Travail 1

IFT3700  
Introduction à la science des données

TRAVAIL PRÉSENTÉ À   
Alain Tapp

13 Novembre 2022  
Par : Hugo Carrier 20197563

# Adult

## Les données

Pour obtenir les données de l’ensemble adulte, nous avons télécharger celle-ci sur <https://archive.ics.uci.edu/ml/datasets/adult>. Selon la description, le but de ces données est de terminer si une personne fait plus de 50 000 $ par années. L’ensemble de données adultes fournis différente statistique sur des gens. Nous avons 14 colonnes d’information afin de déterminer la tranche de salaire. Donc 15 colonnes au totales. Certaines colonnes sont descriptive et d’autre contient de l’information continues. Les données contiennent quelques trous, mais ces décrite comme assez propre.

## Notions de similarité proposée.

### Explication de nos choix

Nous avons approché le problème avec le désir de considéré le plus d’information possible sans en inférer. C’est-à-dire que nous voulions inclure l’information des 14 colonnes mais ne pas devoir boucher les trous s’il en a. Après quelque recherche, nous somme tomber sur la méthode Jaccard que nous avons trouvé intéressante. Or, afin de la rendre utilisable avec nos données et en même temps, proposé une version originale de cette similarité, nous avons dû modifier l’approche.

### Explication de la méthodologie

La méthode de Jaccard peut être facilement utiliser lorsque les données sont binaires ou sont des ensembles. Les nôtres ne le sont pas, alors nous devions penser à une idée intéressante pour résoudre ce problème. Initialement, nous avions pensé transformer les colonnes descriptives en binaire. Par exemple, la colonne sexe serait devenu deux colonnes : est homme et est femme. L’étape suivante aurait été de simplement comparer les colonnes descriptives et de divisé la somme des valeur identique par le nombre de colonnes. Après réflexion et quelques essaies, on s’est rendu que nous avions rendu le problème plus que complexe que nécessaire et que les colonnes non modifiées contenant beaucoup de valeurs descriptives fausseraient nos résultats pas leur transformation en de nombreuses sous-colonnes. Au lieu de cela, nous implémenté notre première originalité à l’approche. Nous avons simplement comparé les colonnes si elles sont identiques ou non et nous avons divisé par le nombre total de colonnes.

Cette façon considère que chacune des colonnes apportent la même valeur à notre similarité. Cependant, nous n’avons pas encore traiter les colonnes continues. Afin de les considérer, nous avons pensé à une autre idée originale pour résoudre le problème. Nous avons simplement considéré la distance absolue et rapporter cette valeur en pourcentage.

Pour y arriver, nous avons trouvé le maximum et le minimum de chacune des colonnes. Par la suite nous calculons la distance absolue entre nos points puis nous divisons par la distance des extremums de cette colonne. Ce résultat donne l’inverse de similarité. Si deux chiffres sont identiques, la valeur obtenue sera 0. Pour contrer ce problème nous avons simplement fait 1 moins la valeur obtenue. Ainsi nous avons une valeur qui se retrouve entre 0 et 1, ce qui nous permet de considérer l’information de chacune des colonnes à part égale.

Au final, le résultat total sera entre 0 et 1 dû à la division du nombre de colonnes. Si deux valeurs sont identiques, nous obtiendrons 1, plus elles seront différentes, plus leur similarité tendra vers 0. Avec cette similarité, nous avons par la suite pue poursuivre l’analyse en l’utilisant avec les diverses méthodologies demander dans le devoir.

## Présentation des résultats

## Discussion de l’approche

Notre approche est loin d’être parfait. Premièrement, en python, elle prend du temps à être calculer. Elle pourrait probablement être optimiser dans des langages plus performants, mais pour python ceci est un gros problème. Un autre problème c’est qu’elle traite chacun des colonnes égales. On peut voir ça comme une qualité car ça apporte une simplicité mais c’est aussi un défaut. Il y a probablement certaines colonnes qui apporte plus d’information que d’autre et notre modèle ignore cette possibilité. Il serait aussi probablement possible d’augmenter la qualité si on augmentait la complexité des ressemblances. Dans notre cas, les valeurs descriptives sont égales ou non. Or, certaines valeurs descriptives ont des ressemblances entre elles. Par exemple le niveau l’éducation ont un ordre. La différence entre une maitrise et un bac n’est pas la même qu’entre une maitrise et une quatrième année,

La plupart de ces problèmes sont causé par un choix que nous avons fait. Nous avons désiré un modèle simple et facilement réplicable. Notre méthodologie peut être facilement répliqué et même utiliser sur d’autres types de données. En science de données, c’est un gros avantage car il en devient facile de vérifier les conclusions que nous avons apporté.

Un autre avantage c’est que nous n’avons pas nécessairement besoin de corrigé les données. Le modèle fonctionne correctement lorsque certaine information est manquante. Les informations manquantes sont vue comme des différence mais étant donnée que le modèle traite les colonnes a part égal, les résultats sont moins influencé par ces erreurs.