Département Sciences du Numérique

Analyse de données et Classification 2

Classification de battements cardiaques

Ayoub Canon, Léopold Lopez

ENSEEIHT
2024-2025

Table des matières

1 Introduction

L'objectif de ce projet est de classifier les rythmes cardiaques à partir d'un jeu de données. Pour cela, nous mettrons en œuvre l'algorithme DTW , ainsi que deux méthodes de classification supervisées et deux méthodes non supervisées. Nous appliquerons ensuite ces mêmes approches après un prétraitement basé sur l'ACP, afin d'étudier l'impact de cette transformation sur la performance des modèles.

2 DTW

Par DTW on obtient une accuracy de 34,7% sur la base de test.

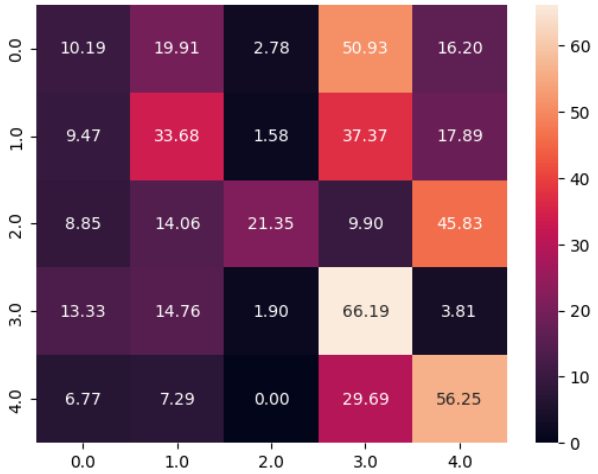

FIGURE 1 – Matrice de confusion pour DTW

Nous avons supposé que l'algorithme DTW a fourni des résultats peu satisfaisants à cause de la trop grande similarité des différentes classes de données.

3 Choix des méthodes utilisées

Tout d'abord, nous étudierons ce problème au regard de la nature des données, de leur distribution ainsi que des objectifs de classification visés, afin de déterminer les approches les plus adaptées pour obtenir des résultats pertinents.

Pour choisir nos méthodes de classifications, nous avons jugé bon de procéder à une représentations de nos données par ACP en trois dimensions.

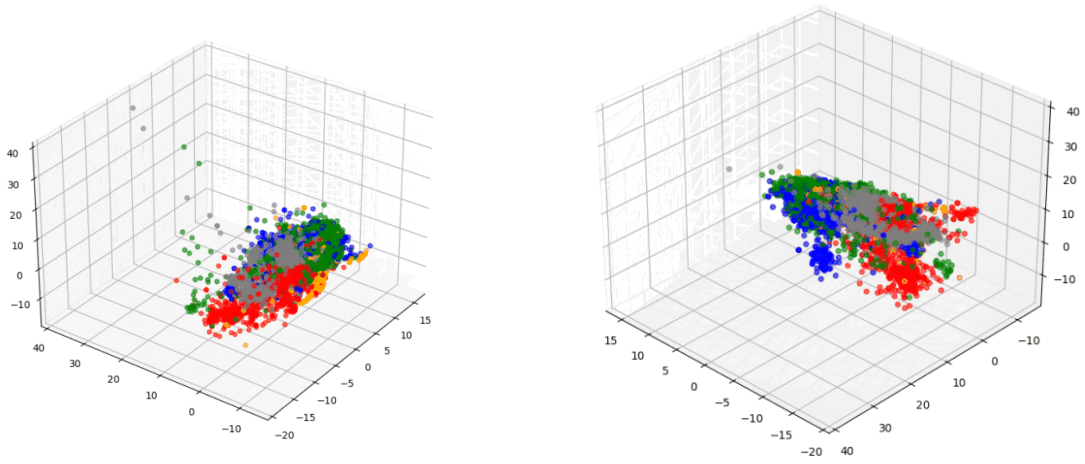

FIGURE 2 – Représentations des données labellisées

Nous avons choisi d'implémenter les méthodes supervisées Random Forest ainsi que SVM. Random Forest présente l'avantage d'être performant et robuste, l'algorithme gère bien les données bruitées. SVM quant à lui, permet de trouver une séparation même dans des données complexes, et a l'avantage d'être robuste au surapprentissage.

Nous avons ensuite décidé d'implémenter les méthodes non-supervisées K-Means et Classification Hiérarchique. K-Means présente l'avantage d'être un algorithme simple et rapide à exécuter permettant d'obtenir un cas de base à partir duquel comparer avec les autres méthodes. De plus, K-Means excelle lorsque les groupes sont "sphériques" et nos données, bien que ne répondant pas à cette exigence, se regroupent tout de même entre elles en cluster. La méthode de Classification Hiérarchique, elle, permet de pouvoir identifier des clusters plus complexes que ceux obtenus par la méthode des K-means.

4 Sans prétraitement

4.1 Random Forest

4.1.1 Études sur les paramètres

Nous avons testé l'algorithme de Random Forest avec un nombre d'arbre allant de 5 à 200 et nous avons trouvé que les meilleurs résultats étaient obtenus à 195 arbres.

4.1.2 Résultats

On obtient une accuracy de 92.5%.

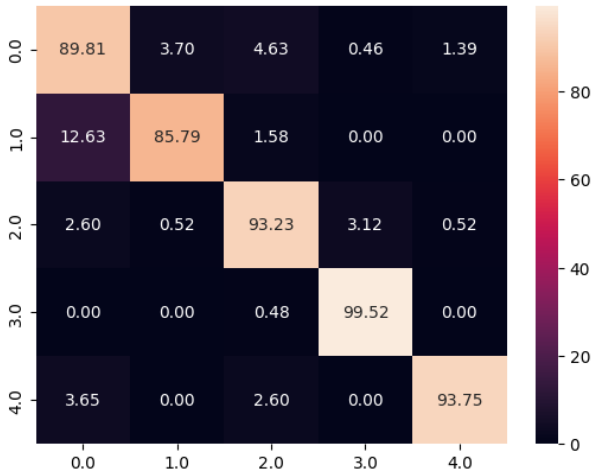

FIGURE 3 – Matrice de confusion pour Random Forest

4.2 SVM

4.2.1 Études sur les paramètres

Nous avons étudié trois paramètres pour cette méthode : le choix du kernel, du paramètre de régularisation, et du gamma.

Le kernel que nous avons choisi est 'rbf' car c'est celui qui est le plus pertinent dans le cas de données non linéarelement séparables, ce qui est notre cas.

Ensuite, nous avons testé la méthode avec un paramètre de régularisation allant de 0.1 à 1000 (en multipliant par 10 à chaque essai), pour un gamma soit à 'scale' soit à 'auto'. Nous avons trouvé un résultat optimal pour C = 1000 et gamma = 'auto'

4.2.2 Résultats

Nous avons obtenu pour cette méthode une accuracy de 92.5% également.

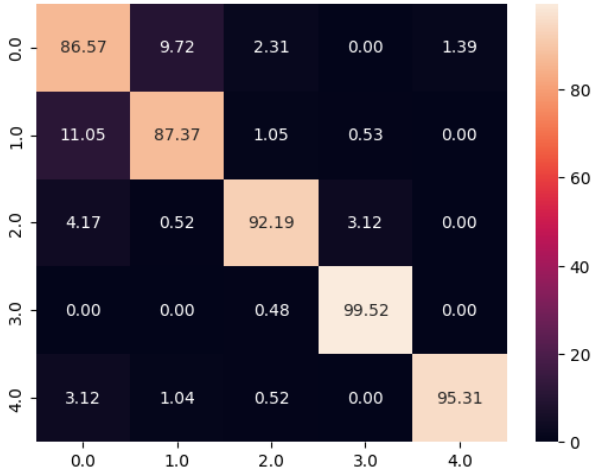

FIGURE 4 – Matrice de confusion pour SVM

Nous pensions qu'avec un paramètre de régularisation aussi élevé il y aurait un problème de sur-apprentissage, cependant cela ne semble pas cela ne semble pas poser de problème sur ce jeu de données. Peut-être que le problème se révélerait avec un jeu de données plus conséquent.

4.3 K-means

4.3.1 Résultats

La méthode des K-means fournit une accuracy de 40,42%

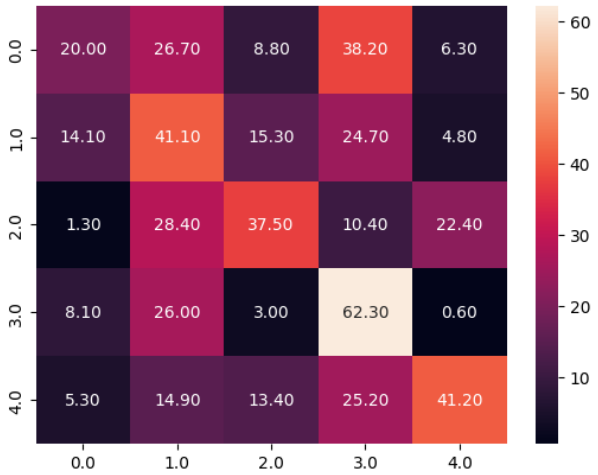

FIGURE 5 – Matrice de confusion pour les K-means

4.4 Classification Hiérarchique

4.4.1 Études sur les paramètres

Contrairement à la méthodes des K-means on ne fixe pas de nombre de classes et on ne donne pas de configuration initiale. Nous avons donc testé avec toutes les mesures et tous les critères de liaisons fournis par sklearn et nous avons trouvé un résultat optimal pour la mesure euclidienne et le critère de Ward qui permet de minimiser la variance intra-cluster.

4.4.2 Résultats

L'accuracy obtenue est de 50,48% par classification hiérarchique.

FIGURE 6 – Matrice de confusion pour la Classification Hiérarchique

4.5 Analyse des résultats

Sans prétraitement nous obtenons de bien meilleurs résultats à l'aide des méthodes supervisées, ce qui semble plus cohérent compte tenu de la nature très imbriquée de nos données, comme le montre l'ACP. En effet, les méthodes non supervisées ont plus de difficultés à discerner les différences que l'on souhaite exhiber dans cet amas de points très proches.

5 Avec prétraitement par ACP

Nous avons décidé de garder les mêmes méthodes de classifications : les méthodes supervisées ayant déjà donné d'excellents résultats, il paraissait cohérent de les garder, et nous avons gardé les méthodes non supervisées car les arguments sur la structure des données restent vérifiables sur les données réduites.

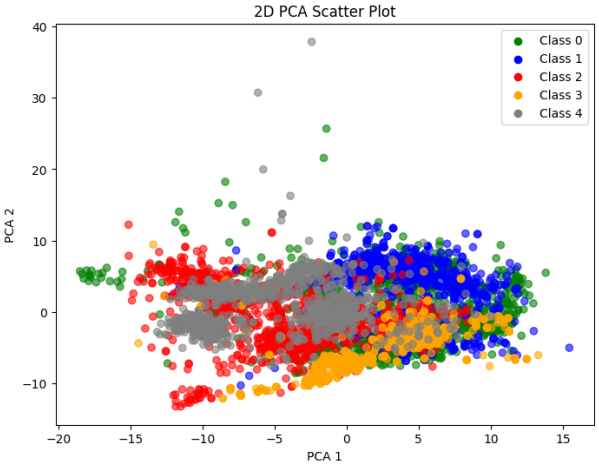

FIGURE 7 – ACP 2D

5.1 Random Forest

5.1.1 Résultats

Nombre d'arbres : 195 Accuracy sur base de test : 70.5

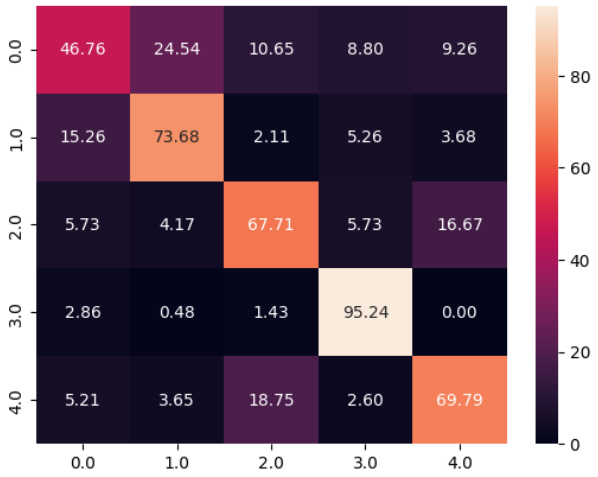

FIGURE 8 – Matrice de confusion pour Random Forest

Le nombre d'arbre optimal s'est révelé être le même mais on note un résultat bien plus faible. C'est à attendre cependant de par la perte d'information dans les données.

5.2 SVM

5.2.1 Résultats

Paramètre de régularisation : 1000 gamma : "auto" Accuracy sur base de test : 66.8

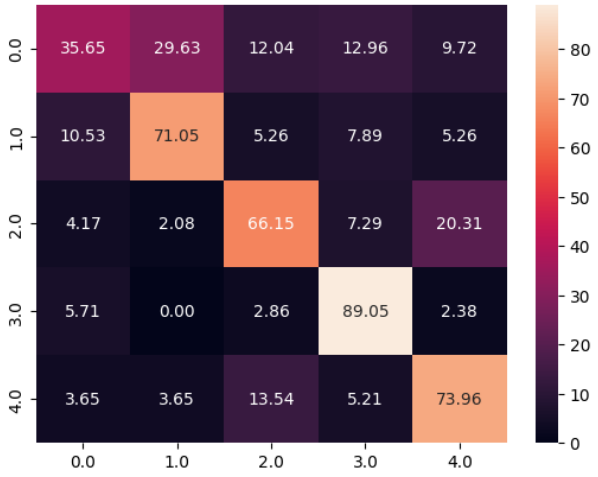

FIGURE 9 – Matrice de confusion pour SVM

De même, les paramètres optimaux restent les mêmes et l'accuracy baisse, ce qui est également attendu pour cette méthode pour les mêmes raisons.

5.3 K-means

5.3.1 Résultats

Accuracy sur base de test : 49.72

FIGURE 10 – Matrice de confusion pour K-Means

Avec cette méthode cependant, on note une amélioration de l'accuracy. Cela s'explique par le fait qu'en réduisant à seulement deux dimensions, cela a créé des groupes plus rapprochés. Par conséquent, tenter de trouver les classes par voisins en se basant sur la moyenne des points de chaque cluster est légèrement plus efficace.

5.4 Classification Hiérarchique

5.4.1 Résultats

Distance : euclidienne Critère de liaison : ward Accuracy sur base de test : 49.24

FIGURE 11 – Matrice de confusion pour la classification hiérachique

De même, on remarque une amélioration de l'accuracy pour la même raison : les données similaires sont légèrement plus rapprochées entre elles.

6 Conclusion

Les méthodes supervisées (Random Forest et SVM) offrent les meilleures performances sur les données brutes, avec des taux de bonne classification avoisinant 92,5%. L'application d'une ACP a toutefois dégradé leurs résultats en raison de la perte d'information lors de la réduction de dimensions. Les méthodes non supervisées (K-Means et Classification Hiérarchique) se sont montrées moins efficaces que les approches supervisées, mais affichent une légère amélioration ou une performance comparable lorsqu'on projette en dimensions plus faibles, bien que cela ne permettent toujours pas d'obtenir des résultats satisfaisants. Ainsi, on privilégiera un algorithme supervisé sur l'ensemble des données pour obtenir des résultats efficaces. Les performances en terme de calculs sont déjà très satisfaisantes en Random Forest sans avoir besoin de procéder à une réduction ACP.