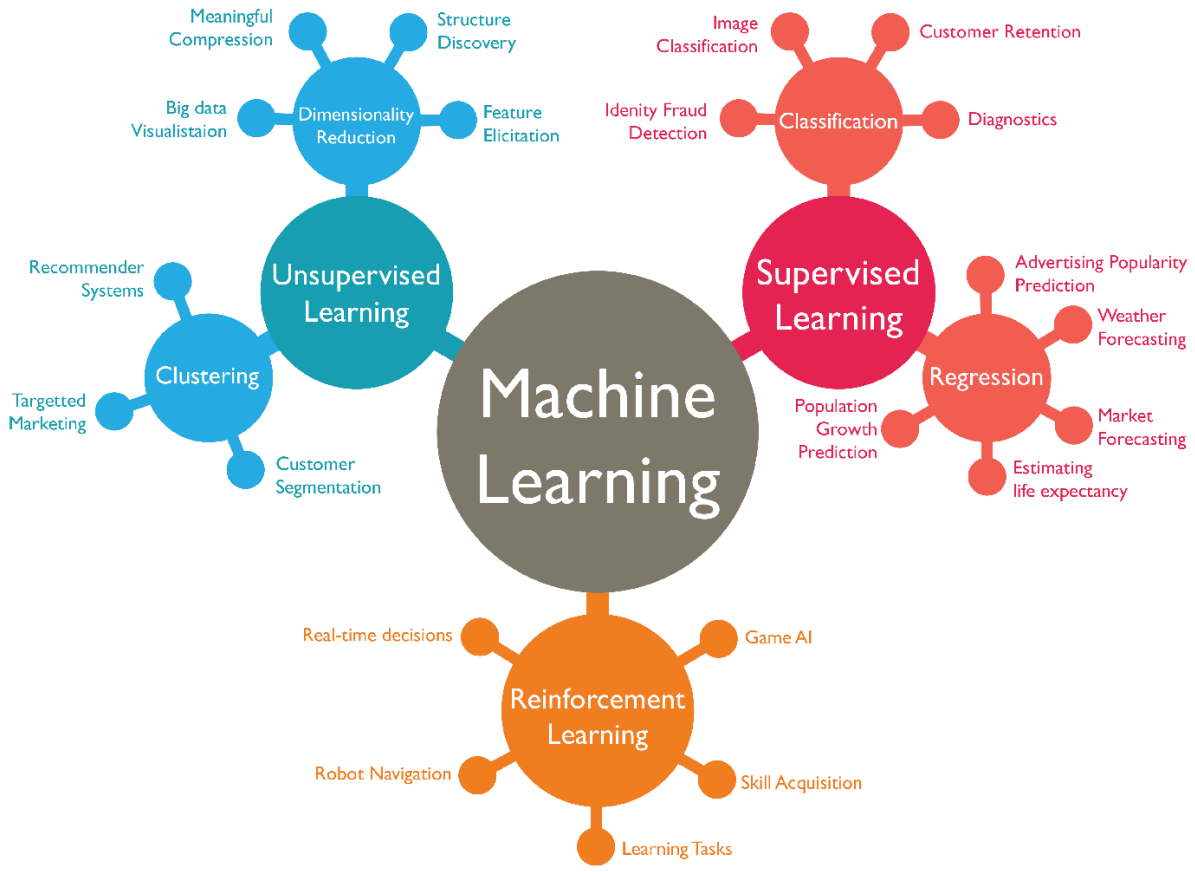

Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Le rapport ici, présent est une application des techniques apprentissages automatiques qui nous ont étés enseignées durant le semestre. En effet, nous avons été introduits au *Machine Learning* durant le semestre et avons appris à faire des apprentissages supervisées et non supervisées. Ceci dit durant tout le long de rapport, je m'attèlerai à vous faire par de toutes les étapes de cet apprentissage tout en y incluant des explications et des conclusions.

LA CLASSIFICATION NON SUPERVISEE

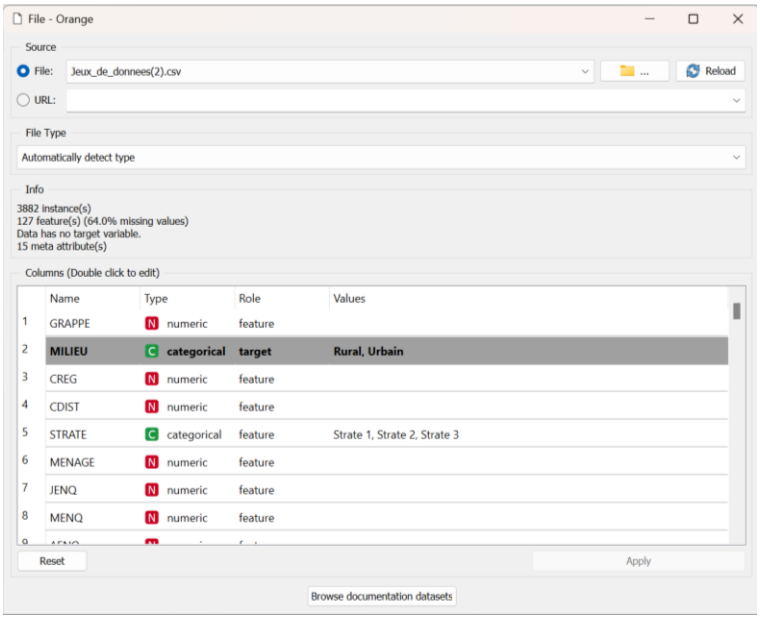
Le jeu de données choisi, présente une enquête post-Campagne contre la rougeole à Madagascar en 2019 nommée « EPC-ROUGEOLE 2019 » portées sur des enfants de 6 à 9 ans révolus. Ces données m'ont été fourni par un parent cependant toutes les informations qui y sont reliées seront en annexes.

Tout d'abord, on ouvre orange, on fait *File* et on charge le jeu de données, **Jeux_de_donnees(2)** . Il contient 3882 instances dont :

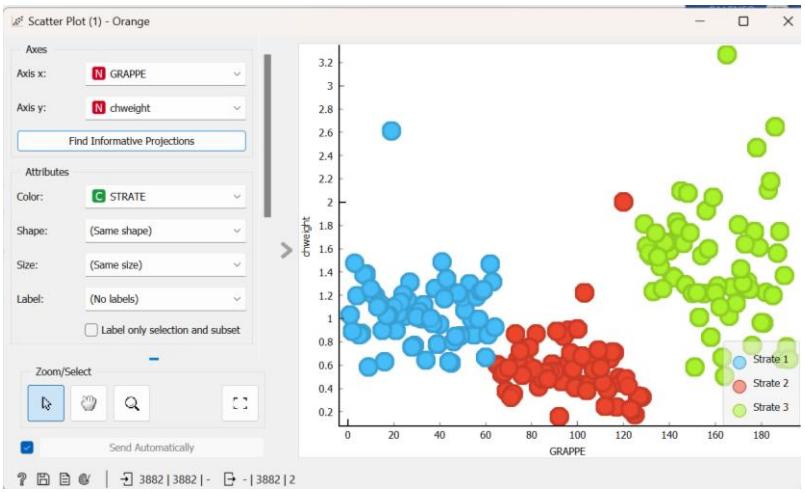
*des variables textuelles,

*des variables catégorielles dont uniquement celle nommée *Strate* sera utilisée comme Target,

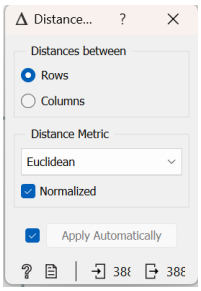
*des variables nominales dont uniquement le *chweight* (poids enfants) et Grappe seront utilisés.



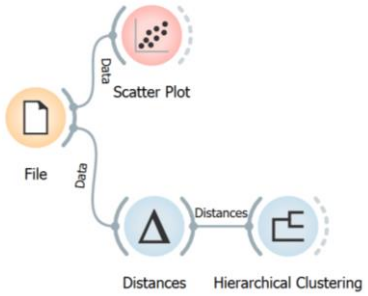
On cherche à visualiser le poids des enfants en fonctions de la *Strate*. On compte 3 *Strates* dans le jeu de données. La *Strate* correspond à une phase de la campagne de vaccination à savoir *Janvier 2019, Février 2019 et Mars/Avril 2019*. Ensuite, on ajoute à la sortie de *File* un *Scatter Plot* et on l'ouvre.



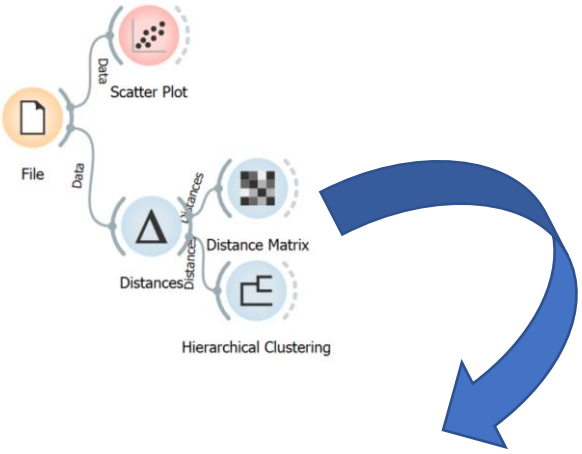
Une fois encore à la sortie de *File*, on ajoute *Distance* et régler les paramètres.



Ensuite, la sortie de *Distance* prend *Hierarchical Cluster*



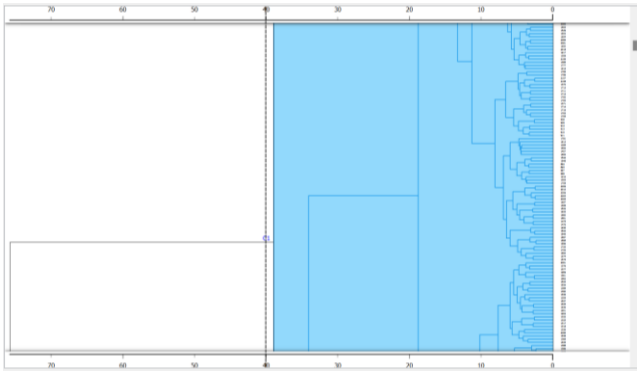
On ouvre le *Hierarchical Cluster* et on observe les clusters :

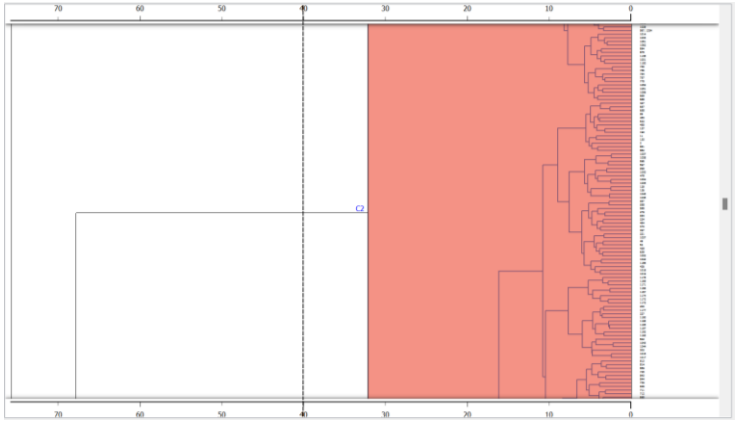


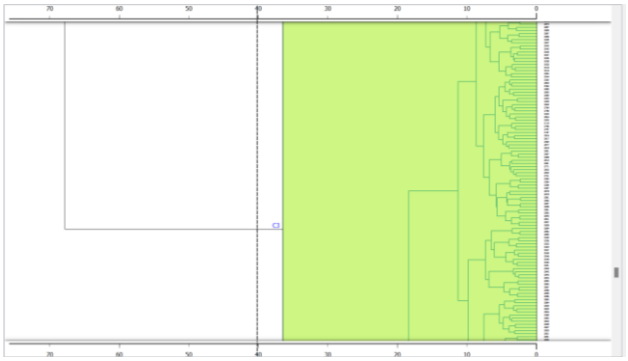
Distance Matrix - Orange

	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate 1	Strate
Strate 1		4,692	3,856	4,317	4,236	3,959	3,856	4,432	4,702	5,094	5,388	4,249	4,191	4,650	4,105
Strate 1	4,692		4,796	4,619	5,462	4,895	4,717	5,039	5,130	5,344	4,833	5,431	5,069	5,279	5,340
Strate 1	3,856	4,796		3,027	3,940	4,472	3,825	4,348	4,729	5,218	5,335	4,026	4,346	4,465	4,446
Strate 1	4,317	4,619	3,027		4,200	4,241	4,121	4,341	4,056	4,166	4,614	4,567	3,912	4,335	4,372
Strate 1	4,236	5,462	3,940	4,200		4,297	4,187	4,565	4,704	5,180	5,317	4,248	4,300	4,599	3,903
Strate 1	3,959	4,895	4,472	4,241	4,297		3,609	3,962	4,075	4,222	4,748	4,501	4,153	3,894	3,892
Strate 1	3,856	4,717	3,825	4,121	4,187	3,609		3,621	4,009	4,405	4,819	3,757	3,797	3,729	4,177
Strate 1	4,432	5,039	4,348	4,341	4,565	3,962	3,621		2,531	3,296	4,626	4,215	3,917	4,064	4,604
Strate 1	4,702	5,130	4,729	4,056	4,704	4,075	4,009	2,531		2,540	4,398	4,399	3,711	4,222	4,618
Strate 1	5,094	5,344	5,218	4,166	5,180	4,222	4,405	3,296	2,540		4,532	4,833	3,941	4,523	4,981
Strate 1	5,388	4,833	5,335	4,614	5,317	4,748	4,819	4,626	4,398	4,532		5,511	4,908	4,348	5,216
Strate 1	4,249	5,431	4,026	4,567	4,248	4,501	3,757	4,215	4,399	4,833	5,511		3,074	4,357	4,393
Strate 1	4,191	5,069	4,346	3,912	4,300	4,153	3,797	3,917	3,711	3,941	4,908	3,074		3,949	4,211
Strate 1	4,650	5,279	4,465	4,335	4,599	3,894	3,729	4,064	4,222	4,523	4,348	4,357	3,949		4,431
Strate 1	4,105	5,340	4,446	4,372	3,903	3,892	4,177	4,604	4,618	4,981	5,216	4,393	4,211	4,431	

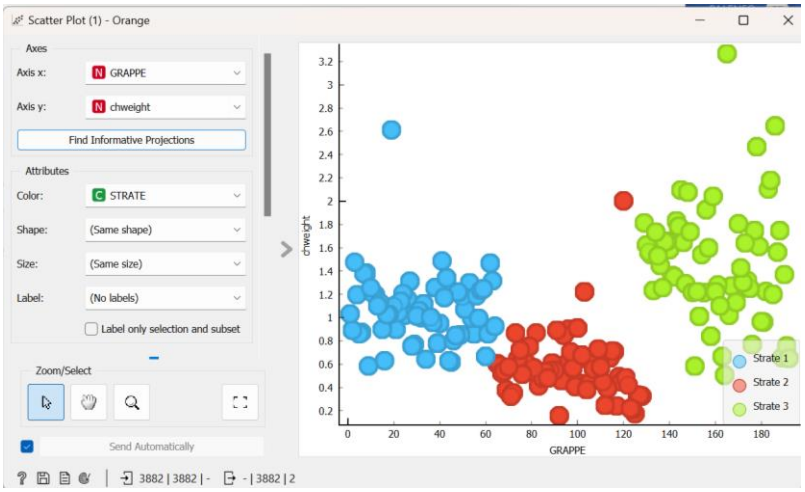
On observe des clusters avec de minimes différences si on place le curseur à partir de 40.







Ensuite à la sortie *du hierarichic clustering*, on rajoute un *Scatter Plot* et l'entrée de ce dernier prend le File. On ouvre *Scatter Pot* et on observe :



Ainsi chaque cluster correspond à une couleur et on peut conclure que le poids des enfants en fonctions des Grappes (environnements d'habitations comme les ilots) sont plutôt similaires durant toutes les phases sans oublier que la strate verte est un peu plus élevée.

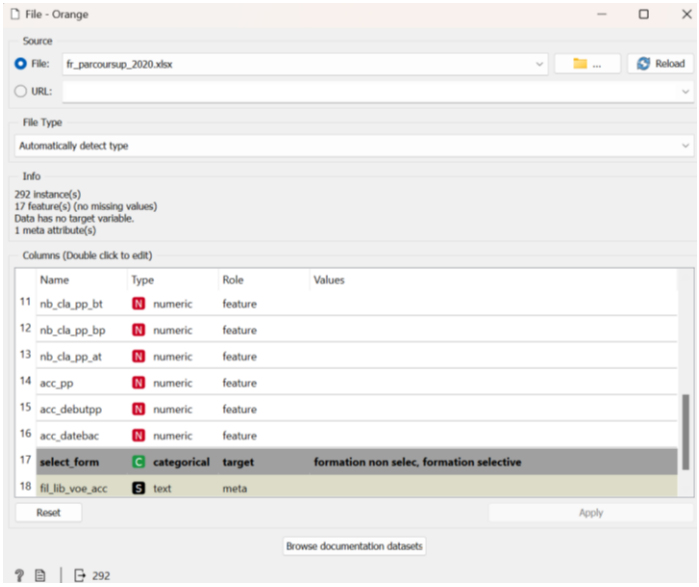
LA CLASSIFICATION SUPERVISEE

Le jeu de données choisi, présente les vœux de poursuite d'études et de réorientation dans l'enseignement supérieur ainsi que les propositions des établissements pour chaque formation hors apprentissage à la fin du processus d'affectation de la plateforme Parcoursup pour la session 2020, recueilli sur le site *Data.gouv* (lien au bas de la page). L'application Parcoursup est la plateforme nationale de préinscription mise en place par le Ministère de l'Enseignement supérieur, de la recherche et de l'innovation permettant aux élèves de candidater à l'entrée dans l'enseignement supérieur. Le jeu de données utilisé pour cette partie a été préalablement été nettoyé. On ouvre orange, on fait *file* et on importe le jeu de données nettoyé. Il contient 18 variables :

* une variable textuelle (*form_lib_voe_acc*) contenant le détail des filières de formation,

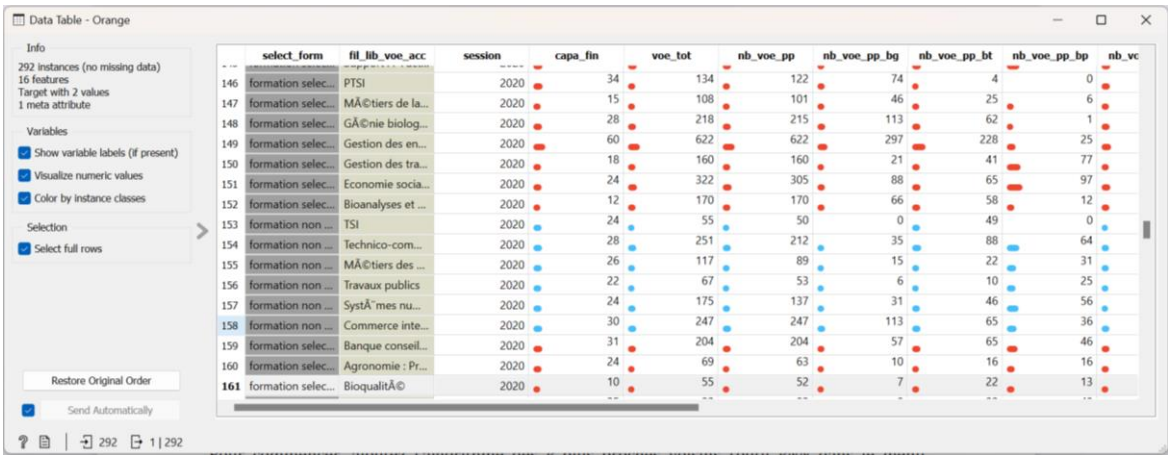
* 16 variables quantitatives caractérisant les effectifs candidats à ces différentes filières,

* 1 variable nominale (*select_form*) caractérisant les formations sélectives ou non.



Ajoutez Data Table en sortie de File et parcourez les données de façon à bien les comprendre :

https://www.data.gouv.fr/fr/datasets/parcoursup-2020-voeux-de-poursuite-detudes-et-de-reorientation-dans-lenseignement-superieur-et-reponses-des-etablissements/#resources

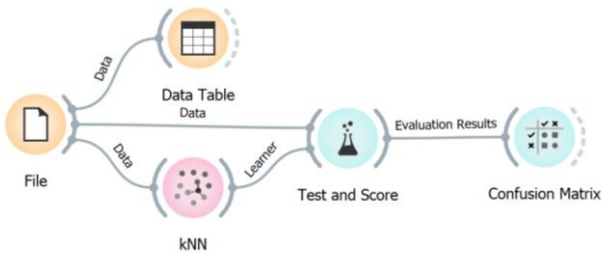


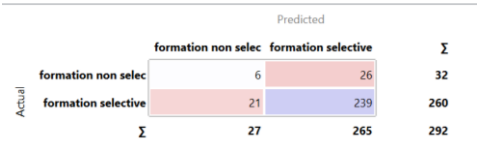
	select_form	fil_lib_voe_acc	session	capa_fin	voe_tot	nb_voe_pp	nb_voe_pp_bg	nb_voe_pp_bt	nb_voe_pp_bp
146	formation selec...	PTSI	2020	34	134	122	74	4	0
147	formation selec...	MÃ©tiers de la...	2020	15	108	101	46	25	6
148	formation selec...	GÃ©nie biolog...	2020	28	218	215	113	62	1
149	formation selec...	Gestion des en...	2020	60	622	622	297	228	25
150	formation selec...	Gestion des tra...	2020	18	160	160	21	41	77
151	formation selec...	Economie socia...	2020	24	322	305	88	65	97
152	formation selec...	Bioanalyses et ...	2020	12	170	170	66	58	12
153	formation non ...	TSI	2020	24	55	50	0	49	0
154	formation non ...	Technico-com...	2020	28	251	212	35	88	64
155	formation non ...	MÃ©tiers des ...	2020	26	117	89	15	22	31
156	formation non ...	Travaux publics	2020	22	67	53	6	10	25
157	formation non ...	SystÃ¨mes nu...	2020	24	175	137	31	46	56
158	formation non ...	Commerce inte...	2020	30	247	247	113	65	36
159	formation selec...	Banque conseil...	2020	31	204	204	57	65	46
160	formation selec...	Agronomie : Pr...	2020	24	69	63	10	16	16
161	formation selec...	BioqualitÃ©	2020	10	55	52	7	22	13

L'objectif ici, est d'entrainer le modèle à reconnaitre, à partir des variables quantitatives, l'étiquette contenue dans la variable *select_form.* Autrement dit, on cherche à créer des modèles classants les filières en formation sélective ou non. Et une fois entrainés, on cherche à utiliser les modèles sur un jeu de données ne contenant pas la variable *select_form.*

Ainsi, ajoutez l'algorithme des *K* plus proches voisins *KNN* (menu *Model*) en sortie de *File.* Ensuite, ajoutez l'outil *Test end Score* (menu *Evaluate*) qui sera sortie du *KNN* et prendra en entrée les données issues de *File*. Cela permet de générer une **validation croisée** ou **cross-validation** qui est une méthode statistique utilisées pour tester l'efficacité d'une *machine Learning.* En sortie du Test Score, ajoutez une *matrice de confusion (*menu *Evaluate*) en

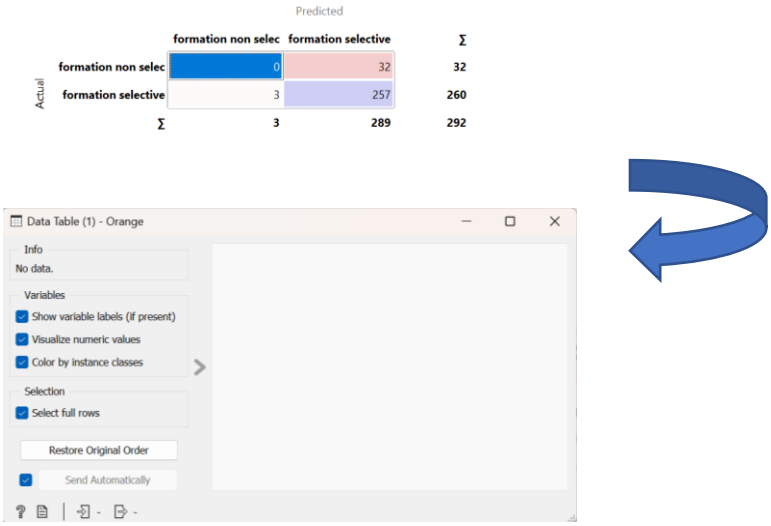
modifiant les paramètres du *KNN* jusqu'à observer l'algorithme de *K* suivant et vous observerez :





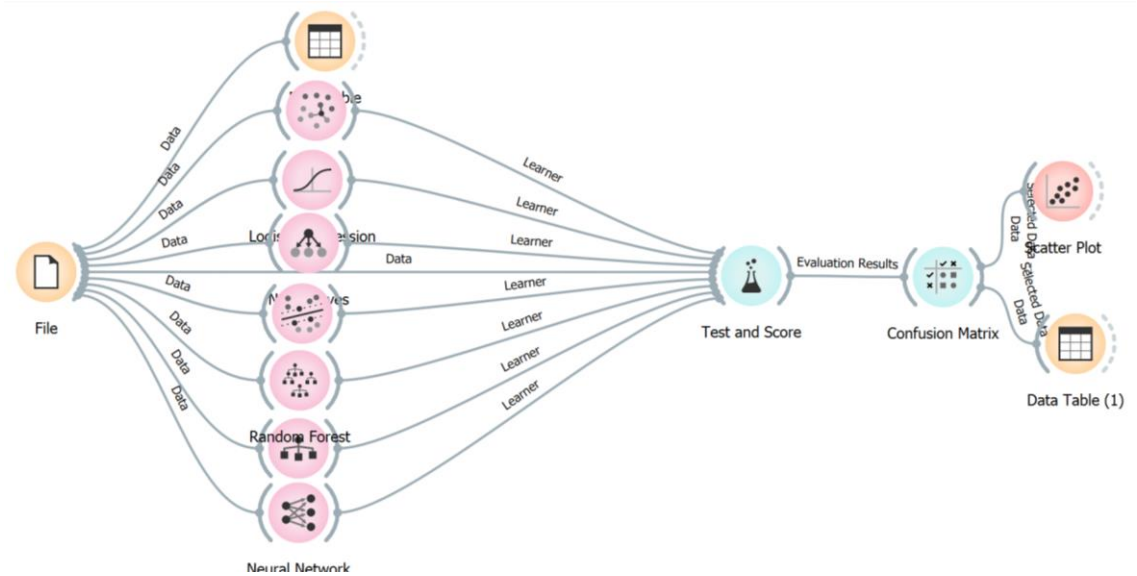
		Predicted		
		formation non selec	formation selective	Σ
Actual	formation non selec	6	26	32
	formation selective	21	239	260
	Σ	27	265	292

Ensuite, ouvrez *Confusion Matrix* et sélectionnez les 0 formations non sélectives observez sur une *Table* vide ou autre et vous observerez une autre *Table* :



Il s'agit des différentes classifications non sélectives issues de la validation croisée.

Ensuite vous allez comparez le KNN avec les autres modèles de classification vus en cours.

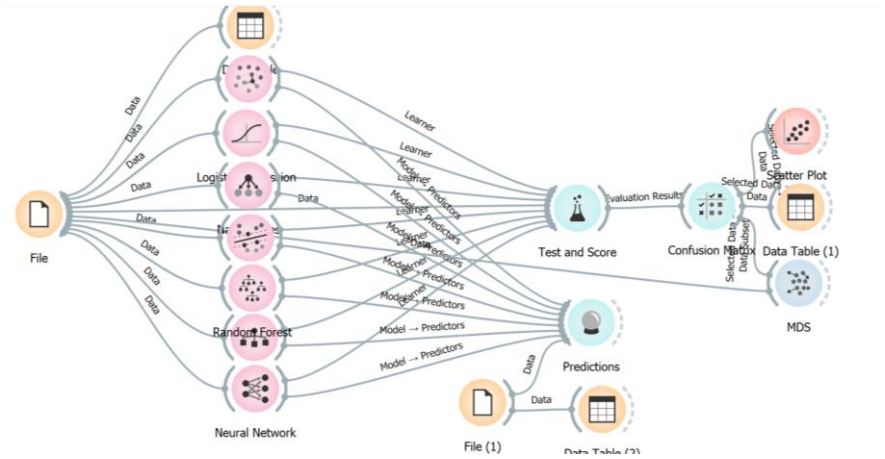


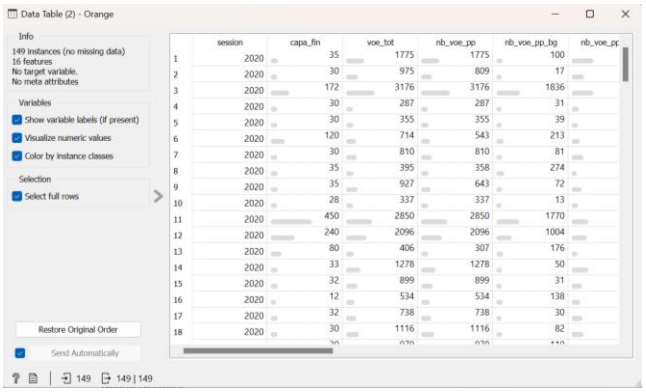
Ouvrez Matrix Confusion et on verra :

Predicted

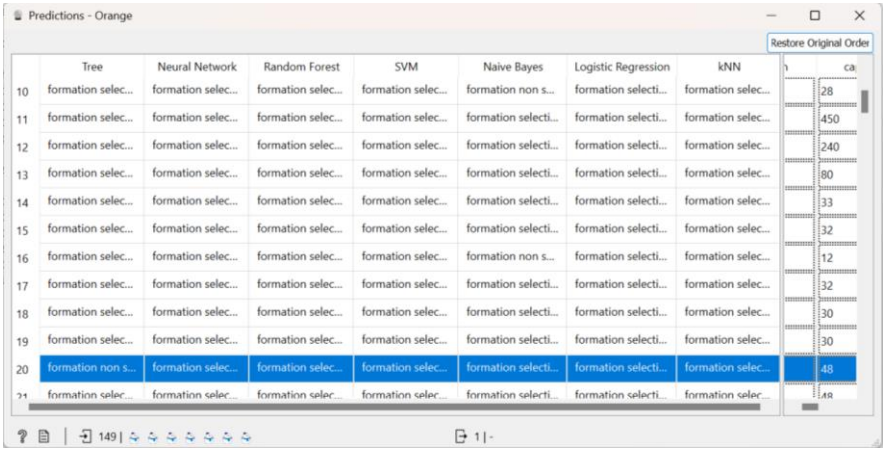
Actual	formation non selec	formation selective	Σ
formation non selec	6	26	32
formation selective	21	239	260
Σ	27	265	292

On peut ainsi voir que l'ajout des autres modèles à améliorer pour le mieux la matrice et fourni de meilleurs résultats. Ces modèles permettront de faire des **prédictions.** Pour commencer, on charge le fichier ***fr_parcoursup_2020*** et on le charge dans le *File(1)* qui prend en sortie de *Prédictions* et *Data Table et* on visualise le contenu de la table :





Comme on peut le voir le fichier contient 149 candidatures pour lesquels les 16 variables caractéristiques *Sélective/Non Sélective*. On pourra ensuite déterminer si une ligne est une fomation sélectives ou non.



On remarque alors que la différence se voit au niveau du *Tree* qui qualifie une formation de non sélective alors que les autres méthodes soutiennent le contraire. On pourrait alors conclure que les méthodes *Naive Bayes, SVM, Random Forest,Logistic Regression et Neutral Network* donnent de meilleures résultats que *Tree.*