ESISA ECOLE SUPÉRIEURE D'INGÉNIERIE EN SCIENCES APPLIQUÉES

AVENIR INGENIEUR

Analyse musicale des chansons sur Spotify

Projet réalisé par : Oukhdouch abdelaali et Ammari Salma

1. Description des données fournit par tanagra :

1513 chansons répartis / 9 classes et de 16 variable :

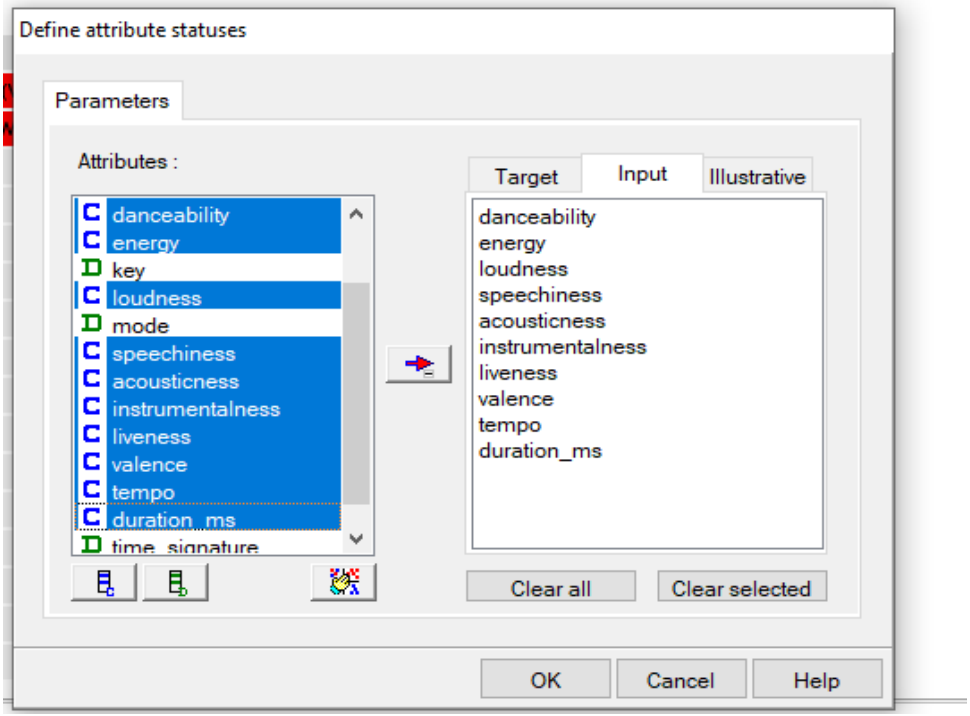
Attribute	Category	Informations
artiste	Discrete	(WARNING !!!) 975 values
titre	Discrete	(WARNING !!!) 1407 values
genre	Discrete	9 values
danceability	Continue	-
energy	Continue	-
key	Discrete	12 values
loudness	Continue	-
mode	Discrete	2 values
speechiness	Continue	-
acousticness	Continue	-
instrumentalness	Continue	-
liveness	Continue	-
valence	Continue	-
tempo	Continue	-
duration_ms	Continue	-
time_signature	Discrete	4 values

Description des données del'énonce :

Nom de la variable	Description	Type de variable
Artiste	Artiste de la chanson	Identifiant
Titre	Titre de la chanson	Identifiant
Genre	Genre de la chanson. Défini en fonction de la playlist d'origine.	Qualitative nominale
Danceability	Indice décrivant si un morceau convient à la danse sur la base d'une combinaison d'éléments musicaux (tempo, stabilité du rythme, ...), entre 0 et 1 (du moins dansant au plus dansant)	Quantitative
Energy	Indice décrivant si un morceau semble intense et énergique sur la base d'éléments tels que la dynamique du morceau, le volume sonore perçu. Entre 0 et 1 (du moins énergique au plus énergique)	Quantitative
Key	Tonalité du morceau par demi-tons, entre 0 (Do) et 12 (Si). La tonalité 4 sera par exemple Ré#/Mib.	Qualitative
Loudness	Intensité sonore globale d'une piste en décibels (dB).	Quantitative
Mode	Indique la modalité majeure (1) /mineure (0) d'une piste. Les chansons majeure "sonnent" plus joyeuses que les chansons mineures.	Qualitative
Speechiness	Détecte la présence de paroles dans une piste. Plus l'enregistrement contient de voix, plus la valeur sera proche de 1.	Quantitative
Acousticness	Indice déterminant si la chanson est acoustique (instrument non électrique), entre 0 et 1, du moins au plus acoustique.	Quantitative
Instrumentalness	Indice détectant la présence d'instruments, entre 0 et 1, du moins instrumental au plus instrumental.	Quantitative
Liveness	Détecte la présence d'un public dans l'enregistrement, entre 0 et 1 (plus probable que la chanson soit enregistrée en live).	Quantitative
Valence	Mesure de 0 à 1 évaluant la positivité musicale d'une piste. Les pistes à valence élevée ont un son plus positif.	Quantitative
Tempo	Tempo global estimé d'une piste en battements par minutes (bpm).	Quantitative
Duration time	Durée de la piste en millisecondes.	Quantitative
Time signature	Une estimation de la signature temporelle globale d'une piste en temps par mesure. Par exemple, une chanson pop moderne aura très certainement 4 temps par mesure, là où la valse a distinctement 3 temps par mesure.	Qualitative

Table 1 : Description du tableau de donnée

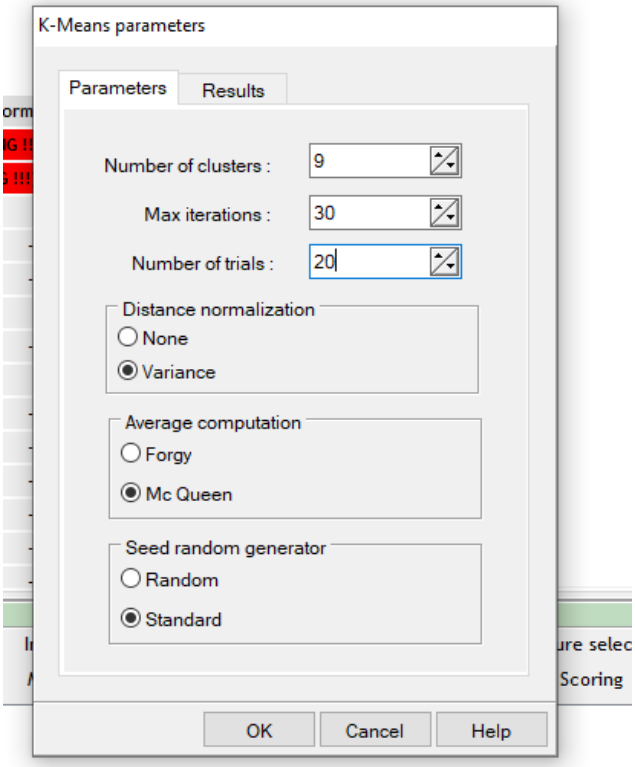
L’ensemble des données continues en entree :

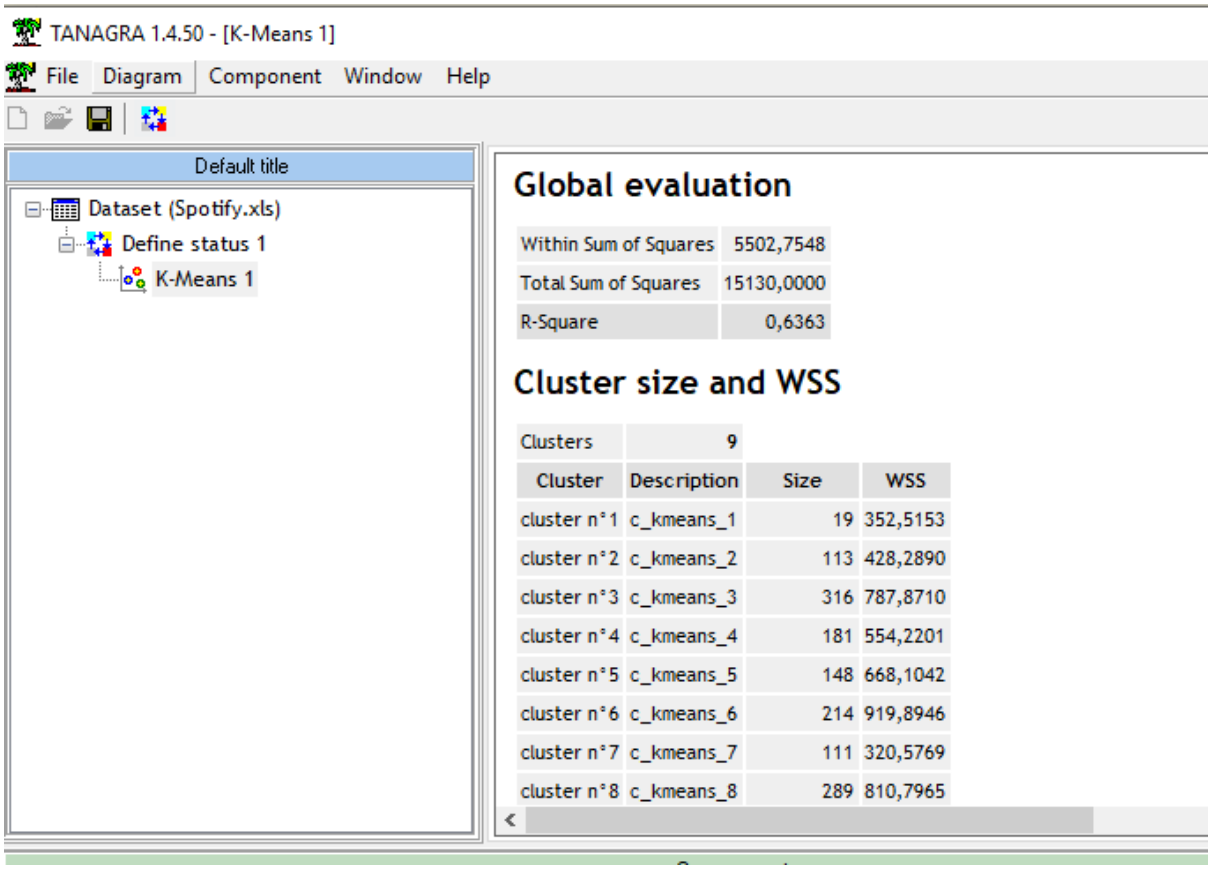


Choix de méthode de classification :

Compte tenu du nombre d'individus = 1513,

la méthode des k-moyennes (classification hiérarchique ascendante)





TANAGRA nous annonce qu'il y a respectivement 19 et 113, 316… observations dans chaque groupe. La partition (R-square) explique 63,63% de l'inertie totale.

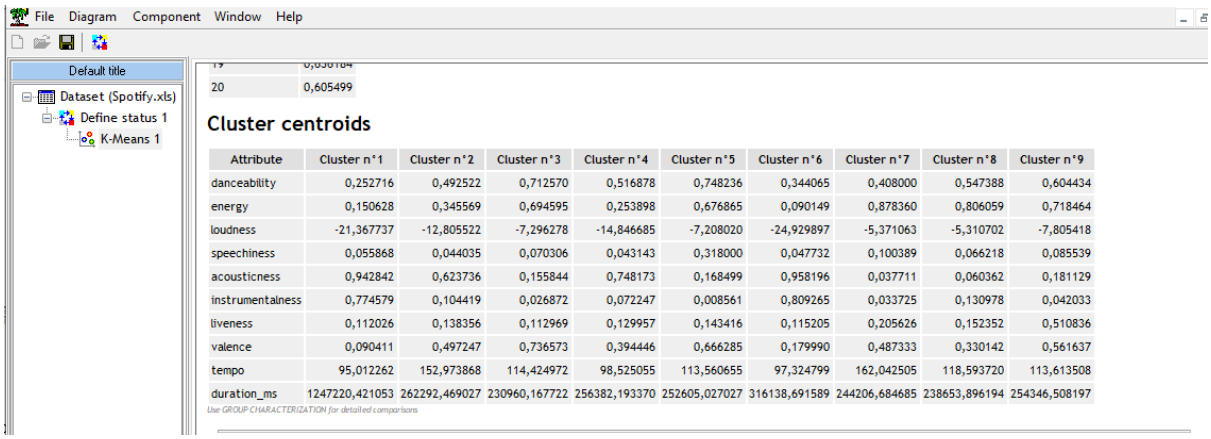
Dans la partie basse de la fenêtre de visualisation, On' a choisi 9 classes qui présentent les effectifs. Tanagra affiche les moyennes conditionnelles sur les variables ayant participé à la construction de la partition. Elles sont donc calculées sur les données centrées et réduites. Elles permettent de comprendre les différenciations entre les groupes, elles ne sont pas vraiment utilisables pour l'interprétation.

il y a respectivement 19,113, 316.

Carré R 63,63% de l'inertie totale.

9 classes qui ont été choisies qui présentent l'inscription. Tanagra affiche les moyennes conditionnelles sur les variables qui ont participé à la construction de la partition. Ils sont donc calculés sur les données centrées et réduites. Ils permettent de comprendre les différenciations entre groupes, ils ne sont pas vraiment utilisables pour l'interprétation.

Elles sont donc calculées sur les données centrées et réduites. Elles permettent de comprendre les différenciations entre les groupes, elles ne sont pas vraiment utilisables pour l'interprétation.

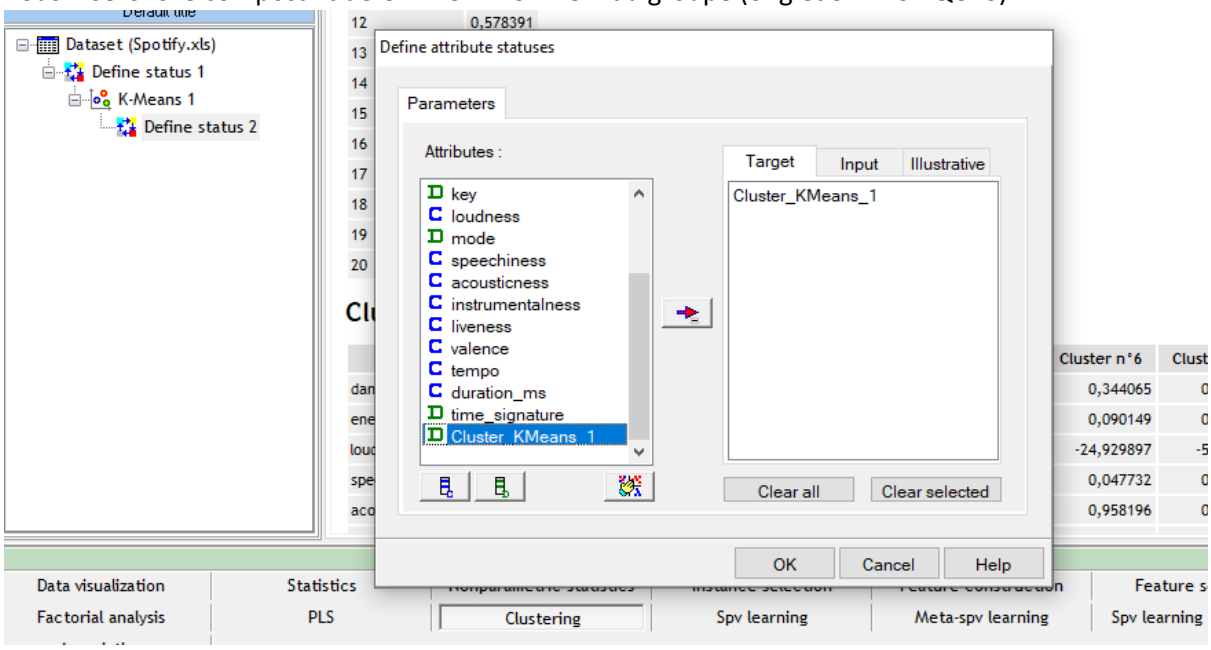


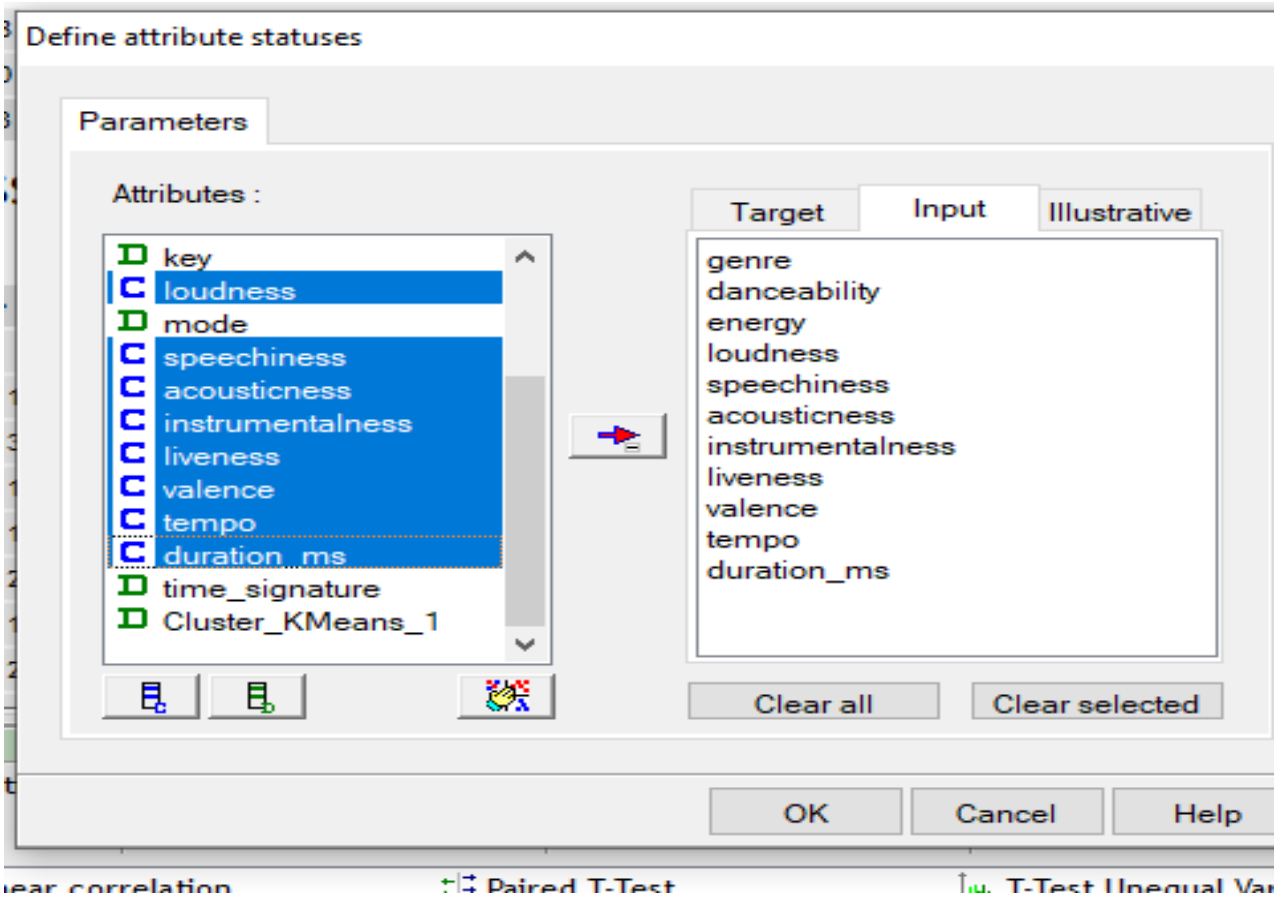
Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6	Cluster n°7	Cluster n°8	Cluster n°9
danceability	0,252716	0,492522	0,712570	0,516878	0,748236	0,344065	0,408000	0,547388	0,604434
energy	0,150628	0,345569	0,694595	0,253898	0,676865	0,090149	0,878360	0,806059	0,718464
loudness	-21,367737	-12,805522	-7,296278	-14,846685	-7,208020	-24,929897	-5,371063	-5,310702	-7,805418
speechiness	0,055868	0,044035	0,070306	0,043143	0,318000	0,047732	0,100389	0,066218	0,085539
acousticness	0,942842	0,623736	0,155844	0,748173	0,168499	0,958196	0,037711	0,060362	0,181129
instrumentalness	0,774579	0,104419	0,026872	0,072247	0,008561	0,809265	0,033725	0,130978	0,042033
liveness	0,112026	0,138356	0,112969	0,129957	0,143416	0,115205	0,205626	0,152352	0,510836
valence	0,090411	0,497247	0,736573	0,394446	0,666285	0,179990	0,487333	0,330142	0,561637
tempo	95,012262	152,973868	114,424972	98,525055	113,560655	97,324799	162,042505	118,593720	113,613508
duration_ms	1247220,421053	262292,469027	230960,167722	256382,193370	252605,027027	316138,691589	244206,684685	238653,896194	254346,508197

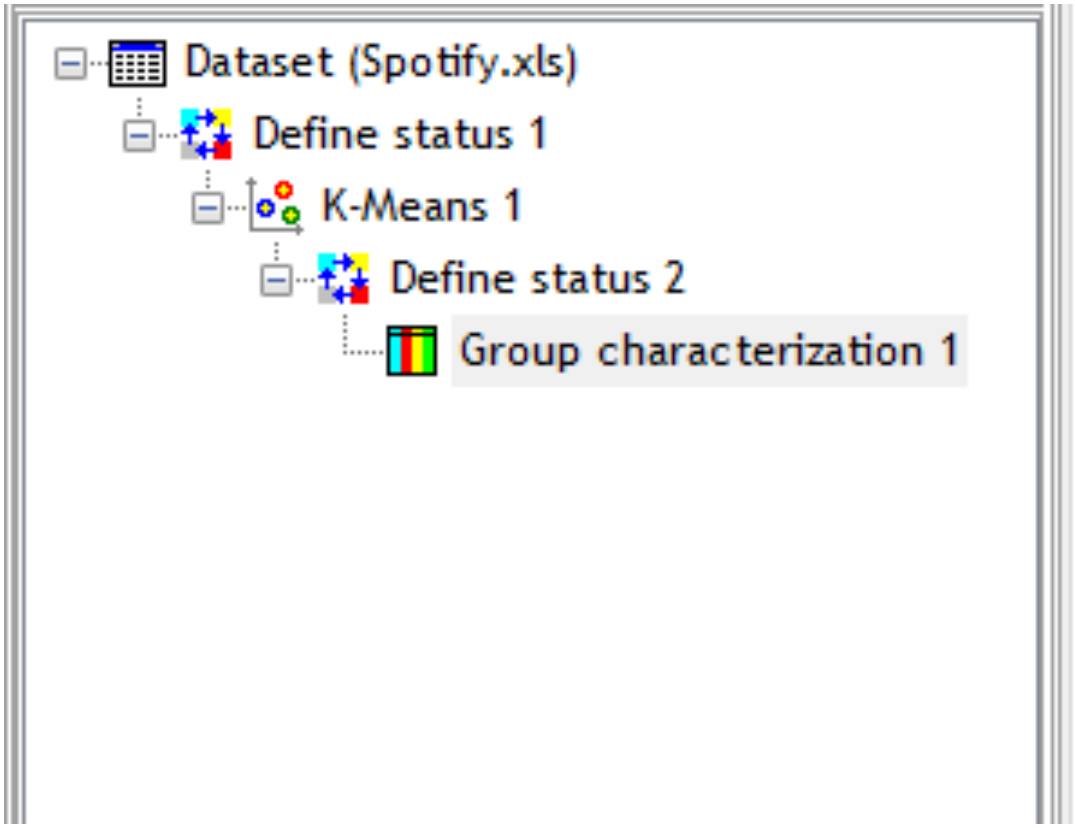
Use GROUP CHARACTERIZATION for detailed comparisons

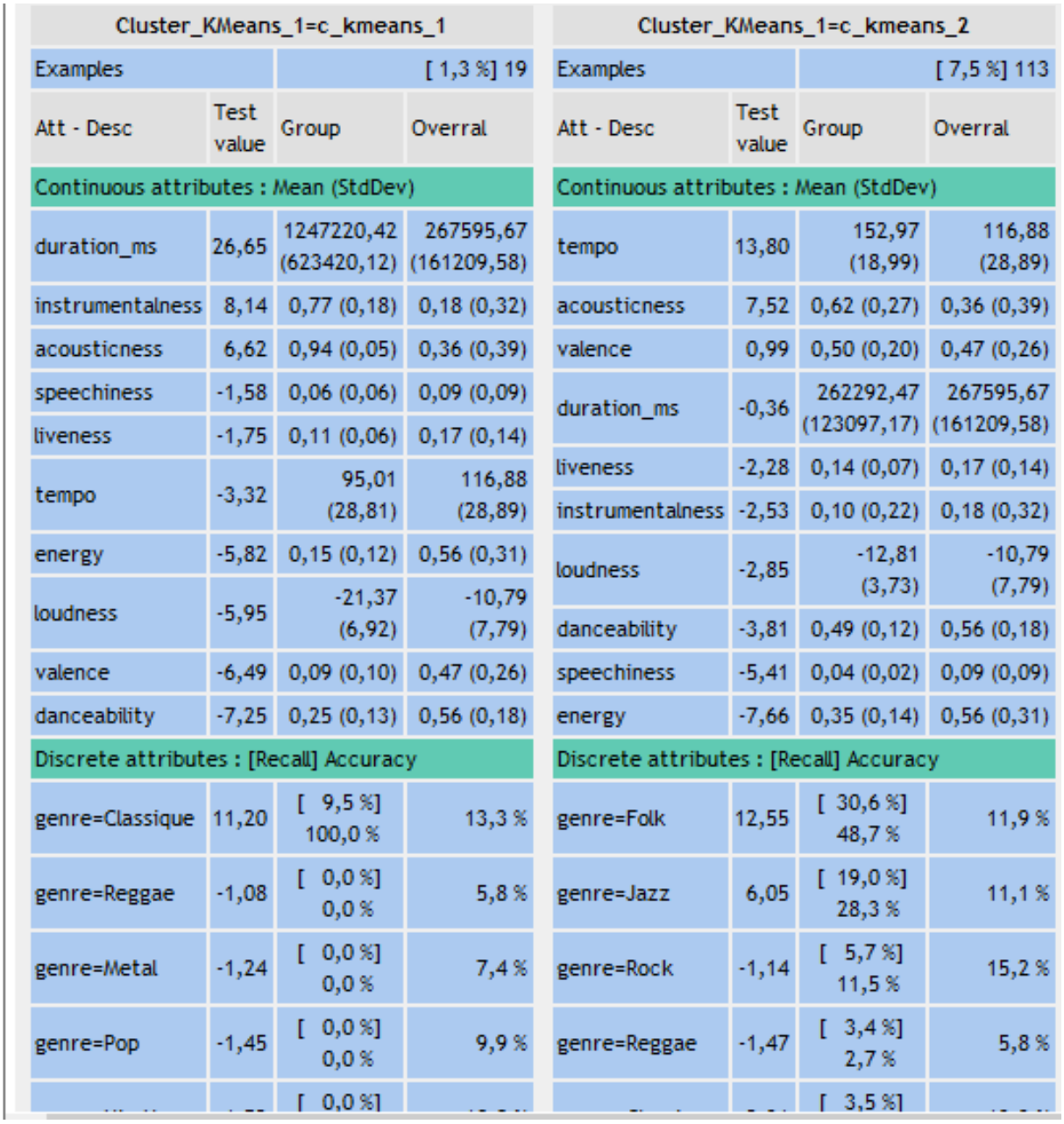
3. Statistiques descriptives comparatives

Nous présentons maintenant le composant DEFINE STATUS dans le diagramme. Nous plaçons dans TARGET la variable désignant les classes CLUSTER_KMEANS_1, dans INPUT les variables qualitatives et quantitatives de dansabilité, énergie, clé, mode,… et la variable pour expliquer le sexe. Ensuite, nous insérons le composant de CARACTÉRISATION du groupe (onglet STATISTIQUES).




Define attribute statuses
Parameters
Attributes :
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
time_signature
Cluster_KMeans_1
Target
Input
Illustrative
genre
danceability
energy
loudness
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
Clear all
Clear selected
OK
Cancel
Help


Dataset (Spotify.xls)
Define status 1
K-Means 1
Define status 2
Group characterization 1

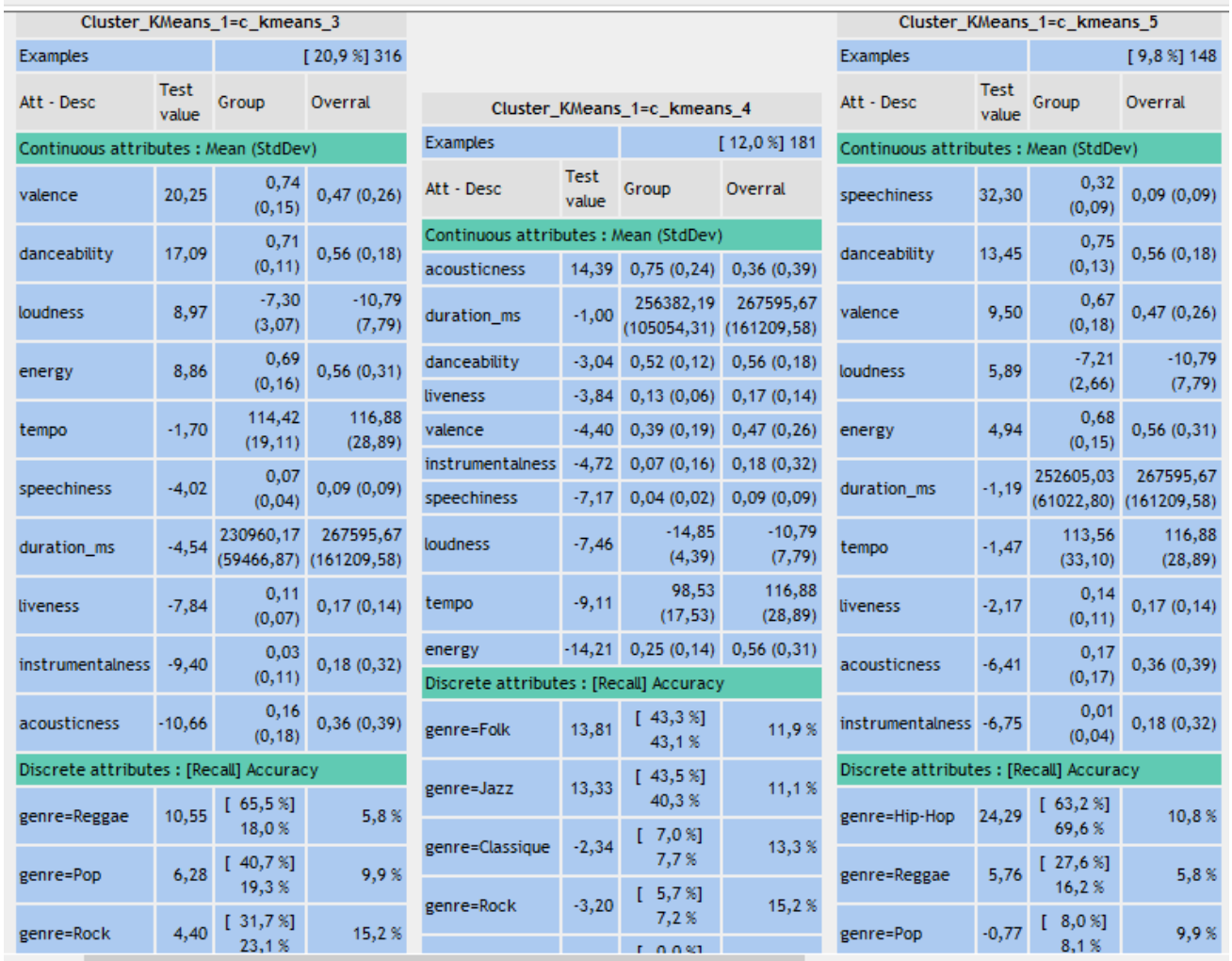


Cluster_KMeans_1=c_kmeans_1			
Examples			[1,3 %] 19
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
duration_ms	26,65	1247220,42 (623420,12)	267595,67 (161209,58)
instrumentalness	8,14	0,77 (0,18)	0,18 (0,32)
acousticness	6,62	0,94 (0,05)	0,36 (0,39)
speechiness	-1,58	0,06 (0,06)	0,09 (0,09)
liveness	-1,75	0,11 (0,06)	0,17 (0,14)
tempo	-3,32	95,01 (28,81)	116,88 (28,89)
energy	-5,82	0,15 (0,12)	0,56 (0,31)
loudness	-5,95	-21,37 (6,92)	-10,79 (7,79)
valence	-6,49	0,09 (0,10)	0,47 (0,26)
danceability	-7,25	0,25 (0,13)	0,56 (0,18)
Discrete attributes : [Recall] Accuracy			
genre=Classique	11,20	[9,5 %] 100,0 %	13,3 %
genre=Reggae	-1,08	[0,0 %] 0,0 %	5,8 %
genre=Metal	-1,24	[0,0 %] 0,0 %	7,4 %
genre=Pop	-1,45	[0,0 %] 0,0 %	9,9 %
		[0,0 %]	

Cluster_KMeans_1=c_kmeans_2			
Examples			[7,5 %] 113
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
tempo	13,80	152,97 (18,99)	116,88 (28,89)
acousticness	7,52	0,62 (0,27)	0,36 (0,39)
valence	0,99	0,50 (0,20)	0,47 (0,26)
duration_ms	-0,36	262292,47 (123097,17)	267595,67 (161209,58)
liveness	-2,28	0,14 (0,07)	0,17 (0,14)
instrumentalness	-2,53	0,10 (0,22)	0,18 (0,32)
loudness	-2,85	-12,81 (3,73)	-10,79 (7,79)
danceability	-3,81	0,49 (0,12)	0,56 (0,18)
speechiness	-5,41	0,04 (0,02)	0,09 (0,09)
energy	-7,66	0,35 (0,14)	0,56 (0,31)
Discrete attributes : [Recall] Accuracy			
genre=Folk	12,55	[30,6 %] 48,7 %	11,9 %
genre=Jazz	6,05	[19,0 %] 28,3 %	11,1 %
genre=Rock	-1,14	[5,7 %] 11,5 %	15,2 %
genre=Reggae	-1,47	[3,4 %] 2,7 %	5,8 %
		[3,5 %]	

On constate que le premier groupe C_K_MEANS_1 correspond davantage aux genres classiques. Notez que la musique du genre musical classique utilise une durée de piste en millisecondes ou instrumentale ou acoustique, ce qui implique que la classification est correcte.

Pour la deuxième classe C_K_MEANS_2 est associée au genre folk et jazz, ce genre utilise les variables de tempo et acoustiques, donc la classification est correcte.



Cluster_KMeans_1=c_kmeans_3

Examples		[20,9 %] 316	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
valence	20,25	0,74 (0,15)	0,47 (0,26)
danceability	17,09	0,71 (0,11)	0,56 (0,18)
loudness	8,97	-7,30 (3,07)	-10,79 (7,79)
energy	8,86	0,69 (0,16)	0,56 (0,31)
tempo	-1,70	114,42 (19,11)	116,88 (28,89)
speechiness	-4,02	0,07 (0,04)	0,09 (0,09)
duration_ms	-4,54	230960,17 (59466,87)	267595,67 (161209,58)
liveness	-7,84	0,11 (0,07)	0,17 (0,14)
instrumentalness	-9,40	0,03 (0,11)	0,18 (0,32)
acousticness	-10,66	0,16 (0,18)	0,36 (0,39)
Discrete attributes : [Recall] Accuracy			
genre=Reggae	10,55	[65,5 %] 18,0 %	5,8 %
genre=Pop	6,28	[40,7 %] 19,3 %	9,9 %
genre=Rock	4,40	[31,7 %] 23,1 %	15,2 %

Cluster_KMeans_1=c_kmeans_4

Examples		[12,0 %] 181	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
acousticness	14,39	0,75 (0,24)	0,36 (0,39)
duration_ms	-1,00	256382,19 (105054,31)	267595,67 (161209,58)
danceability	-3,04	0,52 (0,12)	0,56 (0,18)
liveness	-3,84	0,13 (0,06)	0,17 (0,14)
valence	-4,40	0,39 (0,19)	0,47 (0,26)
instrumentalness	-4,72	0,07 (0,16)	0,18 (0,32)
speechiness	-7,17	0,04 (0,02)	0,09 (0,09)
loudness	-7,46	-14,85 (4,39)	-10,79 (7,79)
tempo	-9,11	98,53 (17,53)	116,88 (28,89)
energy	-14,21	0,25 (0,14)	0,56 (0,31)
Discrete attributes : [Recall] Accuracy			
genre=Folk	13,81	[43,3 %] 43,1 %	11,9 %
genre=Jazz	13,33	[43,5 %] 40,3 %	11,1 %
genre=Classique	-2,34	[7,0 %] 7,7 %	13,3 %
genre=Rock	-3,20	[5,7 %] 7,2 %	15,2 %

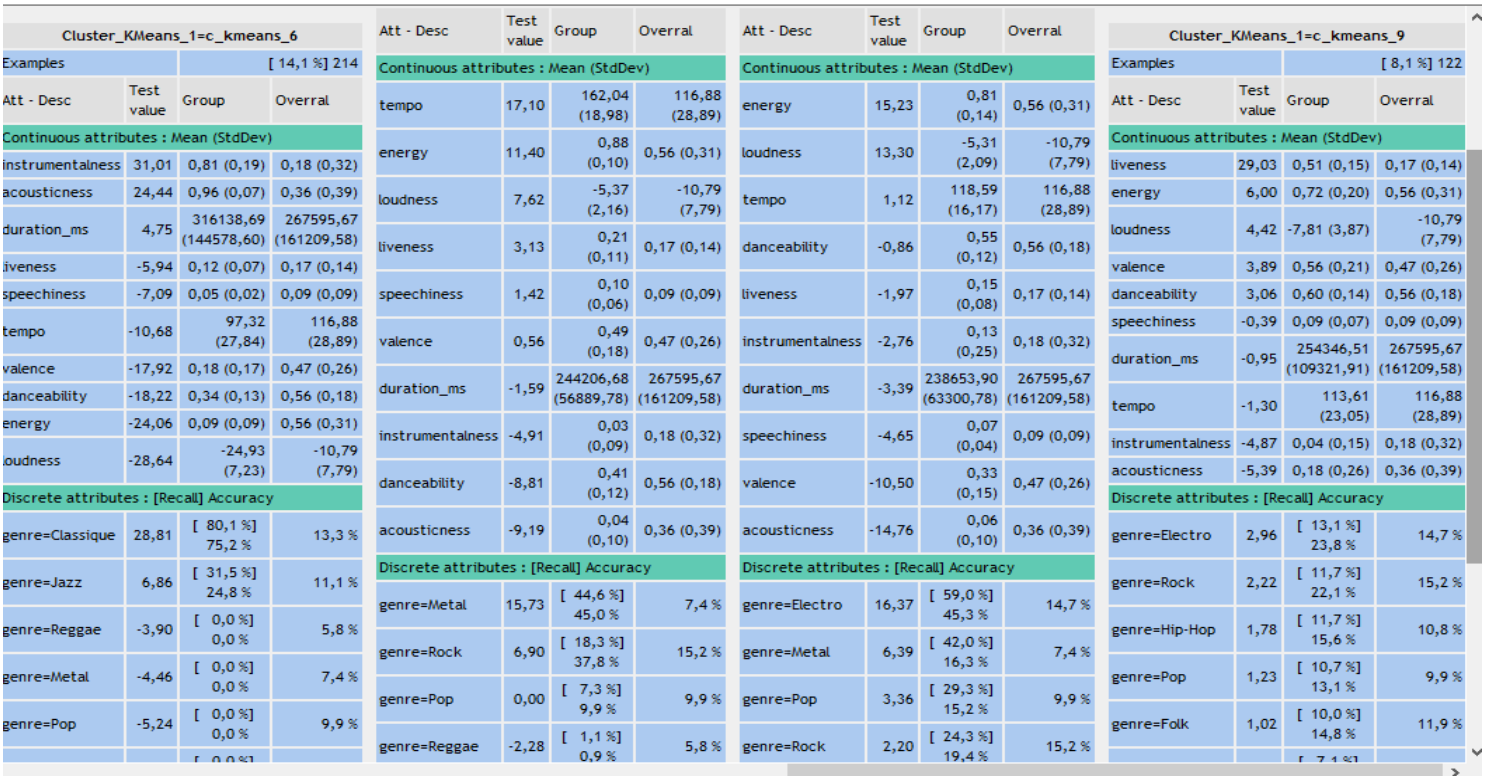
Cluster_KMeans_1=c_kmeans_5

Examples		[9,8 %] 148	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
speechiness	32,30	0,32 (0,09)	0,09 (0,09)
danceability	13,45	0,75 (0,13)	0,56 (0,18)
valence	9,50	0,67 (0,18)	0,47 (0,26)
loudness	5,89	-7,21 (2,66)	-10,79 (7,79)
energy	4,94	0,68 (0,15)	0,56 (0,31)
duration_ms	-1,19	252605,03 (61022,80)	267595,67 (161209,58)
tempo	-1,47	113,56 (33,10)	116,88 (28,89)
liveness	-2,17	0,14 (0,11)	0,17 (0,14)
acousticness	-6,41	0,17 (0,17)	0,36 (0,39)
instrumentalness	-6,75	0,01 (0,04)	0,18 (0,32)
Discrete attributes : [Recall] Accuracy			
genre=Hip-Hop	24,29	[63,2 %] 69,6 %	10,8 %
genre=Reggae	5,76	[27,6 %] 16,2 %	5,8 %
genre=Pop	-0,77	[8,0 %] 8,1 %	9,9 %

Pour le groupe C_K_MEANS 3 correspond aux genres raggae, pop et rock. Se genre de music utilise le valence, danceability, loudness et energy ce qui est tout à fait correcte.

Pour la classe C_K_MEANS 4 est associé au genre folk et jazz, ce genre utilise la variable acousticness, alors la classification est correcte.

Pour la classe C_K_MEANS 5 est associé au genre hip-hop et raggae, ce genre utilise les chansons speechiness, loudness, valence, danceability et energy, alors la classification est correcte.



Cluster_KMeans_1=c_kmeans_6			
Examples			[14,1 %] 214
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
instrumentalness	31,01	0,81 (0,19)	0,18 (0,32)
acousticness	24,44	0,96 (0,07)	0,36 (0,39)
duration_ms	4,75	316138,69 (144578,60)	267595,67 (161209,58)
liveness	-5,94	0,12 (0,07)	0,17 (0,14)
speechiness	-7,09	0,05 (0,02)	0,09 (0,09)
tempo	-10,68	97,32 (27,84)	116,88 (28,89)
valence	-17,92	0,18 (0,17)	0,47 (0,26)
danceability	-18,22	0,34 (0,13)	0,56 (0,18)
energy	-24,06	0,09 (0,09)	0,56 (0,31)
loudness	-28,64	-24,93 (7,23)	-10,79 (7,79)
Discrete attributes : [Recall] Accuracy			
genre=Classique	28,81	[80,1 %] 75,2 %	13,3 %
genre=Jazz	6,86	[31,5 %] 24,8 %	11,1 %
genre=Reggae	-3,90	[0,0 %] 0,0 %	5,8 %
genre=Metal	-4,46	[0,0 %] 0,0 %	7,4 %
genre=Pop	-5,24	[0,0 %] 0,0 %	9,9 %

Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
tempo	17,10	162,04 (18,98)	116,88 (28,89)
energy	11,40	0,88 (0,10)	0,56 (0,31)
loudness	7,62	-5,37 (2,16)	-10,79 (7,79)
liveness	3,13	0,21 (0,11)	0,17 (0,14)
speechiness	1,42	0,10 (0,06)	0,09 (0,09)
valence	0,56	0,49 (0,18)	0,47 (0,26)
duration_ms	-1,59	244206,68 (56889,78)	267595,67 (161209,58)
instrumentalness	-4,91	0,03 (0,09)	0,18 (0,32)
danceability	-8,81	0,41 (0,12)	0,56 (0,18)
acousticness	-9,19	0,04 (0,10)	0,36 (0,39)
Discrete attributes : [Recall] Accuracy			
genre=Metal	15,73	[44,6 %] 45,0 %	7,4 %
genre=Rock	6,90	[18,3 %] 37,8 %	15,2 %
genre=Pop	0,00	[7,3 %] 9,9 %	9,9 %
genre=Reggae	-2,28	[1,1 %] 0,9 %	5,8 %

Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
energy	15,23	0,81 (0,14)	0,56 (0,31)
loudness	13,30	-5,31 (2,09)	-10,79 (7,79)
tempo	1,12	118,59 (16,17)	116,88 (28,89)
danceability	-0,86	0,55 (0,12)	0,56 (0,18)
liveness	-1,97	0,15 (0,08)	0,17 (0,14)
instrumentalness	-2,76	0,13 (0,25)	0,18 (0,32)
duration_ms	-3,39	238653,90 (63300,78)	267595,67 (161209,58)
speechiness	-4,65	0,07 (0,04)	0,09 (0,09)
valence	-10,50	0,33 (0,15)	0,47 (0,26)
acousticness	-14,76	0,06 (0,10)	0,36 (0,39)
Discrete attributes : [Recall] Accuracy			
genre=Electro	16,37	[59,0 %] 45,3 %	14,7 %
genre=Metal	6,39	[42,0 %] 16,3 %	7,4 %
genre=Pop	3,36	[29,3 %] 15,2 %	9,9 %
genre=Rock	2,20	[24,3 %] 19,4 %	15,2 %

Cluster_KMeans_1=c_kmeans_9			
Examples			[8,1 %] 122
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
liveness	29,03	0,51 (0,15)	0,17 (0,14)
energy	6,00	0,72 (0,20)	0,56 (0,31)
loudness	4,42	-7,81 (3,87)	-10,79 (7,79)
valence	3,89	0,56 (0,21)	0,47 (0,26)
danceability	3,06	0,60 (0,14)	0,56 (0,18)
speechiness	-0,39	0,09 (0,07)	0,09 (0,09)
duration_ms	-0,95	254346,51 (109321,91)	267595,67 (161209,58)
tempo	-1,30	113,61 (23,05)	116,88 (28,89)
instrumentalness	-4,87	0,04 (0,15)	0,18 (0,32)
acousticness	-5,39	0,18 (0,26)	0,36 (0,39)
Discrete attributes : [Recall] Accuracy			
genre=Electro	2,96	[13,1 %] 23,8 %	14,7 %
genre=Rock	2,22	[11,7 %] 22,1 %	15,2 %
genre=Hip-Hop	1,78	[11,7 %] 15,6 %	10,8 %
genre=Pop	1,23	[10,7 %] 13,1 %	9,9 %
genre=Folk	1,02	[10,0 %] 14,8 %	11,9 %

Pour le groupe C_K_MEANS_6 est associé aux genres classique et jazz ce genre de music utilise une durée de piste en millisecondes, instrumentales ou acousticness, ce qui implique que la classification est correcte.

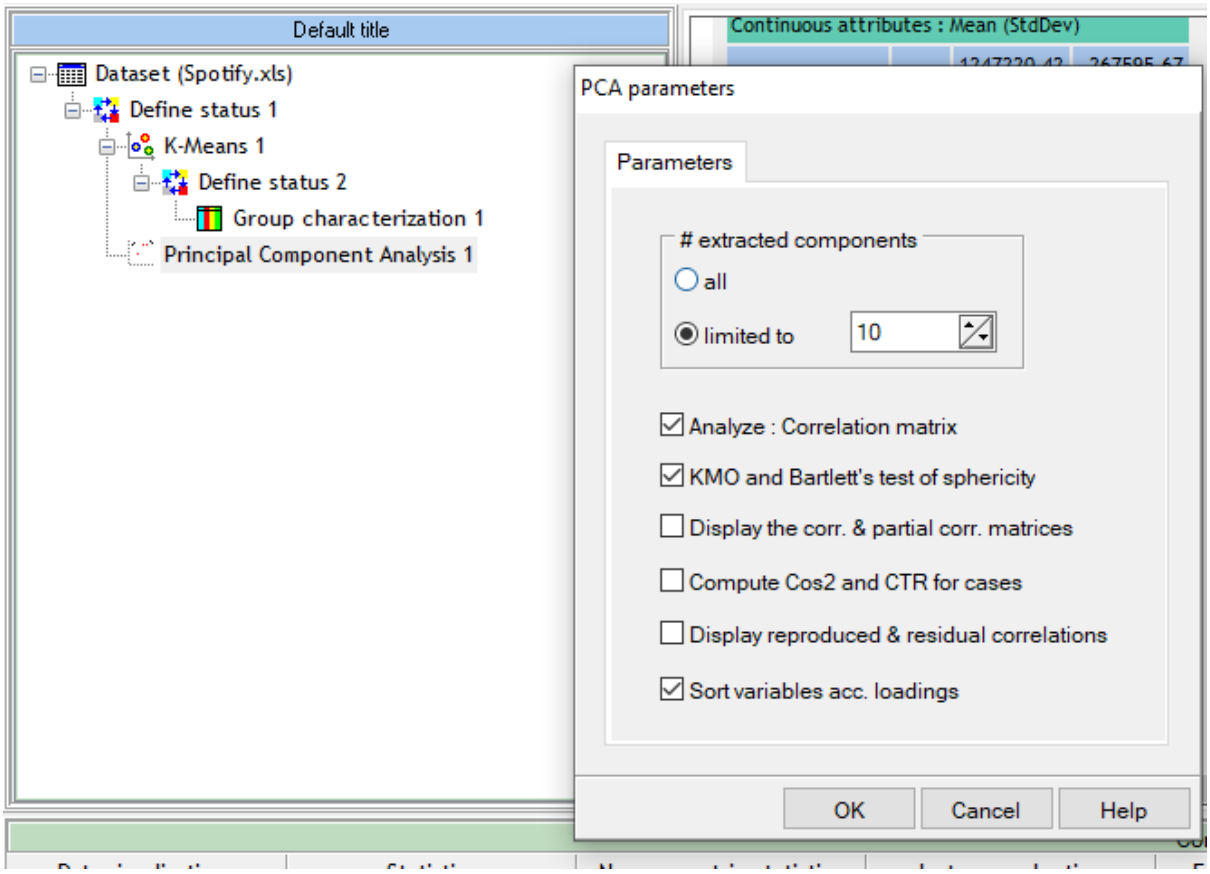
Pour la classe C_K_MEANS_7 est associé aux genres metal et rock et pop, ce genre utilise les variables tempo, loudness, liveness, et energy , alors la classification est correcte.

Pour la classe C_K_MEANS_8 est associé aux genres electro, metal et pop et rock, ce genre utilise les variables loudness et energy, alors la classification est correcte.

Pour la classe C_K_MEANS_9 est associé aux genres rock et electro, ce genre utilise les variables loudness, liveness, et energy,valence,danceability, alors la classification est correcte.

Le principal intérêt de GROUP CHARACTERIZATION est qu'il permet d'introduire à la fois les variables explicatives et à expliquer, qu'elles soient quantitatives ou qualitatives.

4.2.1Analyse en composantes principales



La courbe de la proportion de variance cumulée aide au choix du nombre de facteurs.

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	4,125327	2,848604	41,25 %		41,25 %
2	1,276724	0,255575	12,77 %		54,02 %
3	1,021149	0,111053	10,21 %		64,23 %
4	0,910096	0,108331	9,10 %		73,33 %
5	0,801765	0,094638	8,02 %		81,35 %
6	0,707127	0,220678	7,07 %		88,42 %
7	0,486449	0,117007	4,86 %		93,29 %
8	0,369442	0,150099	3,69 %		96,98 %
9	0,219342	0,136763	2,19 %		99,17 %
10	0,082579	-	0,83 %		100,00 %
Tot.	10,000000	-	-	-	-

7 axes (93%) pour ne pas perdre l'information.

Les variables présentant une corrélation supérieure à 0.5 en valeur absolue, le reste des variables sont triées sur le second axe, avec la même contrainte, etc.

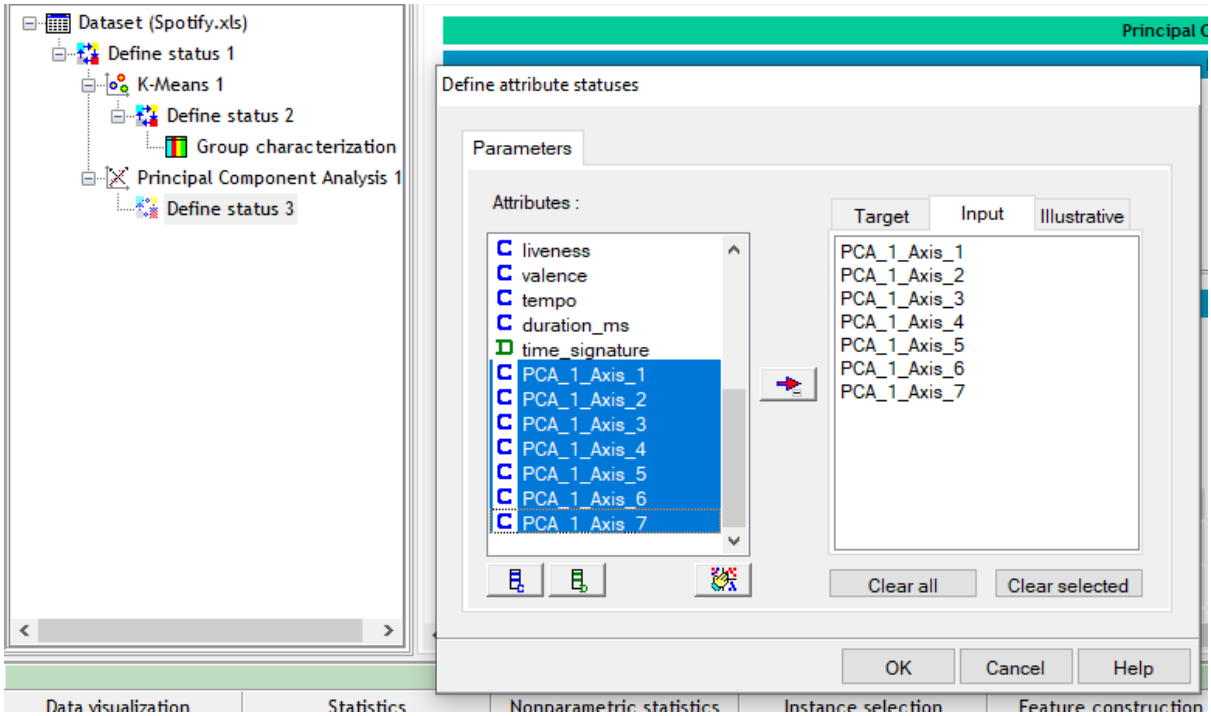
Factor Loadings [Communality Estimates]

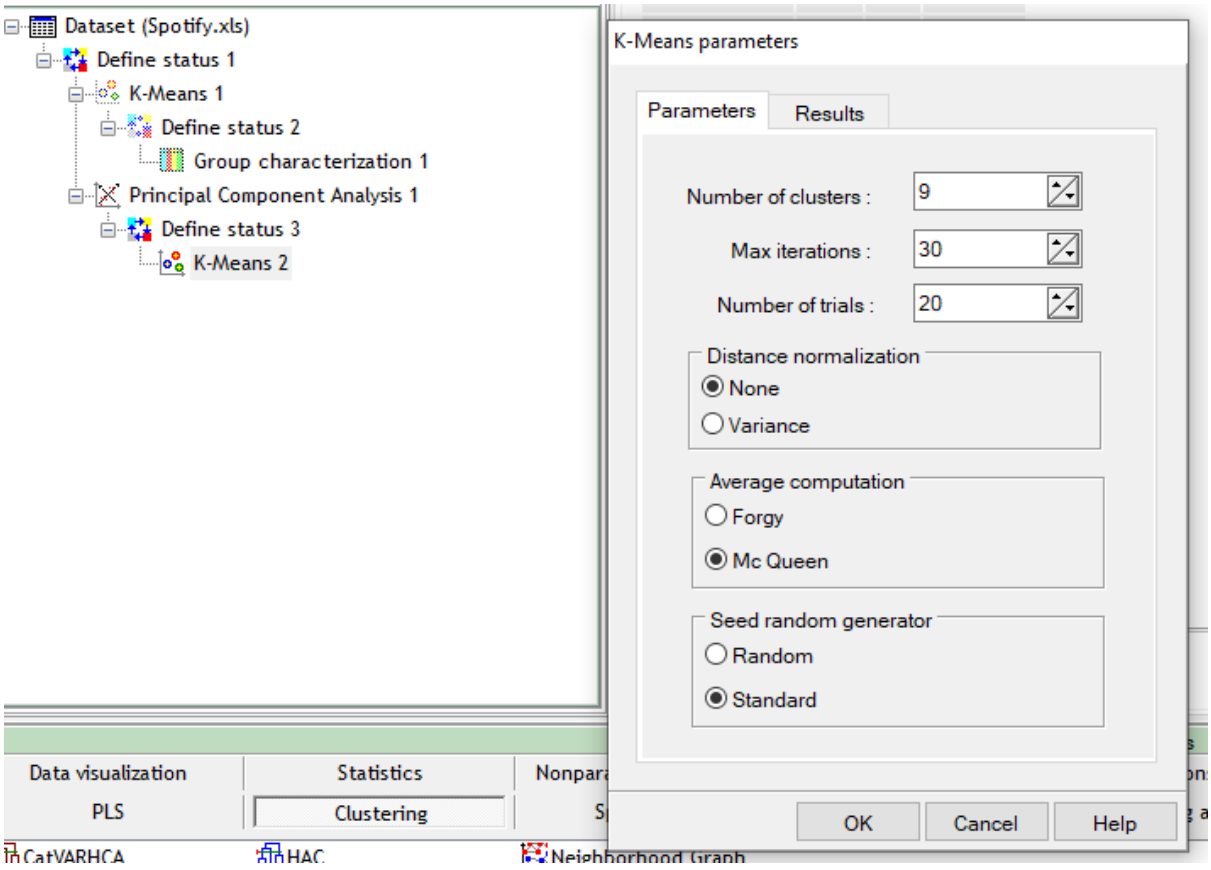
Attribute	Axis_1		Axis_2		Axis_3		Axis_4		Axis_5	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
loudness	0,88574	78 % (78 %)	0,19879	4 % (82 %)	0,01158	0 % (82 %)	-0,08097	1 % (83 %)	-0,21856	5 % (88 %)
energy	0,86378	75 % (75 %)	0,29672	9 % (83 %)	0,07594	1 % (84 %)	-0,09928	1 % (85 %)	-0,18930	4 % (89 %)
acousticness	-0,85949	74 % (74 %)	-0,20130	4 % (78 %)	-0,04144	0 % (78 %)	0,08250	1 % (79 %)	0,22612	5 % (84 %)
instrumentalness	-0,76217	58 % (58 %)	0,02811	0 % (58 %)	0,05052	0 % (58 %)	-0,08645	1 % (59 %)	0,04312	0 % (59 %)
danceability	0,64665	42 % (42 %)	-0,57822	33 % (75 %)	0,01400	0 % (75 %)	0,03354	0 % (75 %)	-0,04661	0 % (76 %)
valence	0,64262	41 % (41 %)	-0,40718	17 % (58 %)	-0,08996	1 % (59 %)	0,04634	0 % (59 %)	0,19962	4 % (63 %)
tempo	0,33108	11 % (11 %)	0,54636	30 % (41 %)	-0,33856	11 % (52 %)	-0,30856	10 % (62 %)	0,57211	33 % (95 %)
liveness	0,20705	4 % (4 %)	0,36558	13 % (18 %)	0,60771	37 % (55 %)	0,60974	37 % (92 %)	0,24307	6 % (98 %)
duration_ms	-0,37009	14 % (14 %)	0,13132	2 % (15 %)	0,57004	32 % (48 %)	-0,56107	31 % (79 %)	-0,19028	4 % (83 %)
speechiness	0,39302	15 % (15 %)	-0,39790	16 % (31 %)	0,44025	19 % (51 %)	-0,30714	9 % (60 %)	0,44783	20 % (80 %)
Var. Expl.	4,12533	41 % (41 %)	1,27672	13 % (54 %)	1,02115	10 % (64 %)	0,91010	9 % (73 %)	0,80176	8 % (81 %)

Le but est de mettre en valeur les groupes. Ici, on note que (volume, énergie, acoustique, instrumentalité, danse, valence) sont associés au premier facteur, tandis que (vivacité, durée_ms) au troisième.

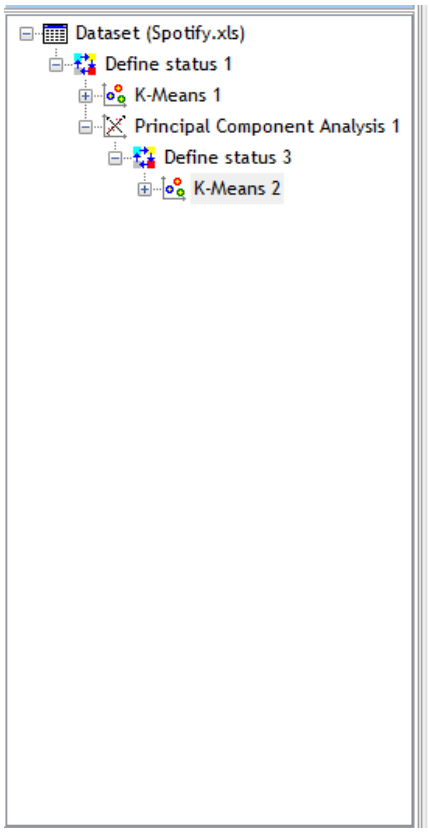
3. K-Means

Il faut préciser à TANAGRA que ce sont ces variables transformées qui seront utilisées pour les calculs. Nous insérons un nouveau DEFINE STATUS, nous plaçons les variables PCA_1_AXIS_1 à PCA_1_AXIS_7 dans INPUT.





==========



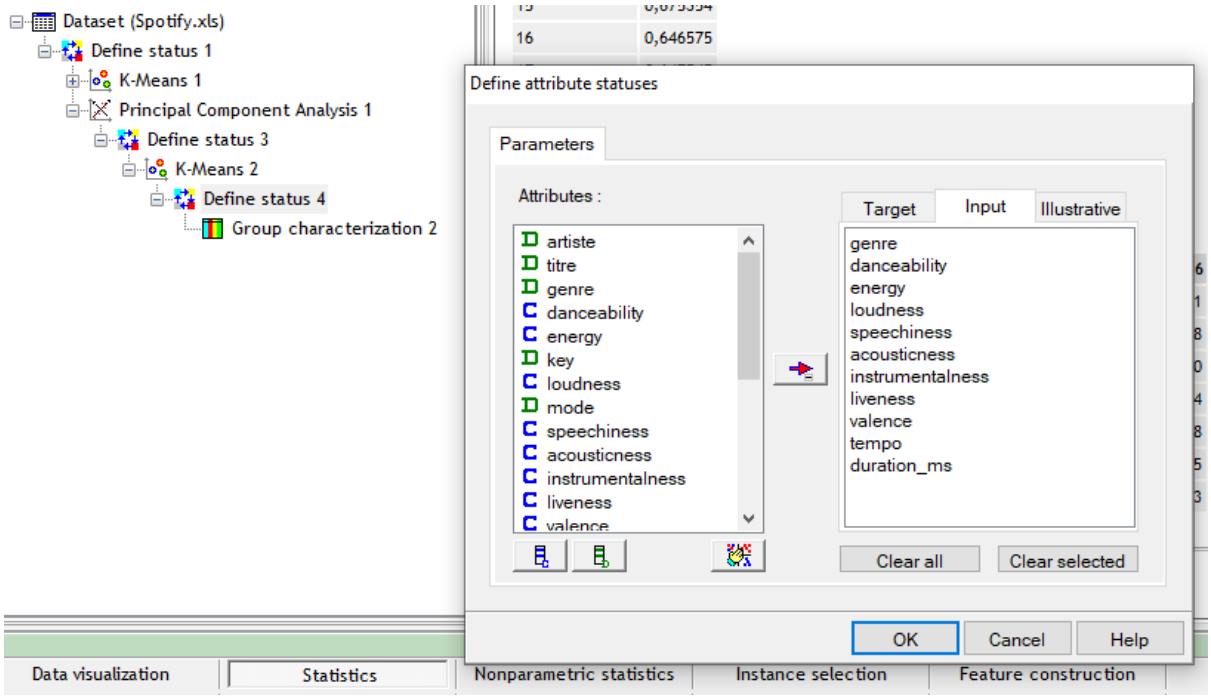
Cluster size and WSS

Clusters	9		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	15	268,5948
cluster n°2	c_kmeans_2	113	354,2263
cluster n°3	c_kmeans_3	319	598,4132
cluster n°4	c_kmeans_4	179	423,8038
cluster n°5	c_kmeans_5	152	598,3526
cluster n°6	c_kmeans_6	220	811,9540
cluster n°7	c_kmeans_7	106	263,1412
cluster n°8	c_kmeans_8	288	638,7120
cluster n°9	c_kmeans_9	121	578,2451

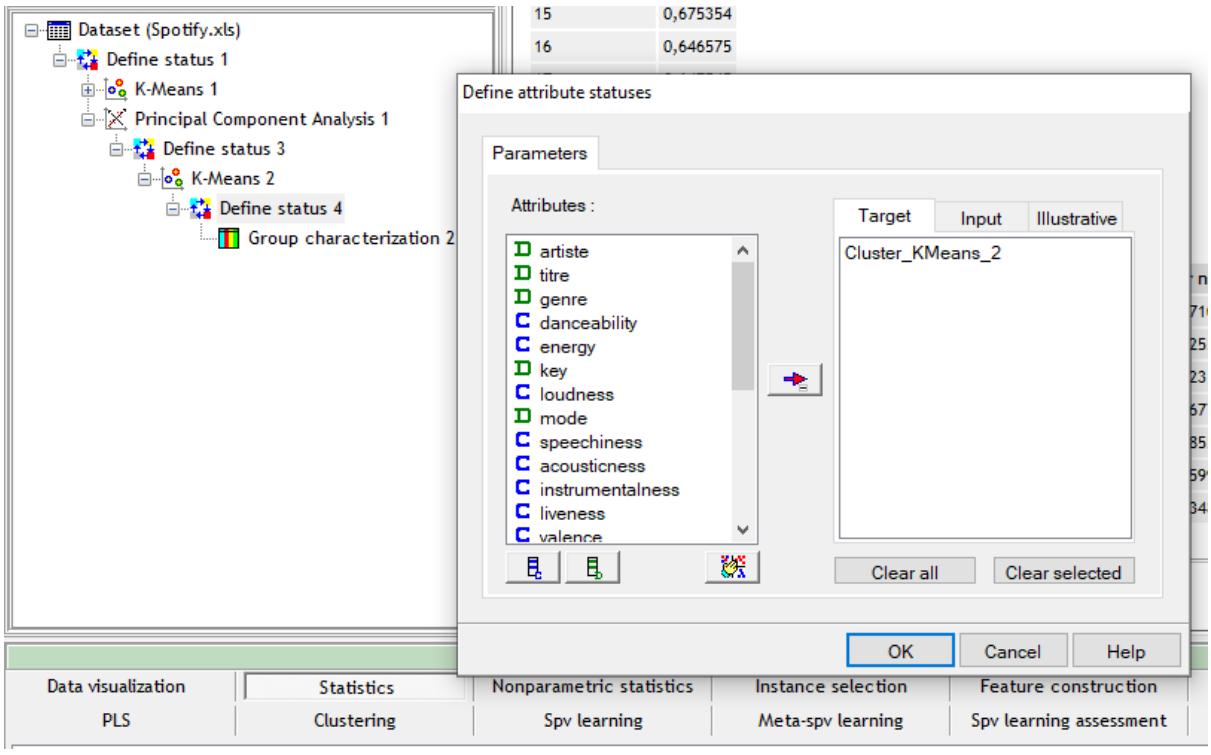
R-Square for each attempt

Number of trials	20
Trial	R-square
1	0,647609
2	0,650831
3	0,646471
4	0,678662
5	0,674617
6	0,676360
7	0,642895

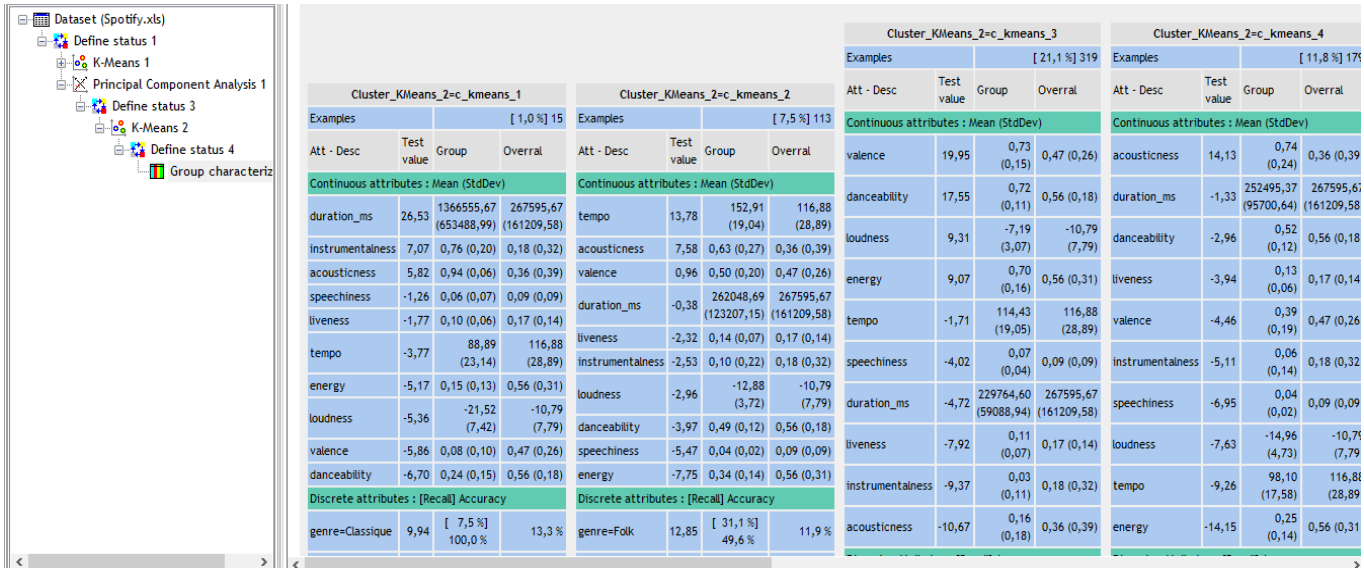
la caractérisation sur k-means.



cette fois, nous mettrons les variables en entrée pour qu'elles puissent les interpréter



Même résultat



Cluster_KMeans_2=c_kmeans_1			
Examples			[1,0 %] 15
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
duration_ms	26,53	1366555,67 (653488,99)	267595,67 (161209,58)
instrumentalness	7,07	0,76 (0,20)	0,18 (0,32)
acousticness	5,82	0,94 (0,06)	0,36 (0,39)
speechiness	-1,26	0,06 (0,07)	0,09 (0,09)
liveness	-1,77	0,10 (0,06)	0,17 (0,14)
tempo	-3,77	88,89 (23,14)	116,88 (28,89)
energy	-5,17	0,15 (0,13)	0,56 (0,31)
loudness	-5,36	-21,52 (7,42)	-10,79 (7,79)
valence	-5,86	0,08 (0,10)	0,47 (0,26)
danceability	-6,70	0,24 (0,15)	0,56 (0,18)
Discrete attributes : [Recall] Accuracy			
genre=Classique	9,94	[7,5 %] 100,0 %	13,3 %

Cluster_KMeans_2=c_kmeans_2			
Examples			[7,5 %] 113
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
tempo	13,78	152,91 (19,04)	116,88 (28,89)
acousticness	7,58	0,63 (0,27)	0,36 (0,39)
valence	0,96	0,50 (0,20)	0,47 (0,26)
duration_ms	-0,38	262048,69 (123207,15)	267595,67 (161209,58)
liveness	-2,32	0,14 (0,07)	0,17 (0,14)
instrumentalness	-2,53	0,10 (0,22)	0,18 (0,32)
loudness	-2,96	-12,88 (3,72)	-10,79 (7,79)
danceability	-3,97	0,49 (0,12)	0,56 (0,18)
speechiness	-5,47	0,04 (0,02)	0,09 (0,09)
energy	-7,75	0,34 (0,14)	0,56 (0,31)
Discrete attributes : [Recall] Accuracy			
genre=Folk	12,85	[31,1 %] 49,6 %	11,9 %

Cluster_KMeans_2=c_kmeans_3			
Examples			[21,1 %] 319
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
valence	19,95	0,73 (0,15)	0,47 (0,26)
danceability	17,55	0,72 (0,11)	0,56 (0,18)
loudness	9,31	-7,19 (3,07)	-10,79 (7,79)
energy	9,07	0,70 (0,16)	0,56 (0,31)
tempo	-1,71	114,43 (19,05)	116,88 (28,89)
speechiness	-4,02	0,07 (0,04)	0,09 (0,09)
duration_ms	-4,72	229764,60 (59088,94)	267595,67 (161209,58)
liveness	-7,92	0,11 (0,07)	0,17 (0,14)
instrumentalness	-9,37	0,03 (0,11)	0,18 (0,32)
acousticness	-10,67	0,16 (0,18)	0,36 (0,39)

Cluster_KMeans_2=c_kmeans_4			
Examples			[11,8 %] 17
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
acousticness	14,13	0,74 (0,24)	0,36 (0,39
duration_ms	-1,33	252495,37 (95700,64)	267595,6 (161209,58
danceability	-2,96	0,52 (0,12)	0,56 (0,18
liveness	-3,94	0,13 (0,06)	0,17 (0,14
valence	-4,46	0,39 (0,19)	0,47 (0,26
instrumentalness	-5,11	0,06 (0,14)	0,18 (0,32
speechiness	-6,95	0,04 (0,02)	0,09 (0,09
loudness	-7,63	-14,96 (4,73)	-10,79 (7,79
tempo	-9,26	98,10 (17,58)	116,88 (28,89
energy	-14,15	0,25 (0,14)	0,56 (0,31

On déduit alors que c'est une bonne caractérisation, et on remarque qu'il y a plusieurs relations entre variables :

Energy et loudness.

La danceability et valence.

Instrumentalness et duration-ms.

Tempo et liveness.

Tempo et energy.

Acoustecness et Instrumentalness

Speechiness et liveness

Speechiness et danceability

Et les chansons appartenant à un même genre musical sont :

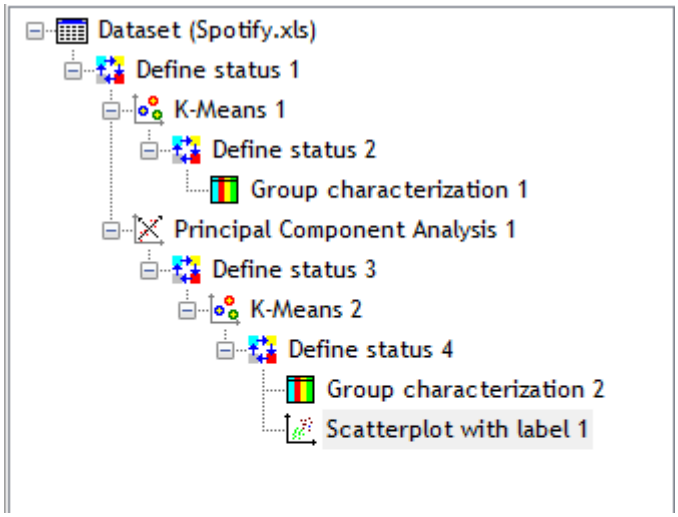
Tempo et acostecness appartenant aux genres Jazz et Falk.

Duration-ms et instrumentalness faites partie de la music classique.

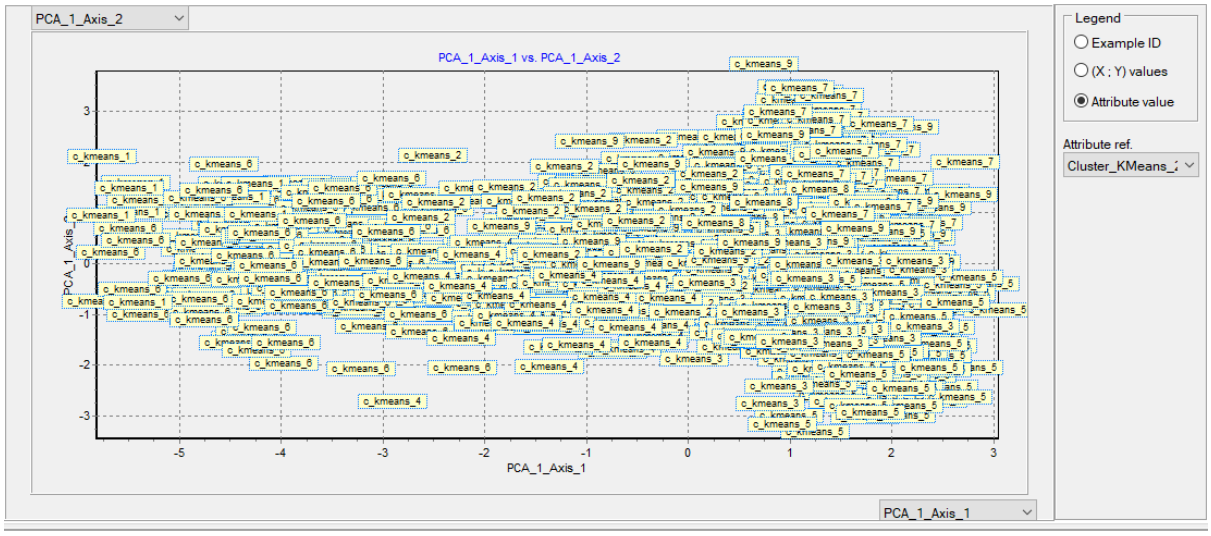
Valence et danceability appartenant a la musique genre Raggae….

Passons maintenant à la visualisation des classes :

Plans factoriels :



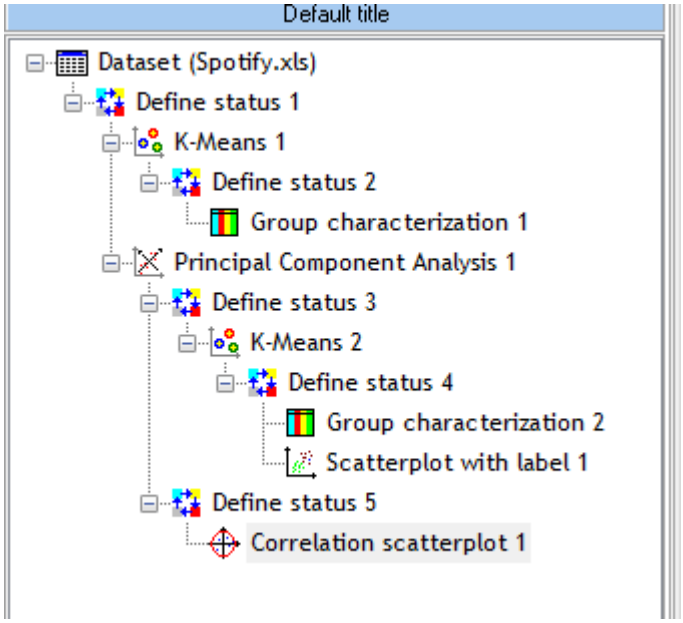
Nous cliquons sur VIEW et nous le paramétrons de manière à avoir en abscisse le premier facteur, en ordonnée le second facteur. Notons qu'il est très aisé de passer d'un plan factoriel à un autre.



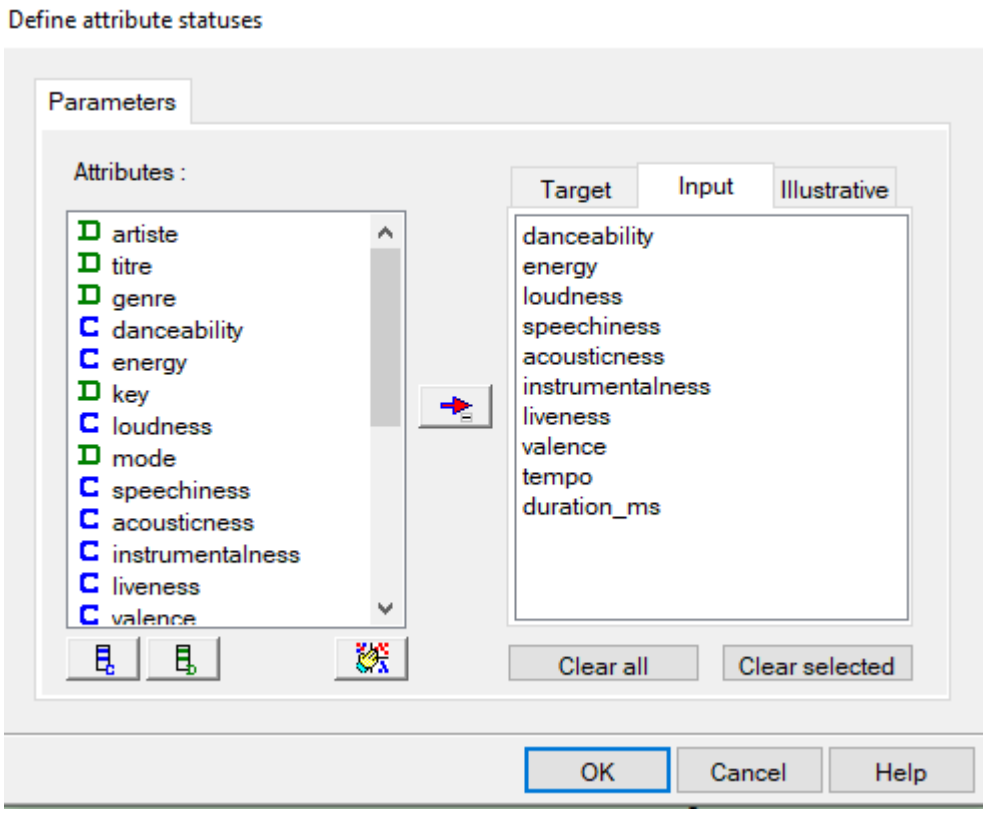
On voit les classe 7, 9, 3, 5 qui sont entrelacés les uns avec les autres, pareil pour les classes 2 et 4, Cette situation est due au genres musicaux qui se partagent les mêmes variables.

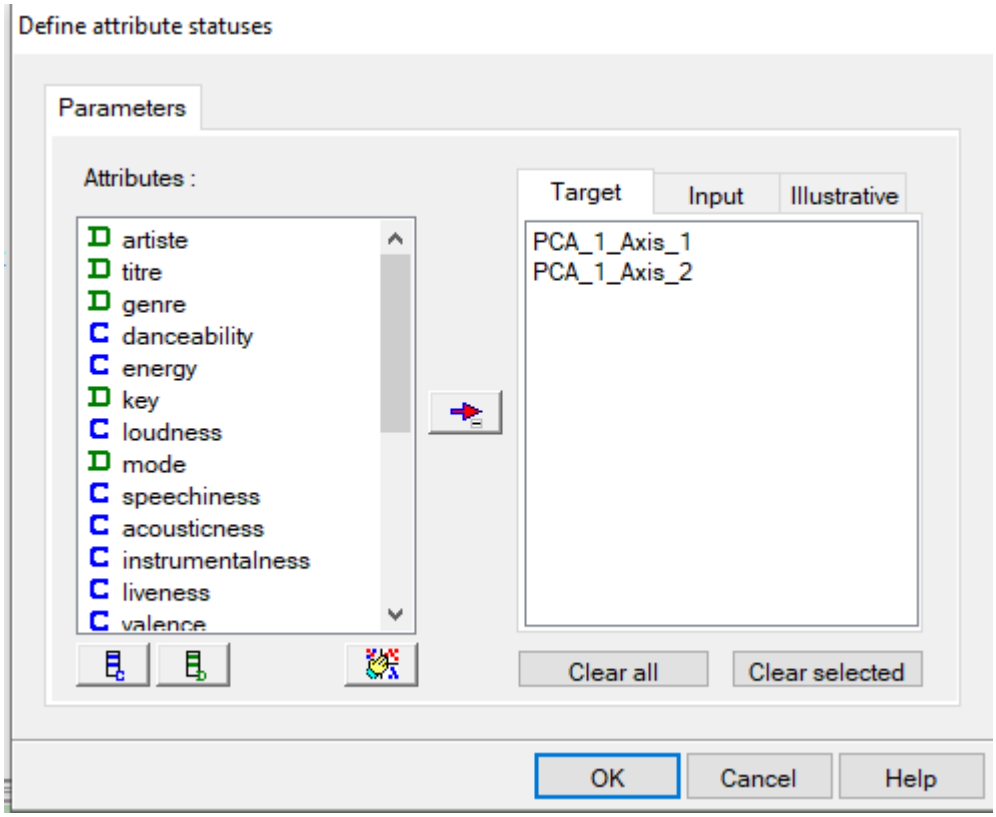
Cercle des corrélations et variables illustratives quantitatives :

Le cercle de corrélations est un outil graphique qui permet de comprendre la nature des axes. Il sert à interpréter les axes.

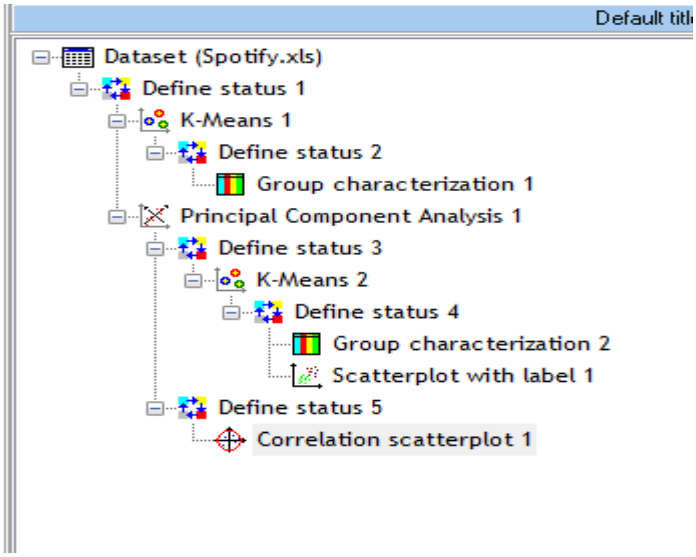


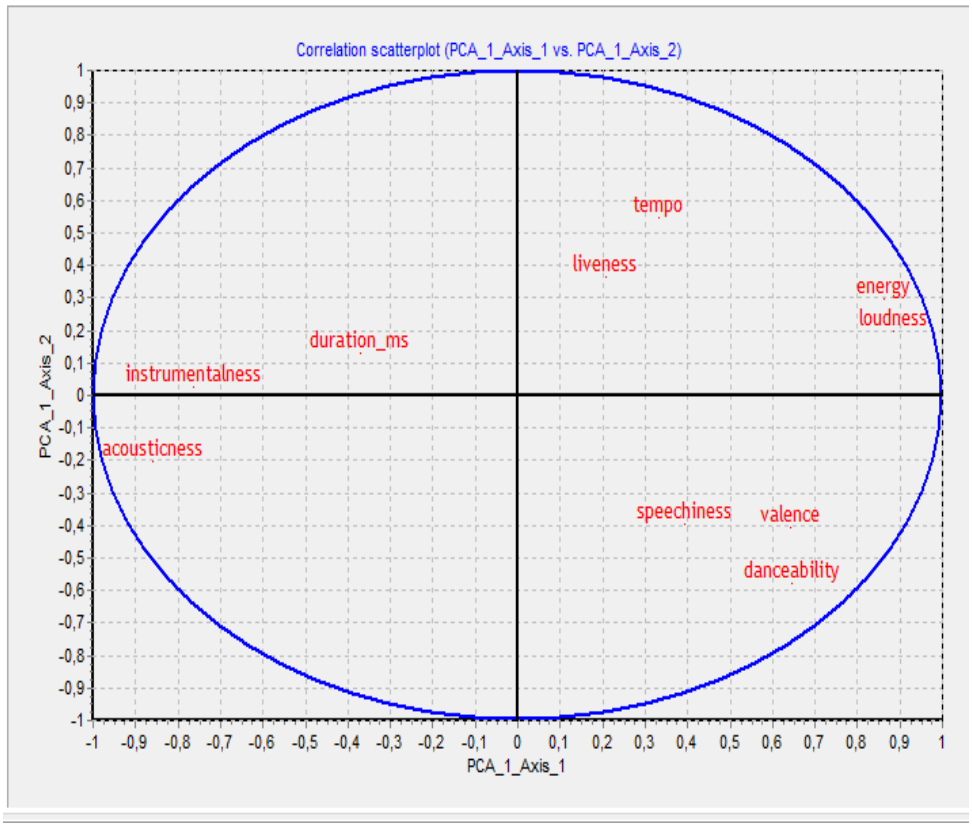
Nous définissons les deux premiers axes comme TARGET, les variables continues et les variables discrètes sont placées en INPUT.





Dans un deuxième temps, nous ajoutons le composant CORRELATION SCATTERPLOT dans le diagramme. Nous obtenons le cercle des corrélations.



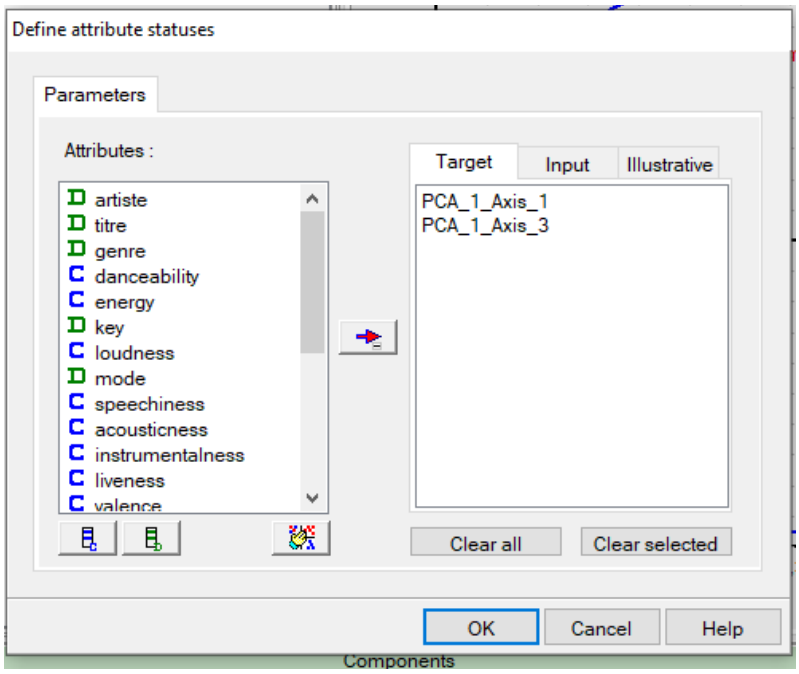


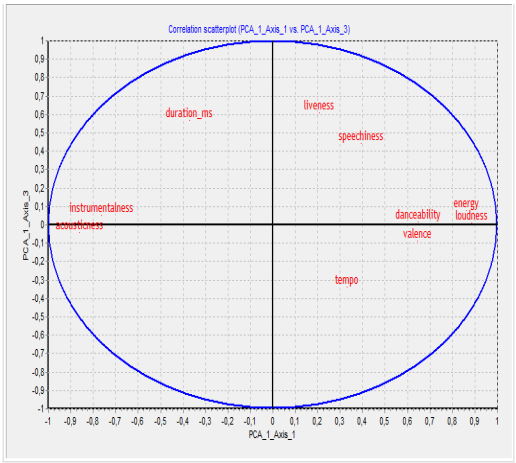
Les variables bien représentées:

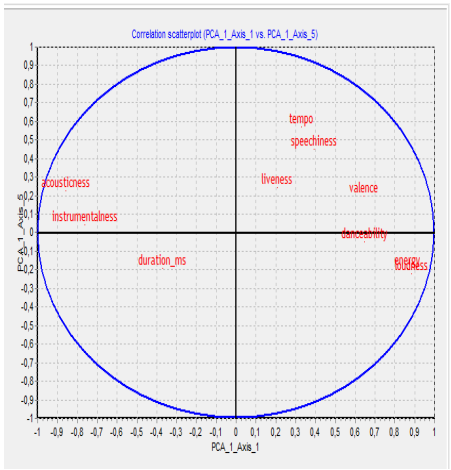
+ : loudness, danceability energy.

_ : acoustecness

Les variables mal représentées : instrumentaless, duration-ms, liveness, speechiness, tempo, valence.



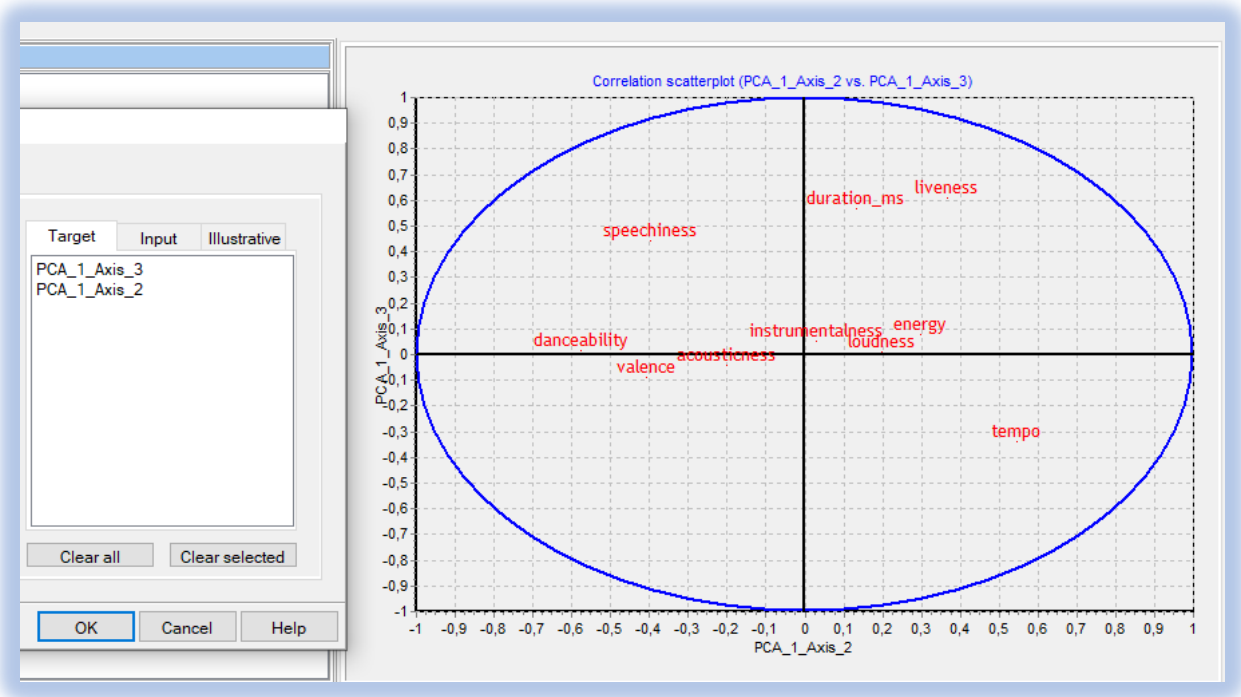




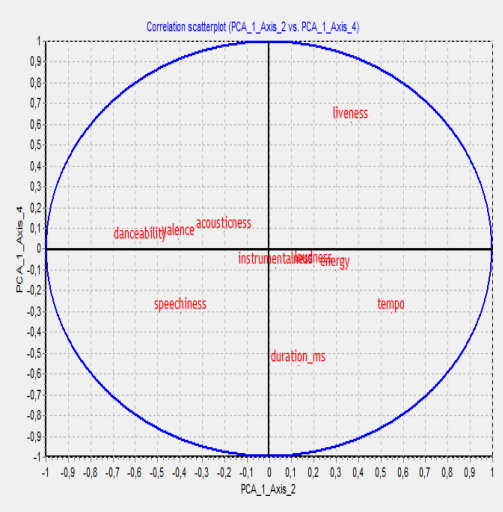
selon l'axe 1 :

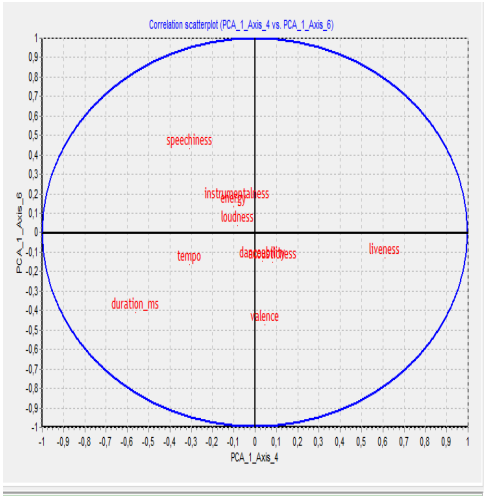
élimination de danceability

bien représentés > energy, loudness Acousticness.



variables mal présentées.

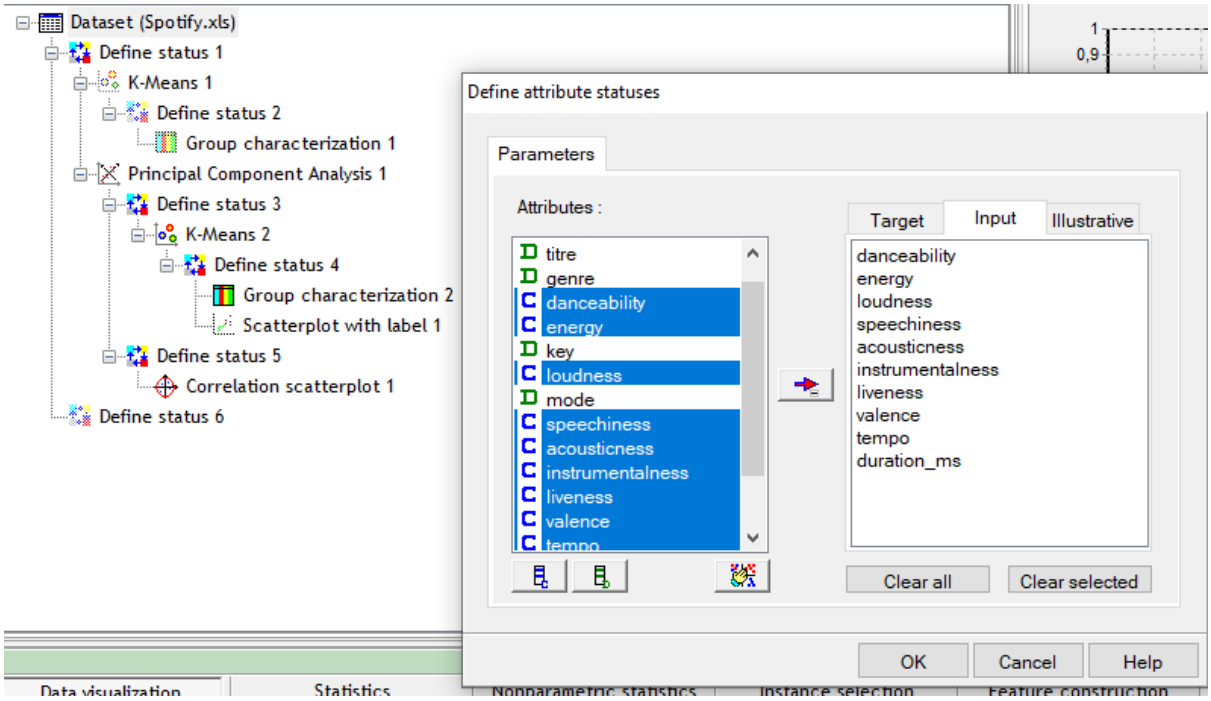


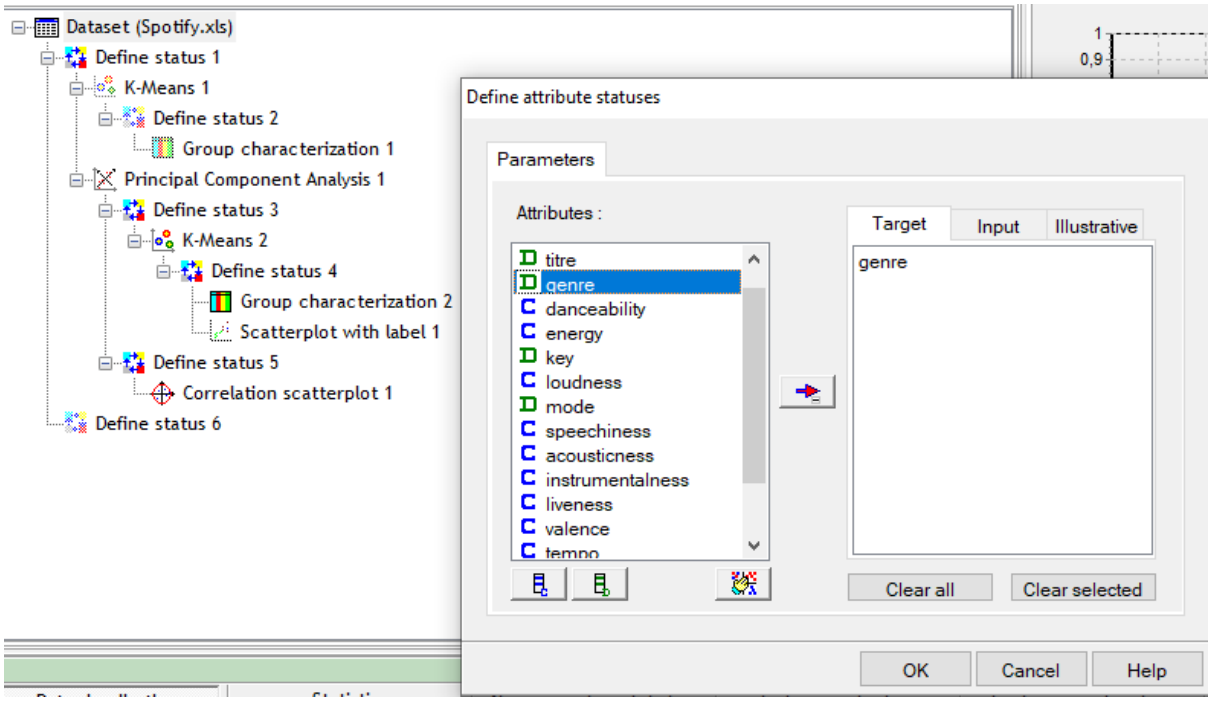


variables mal présentées.

Analyse discriminante linéaire :

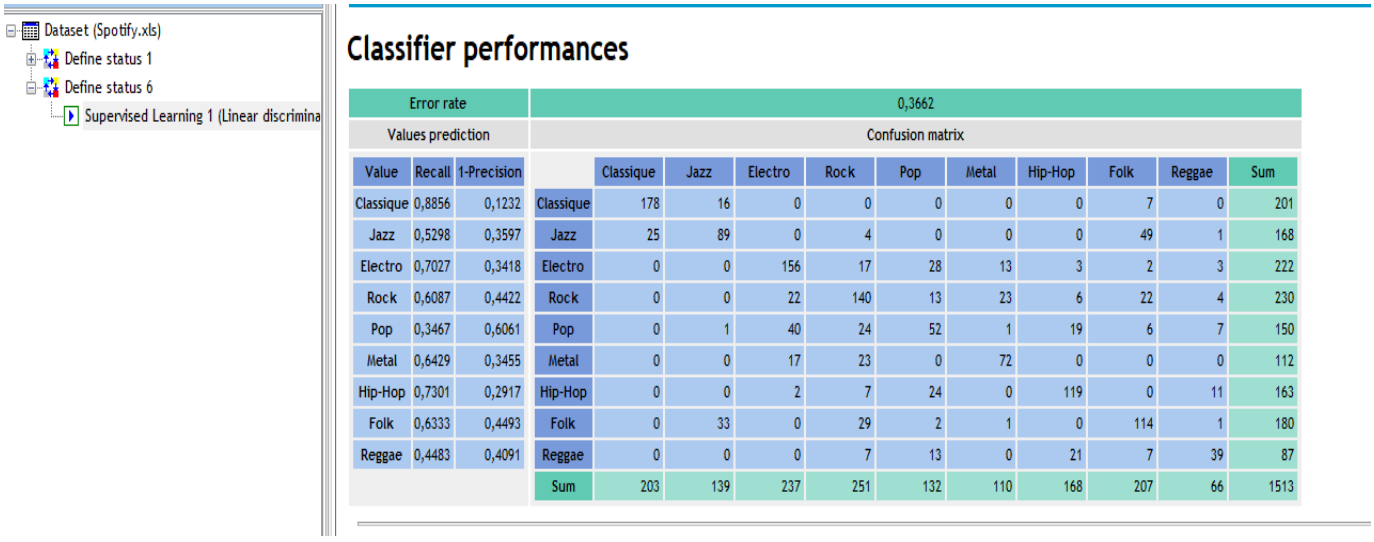
Nous utilisons le composant DEFINE STATUS pour indiquer le rôle des variables: GENRE est la cible (TARGET), les autres (DANCEABILITY….DURATION) sont les prédictives(INPUT)





Nous plaçons le composant LINEAR DISCRIMINANT ANALYSIS (onglet SPV LEARNING).

Nous cliquons sur VIEW pour accéder aux résultats.



Classifier performances

Error rate	0,3662

Values prediction

Value	Recall	1-Precision
Classique	0,8856	0,1232
Jazz	0,5298	0,3597
Electro	0,7027	0,3418
Rock	0,6087	0,4422
Pop	0,3467	0,6061
Metal	0,6429	0,3455
Hip-Hop	0,7301	0,2917
Folk	0,6333	0,4493
Reggae	0,4483	0,4091

Confusion matrix

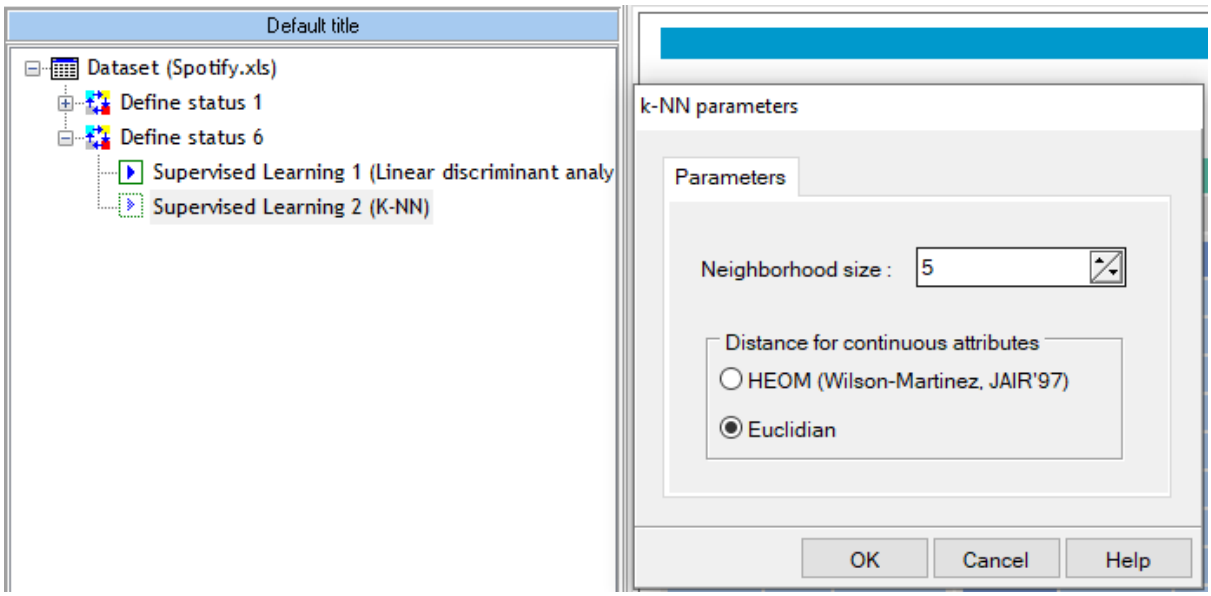
	Classique	Jazz	Electro	Rock	Pop	Metal	Hip-Hop	Folk	Reggae	Sum
Classique	178	16	0	0	0	0	0	7	0	201
Jazz	25	89	0	4	0	0	0	49	1	168
Electro	0	0	156	17	28	13	3	2	3	222
Rock	0	0	22	140	13	23	6	22	4	230
Pop	0	1	40	24	52	1	19	6	7	150
Metal	0	0	17	23	0	72	0	0	0	112
Hip-Hop	0	0	2	7	24	0	119	0	11	163
Folk	0	33	0	29	2	1	0	114	1	180
Reggae	0	0	0	7	13	0	21	7	39	87
Sum	203	139	237	251	132	110	168	207	66	1513

3.2.1Matrice de confusion et le taux d’erreur :

En appliquant le modèle sur les données d’apprentissage, nous obtenons la matrice de confusion et le taux d’erreur associé : 36.62 %.

Passons maintenant à la solution par la méthode des K plus proches voisins. Au moins avec cette méthode on aura la possibilité de gérer les distances et le nombre de voisins.

On choisit 5 comme nombre de voisins au départ.



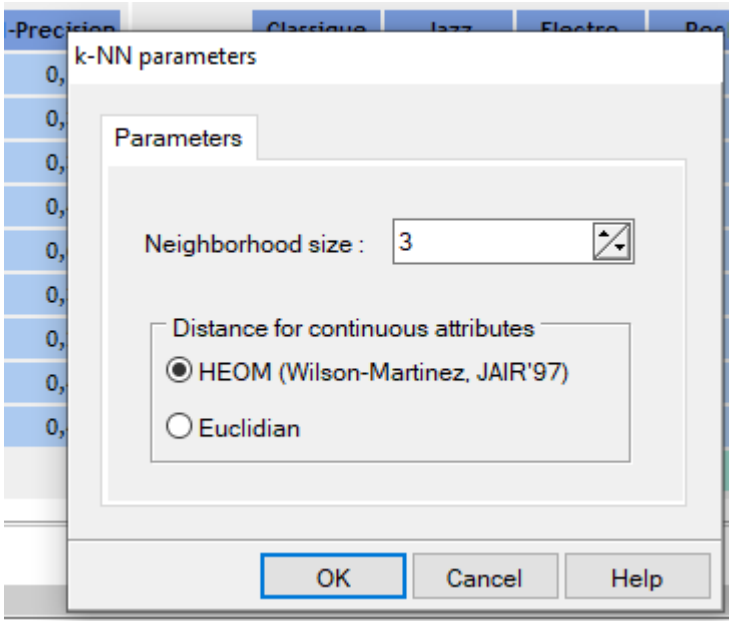
On obtient 54% pour le taux d'erreur, c'est très grand.

Error rate	0,5400
Values prediction	Confusion matrix

Value	Recall	1-Precision
Classique	0,5473	0,4301
Jazz	0,4286	0,5689
Electro	0,5541	0,5040
Rock	0,4348	0,5556
Pop	0,4467	0,5074
Metal	0,3482	0,5938
Hip-Hop	0,4785	0,5806
Folk	0,4111	0,5795
Reggae	0,3793	0,6163

	Classique	Jazz	Electro	Rock	Pop	Metal	Hip-Hop	Folk	Reggae	Sum
Classique	110	26	7	17	9	4	10	12	6	201
Jazz	21	72	12	18	4	5	13	16	7	168
Electro	6	9	123	21	12	9	17	19	6	222
Rock	18	15	30	100	10	17	19	11	10	230
Pop	4	11	16	13	67	6	13	14	6	150
Metal	7	6	13	12	8	39	13	9	5	112
Hip-Hop	10	13	10	18	11	7	78	11	5	163
Folk	14	9	22	17	12	8	16	74	8	180
Reggae	3	6	15	9	3	1	7	10	33	87
Sum	193	167	248	225	136	96	186	176	86	1513

Prenons un nombre plus petit que 5, soit le nombre choisi = 3



Exécutons le résultat par le nombre choisi :

Error rate	0,1963

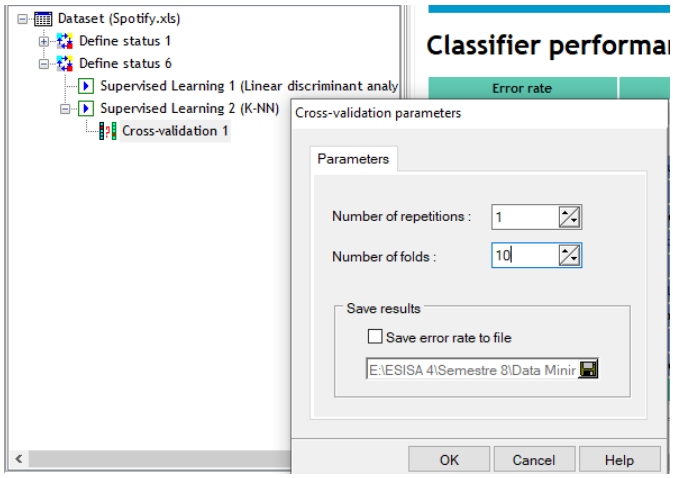
Values prediction

Value	Recall	1-Precision
Classique	0,9353	0,0457
Jazz	0,7321	0,1575
Electro	0,8919	0,1852
Rock	0,7000	0,1990
Pop	0,6133	0,2868
Metal	0,8214	0,2581
Hip-Hop	0,8528	0,1472
Folk	0,8333	0,2958
Reggae	0,8391	0,2474

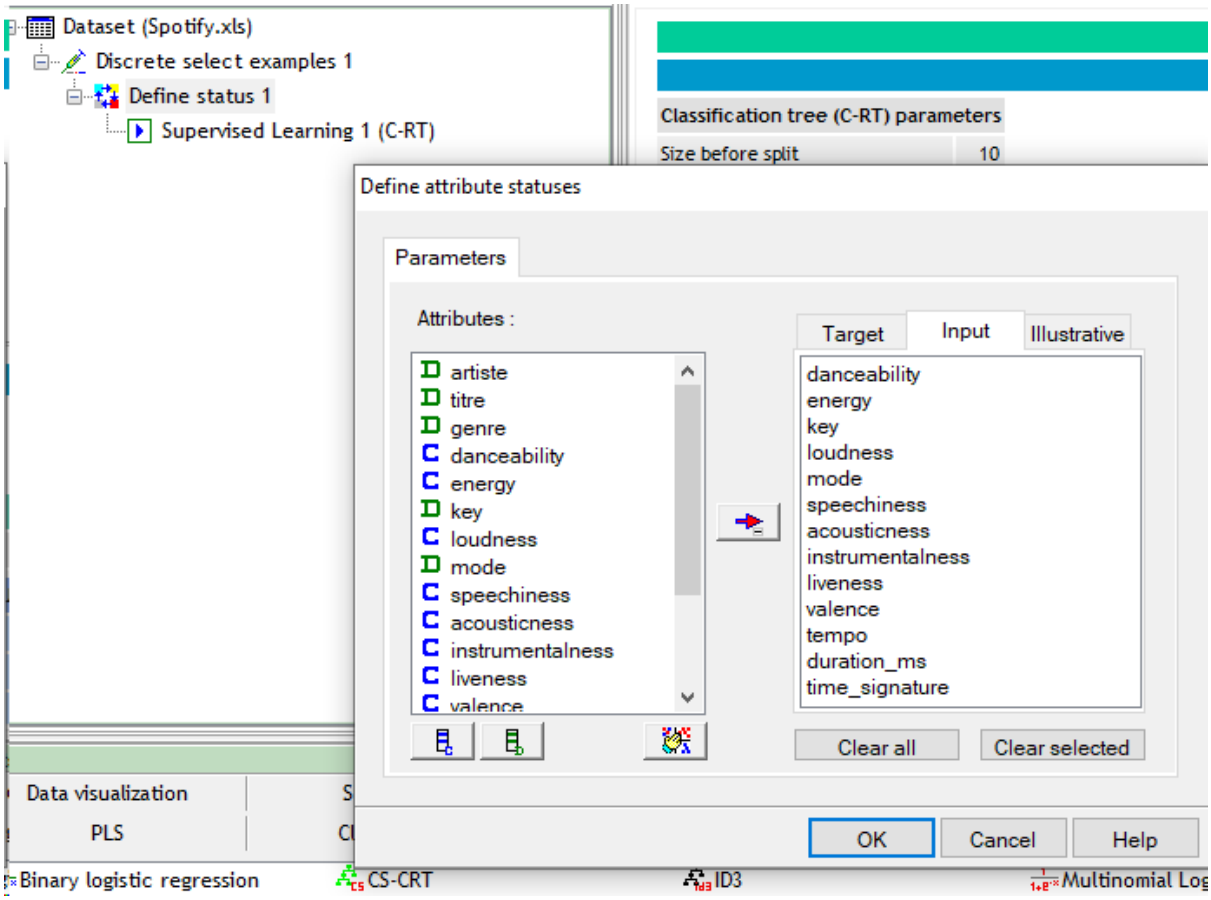
Confusion matrix

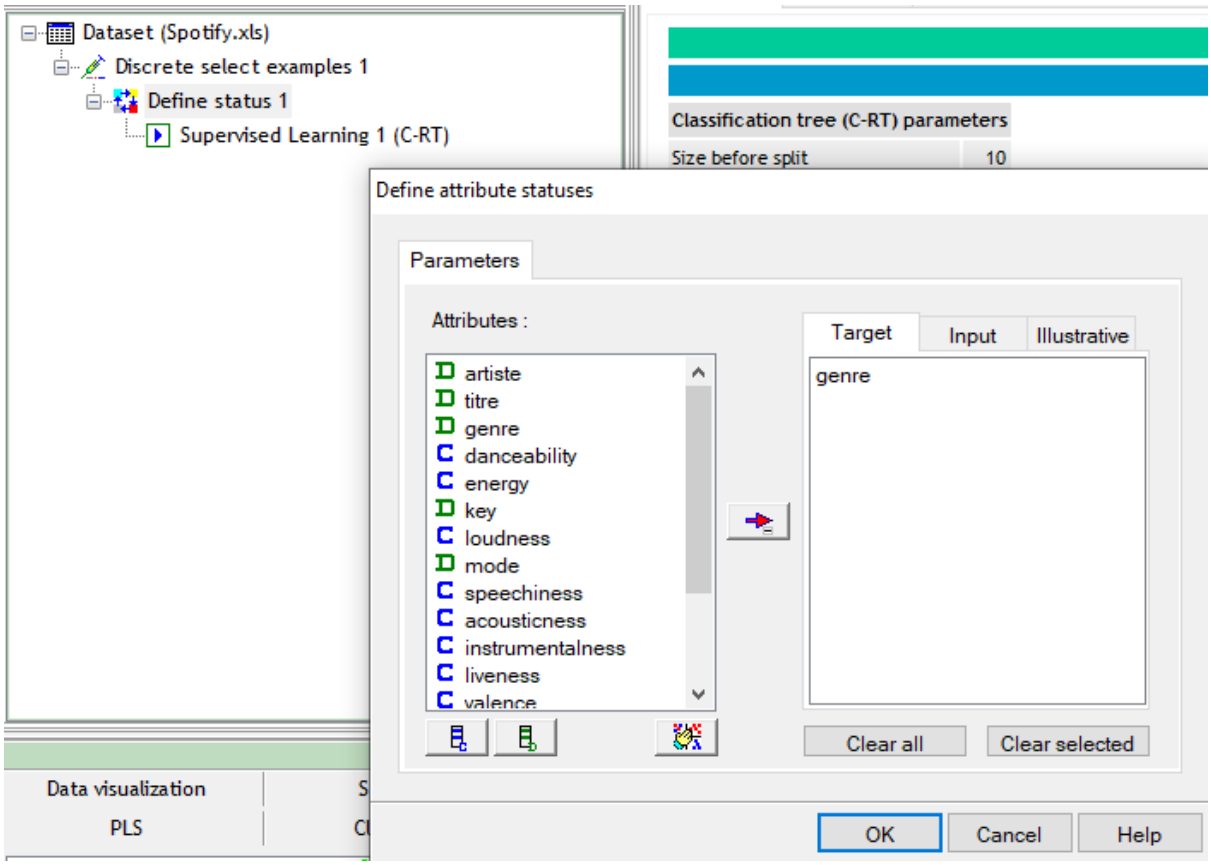
	Classique	Jazz	Electro	Rock	Pop	Metal	Hip-Hop	Folk	Reggae	Sum
Classique	188	6	0	0	0	0	0	7	0	201
Jazz	9	123	1	4	0	0	0	29	2	168
Electro	0	0	198	4	7	8	3	1	1	222
Rock	0	3	16	161	8	19	1	17	5	230
Pop	0	0	21	9	92	3	12	6	7	150
Metal	0	0	4	13	3	92	0	0	0	112
Hip-Hop	0	0	2	2	11	1	139	1	7	163
Folk	0	13	0	7	4	1	3	150	2	180
Reggae	0	1	1	1	4	0	5	2	73	87
Sum	197	146	243	201	129	124	163	213	97	1513

Nous activons le menu PARAMETERS pour définir les paramètres de la méthode.



Cross-validation parameters	
Folds	10
Trials	1


Dataset (Spotify.xls)
Discrete select examples 1
Define status 1
Supervised Learning 1 (C-RT)
Classification tree (C-RT) parameters
Size before split
10
Define attribute statuses
Parameters
Attributes :
artiste
titre
genre
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
Target
Input
Illustrative
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
tempo
duration_ms
time_signature
Clear all
Clear selected
OK
Cancel
Help
Data visualization
PLS
Binary logistic regression
CS-CRT
ID3
Multinomial Log


Dataset (Spotify.xls)
Discrete select examples 1
Define status 1
Supervised Learning 1 (C-RT)
Classification tree (C-RT) parameters
Size before split
10
Define attribute statuses
Parameters
Attributes :
artiste
titre
genre
danceability
energy
key
loudness
mode
speechiness
acousticness
instrumentalness
liveness
valence
Target
Input
Illustrative
genre
Clear all
Clear selected
OK
Cancel
Help
Data visualization
PLS

III.2 Résultats de l'application de l'algorithme CART

III.2.1 Arbre obtenu

D'abord, nous présentons les résultats permettant de choisir la profondeur de l'arbre.

Le tableau ci-dessous montre que l'arbre qui satisfait la règle X-SE Rule = 1 a pour nombre de feuilles 3.

Data partition

Growing set	237
Pruning set	117

Trees sequence

N	# Leaves	Err (growing set)	Err (pruning set)	SE (pruning set)	x
6	1	0,4135	0,4701	0,0461	9,480709
5	2	0,1730	0,2051	0,0373	1,537412
4	3	0,1392	0,1538	0,0334	0,000000
3	4	0,1224	0,1880	0,0361	-
2	6	0,1055	0,2051	0,0373	-
1	7	0,1013	0,1966	0,0367	-

L'arbre obtenu est exprimé sous formes des règles suivantes :

- loudness < -18,4235 then genre = **Classique** (84,03 % of 144 examples)
- loudness >= -18,4235
  - danceability < 0,3745 then genre = **Classique** (70,00 % of 20 examples)
  - danceability >= 0,3745 then genre = Jazz (94,52 % of 73 examples)

III.2 Performances du classifieur

Le taux de rappel et la précision:

Error rate	0,1441
Values prediction	Confusion matrix

Value	Recall	1-Precision
Classique	0,9751	0,1901
Jazz	0,6993	0,0446
Electro	0,0000	1,0000
Rock	0,0000	1,0000
Pop	0,0000	1,0000
Metal	0,0000	1,0000
Hip-Hop	0,0000	1,0000
Folk	0,0000	1,0000
Reggae	0,0000	1,0000

	Classique	Jazz	Electro	Rock	Pop	Metal	Hip-Hop	Folk	Reggae	Sum
Classique	196	5	0	0	0	0	0	0	0	201
Jazz	46	107	0	0	0	0	0	0	0	153
Electro	0	0	0	0	0	0	0	0	0	0
Rock	0	0	0	0	0	0	0	0	0	0
Pop	0	0	0	0	0	0	0	0	0	0
Metal	0	0	0	0	0	0	0	0	0	0
Hip-Hop	0	0	0	0	0	0	0	0	0	0
Folk	0	0	0	0	0	0	0	0	0	0
Reggae	0	0	0	0	0	0	0	0	0	0
Sum	242	112	0	0	0	0	0	0	0	354

Le taux d'erreur 14,41%

97,51% Classique

69,93% jazz

III.2. Validation de l'arbre de décision

Nous utilisons la fonction Test du composant superv.Learning assessment.

genre	
Error rate	0,9905
Values prediction	Confusion matrix

Value	Recall	1-Precision
Classique	0,0000	1,0000
Jazz	0,0105	0,2667
Electro	0,0000	1,0000
Rock	0,0000	1,0000
Pop	0,0000	1,0000
Metal	0,0000	1,0000
Hip-Hop	0,0000	1,0000
Folk	0,0000	1,0000
Reggae	0,0000	1,0000

	Classique	Jazz	Electro	Rock	Pop	Metal	Hip-Hop	Folk	Reggae	Sum
Classique	0	4	7	35	2	35	0	26	0	109
Jazz	0	11	215	195	148	77	163	154	87	1050
Electro	0	0	0	0	0	0	0	0	0	0
Rock	0	0	0	0	0	0	0	0	0	0
Pop	0	0	0	0	0	0	0	0	0	0
Metal	0	0	0	0	0	0	0	0	0	0
Hip-Hop	0	0	0	0	0	0	0	0	0	0
Folk	0	0	0	0	0	0	0	0	0	0
Reggae	0	0	0	0	0	0	0	0	0	0
Sum	0	15	222	230	150	112	163	180	87	1159