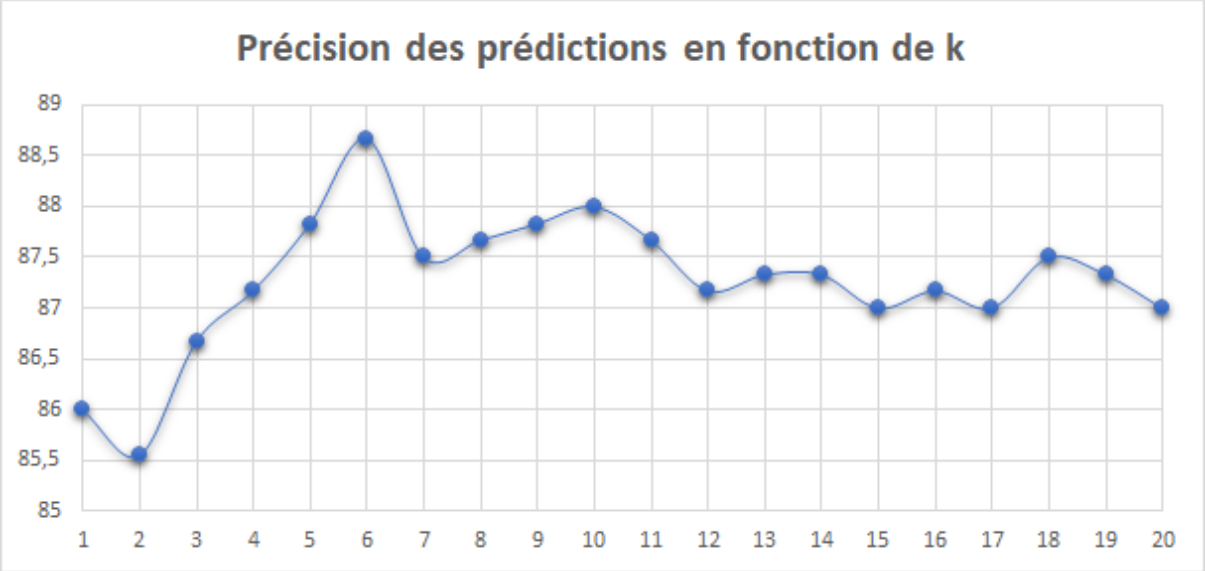
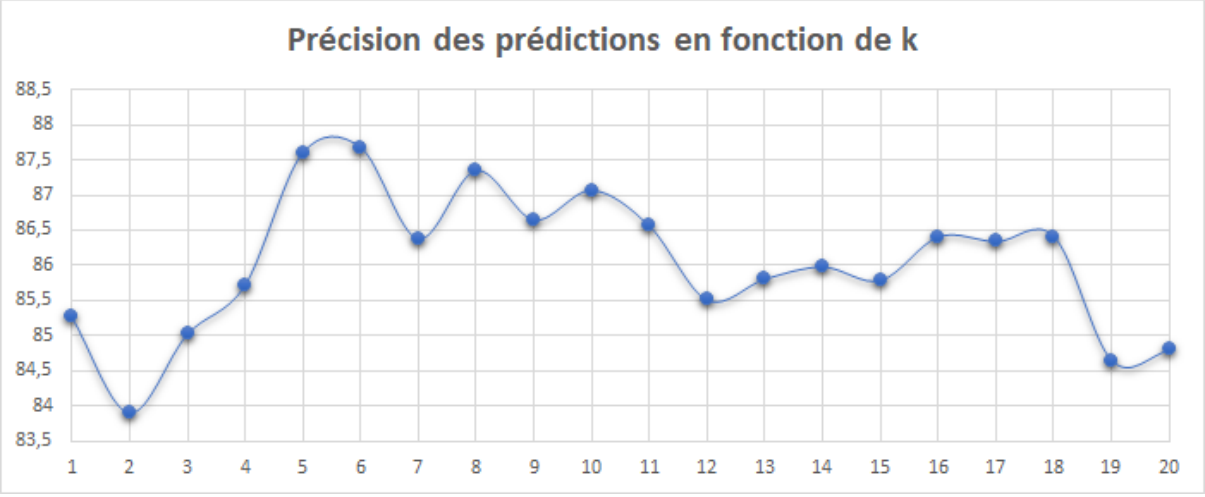
Classification Challenge : KNN

Choix de la valeur de k

Afin de déterminer la meilleure valeur de k, nous avons utilisé le set preTest.csv comme données de test et le set data.csv comme données d'apprentissage. Ensuite, nous avons calculé le pourcentage de valeurs correctement prédites.
Comme le montre le graphe ci-dessous, on trouve un pourcentage de prédictions vraies maximum de 88,67%, lorsque k=6.

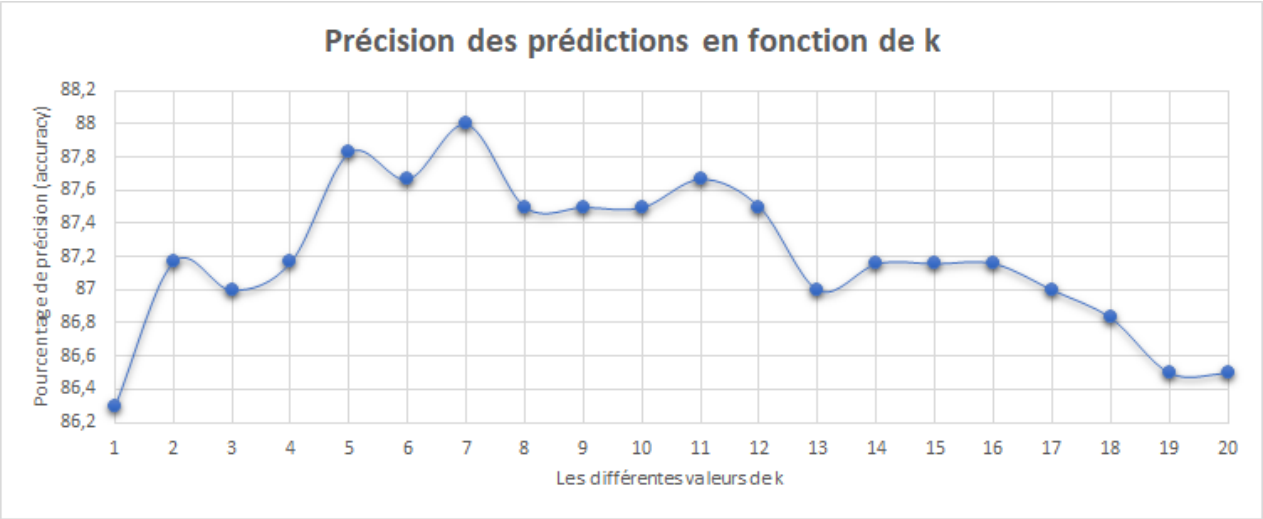


Nous avons également étudié le pourcentage de prédictions vraies en utilisant seulement le set data.csv en divisant les données de manière aléatoire, 80% pour l'apprentissage et 20% pour le test. On observe cependant que les résultats ont une variabilité importante car l'aléatoire prend une part trop grande. Cependant ils coïncident pour la plupart avec le graphique précédent, en nous donnant un pourcentage de prédiction vraie plus élevé lorsque k=6. On remarque aussi que ces résultats oscillent fortement ce qui nuit à leur pertinence (et pourtant les points représentés sont une moyenne de 5 pourcentages).

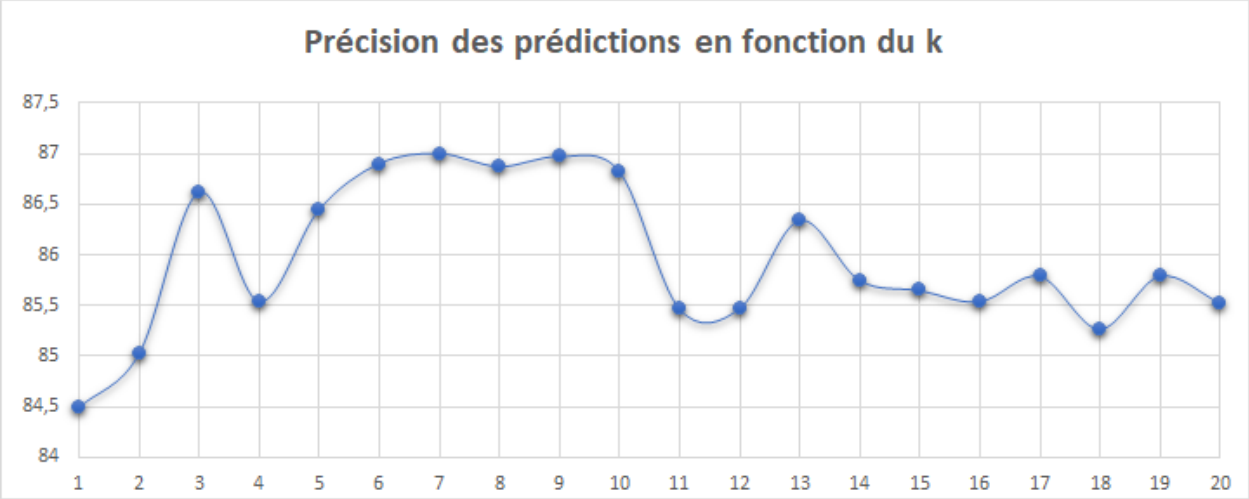


Choix de la distance

Afin de déterminer la proximité des données, nous avons choisi la distance de Manhattan. En effet, nous avons aussi testé notre algorithme en utilisant la distance euclidienne. Les résultats sont moins bons.
Comme vous pouvez le constater sur le graphique ci-dessous, le pourcentage de prédictions vraies le plus élevé est 88% (contre 88,67% lorsque nous utilisons la distance de Manhattan). C'est pourquoi nous avons gardé la distance de Manhattan.



En utilisant la division aléatoire des données de data.csv pour générer les sets d'apprentissage et de test, nous obtenons le graphe ci-dessous. La courbe semble plus régulière que lorsque nous avons utilisé la distance de Manhattan, ce qui pourrait nous laisser penser que l'utilisation de la distance euclidienne nous donne des résultats plus fiables. Cependant, l'aléatoire occupe une place importante dans cet algorithme. C'est pourquoi nous préférons nous fier aux résultats obtenus lorsque nous utilisons 2 sets bien distincts, data.csv pour l'apprentissage et preTest.csv pour les tests.



Classification Challenge

Pour classifier les données de finalTest.csv, nous avons finalement utilisé notre algorithme avec data.csv comme données d'apprentissage, et finalTest.csv comme données test. La distance utilisée pour étudier la proximité des données est la distance de Manhattan, et k prend la valeur de 6.
Avec ces choix, nous espérons donc que notre pourcentage de prédictions vraies s'approche de 88,67% !