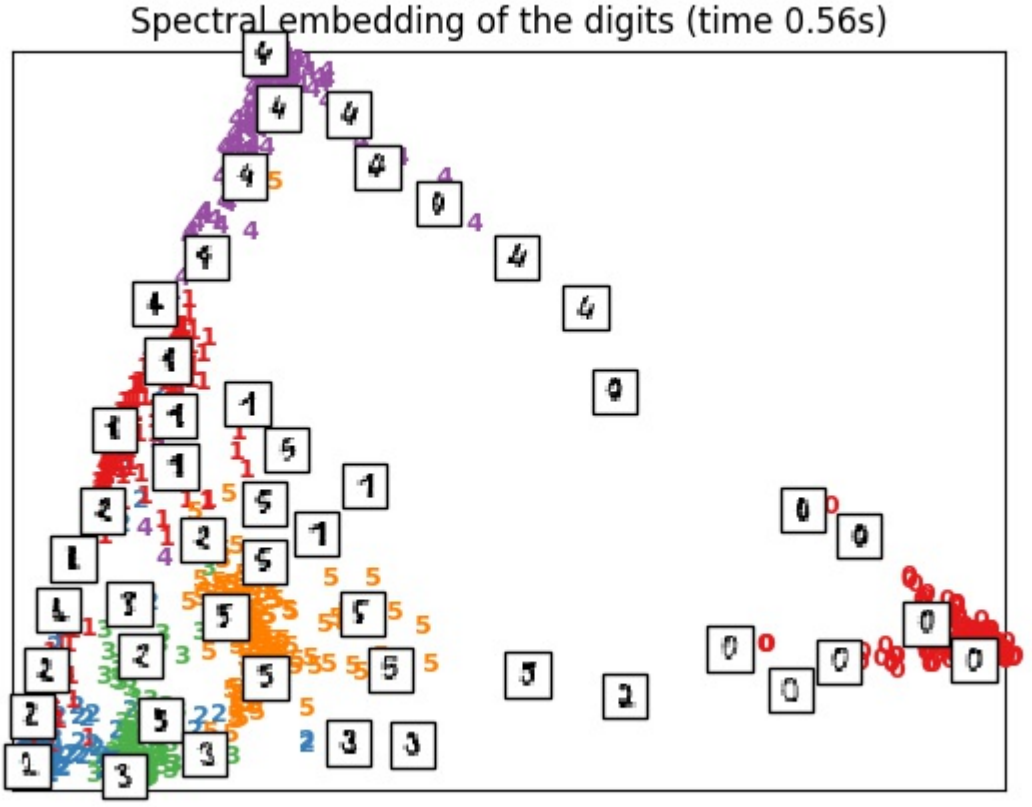


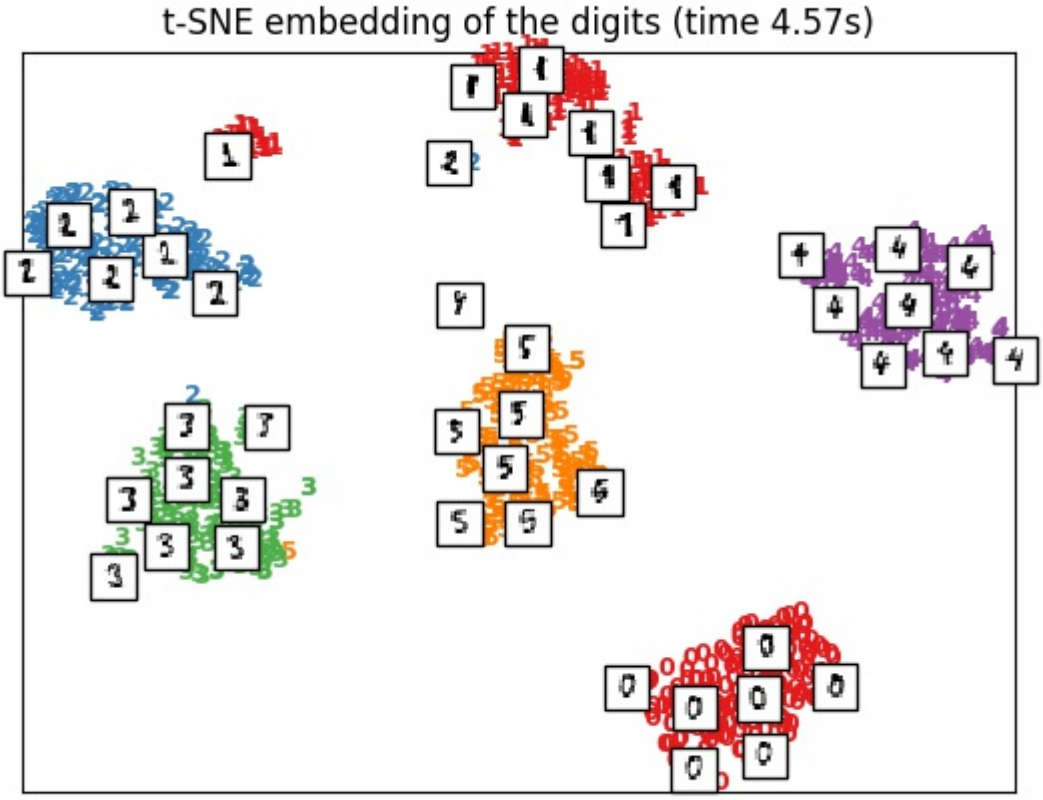


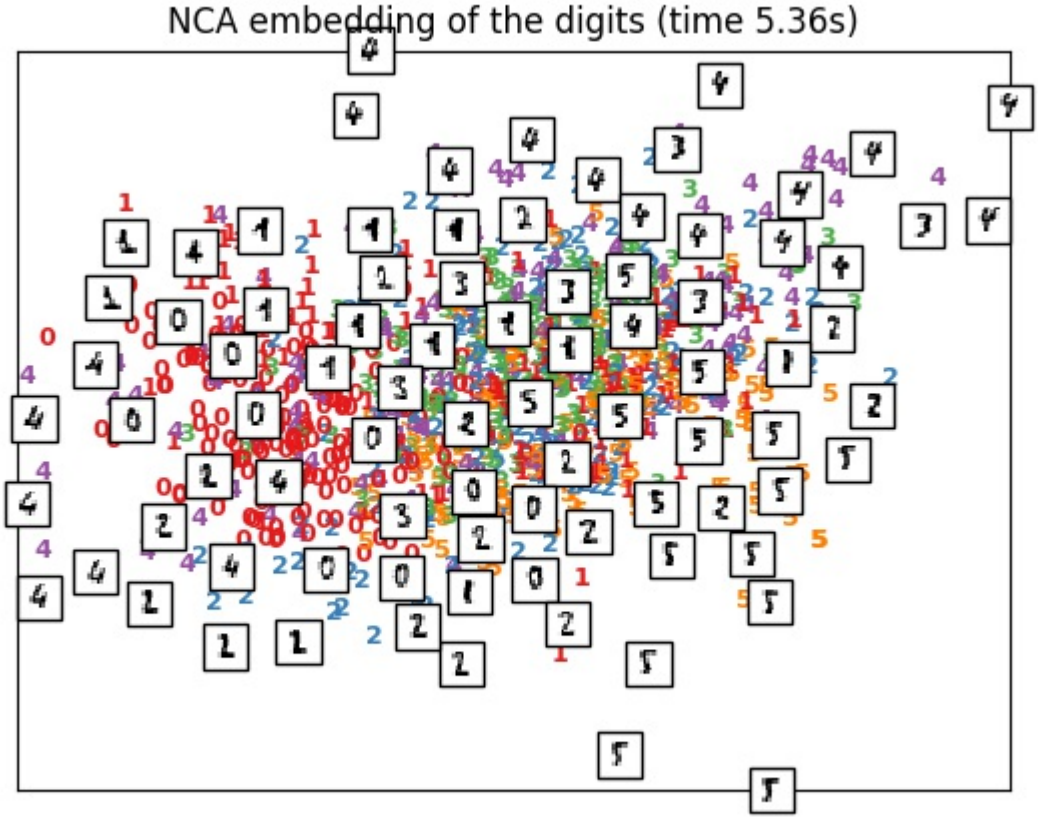










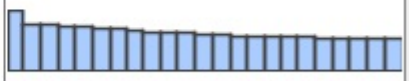


Partie 2

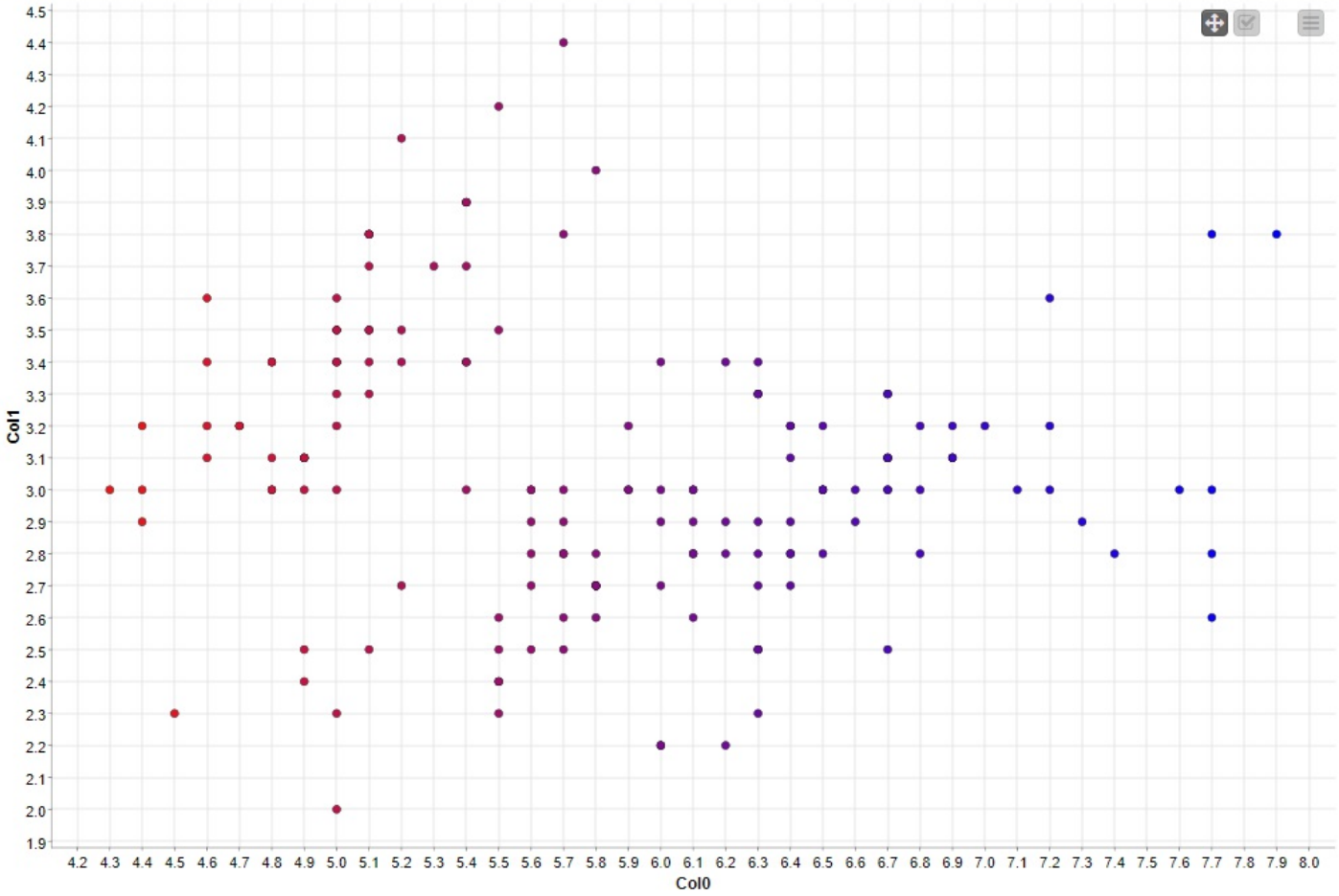
Exercice 1

1. Download the Knime software with all the pacakges v3.0.0
2. Load the Iris dataset using the ‘File Reader’ node in a knime project and tests use several nodes as Statistics and Plot to make a small analysis of the dataset.

- histogramme du noeud "Statistics"

Row ID	Column	No. mis...	Histogram
Col0	Col0	0	
Col1	Col1	0	
Col2	Col2	0	
Col3	Col3	0	
Col4	Col4	0	

- Nuage de points

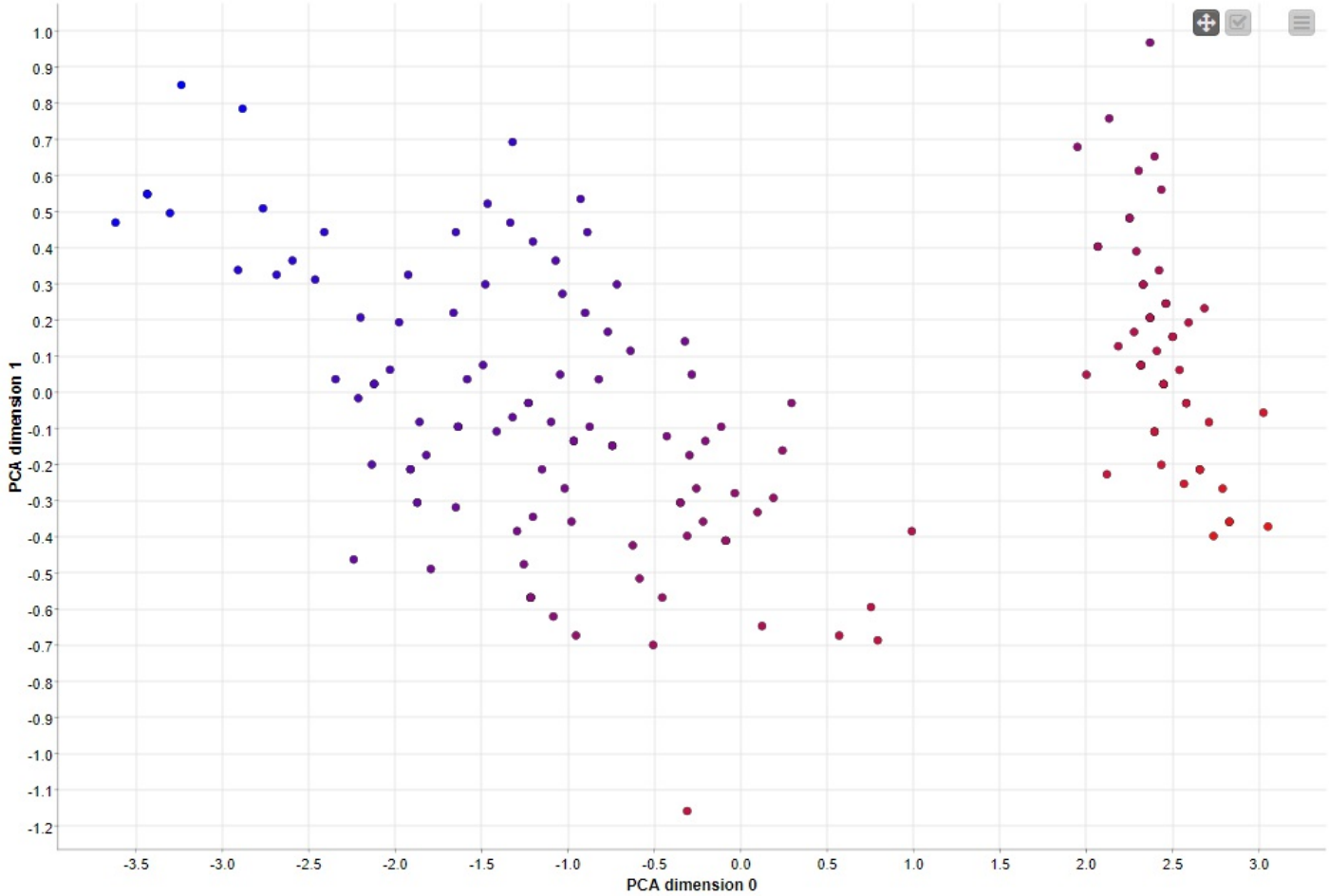


3. Compute the Correlation using the 'Linear Correlation' node and reduce the dimension by using the 'Correlation Filter'. Plot the data using the selected variables.

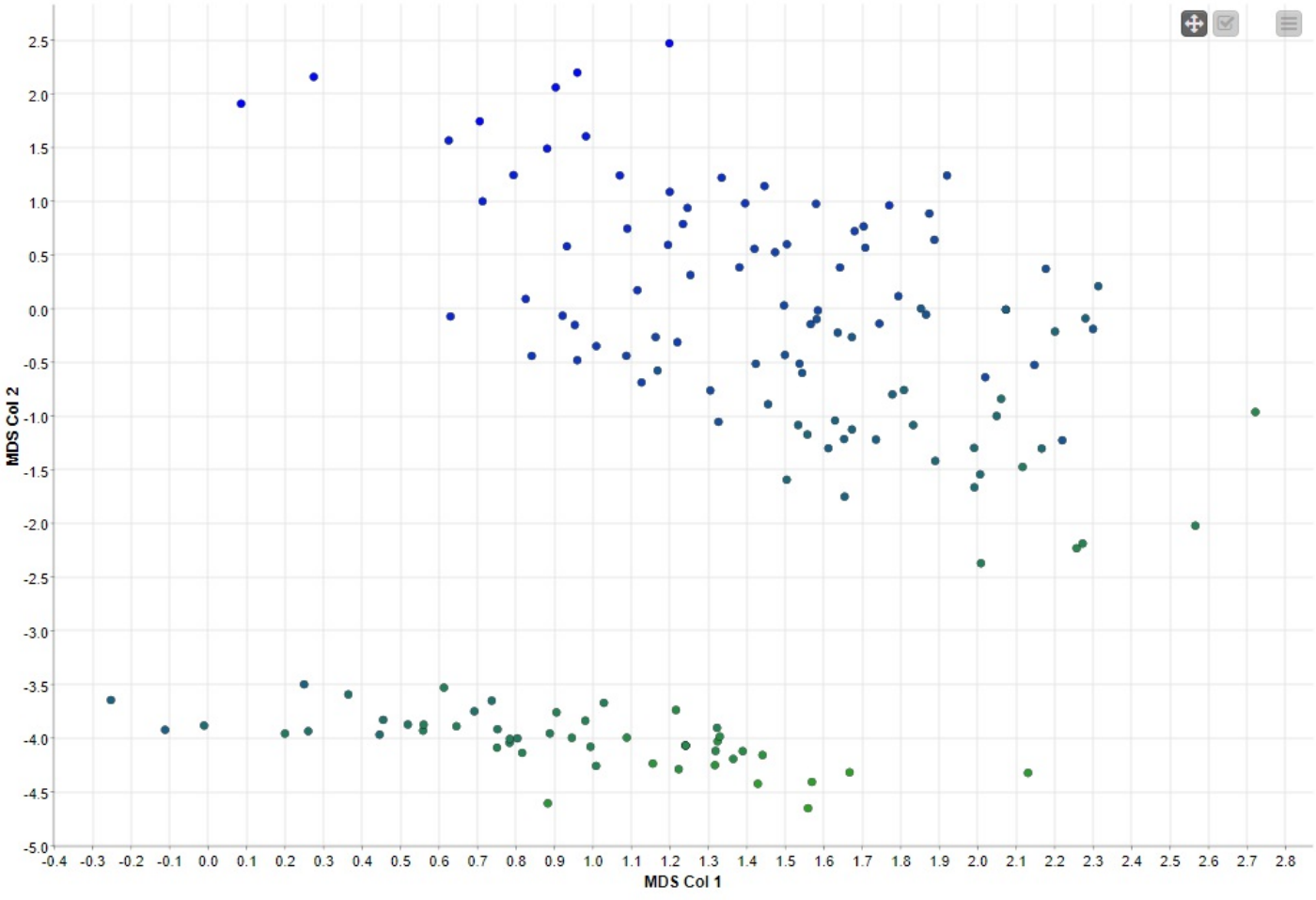
AF

4. Use the PCA and MDS dimensional reduction techniques to reduce the dimension of the data to 2 variables and plot the results. Make an analysis between the result in question 3 and those obtained here.

- Voici le resultat pour le PCA



- Voici le resultat pour le MDS



On peut constater que la méthode MDS regroupe mieux certains point (notamment ceux proche de l'Axe X) par rapport a la méthode PCA.

Exercice 2

- **Reading full small data set** : Read large data set from DB contained nodes
- **Target Selection** : Selection: Allow to rename some columns and then select or them.
- **Baseline Evaluation** : Choose the best algorithm between three classification algorithms: MLP, decision tree and Naïve Bayes.
- **Reduction based on LDA** : Reduces the number of columns in the input data by linear discriminant analysis
- **Auto Encoder based Reduction** :
- **Reduction based on t-SNE** : Create a probability distribution capturing the relationships between points in the high dimensional space and find a low dimensional space that resembles the probability dimension as well as possible
- **Reduction based on High Corr.** : Identifies pairs of columns with a high correlation (i.e. greater than a given threshold), and removes one of the two columns for each identified pair.
- **Tree Ensemble based Reduction** : Generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features.
- **Reduction based on PCA** : Reduce all values to those which are the most accurate for the PCA method.
- **Row Sampling** :
- **Column Splitter** : This node splits the columns of the input table into two output tables.
- **Column Selection by Missing Values** : Remove columns with excessive values.
- **Column Appender** : Takes two tables and quickly combines them by appending the columns of the second table to the first table.
- **Backward Feature Elimination** : Elimine une par une les valeurs les moins pertinente en les comparants 2 par 2.
- **Forward Feature Selection** : Elimine tout et récupère une par une les valeurs pertinente en les comparants 2 par 2.
- **Joiner** : This node joins two tables in a database-like way.
- **ROC Curve** : This node draws ROC curves for two-class classification problems.
- **Positive class probabilities** : Select the value from the class column that stands for the "positive" class
- **Accuracies** :
- **Bar Chart** : Visualizes one or more aggregated metrics for different data partitions with rectangular bars where the heights are proportional to the metric values.

Il y a 233 colonnes et 50 000 lignes

Du a manque de ressource nous n'avons pas pu executer le workflow, il nous est impossible d'analyser les résultats.