UNIVERSITÉ TOULOUSE III PAUL SABATIER

DECEMBRE 2023

Classification de battements cardiaques

RÉALISÉ PAR : NOUH CHELGHAM ET NOURA FAIZ

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

DECEMBRE 2023

Introduction

Dans ce projet, nous analysons les battements cardiaques en utilisant l'apprentissage automatique. Nos données, issues d'applications grand public et d'essais cliniques, sont converties en MFCC pour classer divers bruits cardiaques. L'étude se concentre sur des méthodes d'apprentissage automatique, visant à affiner la détection des anomalies cardiaques.

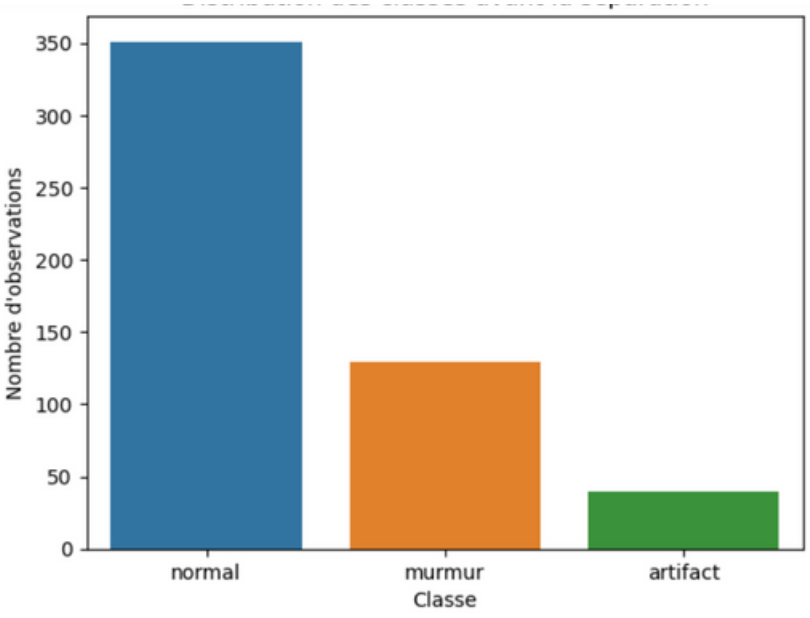
Classification sans prétraitement

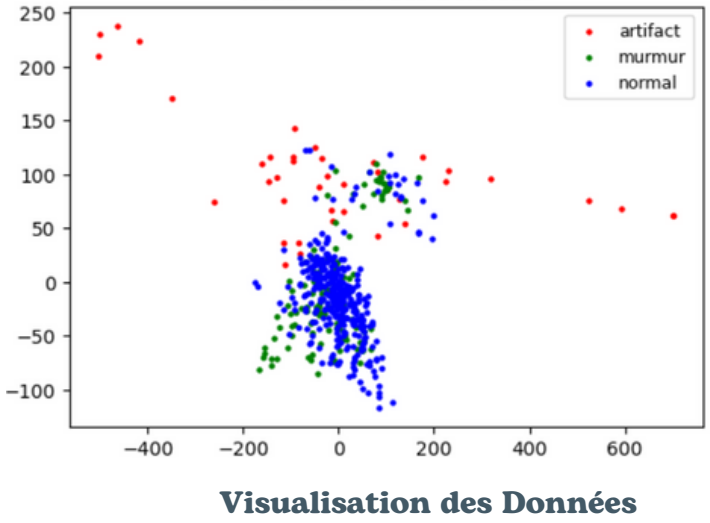
I. Méthode supervisée (KNN : k nearest neighbors)

Le KNN est un algorithme de classification supervisé. Pour chaque observation $x_i = (x_i^1, x_i^1, x_i^2, \ldots x_i^n)$, nous avons une étiquette associée **yi** , qui représente la classe de x_i

Pour classer un nouveau vecteur (x) à l'aide de **KNN** :

- Calculez la distance (souvent euclidienne, mais d'autres métriques peuvent être utilisées) entre x et chaque échantillon xi dans l'ensemble d'entraînement.
- Identifiez les k échantillons xi les plus proches de x .
- La classe attribuée à x sera celle qui est majoritairement représentée parmi ces k voisins les plus proches.



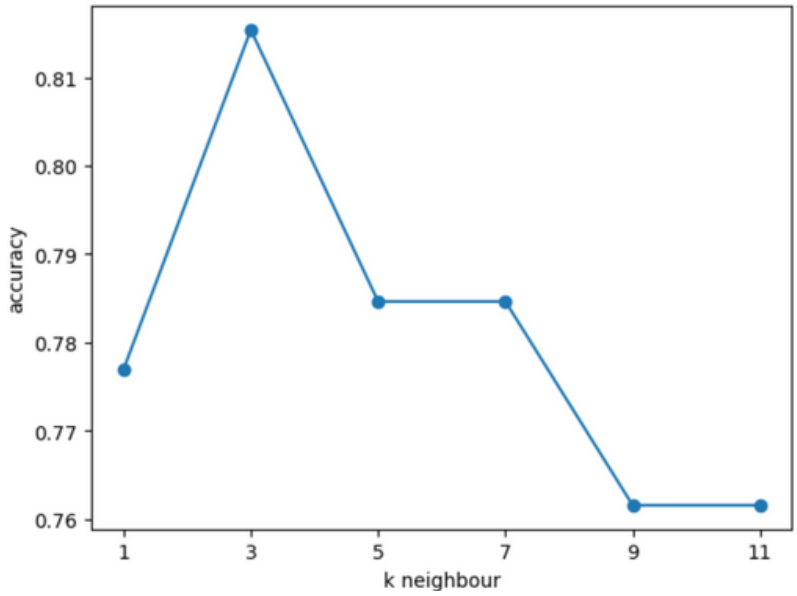


Visualisation des Données

A. Étude sur les paramètres inhérents à la méthode supervisée (k-NN)

1. Impact de la valeur de k sur la précision :

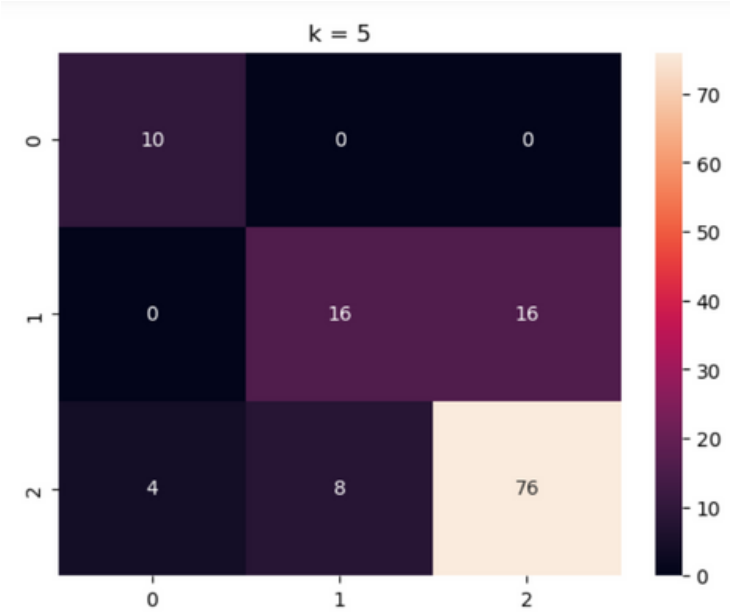
Le graphique ci-dessous illustre comment la précision du modèle k-NN fluctue avec différentes valeurs de k:

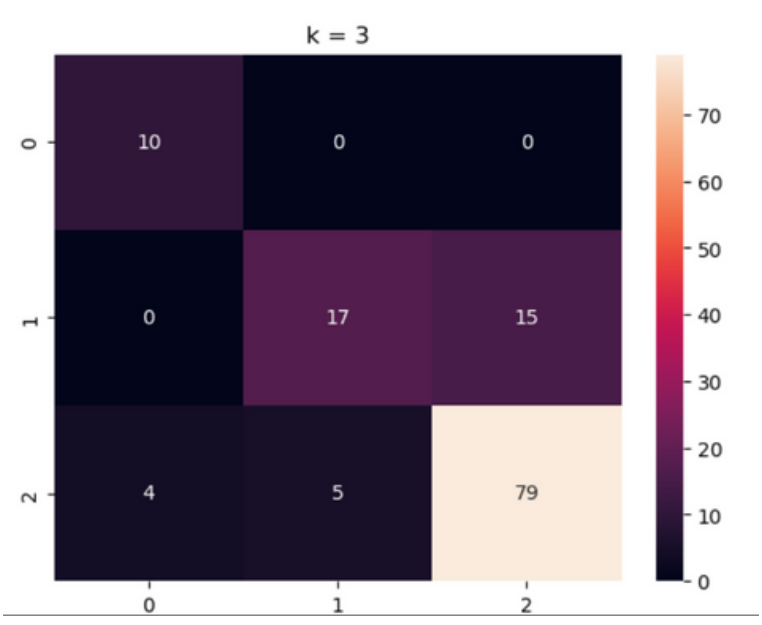


Le modèle KNN a atteint une précision maximale à k = 3 , indiquant que c'est le paramètre optimal pour notre ensemble de données.

2. Matrices de confusion pour différentes valeurs de k :

Elles illustrent la performance du modèle KNN sur la classification des battements cardiaques à différents k.





Interprétation:
Les matrices de confusion montrent que pour k = 3 et k = 5, le modèle KNN confond la classe 1 avec la classe 2. Bien qu'une légère amélioration soit observée à k=3, les erreurs de classification indiquent un **déséquilibre potentiel entre les classes**, suggérant que le modèle pourrait bénéficier d'un **réajustement** ou de l'**application de stratégies d'équilibrage** des classes pour améliorer l'exactitude de la classification.

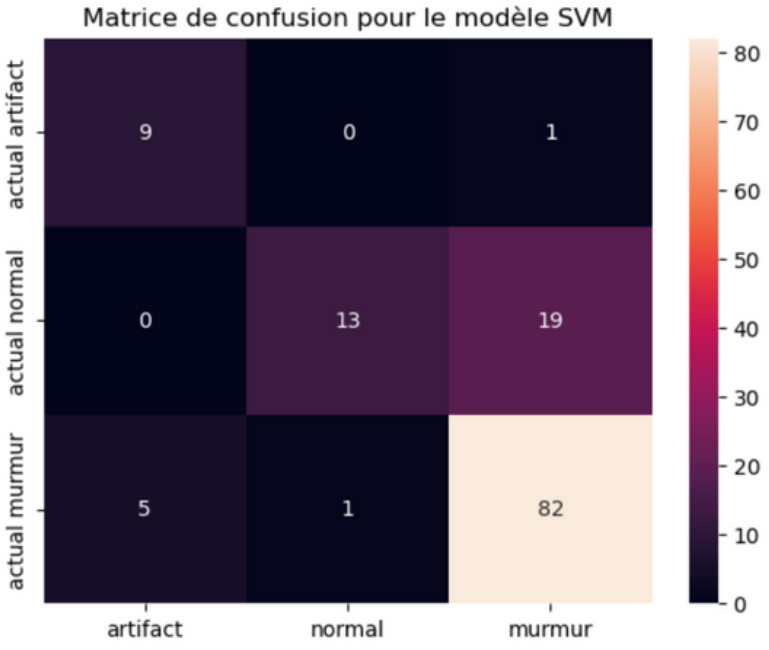
II. Méthode supervisée (SVM : Support Vector Machines)

Le SVM est un algorithme de classification supervisé qui cherche à trouver l'hyperplan qui maximise la marge entre les deux classes. Pour chaque observation , nous avons une étiquette associée (yi), qui représente la classe de (xi).
Pour classer un nouveau vecteur (x) à l'aide de SVM :

1. Construisez un hyperplan ou un ensemble d'hyperplans dans un espace de dimension élevée ou infinie.
2. Choisissez l'hyperplan qui maximise la marge entre les classes d'entraînement.
3. La classe attribuée à (x) sera celle qui est de l'autre côté de la marge par rapport à l'hyperplan.

2. Matrices de confusion et Précision :

Elles illustrent la performance du modèle SVM sur la classification des battements cardiaques .



Interprétation:
La matrice de confusion du modèle SVM démontre une classification précise des "artifact", tandis qu'elle révèle un déséquilibre notable dans la distinction entre "normal" et "murmur".

```
Précision du SVM: 0.8
```

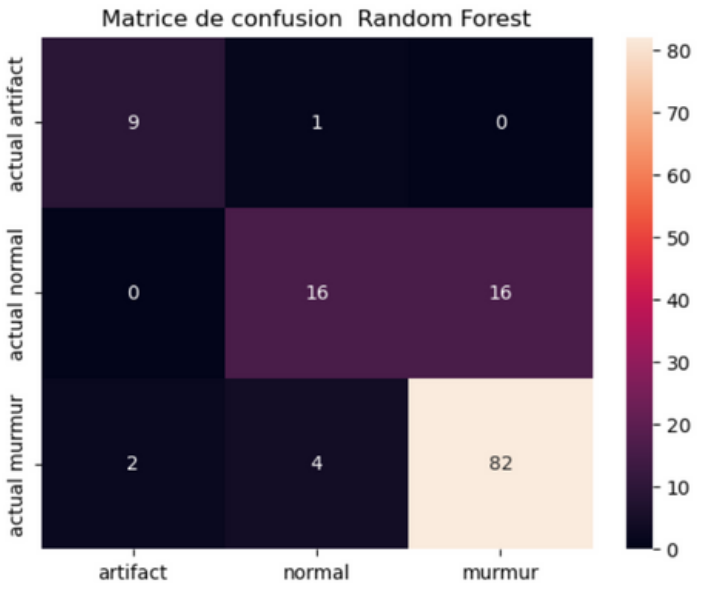
La précision de 0.8 indique que le modèle SVM prédit correctement 80% des cas dans l'ensemble des prédictions.

III. Méthode supervisée (Random Forest)

Le Random Forest est un algorithme qui combine plusieurs arbres de décision pour faire des prédictions plus précises et robustes. Il utilise à la fois des échantillons et des caractéristiques aléatoires pour construire divers arbres, dont les résultats sont ensuite agrégés pour la prédiction finale

2. Matrices de confusion et Précision :

Elles illustrent la performance du modèle sur la classification des battements cardiaques.



Interprétation:
La matrice de confusion pour **Random Forest** montre une bonne identification des "artifact", mais une confusion entre "normal" et "murmur", reflétant un **déséquilibre** dans la classification des classes.

Précision de la Random Forest: 0.823076923076923

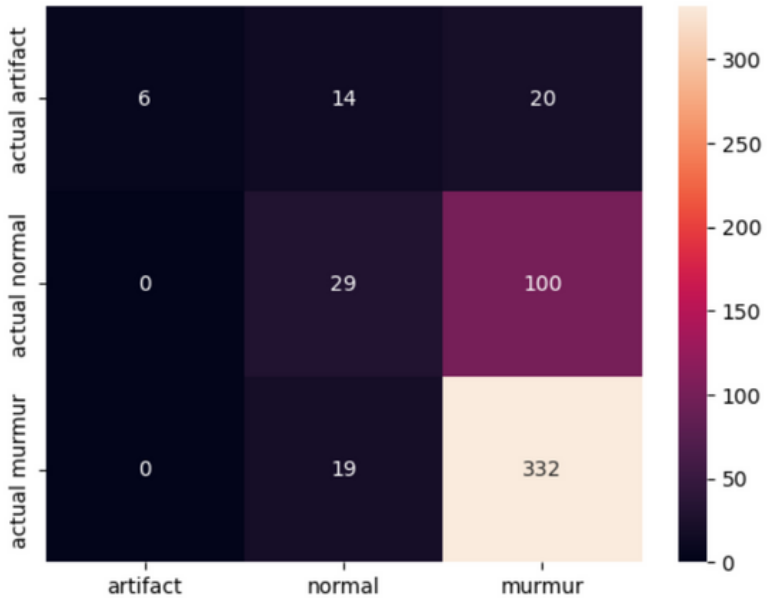
La précision de 0.823 pour le modèle Random Forest signifie que, globalement, environ 82.3% des prédictions du modèle étaient correctes sur l'ensemble des données testées..

IV. Méthode non supervisée (K-means)

K-means est un algorithme de clustering non supervisé qui regroupe les données en k clusters distincts basés sur les caractéristiques similaires. Il attribue les points au centroïde le plus proche et converge quand les centroïdes sont stables.

2. Matrices de confusion et Précision :

Elles illustrent la performance du modèle sur la classification des battements cardiaques.



Interprétation:
La matrice de confusion révèle un déséquilibre de **K-means** dans la distinction entre "artifact" et "normal", avec une meilleure identification des "murmur"..

Score de précision : 0.7057692307692308

La précision de 0.70 pour le modèle K-means signifie que, globalement, environ 70% des prédictions du modèle étaient correctes sur l'ensemble des données testées..

Rééquilibrage des Données

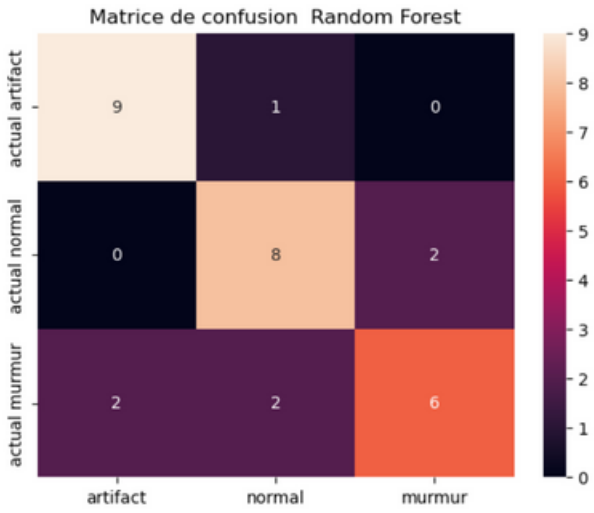
I. Équilibrage des Données par Sous-échantillonnage

C'est une technique de prétraitement des données pour les ensembles de données déséquilibrés. Considérons un ensemble de données avec des classes ($C1, C2, ..., Cn$), où chaque classe Ci a un nombre d'observations ni.
Pour appliquer le sous-échantillonnage :

1. Identifiez la classe avec le moins d'observations, notée par son nombre d'observations $n$min.
2. Pour chaque classe Ci avec $ni > n$**min**, effectuez un échantillonnage aléatoire pour sélectionner $n$min observations.
3. La nouvelle distribution des classes Ci aura alors un nombre d'observations égal ou inférieur à n**min**, conduisant à un ensemble de données équilibré.

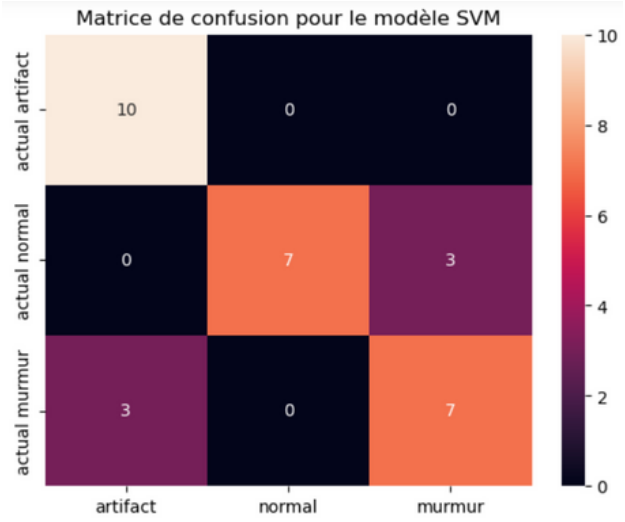
A.Réimplémentation des Méthodes de Classification:

1. RANDOM FOREST :



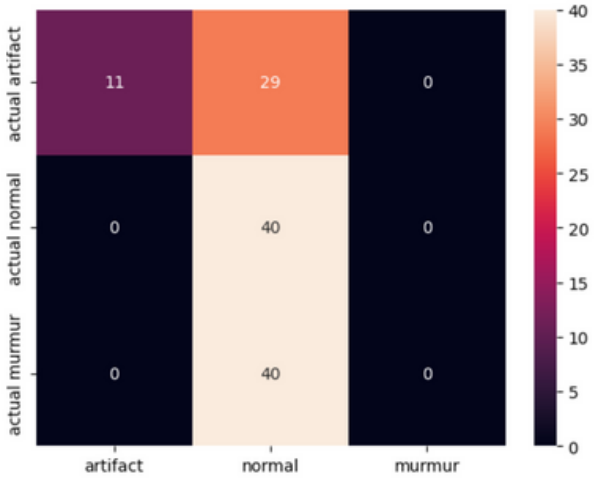
Précision de la Random Forest: 0.76

2. SVM :



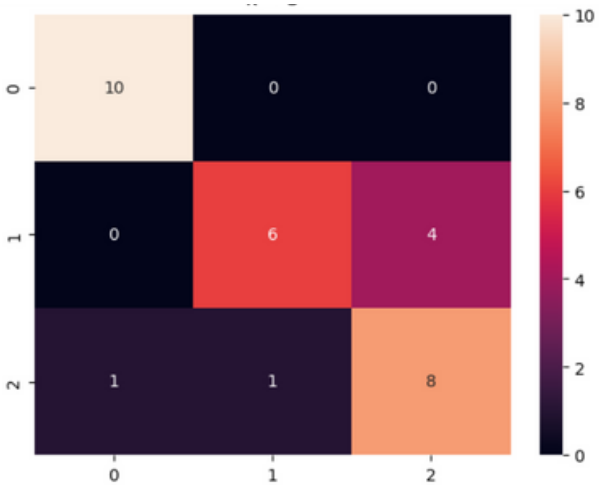
Précision du SVM: 0.8

3. K-MEANS :



Score de précision : 0.425

4. KNN :



k = 3 | accuracy = 0.8

COMPARAISON:

- **KNN** : Avant le rééquilibrage, la précision était d'environ 81.5%. Après le rééquilibrage, la précision reste stable à 80%. Cela suggère que KNN était moins affecté par le déséquilibre des classes que les autres modèles, ou que les caractéristiques discriminantes étaient suffisamment robustes pour maintenir une performance élevée même avant le rééquilibrage.

- **SVM** : Le SVM maintient une précision de 80% avant et après le rééquilibrage. Cela indique que le SVM a été capable de généraliser assez bien malgré le déséquilibre initial des classes, et que le rééquilibrage n'a pas eu d'effet négatif ou positif significatif sur sa performance.

- **Random Forest** : La Random Forest a présenté une légère baisse de précision, passant de 82.3% avant le rééquilibrage à 76% après. Cela peut indiquer que le modèle était initialement biaisé en faveur des classes majoritaires et que le rééquilibrage a révélé des faiblesses dans la capacité du modèle à distinguer entre les classes désormais équilibrées.

- **K-Means** : K-Means a montré une diminution significative de la précision, passant de 70% avant le rééquilibrage à 42.5% après. Comme K-Means est un algorithme de clustering et non de classification, cette baisse pourrait refléter le fait que les clusters formés ne correspondent pas aussi bien aux classes équilibrées, ou que l'algorithme bénéficiait du biais introduit par le déséquilibre des classes.

Classification avec prétraitement

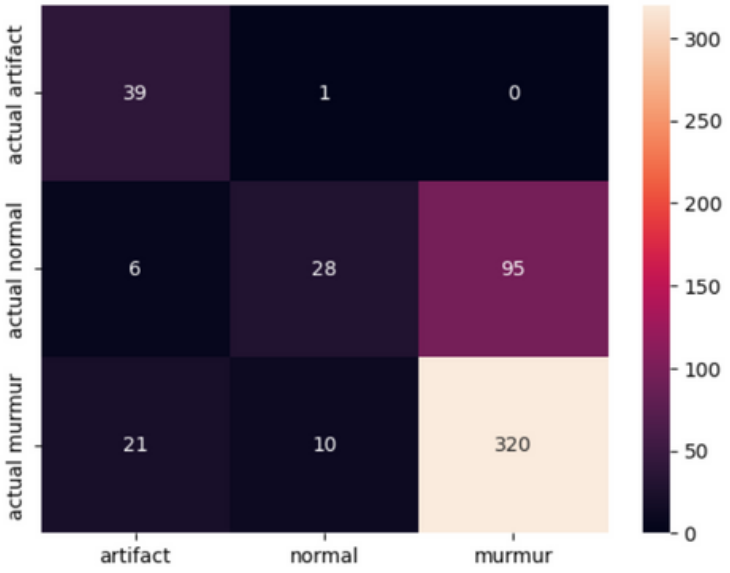
A. Méthodes supervisées et non supervisées avec réduction de dimension par ACP

L'analyse en composantes principales (ACP) est une méthode statistique multivariée qui permet de réduire la complexité d'un ensemble de données en projetant les variables sur un espace de dimensions inférieures. Elle permet de déterminer les relations entre les variables et de les visualiser sous forme de graphiques.

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

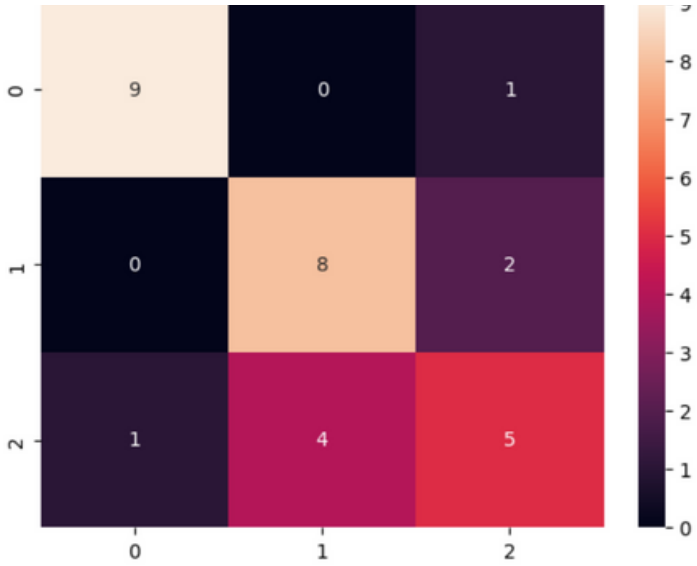
DECEMBRE 2023

1. K-MEANS:



Score de précision : 0.74

2. KNN



k = 1 | accuracy = 0.73

COMPARAISON:

K-MEANS: Avec une précision de 0.74, K-Means montre une amélioration post-ACP par rapport à la précision de 0.70 pré-ACP. Les "murmur" sont bien reconnus, mais des confusions persistent entre "artifact" et "murmur", ainsi qu'entre "normal" et "murmur".

-**KNN**: En revanche, le KNN a vu sa précision baisser de 81.5% pré-ACP à 0.73 post-ACP pour k = 1. Bien que les "artifact" soient correctement identifiés, le modèle se trompe entre "normal" et "murmur". Cette baisse suggère que l'ACP pourrait altérer certaines caractéristiques cruciales pour le KNN, ou que le paramètre k n'était pas idéalement choisi après l'ACP.

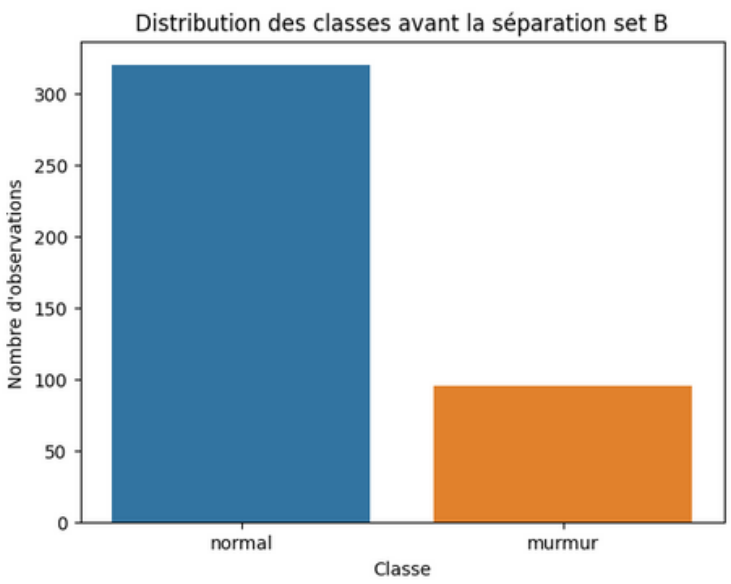
Etude comparative entre A (DataMFCC_SetA.csv) et B (DataMFCC_SetB.csv)

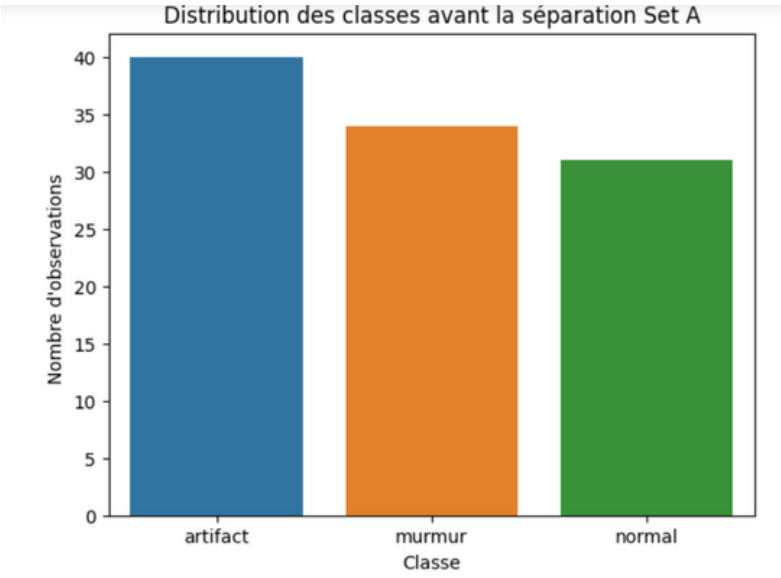
Ces données ont été recueillies auprès de deux sources :

- (A) auprès du grand public via une application de smartphone,
- (B) dans le cadre d'un essai clinique dans des hôpitaux utilisant le stéthoscope numérique.

Les enregistrements de ces 2 sources étant de durées différentes, ils ont été ensuite transformés en MFCC pour extraire le contenu fréquentiel de ces données.

a. Visualisation des Données:



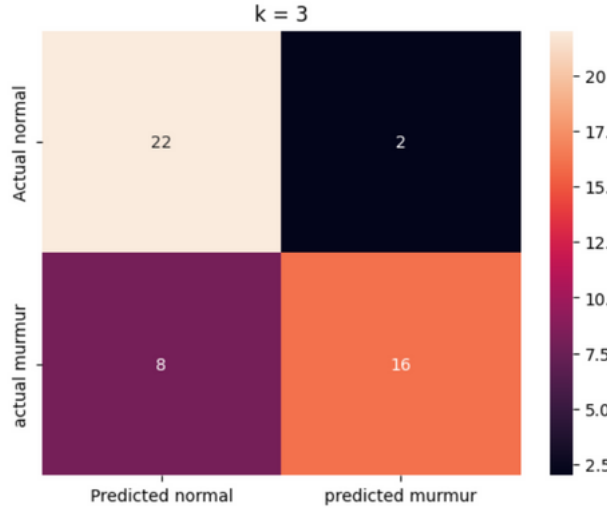


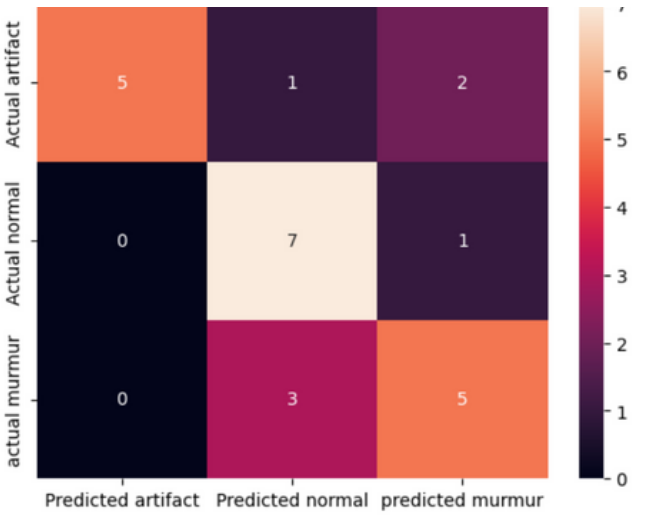
Le graphe illustre une distribution des classes où seulement "normal" et "murmur" sont présents, avec une absence notable de la classe "artifact" sur la base ste B.

I. Méthode supervisée (KNN : k nearest neighbors)

a. Matrices de confusion et Précision :

Elles illustrent la performance du modèle sur la classification des battements cardiaques.



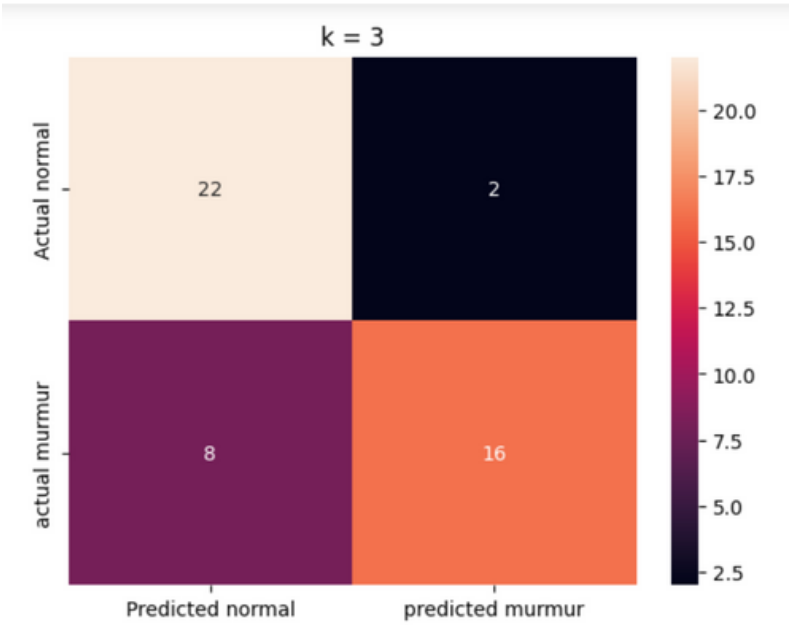



```
k = 1  |  accuracy = 0.68
k = 3  |  accuracy = 0.79
k = 5  |  accuracy = 0.77
k = 7  |  accuracy = 0.72
```

```
k = 5  |  accuracy = 0.7
k = 7  |  accuracy = 0.73
k = 9  |  accuracy = 0.73
k = 11 |  accuracy = 0.73
```

Interprétation:
Set A, bien que mieux équilibré, montre une confusion entre "normal" et "murmur" à k = 5 , tandis que Set B, déséquilibré, ne présente pas de classe "artifact" et continue de confondre "normal" et "murmur".

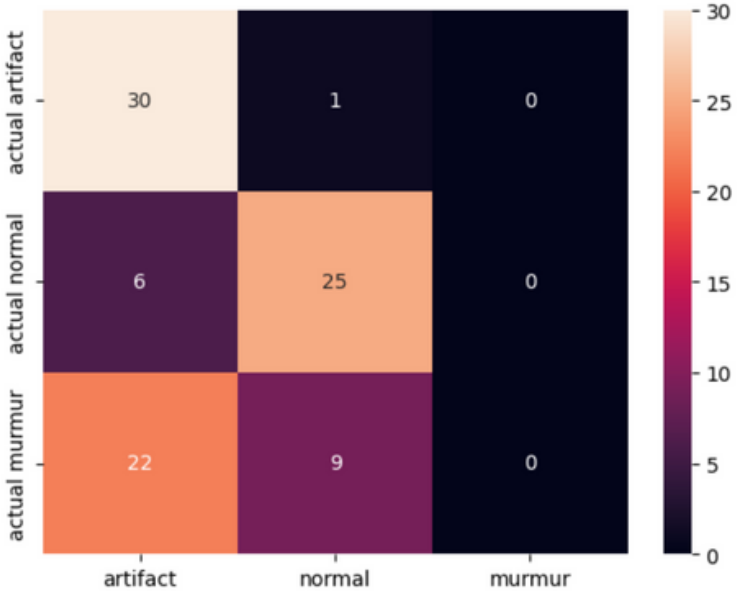
a. Équilibrage de SEt B par Sous-échantillonnage:



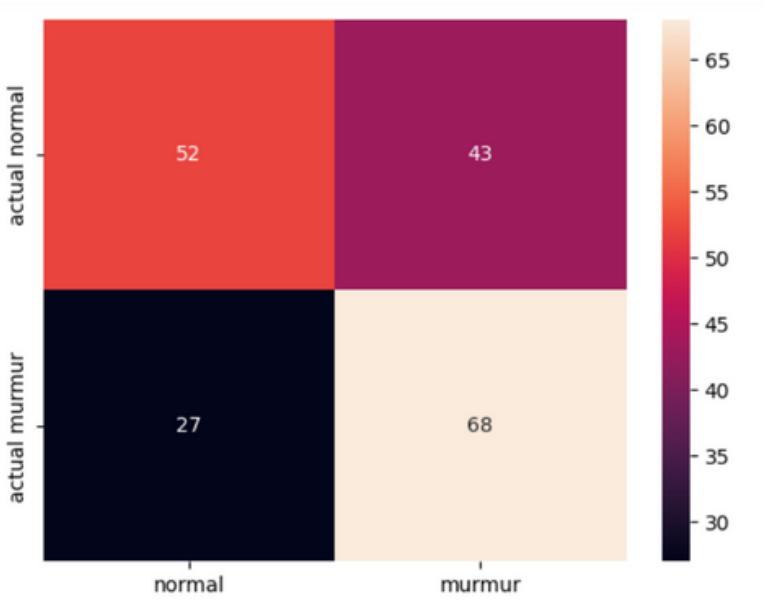
Score de précision : 0.74

b. Prétraitement Avancé : Application de l'ACP

SET A



SET B



Interprétation:

Après l'**ACP**, la confusion entre les classes s'est accentuée dans le **set A**, en particulier pour les "artifact" et "murmur". Pour le **set B**, l'ACP n'a pas amélioré la distinction entre "normal" et "murmur". En comparaison, les résultats avant l'ACP étaient généralement meilleurs, suggérant que l'ACP pourrait ne pas être bénéfique pour ces ensembles de données.

CONCLUSION:

Ce projet a souligné que pour la classification des battements cardiaques, l'équilibrage des classes est plus déterminant que la réduction de dimensionnalité via l'ACP. Les techniques de rééquilibrage ont amélioré la précision des modèles, tandis que l'ACP n'a pas montré d'avantages clairs. Ces découvertes orientent vers une optimisation des prétraitements pour de futures analyses dans le domaine médical.