Cassandra Sénécaille 22100082

Sciences des Données

Projet individuel

Table des matières

Cassandra Sénécaille 22100082

1- Apprentissage non supervisé

Comment évolue le nombre de morts par la malnutrition, en 2019, par pays ?

1.1 – Jeu de données

Utilisation d'un fichier CSV trouvé sur le site « *https://ourworldindata.org/famines* ».
C'est donc sur les pays touchés par la famine que je me suis basée. Il y a donc plusieurs catégories :

- Le nom du pays (Entity),
- le nombre de morts par la famine pour les plus de 70 ans (Deaths – Age : 70+ Years),
- le nombre de morts par la famine pour les 50-69 ans (Deaths – Age : 50-69 Years),
- le nombre de morts par la famine pour les 15-49 ans (Deaths – Age : 15-49 Years),
- le nombre de morts par la famine pour les 5-14 ans (Deaths – Age : 5-14 Years),
- et le nombre de morts par la famine pour les moins de 5 ans (Deaths – Age : Under 5 Years).
- Une colonne total regroupant le total des morts par pays(Total)

1.2 – Nettoyage et prétraitement des données

Afin de pouvoir voir plus clairement les données, j'ai décidé de me baser sur une seule année : 2019, étant donné qu'il y a plus de 229 pays.
J'ai pris la décision de ne pas prendre 5 lignes, regroupant les données du Monde et des banques mondiales. En effet, ces chiffres regroupent des pays déjà exploitées.

Également, j'ai supprimé le G20 qui est un regroupement de différents pays comme le Canada, les États-Unis, le Japon, la Russie ou encore l'Union Européenne, l'Argentine, l'Australie,...

J'ai donc créé un nouveau fichier excel pour pré-traiter les données et donc prendre celles qui sont exploitables.
J'ai aussi réduit le nom des colonnes, afin d'avoir quelque chose de plus simple à lire. Nous sommes donc passé, par exemple, de:

Deaths - Protein-energy malnutrition - Sex: Both - Age: 70+ years (Number)
à
Deaths – Age : 70+ years (Number)

Ainsi, il sera plus compréhensible de voir que nous parlons de morts par pays et par tranches d'âge.
J'ai également ajouté une colonne à la fin, regroupant le total des morts par pays toutes tranches d'âge confondues.

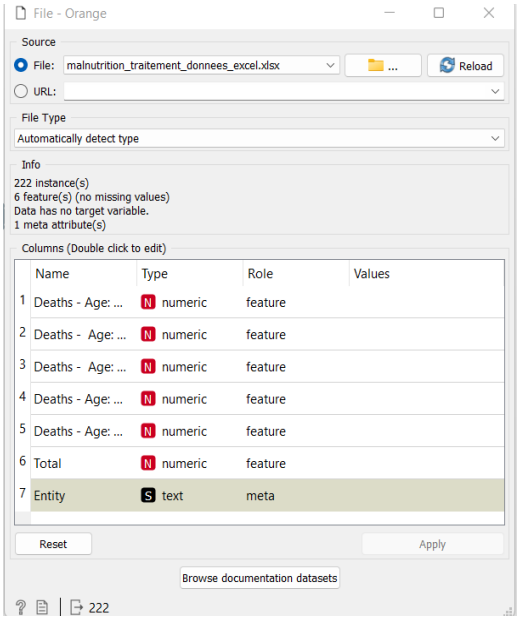
Pour plus de lisibilité, j'ai par la suite supprimé la colonne 'Code' qui correspondait aux codes des pays, car ces derniers ne possédaient pas tous de codes.
Puis, j'ai supprimé la colonne année,car nous nous concentrons exclusivement sur 2019. Cette colonne ne servait donc plus à grand-chose dans mon traitement de données.
On passe donc d'un fichier difficilement compréhensible à un autre facilement lisible et exploitable.

	A	B	C	D	E	F	G
1	Entity	Deaths - Age: 70+ years (Number)	Deaths - Age: 50-69 years (Number)	Deaths - Age: 15-49 years (Number)	Deaths - Age: 5-14 years (Number)	Deaths - Sex: Both - Age: Under 5 (Number)	Total
2	Afghanistan	12	35	163	90	911	1211
3	African Regio	13698	7628	6760	4156	62539	94781
4	Albania	1	0	0	0	2	3
5	Algeria	45	8	9	4	25	91
6	American Sa	1	0	0	0	0	1
7	Andorra	0	0	0	0	0	0
8	Angola	372	313	254	123	1844	2906
9	Antigua and	2	0	0	0	0	2
10	Argentina	1001	147	48	10	49	1255
11	Armenia	0	0	1	0	0	1

1.3 – Clustering

a) Description de la chaîne de traitement

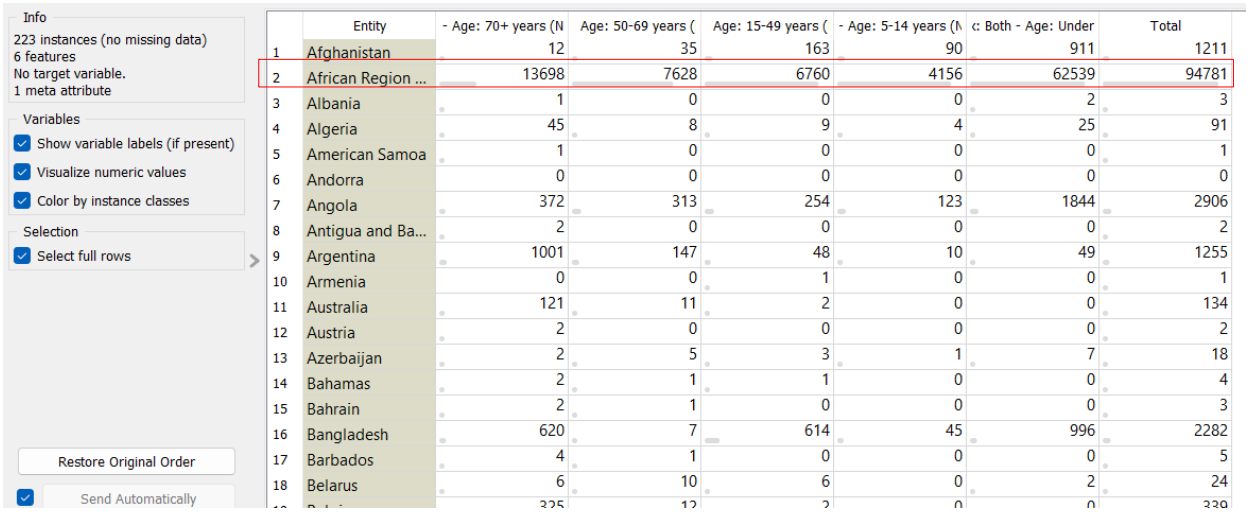
J'ai récupéré mon fichier nettoyé et pré-traité : « *malnutrition_traitement_donnees_excel.xlsx* ». Par la suite, je l'ai importé grâce à l'outil 'File'.



On voit donc que Orange retrouve nos 222 pays, et 7 variables (6 quantitatives et 1 qualitative).

Je vais donc par la suite ajouter une Data Table, afin de vérifier que les données s'affichent correctement.

De plus, Orange nous permet de voir visuellement les pays qui sont les plus touchés via ses « barres de progression. » Par exemple, on peut voir ici, sur notre feuille, que le pays le plus touché par les morts dus à la malnutrition est la région Africaine avec un total de 94 781 morts.



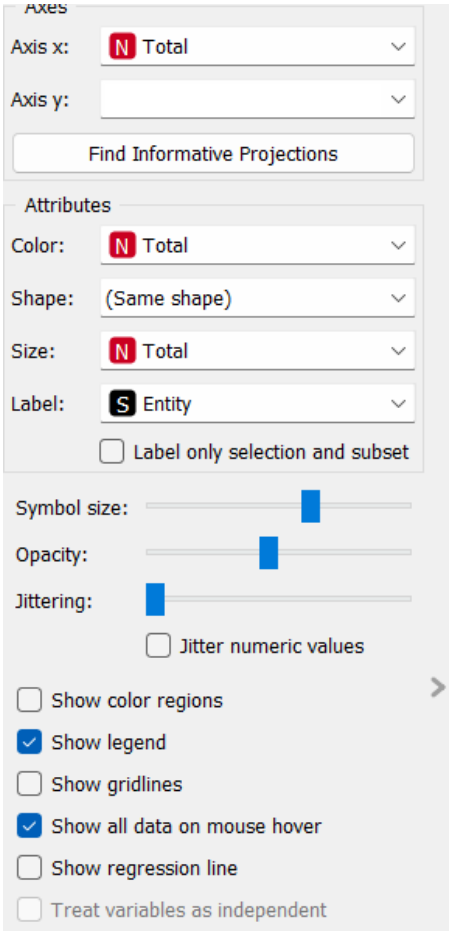
	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	c: Both - Age: Under	Total
1	Afghanistan	12	35	163	90	911	1211
2	African Region ...	13698	7628	6760	4156	62539	94781
3	Albania	1	0	0	0	2	3
4	Algeria	45	8	9	4	25	91
5	American Samoa	1	0	0	0	0	1
6	Andorra	0	0	0	0	0	0
7	Angola	372	313	254	123	1844	2906
8	Antigua and Ba...	2	0	0	0	0	2
9	Argentina	1001	147	48	10	49	1255
10	Armenia	0	0	1	0	0	1
11	Australia	121	11	2	0	0	134
12	Austria	2	0	0	0	0	2
13	Azerbaijan	2	5	3	1	7	18
14	Bahamas	2	1	1	0	0	4
15	Bahrain	2	1	0	0	0	3
16	Bangladesh	620	7	614	45	996	2282
17	Barbados	4	1	0	0	0	5
18	Belarus	6	10	6	0	2	24

Analysons ensemble un pays pour comprendre comment le fichier fonctionne. Ce n'est pas compliqué, par exemple, on peut lire que l'Australie a été victime de malnutrition en 2019, causant en tout 134 morts. Il n'y a eu aucun décès pour les moins de 5ans, ainsi que pour les 5-14 ans. Il y a eu 2 morts pour la tranche d'âge des 15-49 ans et 11 pour les 50-69 ans. Enfin, il y a eu 121 décès pour les plus de 70 ans. Cette dernière tranche d'âge est la plus touchée par la malnutrition.

Par la suite, j'ai relié un 'Scatter Plot' à ma base de données afin de voir graphiquement celles-ci.

Afin de voir différentes tailles, j'ai attribué pour le paramètre 'size' la variable Total. Puis j'ai fait afficher le nom des Pays, afin de voir quels sont ceux qui seront les plus touché par la malnutrition.

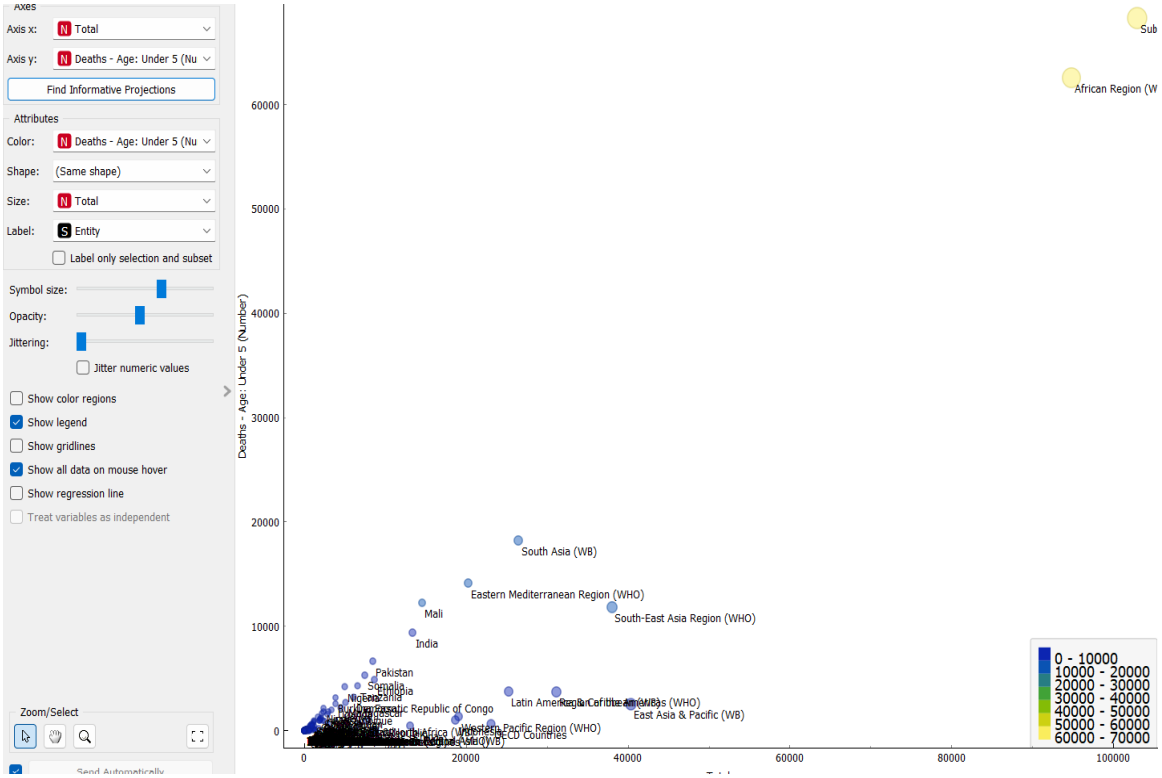
Dans un premier temps, j'ai décidé de me concentrer sur le Total pour l'axe des X. Et ainsi avoir un axe identique pour les premières analyses.



(1) Première analyse : Les moins de 5 ans :

Dans un premier temps, j'ai attribué à l'axe des Y, les morts par la malnutrition des les moins de 5 ans.

Maintenant, concentrons nous sur la légende, qui correspond au paramètre 'Color' auquel nous avons attribué la variable 'Deaths – Age : Under 5 Years'. Les pays les moins touchés par la malnutrition seront en bleu foncé, puis au fur et à mesure, on dérivera vers le vert puis vers le jaune pour les plus touchés avec plus de 60 000 morts.



On remarque donc que 2 pays se démarquent des autres : *African Region (WHO) et Sub-Saharan Africa (WB).*

Le Sub-Saharan Africa est jaune, ce qui signifie que son total, de morts par la malnutrition, se trouve dans la tranche des 60 000 – 70 000 morts. Puis quand on regarde sur notre fichier, on se rend compte que pour les moins de 5 ans, il y a eu 68 272 décès, en 2019.

Par la suite, nous pouvons voir que l'African Region se situe également dans le jaune, ce qui signifie que ses données seront également supérieures à 60 000 décès, en 2019. Regardons maintenant dans le fichier quel est son nombre exact. On trouve donc : 62 539 morts pour cette tranche d'âge.

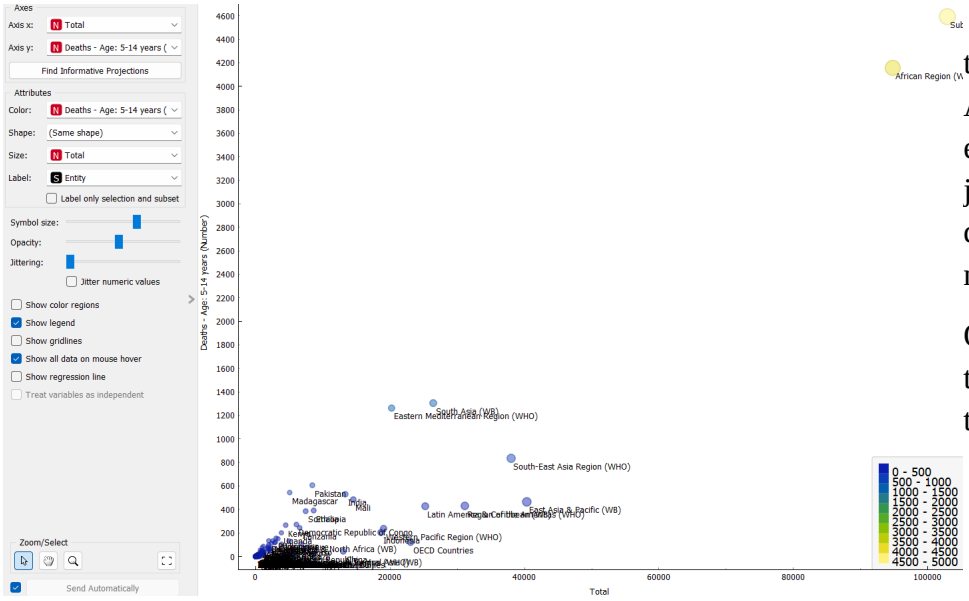
	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
188	Sub-Saharan Africa (WB)	14320	8247	7459	4591	68272	102889
2	African Region (WHO)	13698	7628	6760	4156	62539	94781
182	South Asia (WB)	2646	2558	1784	1302	18181	26471
56	Eastern Mediterranean R...	1924	1487	1470	1264	14128	20273
116	Mali	1000	393	378	483	12281	14535
185	South-East Asia Region (...	17794	5336	2163	837	11861	37991
88	India	1198	1767	506	531	9382	13384
148	Pakistan	351	515	370	605	6611	8452

Les autres données se retrouvent dans la zone bleue foncée, voire bleue claire, ce qui veut dire qu'ils seront entre 0 et 20 000 morts par la malnutrition en 2019.

Et quand on va sur la Data Table, et qu'on range la colonne des moins de 5 ans par ordre décroissant, le troisième pays le plus touché est South Asia avec 18 181 décès.

(2) Deuxième analyse : Les 5 – 14 ans :

Après avoir adapté les paramètres, nous voila avec une nouvelle légende.



On remarque que c'est toujours le Sub-Saharan Africa qui est le plus élevé, car il est dans le jaune clair. Cela veut dire qu'il y a eu plus de 4 500 morts.

Quand on regarde sur la table des données, on trouve 4 591 morts.

En seconde position, c'est l'African Region qui a entre 4 000 et 4 500 morts. En effet, on trouve la données de 4 156 décès par la malnutrition.

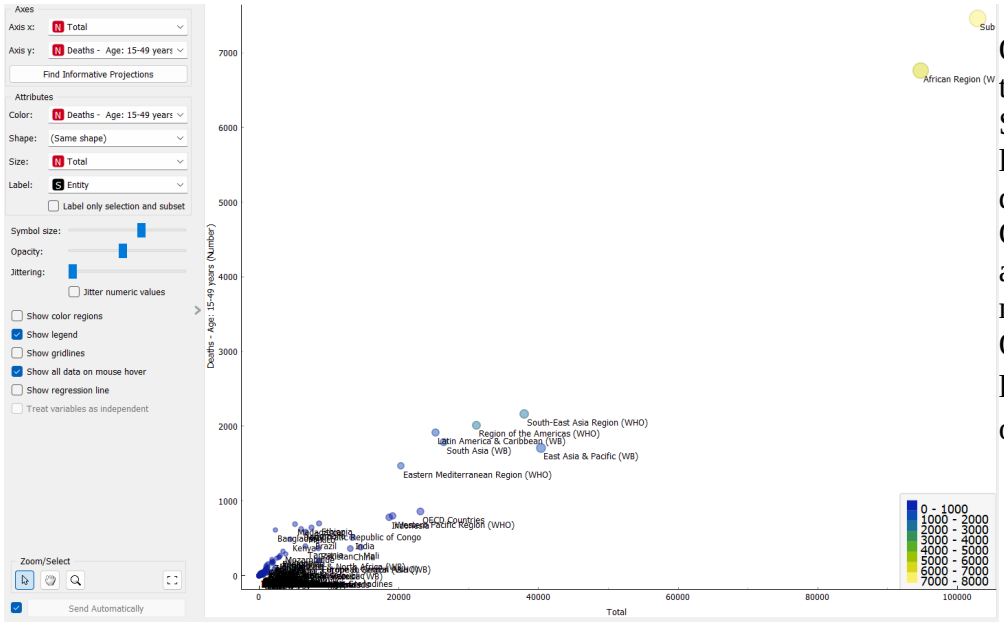
	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
188	Sub-Saharan Africa (WB)	14320	8247	7459	4591	68272	102889
2	African Region (WHO)	13698	7628	6760	4156	62539	94781
182	South Asia (WB)	2646	2558	1784	1302	18181	26471
56	Eastern Mediterranean R...	1924	1487	1470	1264	14128	20273
185	South-East Asia Region (...	17794	5336	2163	837	11861	37991
148	Pakistan	351	515	370	605	6611	8452
112	Madagascar	590	580	694	541	2698	5103

Comme sur la précédente analyse, on se rend compte qu'il y a un grand écart entre les deux premières données et le reste.

On remarque que South Asia est encore à la troisième place, avec 1 302 morts.

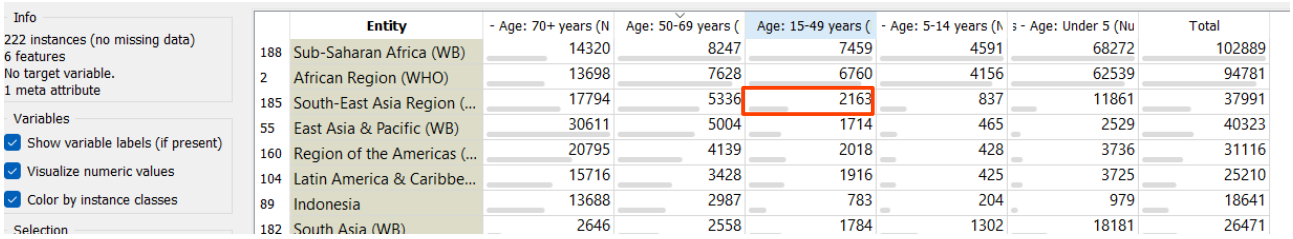
(3) Troisième analyse : Les 15 – 49 ans :

Après avoir adapté les paramètres, nous voila avec une nouvelle légende.



On remarque que c'est toujours le Sub-Saharan Africa qui est le plus élevé, car il est dans le jaune clair. Cela veut dire qu'il y a eu plus de 7 000 morts.
Quand on regarde sur la table des données, on trouve 7 459 décès.

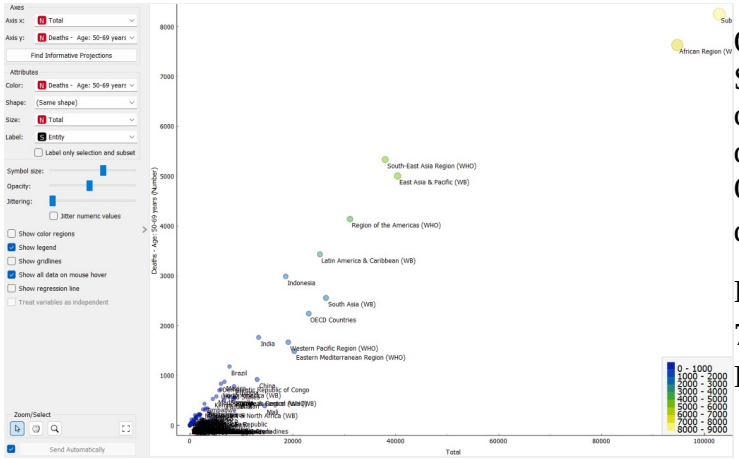
Pour African Region, on remarque qu'il est dans la zone jaune foncé, donc il se trouve dans la tranche des 6 000 à 7 000 morts. Et quand on vérifie sur la table des données, on peut lire qu'il y a eu 6 760 décès par la malnutrition en 2019.



	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
188	Sub-Saharan Africa (WB)	14320	8247	7459	4591	68272	102889
2	African Region (WHO)	13698	7628	6760	4156	62539	94781
185	South-East Asia Region (...	17794	5336	2163	837	11861	37991
55	East Asia & Pacific (WB)	30611	5004	1714	465	2529	40323
160	Region of the Americas (...	20795	4139	2018	428	3736	31116
104	Latin America & Caribbe...	15716	3428	1916	425	3725	25210
89	Indonesia	13688	2987	783	204	979	18641
182	South Asia (WB)	2646	2558	1784	1302	18181	26471

(4) Quatrième analyse : Les 50 – 69 ans :

Après avoir adapté les paramètres, nous voila avec une nouvelle légende.



On remarque que c'est toujours le Sub-Saharan Africa qui est le plus élevé, car il est dans le jaune clair. Cela veut dire qu'il y a eu plus de 8 000 morts.
Quand on regarde sur la table des données, on lit 8 247 décès.

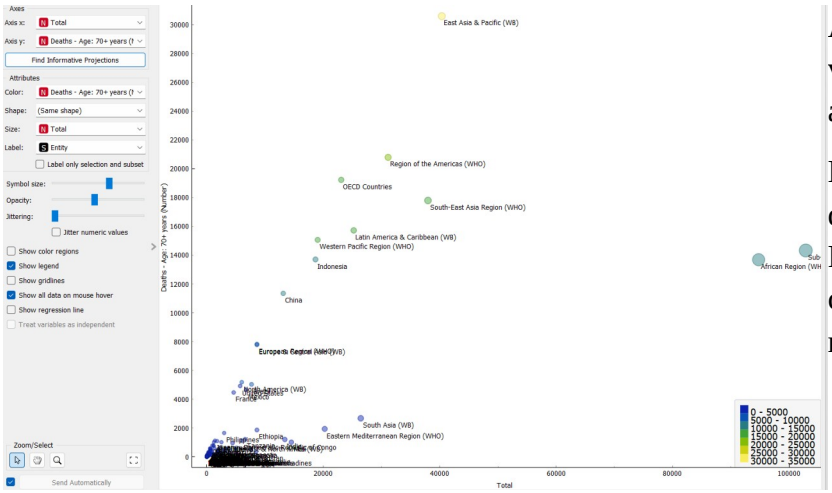
Par la suite, dans la zone des 7 000 – 8 000 morts, on trouve l'African Region avec 7 628 morts.

On peut aussi remarquer qu'il y a des pays dans la zone verte, comme South-East Asia Region, avec 5 336 morts par la malnutrition, en 2019.

	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
188	Sub-Saharan Africa (WB)	14320	8247	7459	4591	68272	102889
2	African Region (WHO)	13698	7628	6760	4156	62539	94781
185	South-East Asia Region (...	17794	5336	2163	837	11861	37991
55	East Asia & Pacific (WB)	30611	5004	1714	465	2529	40323
160	Region of the Americas (...	20795	4139	2018	428	3736	31116
104	Latin America & Caribbe...	15716	3428	1916	425	3725	25210
89	Indonesia	13688	2987	783	204	979	18641
182	South Asia (WB)	2646	2558	1784	1302	18181	26471
146	OECD Countries	19239	2239	857	124	613	23072
88	India	1198	1767	506	531	9382	13384
219	Western Pacific Region (...	15058	1672	799	236	1325	19090
56	Eastern Mediterranean R...	1924	1487	1470	1264	14128	20273

(5) Cinquième analyse : Les plus de 70 ans :

Après avoir adapté les paramètres, nous voila avec une nouvelle légende.



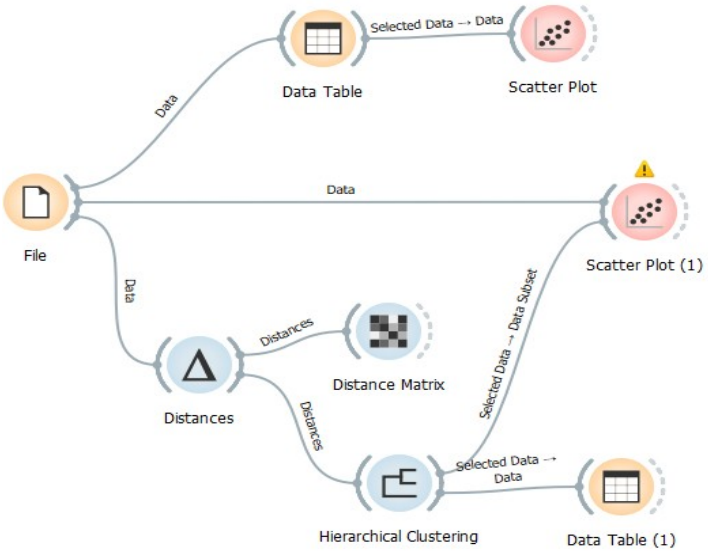
Ainsi, on remarque ici des données vraiment différentes des quatre autres analyses.

Le premier pays qui souffre le plus de la malnutrition est East Asia et Pacific qui est dans la zone jaune claire, il a donc plus de 30 000 morts.

	Entity	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
55	East Asia & Pacific (WB)	30611	5004	1714	465	2529	40323
160	Region of the Americas (...	20795	4139	2018	428	3736	31116
146	OECD Countries	19239	2239	857	124	613	23072
185	South-East Asia Region (...	17794	5336	2163	837	11861	37991
104	Latin America & Caribbe...	15716	3428	1916	425	3725	25210
219	Western Pacific Region (...	15058	1672	799	236	1325	19090
188	Sub-Saharan Africa (WB)	14320	8247	7459	4591	68272	102889
2	African Region (WHO)	13698	7628	6760	4156	62539	94781
89	Indonesia	13688	2987	783	204	979	18641
39	China	11339	916	362	43	439	13099
67	European Region (WHO)	7797	564	202	22	83	8668

Par la suite, j'ai ajouté à 'File' le module 'Distances' qui va regarder les éléments les plus similaires en renvoyant une distance euclidienne. La Distance Matrix servira par la suite à visualiser les distances calculées précédemment.

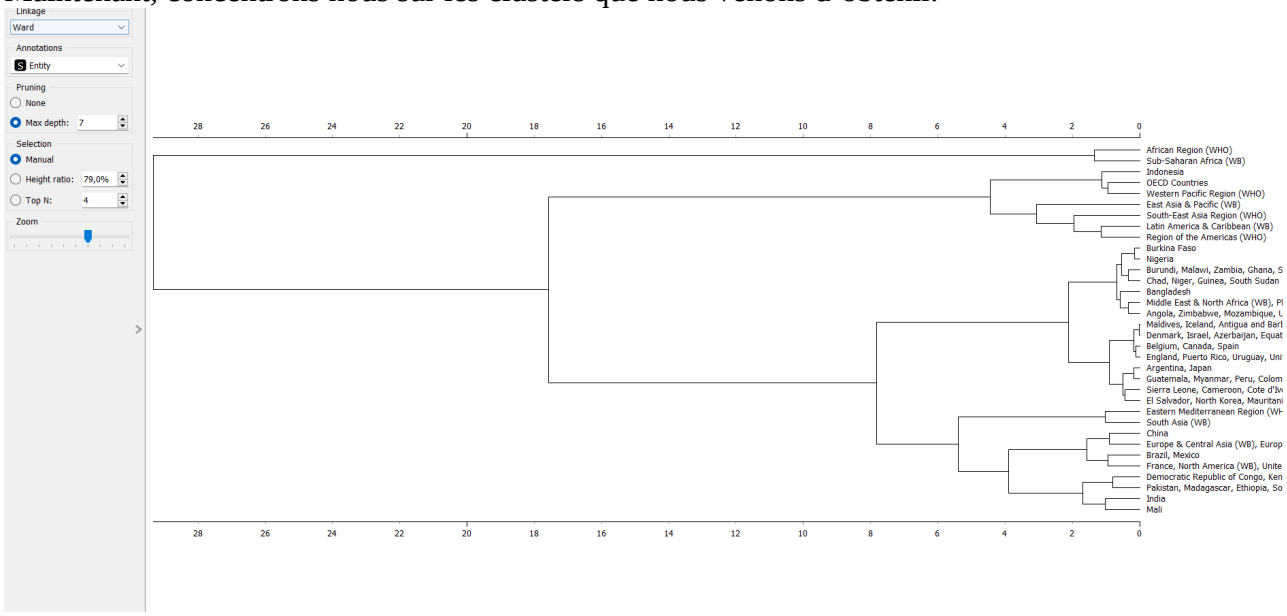
J'ai mit un 'Hierarchical Clustering' qui me permettra de regrouper les éléments suivant leurs distances. Puis, pour les visualiser plus facilement, j'ai ajouté un 'Scatter Plot' qui prend en entrée 'File' ainsi que le cluster que l'on sélectionnera.



Puis pour avoir un accès plus simple aux données que je sélectionnerais, j'ajoute en sortie du cluster une Data Table.

b) Analyse des données

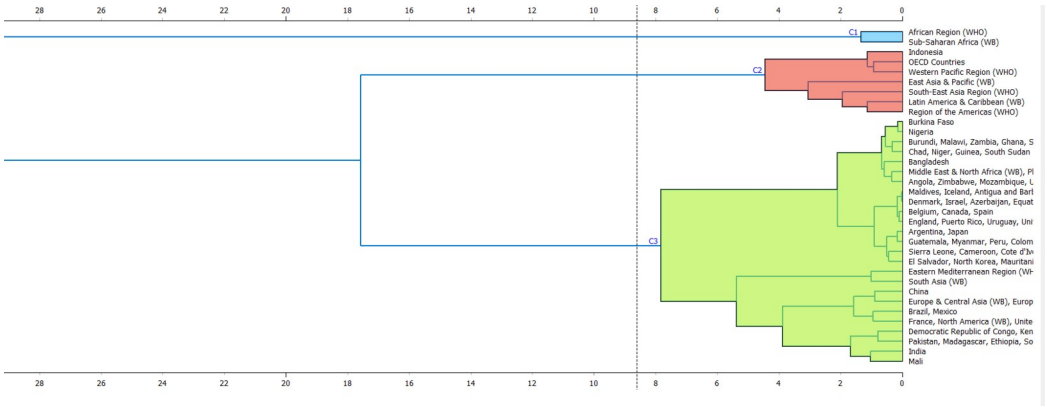
Maintenant, concentrons nous sur les clusters que nous venons d'obtenir.



On peut remarquer qu'il y a trois gros clusters.

(1) Gros cluster 1 :

Ainsi, le premier cluster regroupe African Region et Sub-Saharan Africa. Ces deux pays souffrent le plus de la malnutrition dans la majeur partie des tranches d'âges. Il est donc logique de les retrouver ensemble.



(2) Gros cluster 2 :

Le deuxième cluster regroupe les pays se trouvant dans la moyenne, en effet, ils sont dans la moyenne de toutes les tranches d'âges.

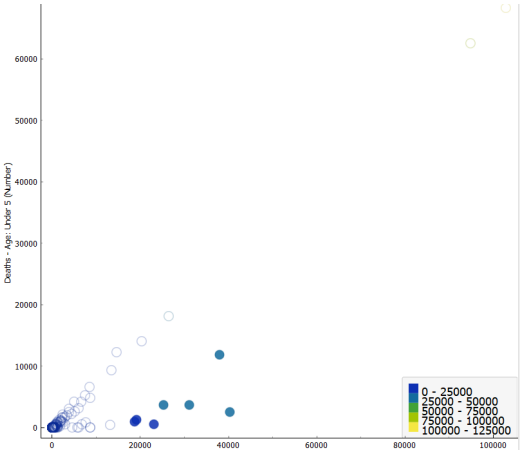


Illustration 2: Moins de 5 ans

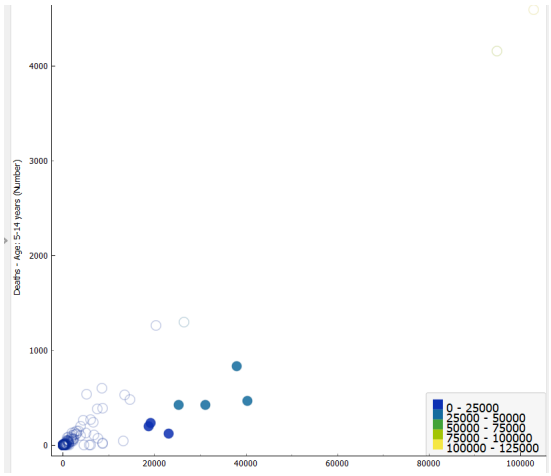


Illustration 1: 5 - 14 ans

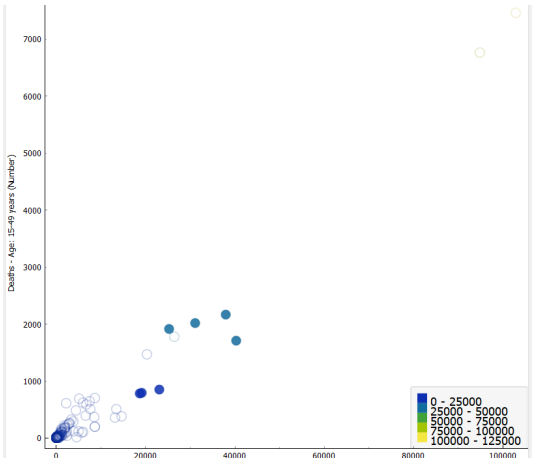


Illustration 4: 15 - 49 ans

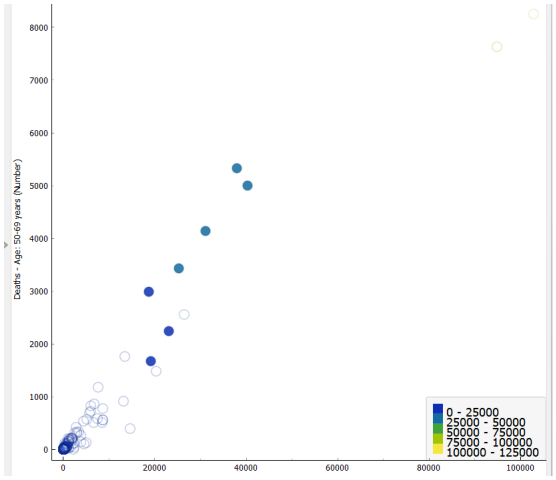


Illustration 3: 50 - 69 ans

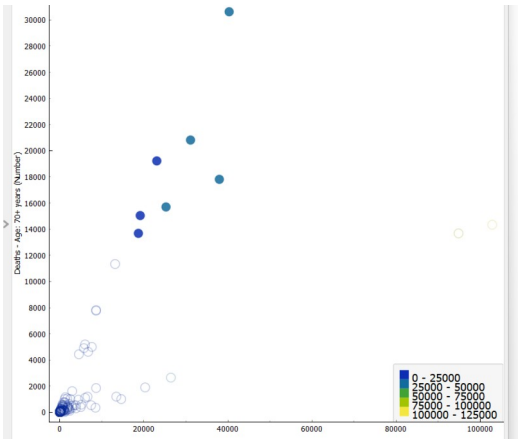
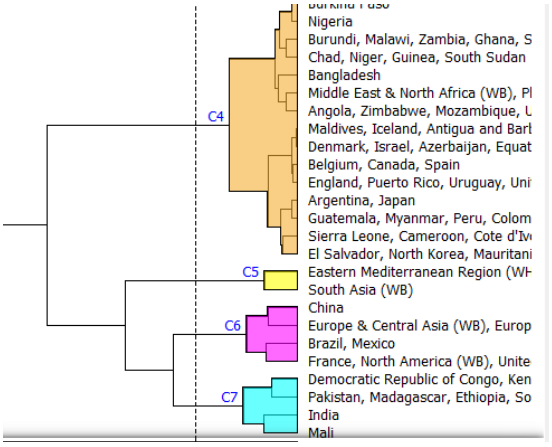


Illustration 5: Plus de 70 ans

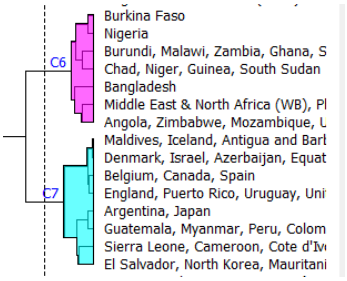
(3) Gros cluster 3 :

Le troisième gros cluster comprend donc tous les pays proches de l'origine, car ils ne sont pas les plus touchés par la malnutrition. Mais ils restent quand même assez présents.

- Dans ce gros cluster 3, nous pouvons voir distinctement 4 clusters moyens.



1. Dans le premier cluster moyen nous retrouvons à son tour deux petits clusters distincts.

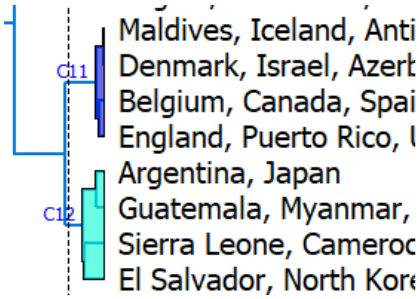


a) Concentrons nous d'abord sur le premier contenant le Burkina Faso, le Nigeria,…

Ces pays sont plutôt proches. En effet, ils se trouvent tous vers l'origine. Par la suite, quand nous regardons leurs données, nous pouvons constater qu'ils sont tous en-dessous de 5 000 morts au total.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
1	Angola	C1	372	313	254	123	1844	2906
2	Bangladesh	C1	620	7	614	45	996	2282
3	Burkina Faso	C1	328	143	125	102	3132	3830
4	Burundi	C1	269	216	187	132	1177	1981
5	Chad	C1	256	123	115	111	1771	2376
6	Ghana	C1	480	189	160	70	960	1859
7	Guinea	C1	158	59	54	68	1346	1685
8	Malawi	C1	417	206	192	99	1066	1980
9	Middle East & ...	C1	986	322	239	135	876	2558
10	Mozambique	C1	536	338	322	152	2014	3362
11	Niger	C1	95	50	44	64	2117	2370
12	Nigeria	C1	410	137	121	130	4199	4997
13	Philippines	C1	1632	336	271	154	618	3011
14	South Africa	C1	363	221	180	56	1216	2036
15	South Sudan	C1	230	119	92	67	1631	2139
16	Uganda	C1	531	274	289	202	2529	3825
17	Zambia	C1	340	206	220	82	980	1828
18	Zimbabwe	C1	476	428	241	114	1596	2855

b) Puis, par la suite, regardons le deuxième cluster, avec les Maldives, la Belgique, le Japon,… Nous pouvons ainsi redécouper le cluster en deux parties.



- Ce premier petit cluster comprend au total 138 pays sachant qu'ils sont tous en-dessous de 340 morts dues la malnutrition en 2019.

La Belgique prend la tête du classement avec 339 morts au total.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
14	Belgium	C1	325	12	2	0	0	339
114	Spain	C1	282	15	3	0	1	301
22	Canada	C1	272	26	3	0	0	301
68	Malaysia	C1	179	38	7	2	4	230
123	Togo	C1	58	31	28	10	92	219
93	Papua New Gui...	C1	66	24	16	9	88	203
32	Djibouti	C1	36	28	21	8	107	200
64	Lesotho	C1	25	28	23	8	105	189
79	Namibia	C1	35	22	16	6	104	183
19	Botswana	C1	19	19	18	6	119	181
61	Laos	C1	77	21	16	7	55	176
51	Iran	C1	108	18	19	6	22	173
78	Morocco	C1	94	29	17	8	24	172
115	Sri Lanka	C1	109	28	7	1	3	148
83	Nicaragua	C1	41	16	11	7	68	143
7	Australia	C1	121	11	2	0	0	134
95	Portugal	C1	119	5	1	0	1	126
81	Netherlands	C1	118	6	1	0	0	125
113	South Korea	C1	108	10	4	0	0	122
92	Panama	C1	46	12	9	5	48	120
132	United Kingdom	C1	97	15	5	0	1	118
30	Czechia	C1	95	19	4	0	0	118
120	Taiwan	C1	91	19	5	0	0	115

- Le deuxième petit cluster comprend 38 pays, sachant qu'ils ont tous un total inférieur à 1750 décès.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
18	Guatemala	C1	1059	225	179	41	239	1743
22	Japan	C1	1128	208	57	1	2	1396
9	Colombia	C1	588	182	112	29	376	1287
2	Argentina	C1	1001	147	48	10	49	1255
7	Central African ...	C1	97	145	118	44	808	1212
1	Afghanistan	C1	12	35	163	90	911	1211
25	Myanmar	C1	749	178	68	18	171	1184
29	Peru	C1	761	157	97	25	131	1171
33	Sierra Leone	C1	87	33	35	36	953	1144
31	Rwanda	C1	215	126	103	56	595	1095
16	Eritrea	C1	154	153	147	46	506	1006
26	Nepal	C1	354	205	124	29	276	988
38	Yemen	C1	158	70	100	78	535	941
35	Thailand	C1	787	110	28	3	5	933
19	Haiti	C1	96	61	60	48	650	915
6	Cameroon	C1	187	84	89	43	496	899

2. Dans le deuxième cluster moyen nous retrouvons South Asia et Eastern Mediterranean Region. Leurs chiffres presque similaire les relient dans les différentes tranches d'âges. Les deux pays tournent autour 2 300 morts au total.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
1	Mediterranean Region (W...	C1	1924	1487	1470	1264	14128	20273
2	ia (WB)	C1	2646	2558	1784	1302	18181	26471

3. Dans le troisième cluster regroupant le Chine, le Brésil, le France,…

On peut supposer que ces pays se regroupent car ils sont tous autour de 6 000 morts, exceptée la Chine qui est à plus de 13 000 décès par la malnutrition en 2019.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
1	Brazil	C1	5002	1177	508	72	890	7649
2	China	C1	11339	916	362	43	439	13099
3	Europe & Centr...	C1	7757	562	201	21	83	8624
4	European Regio...	C1	7797	564	202	22	83	8668
5	France	C1	4455	109	15	0	1	4580
6	Mexico	C1	4624	871	593	107	550	6745
7	North America (...	C1	5170	722	105	3	12	6012
8	United States	C1	4898	697	102	3	11	5711

4. Puis, nous finissons par le dernier cluster contenant l'Inde, le Mali, le Pakistan,..

Ces pays sont tous au alentour des 7 000 morts, sauf le Mali et l'Inde qui ont un total supérieur à 13 500 décès.

	Entity	Cluster	- Age: 70+ years (N	Age: 50-69 years (	Age: 15-49 years (	- Age: 5-14 years (N	- Age: Under 5 (Nu	Total
1	Democratic Re...	C1	1108	834	628	272	3187	6029
2	Ethiopia	C1	1865	779	703	389	4868	8604
3	India	C1	1198	1767	506	531	9382	13384
4	Kenya	C1	957	533	486	265	2221	4462
5	Madagascar	C1	590	580	694	541	2698	5103
6	Mali	C1	1000	393	378	483	12281	14535
7	Pakistan	C1	351	515	370	605	6611	8452
8	Somalia	C1	537	587	646	387	5318	7475
9	Tanzania	C1	1213	506	394	243	4258	6614

On peut donc conclure que les pays les plus touchés par la malnutrition sont ceux du premier gros cluster, avec un total supérieur ou égal à 95 000 morts voire à plus de 102 000 morts, en 2019.
Ainsi, le clustering hiérarchique que nous venons d'analyser regroupe correctement les pays ayant à peu près les mêmes résultats.
On peut en conclure que les pays les plus concernés sont ceux en voie de développement. Ou encore ceux qui sont victimes de surpopulations ou de mauvaises conditions de vies.

2- Apprentissage supervisé

Comment l'algorithme de classification supervisée réussira à différencier les Champignons Toxiques des Non Toxiques ?

2.1 – Jeux de données

J'ai décidé de me lancer dans la classification des champignons toxiques, des non-toxiques. Cette simulation nous aidera à savoir si Orange arrivera à différencier les différentes sortes de champignons.

Sur internet, j'ai trouvé des noms de champignons comestibles comme les cèpes, les morilles, les trompettes de la mort,...
Puis quelques noms de non-comestibles, comme l'amanite tue mouches, les Laccaire améthystes,…

Par la suite, j'ai cherché des images, en essayant de varier les paramètres des photos (exposition au soleil, seul ou en groupe, focus ou éloigné,…). J'ai pris 30 images pour chaque catégorie.
Après avoir rangé les photos dans deux dossiers « Comestibles » ou « Toxiques », j'ai pris d'autres images de ces espèces, que j'ai ajouté à un nouveau dossier, qui servira de Test.

Le premier dossier entraînera donc le système à trier les deux sortes de champignons, puis le deuxième sera celui qui servira de test.

2.2 – Nettoyer et prétraitement des données

Je n'ai pas eu besoin de nettoyer ou de pré-traiter les données, en effet, car j'ai fait moi-même le dossier contenant les photos.

2.3 – Modèles d'apprentissage

a) Description de la chaîne de traitement :

J'ai commencé par importer mon dossier 'ImagesChampignons' avec l'aide d'Import Images. Par la suite avec Image Viewer j'ai vérifié que les 60 images s'affichaient correctement.

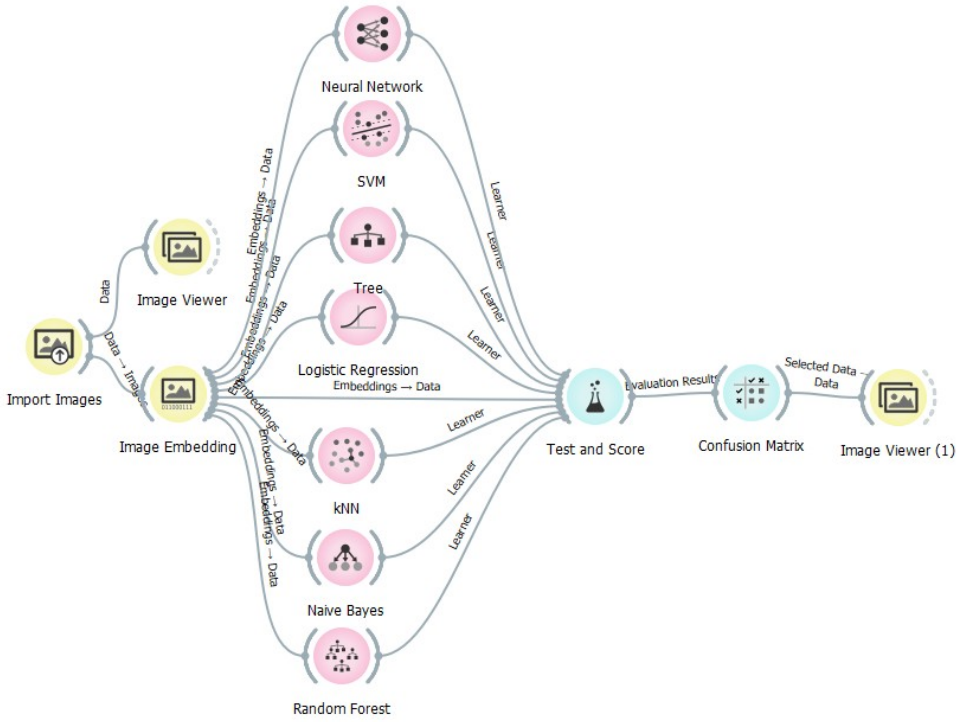
Ensuite, Image Embedding servira à traiter les images. Car Image Embedding permet de transformer des photos en vecteurs de nombres appelés plongement d'image. Il va se baser sur les modèles d'apprentissages profond qui ont été entraînés sur des très grands jeux de données.

J'ai utilisé sept modèles de traitement différents pour prendre le plus performant : Neural Network, SVM, Tree, Logistic Regression, KNN, Naive Bayes et Ramdom Forest.

J'ai décidé de comparer les différents modèles afin de prendre le plus optimal et le plus juste dans ses traitement de données.

J'ai relié les différents modèles à 'Test and Score', qui me permettra de les comparer entre eux suivant leur Accuracy, Précision, Rappel,… Car il permet de générer une validation croisée du model. Test and Score que je relie également à Image Embedding.

A la sortie de Test and Score, j'ai placé une Matrice de confusion afin de voir plus clairement quel modèle à les meilleures données biens placées, donc l'Accuracy. Puis pour pouvoir remarquer les points positifs des probables erreurs faites par les méthodes, j'ai placé Image Viewer en sortie de la matrice.

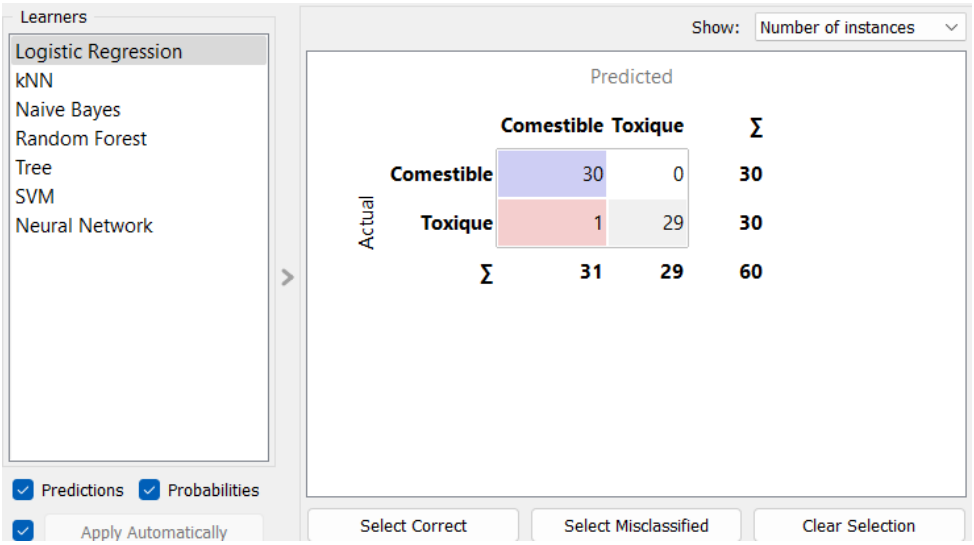


b) Analyse des données

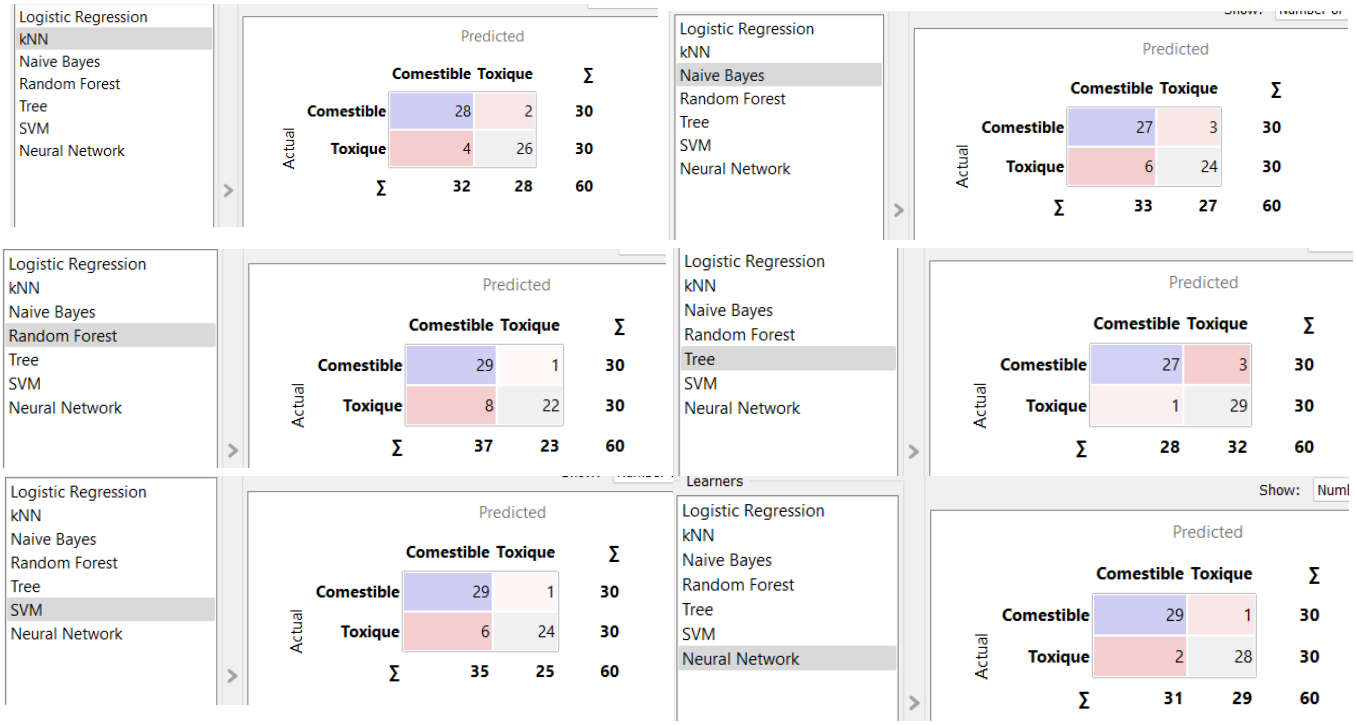
Il a donc fallu trouver le modèle le plus performant, et donc celui qui aurait l'Accuracy le plus proche de 1, ainsi que les meilleurs : rappel, précision, F1,…

Model	AUC	CA	F1	Precision	Recall
KNN	0.944	0.900	0.900	0.902	0.900
Tree	0.880	0.933	0.933	0.935	0.933
SVM	0.973	0.883	0.8825	0.894	0.883
Ramdom Forest	0.943	0.850	0.848	0.870	0.850
Neural Network	0.992	0.950	0.950	0.951	0.950
Naive Bayes	0.968	0.850	0.850	0.854	0.850
Logistic Regression	*0.998*	*0.983*	*0.983*	*0.984*	*0.983*

Quand on regarde la matrice de confusion, on voit clairement que le modèle le plus performant s'est trompé qu'une seule fois.



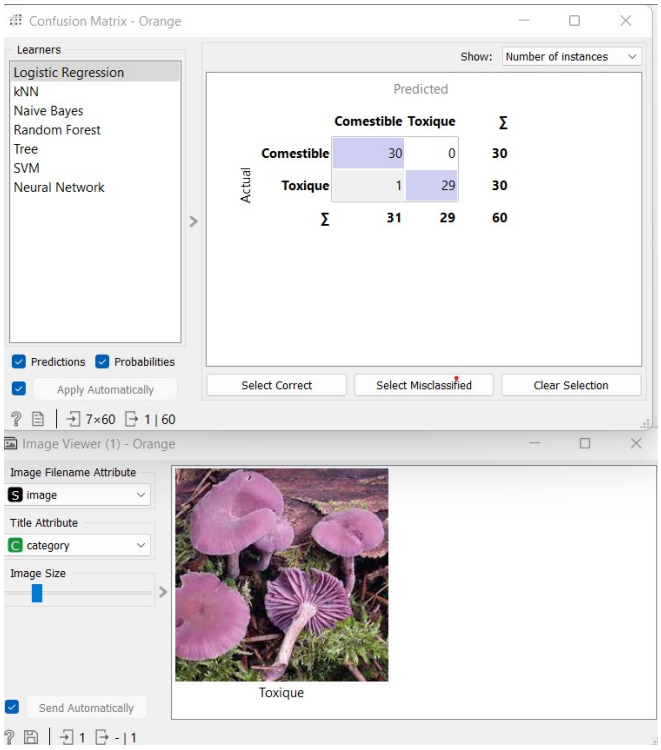
Alors que les six autres modèles sont beaucoup moins performants car on se rend compte qu'ils se sont trompés plus de fois que Logistic Regression.



On va donc se concentrer sur Logictic Regression.

Cassandra Sénécaille 22100082

J’ai d’abord essayé de savoir pourquoi le modèle s’était trompé une fois.



On remarque donc que son erreur est d’avoir classé dans champignon Comestible, un Toxique.

La question que nous pouvons nous poser est : Pourquoi le modèle se trompe sur cette image, alors qu’il a classé correctement les autres images de cette même variété de champignon toxique ?

On va alors se concentrer dans un premier temps sur l’image en elle-même, puis sur les photos pouvant se rapprochées de celle-ci dans le dossier des champignons Comestibles.

Toxique

Toxique

- Sur la photo, on peut se rendre compte que nous pouvons retrouver cinq champignons, de couleur violette et plutôt claire. Un des champignons est couché sur le sol, nous montrant le dessous de sa ‘tête’.

  Quand on compare cette photo avec les autres de la même espèce dans le dossier Champignons toxiques, on peut remarquer qu’un autre cliché ressemble fortement à celui qui nous pose problème.

  Or on peut voir une différence entre les deux photos : la couleur des champignons. Celle qui a été correctement classée est d’une couleur très foncée, par rapport à celle qui a été rangée dans Comestible.

Comestible

Toxique

- On va donc se concentrer sur le dossier des champignons comestibles, et donc voir si certains pourraient ressembler à celui que nous cherchons à bien classer.

  On remarque qu'il pourrait y avoir quelques ressemblances avec des champignons comestibles.

  Par exemple, pour cette photo, on peut dire que la forme de la 'tête' des deux champignons se ressemble. Et donc que cette ressemblance serait interprétée par le modèle comme deux images de la même espèce de champignon, même si la couleur n'est pas identique.

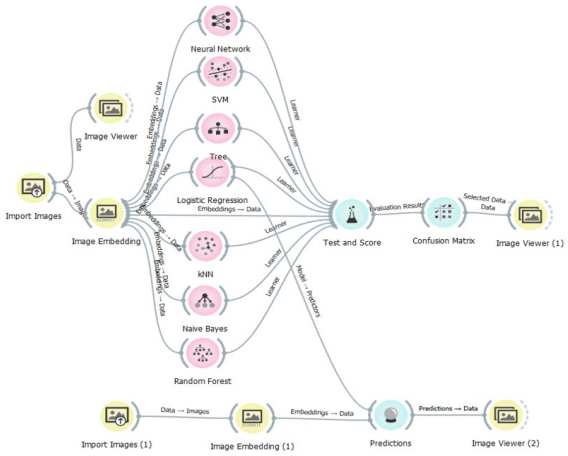
  Cette ressemblance peut donc expliqué pourquoi ce champignon toxique a été classé dans les champignons comestibles.

2.4 – Prédictions

a) Description de la chaîne de données :

Après avoir entraîné le modèle d'apprentissage, j'ai recherché d'autres photos, afin de les mettre dans un fichier test, qui comportera en tout 20 photos. Ce fichier nous permettra de vérifier que les modèles ont bien utilisés le dossier d'apprentissage.
On va donc se baser sur Logistic Regression, car c'est le meilleur modèle d'apprentissage sur nos données.

J'ai donc importé le dossier avec 'Import Image', que j'ai relié à Image Embedding. Ensuite, j'ai relié Logistic Regression, le modèle que nous avons entraîné plus tôt, à l'entrée de Prédictions, se trouvant en sortie de Image Embedding. Puis, pour que cela soit plus compréhensible, j'ai ajouté un Image Viewer à la fin.



b) Analyse des données prédites:

Regardons le résultat de la prédiction entraînée par le modèle Logistic Regression.

Quelles-sont les données **correctement prédites** et celles où ils y a eu une **erreur** ?

	Logistic Regression	image name	image	size	width	height
1	Comestible	Comestible1	Comestible1.jpg	18443	286	176
2	Comestible	Comestible10	Comestible10.jpg	11588	259	194
3	Comestible	Comestible2	Comestible2.jpg	178918	1220	792
4	Comestible	Comestible3	Comestible3.jpg	417900	1600	1200
5	Comestible	Comestible4	Comestible4.jpg	10758	300	168
6	Comestible	Comestible5	Comestible5.jpg	15086	276	183
7	Comestible	Comestible6	Comestible6.jpg	97211	907	605
8	Comestible	Comestible7	Comestible7.jpg	13822	259	194
9	Comestible	Comestible8	Comestible8.jpg	14277	259	194
10	Comestible	Comestible9	Comestible9.jpg	11904	299	169
11	Toxique	Toxique1	Toxique1.jpg	183215	1250	625
12	Comestible	Toxique10	Toxique10.jpg	11443	275	183
13	Toxique	Toxique2	Toxique2.jpg	17395	299	169
14	Toxique	Toxique3	Toxique3.jpg	10274	275	183
15	Toxique	Toxique4	Toxique4.jpg	14097	275	183
16	Toxique	Toxique5	Toxique5.jpg	253927	1600	1067
17	Comestible	Toxique6	Toxique6.jpg	7416	259	194
18	Comestible	Toxique7	Toxique7.jpg	16606	275	183
19	Toxique	Toxique8	Toxique8.jpg	12016	266	189
20	Toxique	Toxique9	Toxique9.jpg	225743	837	570

On remarque donc que le modèle s'est trompé 3 fois sur la prédiction. On va donc regarder les photos et essayer de comprendre les erreurs.

(1) Première erreur:

On remarque que le champignon toxique a été classé dans comestible. Regardons donc pourquoi :
On remarque que cette variété de champignons ressemble fortement aux Cèpes.

En effet, ils ont la même couleur et à peut près la même forme. Ces caractéristiques peuvent donc jouer sur la prédiction.

Par la suite, on peut remarquer que le champignon toxique possède une sorte de creux sur sa ‘tête’, qui n’est pas très visible.

Mais, on remarque que dans les champignons toxiques dans le dossier d’entraînement, il y a des images de la même variété de champignons.

On va donc se demander pourquoi le modèle n’a pas réussi à bien classé cette photo.
On peut supposer que lors de la transformation en vecteur par Image Embedding, tous les détails discrets n’ont pas été repérés, ce qui expliquerait cette mauvaise prédiction.

Toxique10

Illustration 1: Champignon Toxique mal prédit

Comestible

Illustration 2: Cèpe Comestible

Toxique

Illustration 3: Champignon Toxique bien prédit

(2) Deuxième erreur :

Cette deuxième erreur de prédiction est sur un champignon toxique classé dans les champignons comestibles.
Analysons le champignon : On remarque que c'est un champignon de couleur violette, il parait assez grand en hauteur et possède des petites 'têtes' assez arrondies et une tige régulière et pas trop épaisse.

Toxique6

Illustration 6: Champignon toxique mal prédit

On suppose que le modèle n'a pas réussi à l'associer aux champignons toxiques par sa ressemblance avec quelques champignons comestibles. Comme par exemple ces trois photos :

Comestible Comestible Comestible

Illustration 7: Champignons comestibles bien classés

En effet, on remarque que ces trois photos peuvent ressembler à celle mal prédite. On peut voir sur la première photo des tiges assez régulières, pas trop épaisses et avec des têtes arrondies.
Sur la deuxième photo, on peut voir assez rapidement la tête bien ronde du champignon.
Puis, sur la troisième photo, on peut constater une tête arrondie, malgré le fait que la tige soit assez épaisse et non régulière.
On peut donc se demander si ces trois photos ne joueraient pas un rôle sur la mauvaise prédiction de la photo 17.

(3) Troisième erreur :

Regardons maintenant la troisième erreur de prédiction.
La photo 18, un champignon toxique qui à été classé dans les champignons comestibles.

Analysons cette photo rapidement. On peut voir un regroupement de trois champignons, avec des têtes assez arrondies et des tiges régulières et fines. On revient donc à la même analyse que pour la deuxième erreur.

Toxique7

Illustration 8: Champignon toxique mal prédit

En conclusion, on peut donc dire que le modèle Logistic Regression est plutôt performant. Car sur 20 photos il n'a fait que 3 erreurs.