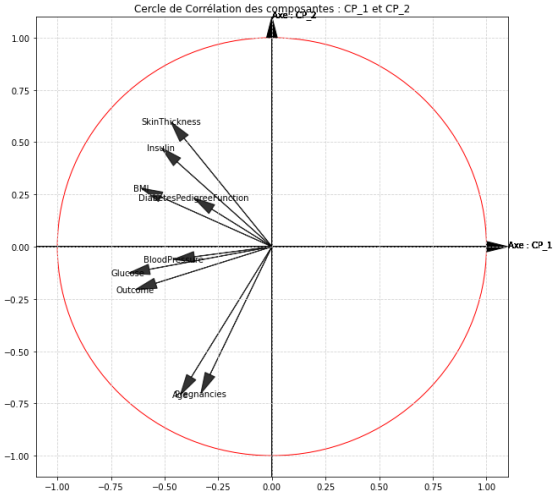
Compte-rendu du TP n°3 – Cartes topographiques

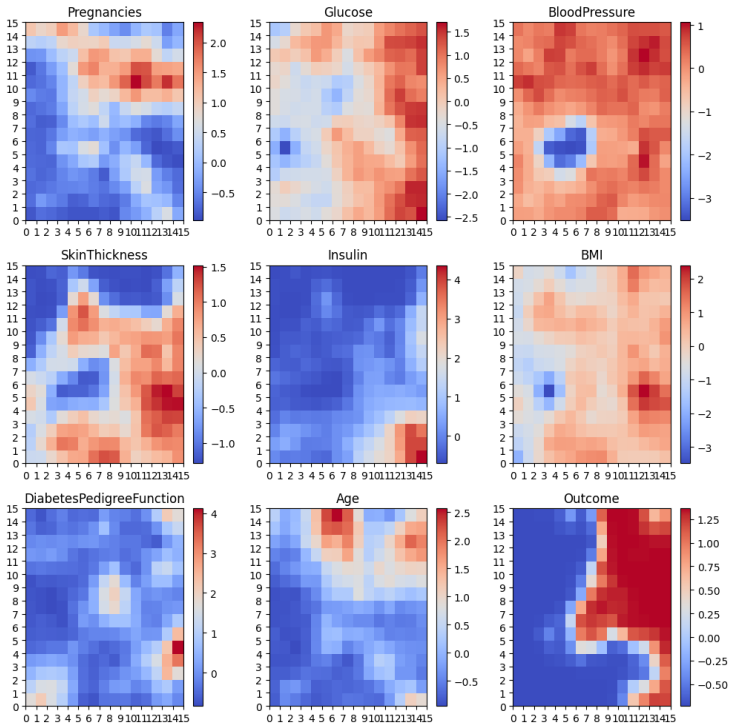
Rappel TP2 : on n'arrive pas à en tirer quelque chose de la matrice de corrélation. C'est l'existence de plusieurs types qui peut expliquer cette matrice si dispersée. Il y a plusieurs types de populations ici donc c'est compliqué d'analyser la matrice de population sur l'ensemble de la population. Au final, à travers nos analyses, on avait déterminé qu'il y avait 2 comportements distincts qu'on constatait bien entre CP1 & CP2 (avec certains individus notamment BloodPressure, Gluscose et Outcome) qui participe aux 2 groupes. Dans ce TP, on va utiliser les cartes topologiques, c'est une représentation des données sous forme de cartes.



Quelle est la force des cartes topographiques ?

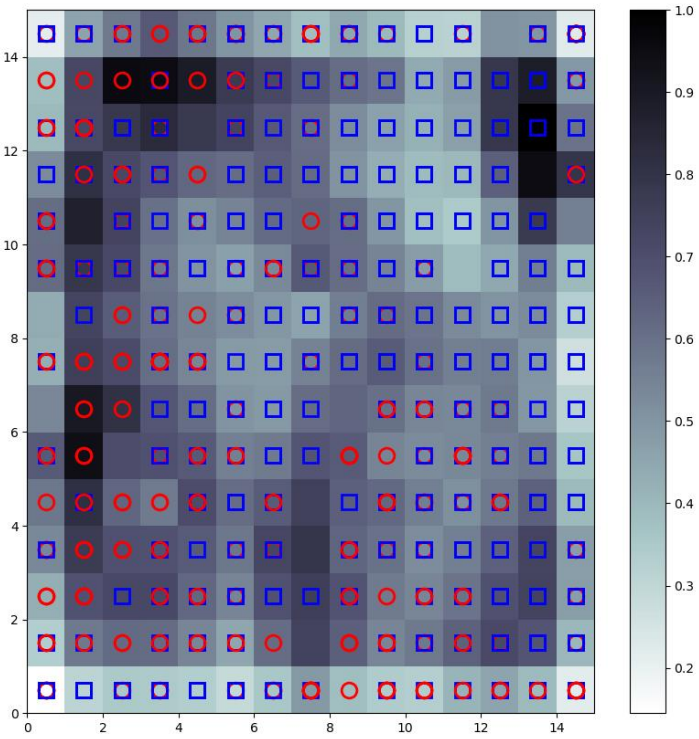
Sa force, contrairement à la matrice de corrélation est qu'on y voit à la fois les corrélations linéaires et non-linéaires = c'est plus puissant que les matrices & cercles de corrélations.

- Sur les cartes topographiques, le size (le zoom) & sigma (dépendance des autres clusters) est très important car en jouant dessus, on prend du recul sur les graphiques analysés. Ici size=15 et sigma=1.5 (c'est le juste milieu après plusieurs tentatives).
- On constate des corrélations. On regarde les zones en rouges et on voit directement des corrélations typiquement avec BloodPressure, BMI + SkinThickness = la petite tâche bleue. Cela démontre la présence d'un sous-comportement d'individus au sein de ces facteurs.
- Diabète insulino-dépendant : on retrouve notre constat avec le TP précédent, on voit que SkinThickness, Insulin, BMI (et DiabetesPedigreeFunction dans une moindre mesure). Ce constat on le voit car on part du bleu en haut à gauche vers le rouge qui part du bas à droite. Cela montre une corrélation entre ces facteurs.
- Diabète lié à l'âge / grossesse : on constate aussi clairement une corrélation entre l'âge et la grossesse car les 2 graphes se ressemblent de part la concentration des individus rouges en haut à droite (les 2 tâches en haut) et la répartition en bleu sur l'ensemble des autres parties de la carte.
- Enfin, autre constat, BloodPressure et Insulin sont anti-corrélés car les couleurs partent de manières opposées.



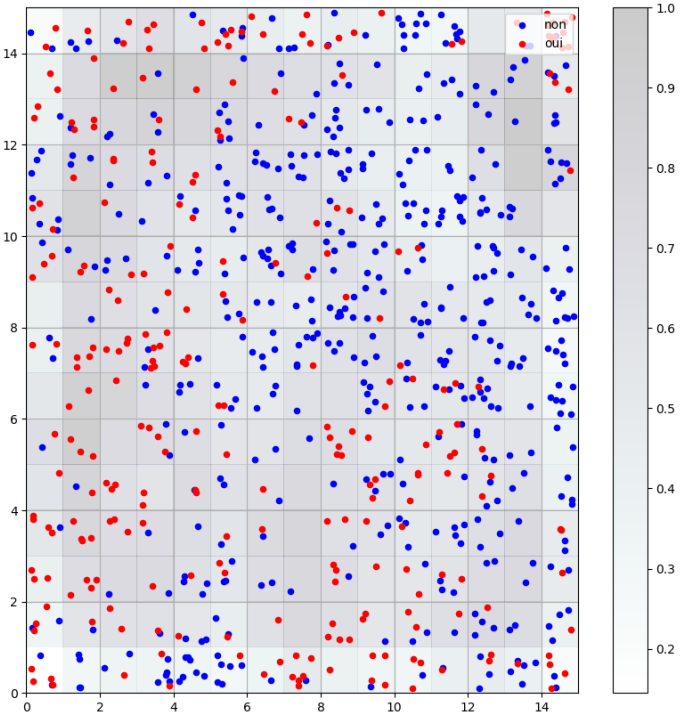
Sur la quantization error, elle représente perte d'informations (1.20 / 3.12) = 38% pertes entre si on met size=15 ou size=1. Cette perte est plus ou moins importante : cela dépend de ce qu'on cherche à analyser.

On va maintenant zoomer sur les Outcome. Sur cette carte, ce qui est intéressant c'est la couleur foncée. Elle représente les zones avec peu d'informations. On voit quelques tâches notamment en haut à droite et sur la longueur à gauche. Ainsi, ce qu'on comprend c'est des zones de vides. On va mettre en perspective avec les prochains graphes.

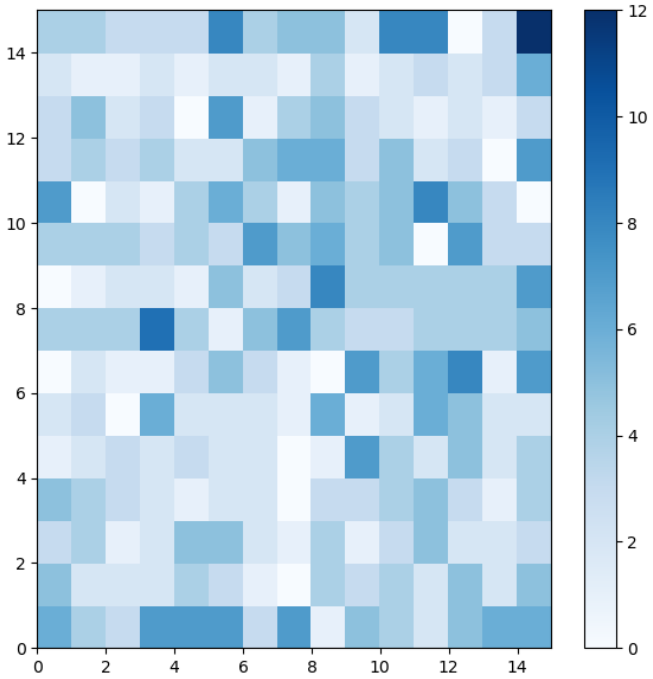


Du coup, on retrouve les zones en noires en haut, sur l'exact même localisation sur cette carte en bas, on voit qu'à ces zones noires, il y a pas/peu de points.

D'ailleurs, vu que les couleurs représentent en terme de couleurs les Outcome, on retrouve exactement la même répartition que les Outcome avec ces points (sur le premier graphe de ce rapport). Les zones ne vides nous en disent beaucoup sur la présence éventuelle de groupes car les clusters sont comme l'espace et attirent les individus vers eux, donc les zones de vides sont comme les galaxies et démontrent la présence d'une délimitation/frontière. On arrive à voir la présence de 2 groupes (2 tâches claires) au centre de la carte délimitée par les zones noires.

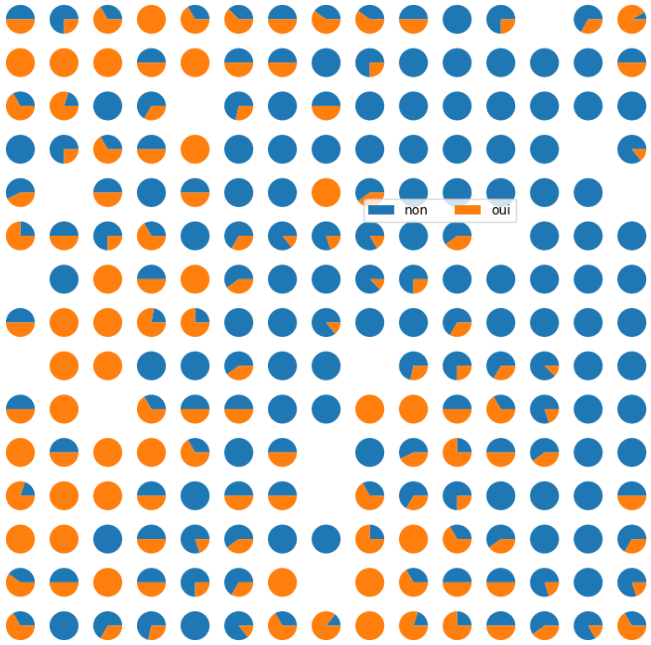


Cette carte montre qu'il y a une forte concentration d'individus. On voit les individus qui sont représentés avec les tâches foncées plus haut. Par exemple, le point tout en haut à droite est très foncé et on voit sur la carte précédente 12 points en haut à droite. Après en elle-même, cette carte ne nous montre pas plus de choses.



Cette nouvelle carte montre la répartition entre oui et non des résultats d'Outcome par rapport aux précédent graphe. Cette carte est très intéressant car si on superpose par-dessus le cercle de

corrélation, on retrouve clairement nos 2 groupes à gauche (présence de oui) en haut à gauche et en bas à gauche de la carte avec un groupe assez dispersé au milieu. Et on retrouve notre hama de personnes saines qui étaient de couleurs violettes (sur le diagramme de nuage de dispersion).



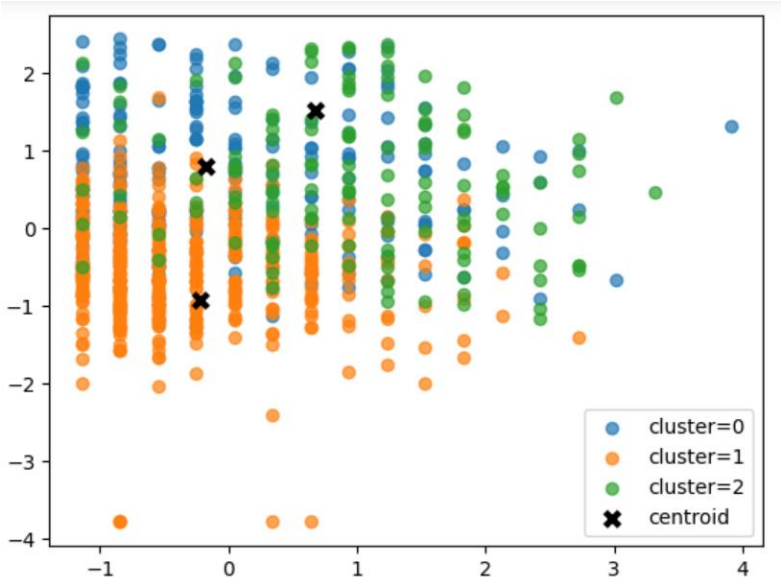
- Sur ce tableau de classification, ce qui est important c'est le F1 Score. Ce score est composé d'un calcul de ratio entre la précision et le recall. La précision c'est la précision du modèle (à quel point ce modèle est bien ou pas) : pour juger si le modèle est bon pour de la prédiction.
- Ainsi, on constate un déséquilibre entre les diabétiques ou non.
- Pour les 0 (non-diabétique), on voit qu'il y a un F1 Score de 0.79. C'est un plutôt bon score (encore tout dépend ce que qu'on cherche).
- Pour les 1 (les diabétiques), on constate qu'il a néanmoins un F1 Score de 0.42, c'est-à-dire pour le coup que le modèle est plutôt mauvais. Il fait plus de 2 fois d'erreurs pour une bonne prédiction.

Lecture : un **F1-score de 50%** équivaut à TP = ½ (FN + FP) et s'interprète donc de la façon suivante : pour une prédiction positive correcte, le modèle fait deux erreurs (faux négatif ou faux positif).

	precision	recall	f1-score	support
0	0.70	0.91	0.79	99
1	0.65	0.31	0.42	55
accuracy			0.69	154
macro avg	0.68	0.61	0.61	154
weighted avg	0.69	0.69	0.66	154

Après plusieurs tests pour rentrer le bon nombre de cluster et le bon nombre de training, j'ai laissé 500 trains + 3 clusters.

J'en arrive à trouver ce graphique :



- On constate qu'avec ce clustering, on retrouve plus ou moins le diagramme de dispersion des points qu'on avait trouvé en ACP/K-Means (et le cercle de corrélations).
- Juste ici on a colorisé les points en fonction des couleurs et qu'on voit les centroid des clusters.
- D'ailleurs, presque dire que le cluster 2 (en vert) est les individus sains à droite, le cluster 0 (en bleu) est les diabétiques insulino-dépendants et le cluster 1 (en orange) avec les diabétiques liés à l'âge et à la grossesse.
- On voit d'ailleurs plusieurs individus qui participent aux 2 clusters à gauche.

Conclusion

Avec les cartes topographiques, on arrive à retrouver presque les mêmes résultats que ACP/K-Means mais avec plus d'avantages car on analyse les données sous forme de cartes (avec différentes granularités : on zoom, zoom ou on change de perspectives) et c'est cela qui est intéressant.