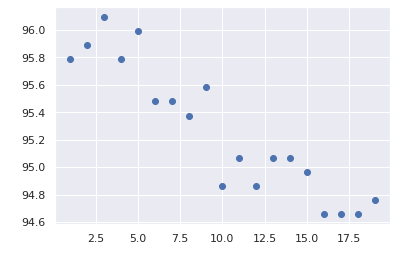
Python For Data Analysis

L’objectif de mon jeu de données était de réussir à identifier à quoi correspondait un bloc d’une page internet. Cette tâche est très importante dans l’analyse de documents. Pour ce faire, j’ai décidé de comparer l’efficacité de plusieurs modèles de Machine Learning. L’objectif est donc de déterminer la classe du bloc étudié. Il y a dix variables et les données se répartissent en cinq classes : texte, ligne horizontale, ligne verticale, graphique, image.

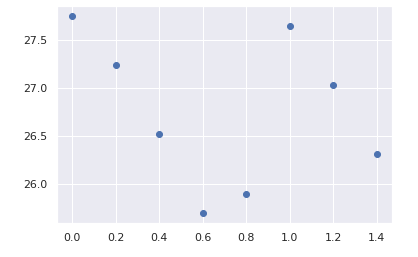
J’ai tout d’abord commencé par réaliser des affichages à l’aide de la librairie seaborn. J’ai ainsi pu observer comment se comportaient les variables en fonction de la classe. La principale observation que j’ai pu faire est que certaines variables semblaient ne pas apporter d’informations. En effet, il n’y avait pas de tendance qui se dégageait en fonction de leur classe :



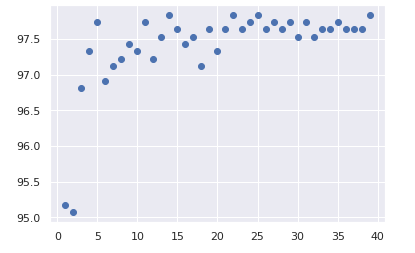
On voit clairement sur ce graphique qu’aucune tendance ne se dégage en fonction de la classe. J’ai donc décidé que, dans un second temps, je devrai tester mes modèles de Machine Learning en ayant enlevé ces variables qui semblent inutiles, voire nuisibles. Cependant, j’ai décidé dans un premier temps de tester des modèles sur les variables brutes, afin de voir s’il y avait une tendance qui se dégageait. J’ai commencé par tester la méthode KNN en faisant varier le nombre de voisin, afin d’avoir une idée de la pertinence de ce modèle en fonction de cet hyperparamètre. J’ai ainsi obtenu les performances suivantes :



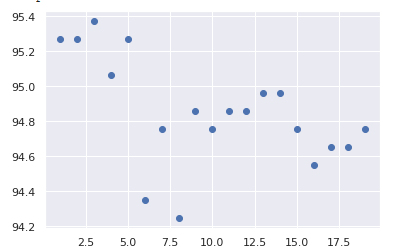
On voit directement que les performances de ce modèle simple sur les données brutes est déjà très bonne. J’ai quand même voulu essayer d’autres modèles, afin de peut-être en trouver un plus efficace. Pour rester dans l’univers du KNN, j’ai essayé la méthode des centroïdes. Cependant, cette fois-ci, la performance a été bien moindre :



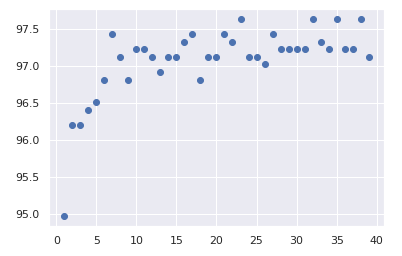
J’ai aussi essayé d’implémenter la méthode des random forest, qui s’est montrée particulièrement efficace.



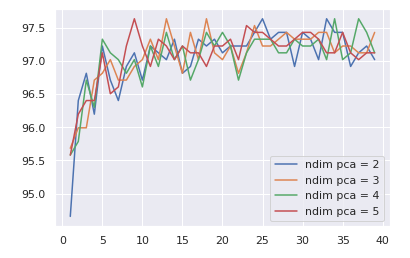
J’ai ensuite essayé d’enlever deux des variables qui semblaient ne pas apporter d’informations. Je pensai alors que cela pourrait être bénéfiques pour les performances de mon modèle. J’ai donc refait le test avec cette fois-ci huit variables au lieu de dix pour chaque donnée. J’ai donc réessayé la méthode de KNN.



On voit, fait surprenant, que pour le premier modèle, la performance du dataset modifié est moindre que celle du dataset brut. Pour vérifier cette tendance, j’ai testé la méthode des random forest sur ce nouveau jeu de données.



On voit une nouvelle fois qu’il s’est avéré peu concluant de supprimer certaines variables du dataset. Enfin, j’ai essayé de faire des réductions de dimensions, puis de tester la méthode des random forest sur les données obtenues. J’obtiens donc à nouveau les performances de la méthode de random forest sur mon jeu de données en fonction du nombre d’arbres qui composent la forêt, mais sur les différents dataset à la dimension réduite.



Une nouvelle fois, les performances ne sont pas meilleures avec le jeu de données modifié qu’avec le jeu de données brut. J’ai donc choisi comme modèle de prendre la random forest avec un nombre d’arbre de 25, et de faire mes prédictions à partir du jeu de données brut.