

Compte Rendu Projet MCS

Pour ce TP nous avons effectué un grand nombre de tests en modifiant à chaque fois des paramètres avec plus ou moins de différence par rapport à notre test initial.

Cela nous a permis d'observer les changements de résultats associés aux changements de valeurs dans les paramètres.

Nous avons donc remarqué que le paramètre le plus impactant était la taille de l'ensemble d'apprentissage dans le cas de l'ACP+kNN.

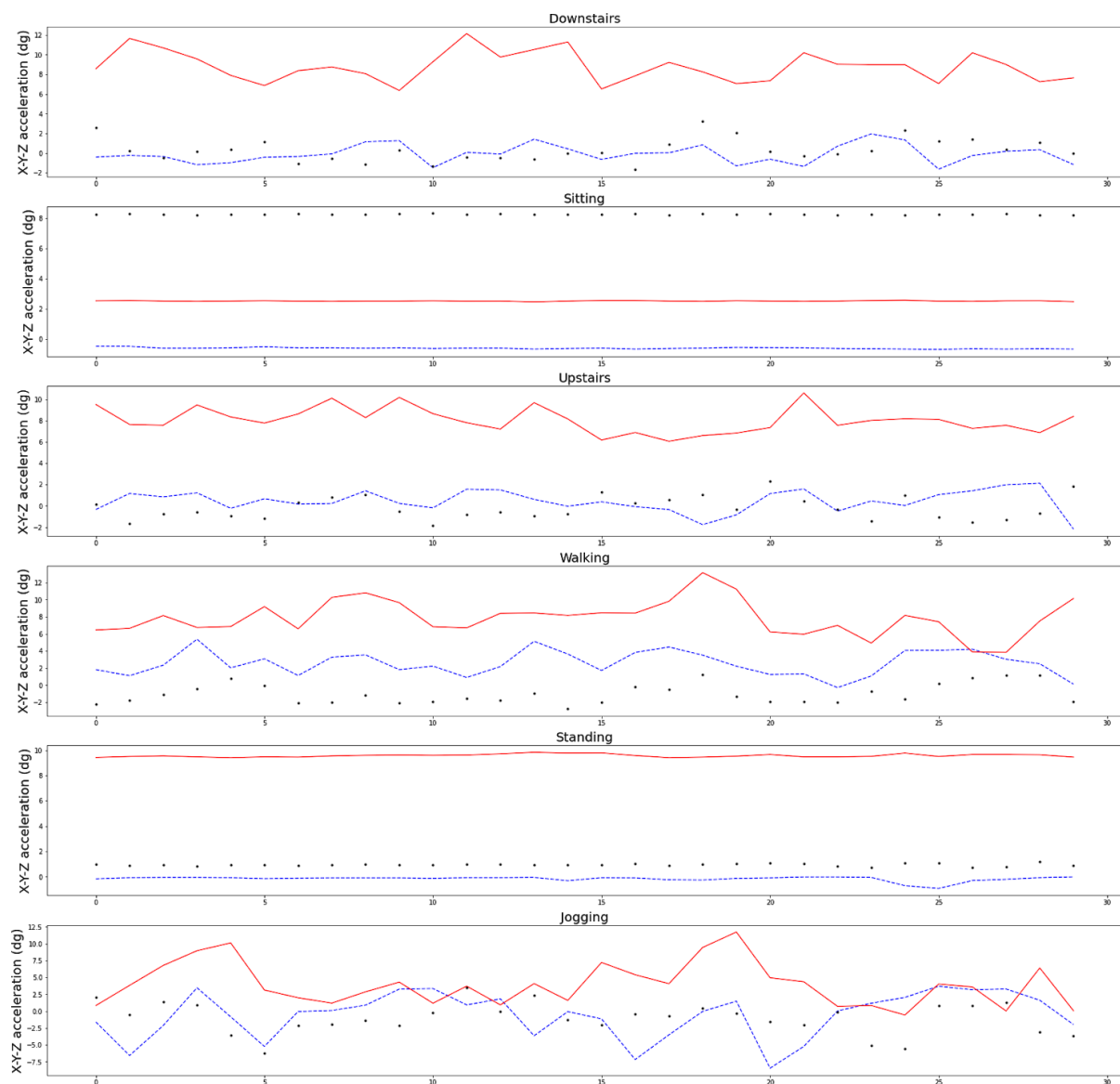
Notre objectif pour ce TP a donc été de chercher le point de surapprentissage de cet ensemble, puis de par la suite faire varier les autres paramètres afin d'affiner encore si possible le pourcentage de bons résultats.

Test initial

Nous sommes partis d'un test initial avec les paramètres suivants :

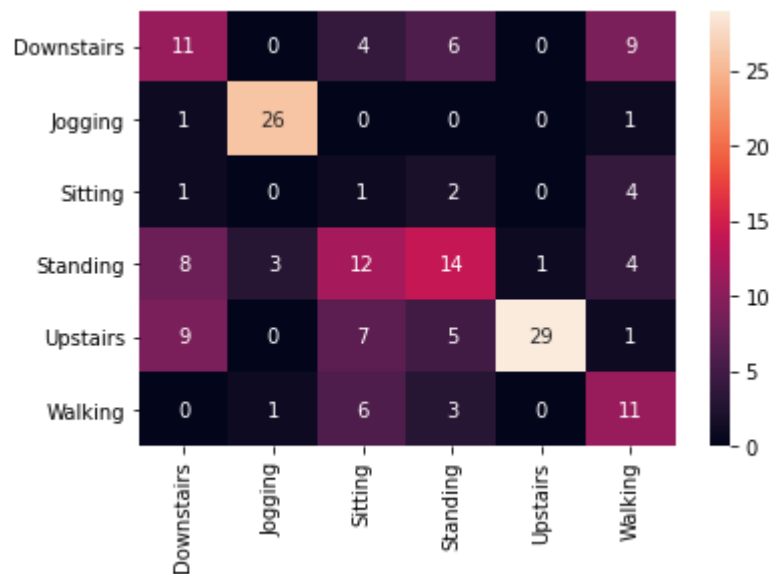
- TIME_STEP = 80
- N_CLASSES = 6
- N_FEATURES = 3
- SEGMENT_TIME_SIZE = 30
- TAILLE_TEST = 30
- TAILLE_APP = 10
- W1 = 1
- W2 = 1
- W3 = 1
- K = 3

Ce test nous a donné les résultats suivants :



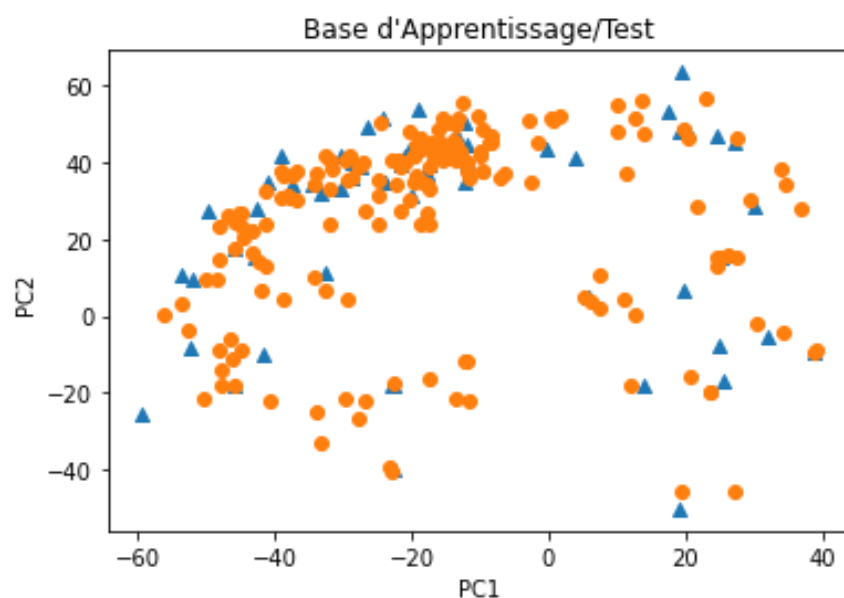
On peut voir que les courbes ne sont pas très constantes et varient beaucoup hormis les courbes concernant les positions debout et assis.

Le score DTW est de 3.6 pour ce test et on a obtenu la matrice de confusion suivante :

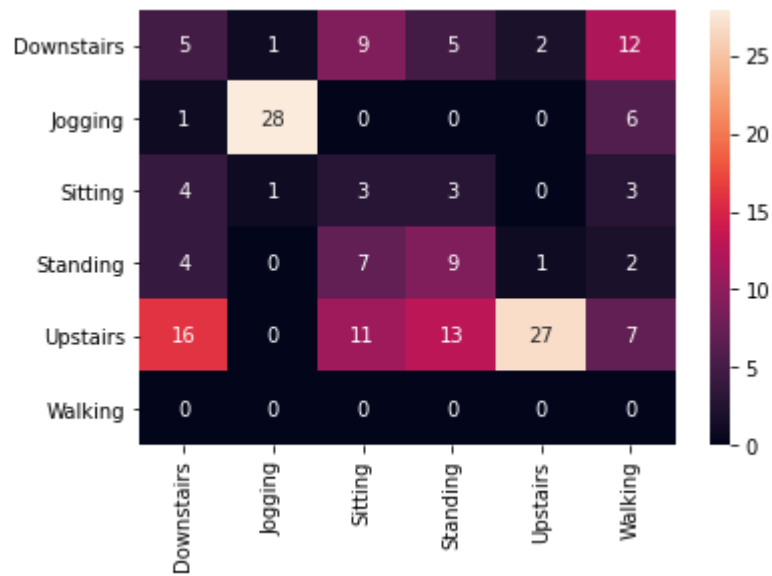


On a une précision de 51,1 % et on peut voir sur la matrice de confusion qu'on obtient des résultats partiellement bons. Notre diagonale est assez chargée mais on a encore des éléments qui sont mal classés comme par exemple des éléments qui sont standing alors qu'ils étaient prédits sitting.

Par la méthode de l'ACP on voit qu'on obtient un contraste de 0.8097 soit un pourcentage d'information conservée de 81% et le graphe suivant :



Par la méthode des k plus proches voisins on obtient la matrice de confusion suivante :



On a une précision de 40% soit plus de 10% de moins que par la méthode de DTW, ce qui se voit sur la matrice pour laquelle on a des cases autres que sur la diagonale indiquant un grand nombre.

On a par exemple beaucoup d'éléments upstairs qui ont été mal prédits à la base.

C'est ce pourcentage de bons résultats que nous allons essayer d'augmenter le plus possible dans la prochaine partie.

Test avec la meilleure taille d'apprentissage

Afin de trouver le point de surapprentissage nous avons donc augmenté petit à petit la taille de l'ensemble d'apprentissage.

L'objet d'étude était de voir pour quelle taille le pourcentage de bonnes prédictions diminuait et par la suite de modifier les autres paramètres afin d'optimiser la bonne détermination des classes.

La diminution du pourcentage de bons résultats aurait indiqué un surapprentissage et la taille, assez précisément, de l'ensemble d'apprentissage à partir de laquelle nous n'obtiendrions pas un meilleur taux de réussite.

Nous avons tout au long de nos tests gardé une taille d'ensemble de tests au moins la moitié supérieure à celle de l'ensemble d'apprentissage afin d'être certains que l'on évalue des éléments qui n'ont pas fait partis de l'ensemble d'apprentissage.

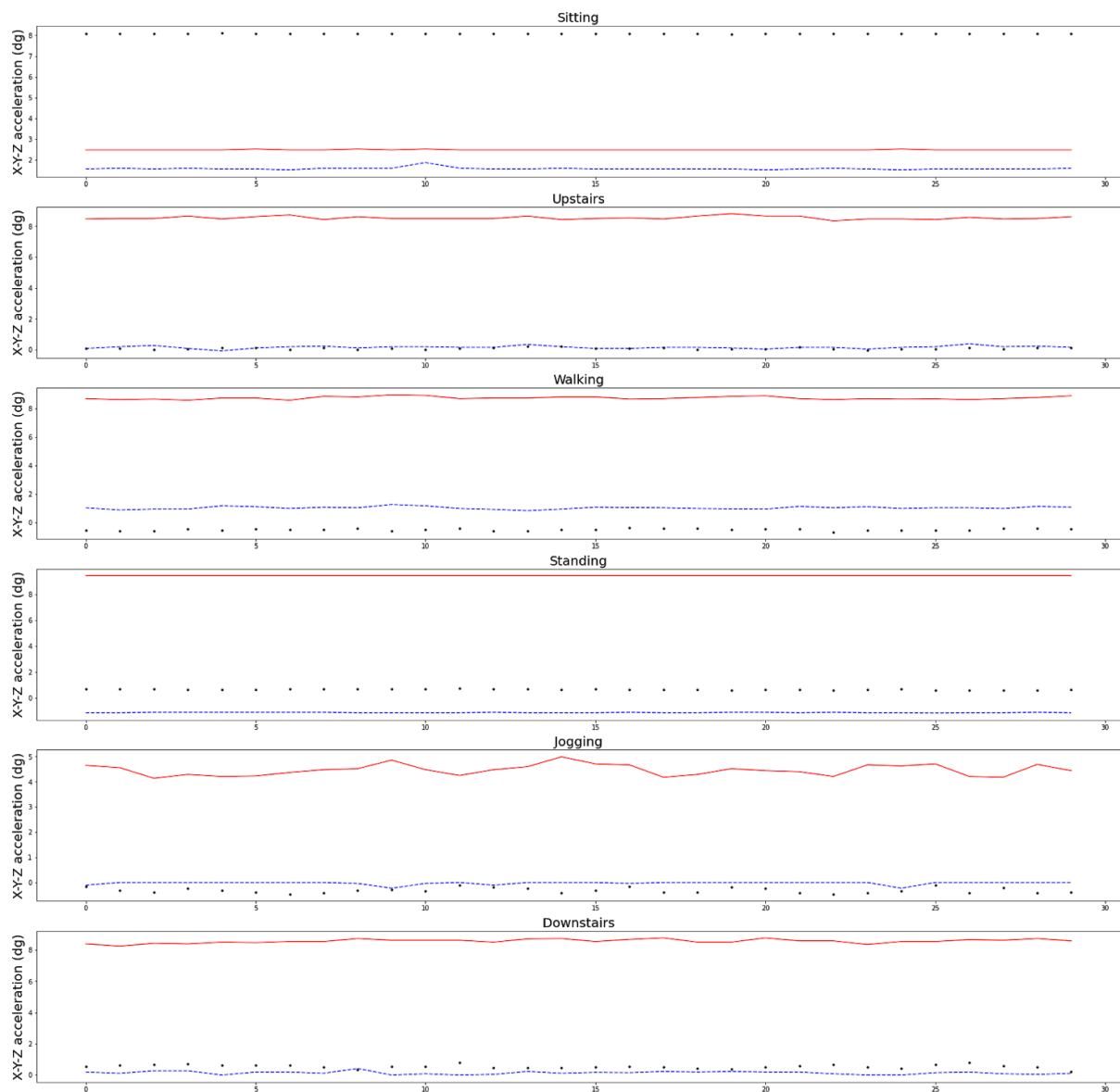
En pratique nous n'avons pas trouvé ce point de surapprentissage car l'ensemble de données est trop grand et plus nous augmentions la taille de l'ensemble d'apprentissage et de tests plus le temps d'exécution était long, jusqu'à durer des heures.

Nous sommes cependant parvenus à obtenir des résultats se rapprochant des 100% pour l'ACP+kNN sur nos derniers tests.

Voici les paramètres du dernier que nous avons réalisé, donc avec la plus grande taille d'ensemble d'apprentissage que nous ayons pu tester :

- TIME_STEP = 80
- N_CLASSES = 6
- N_FEATURES = 3
- SEGMENT_TIME_SIZE = 30
- TAILLE_TEST = 15000
- TAILLE_APP = 10000
- W1 = 1
- W2 = 1
- W3 = 1
- K = 3

Il nous a donné les résultats suivants :



Avec un si grand ensemble d'apprentissage on constate que les courbes pour chaque classe sont bien plus constantes que dans notre test initial.

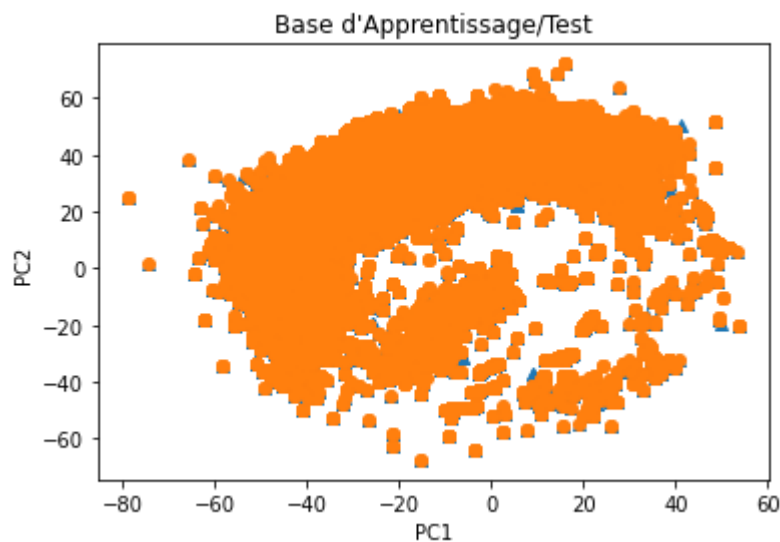
Nous avons obtenu un score DTW de 3.6 et la matrice de confusion suivante :



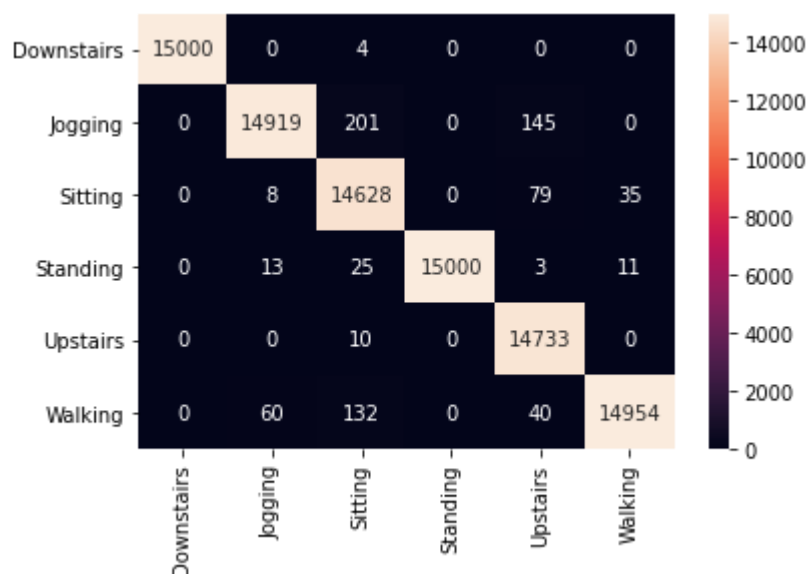
On a une précision de 46,7 % et on peut voir sur la matrice de confusion qu'on obtient des résultats partiellement bons. Notre diagonale est assez chargée mais on a encore des éléments qui sont mal classés comme par exemple des éléments qui sont standing alors qu'ils étaient prédits jogging.

La précision en DTW est assez similaire avec celle du test initial malgré un plus grand nombre de données prises en compte.

Par la méthode de l'ACP on voit qu'on obtient un contraste de 0.6467 soit un pourcentage d'information conservée de 64.7% et le graphe suivant :



Par la méthode des k plus proches voisins on obtient la matrice de confusion suivante :



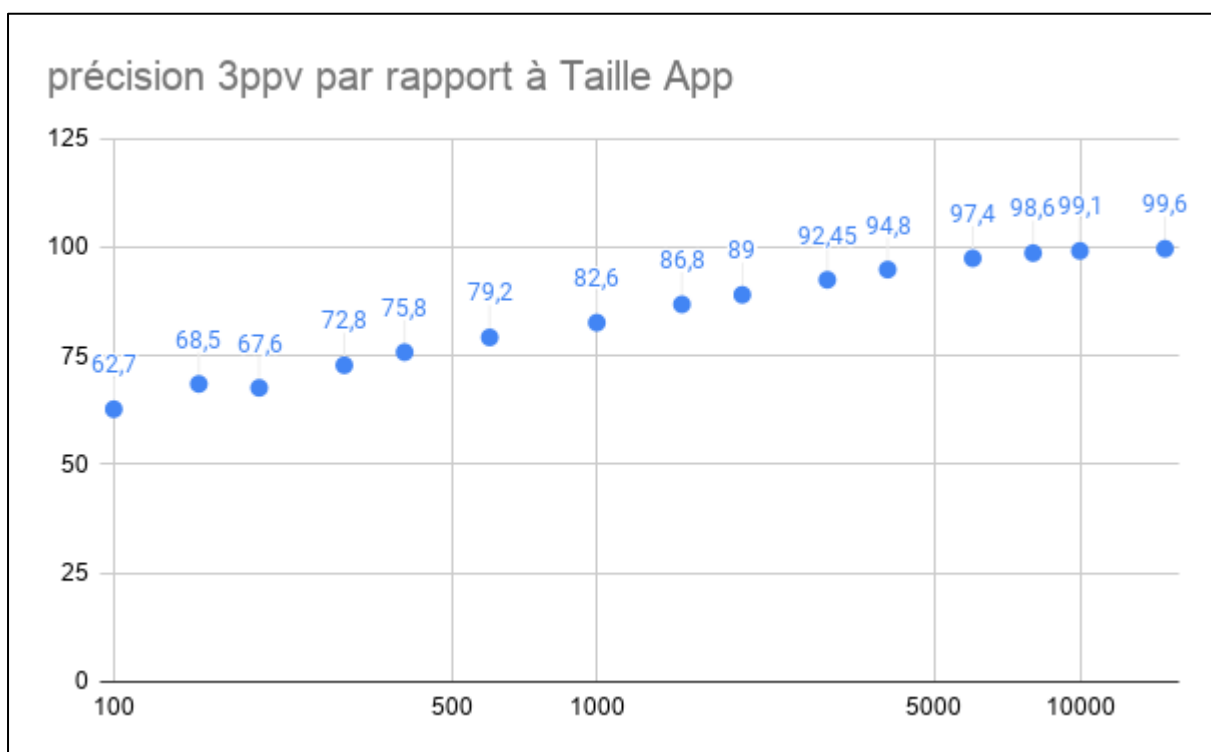
On obtient un pourcentage de bons résultats de 99.1%, ce qu'on constate sur la matrice par le peu de données mal prédites.

On constate donc que plus l'ensemble d'apprentissage est grand plus le pourcentage de bons résultats est élevé.

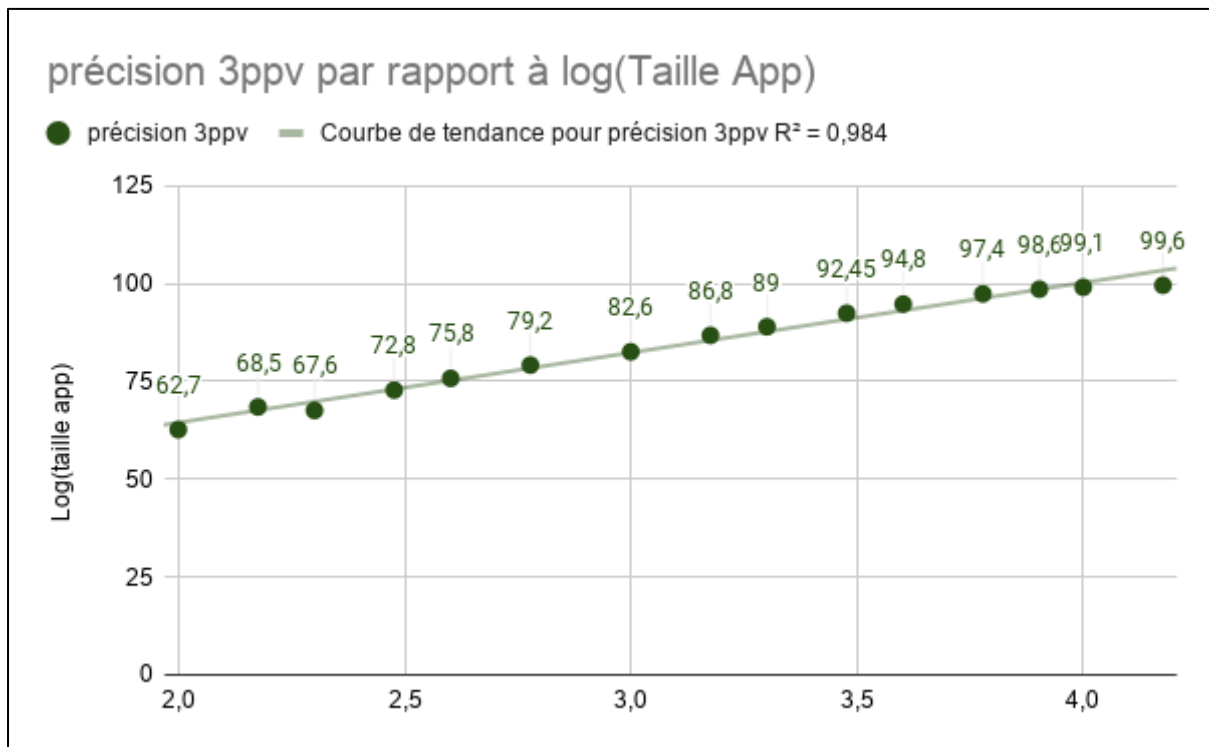
Nous n'avons pas pu tester avec un plus grand ensemble d'apprentissage à cause du temps d'exécution demandé pour chaque test.

Nous devrions cependant théoriquement arriver à un point de surapprentissage en continuant d'augmenter petit à petit la taille de l'ensemble d'apprentissage.

Nous avons fait à partir de nos tests effectués les deux graphiques suivants :



On y observe l'augmentation du pourcentage de bons résultats par rapport à la taille de l'ensemble d'apprentissage.



Ce deuxième graphique montre la même augmentation du pourcentage de bons résultats mais cette fois-ci par rapport au logarithme de la taille d'apprentissage.

On peut aussi y observer la courbe tendance de coefficient $R^2=0.984$ des résultats et le fait que chaque pourcentage obtenu reste proche de la droite d'équation et donc que les résultats sont plutôt homogènes.

En conclusion

En conclusion on observe qu'en influant sur la taille de l'ensemble d'apprentissage le pourcentage de bons résultats obtenus par la méthode DTW ne fluctue pas beaucoup. Il reste assez similaire et n'avons pas de grosse différence entre les différents tests.

La taille de l'ensemble d'apprentissage n'est donc pas un bon paramètre à influencer pour la méthode DTW.

Cependant par la méthode d'ACP+kNN plus on augmente la taille de l'ensemble d'apprentissage plus on obtient un bon pourcentage, mais ce, théoriquement jusqu'au point de surapprentissage.

La taille de l'ensemble d'apprentissage est donc le paramètre le plus influant sur le pourcentage de bons résultats obtenus par cette méthode.