Cédric Evrard
22/02/2020

ANALYSE DE DONNEES : ÉVALUATION 1

CARACTERISATION DES DONNEES

Le dataset présente le Human Development Index (1990-2015) de 188 pays. L'Human Development Index est une mesure qui se base sur différents critères : la longueur de la vie ainsi que la santé, les connaissances et le standard de vie et est compris entre 0 et 1 (1 étant la valeur la plus élevée).

Grace à un Scatter Plot, on peut remarquer plusieurs tendances dans celui-ci :

- Plus la durée de vie augmente, plus l'HDI est haut ;
- Plus la moyenne d'année à l'école est haute, plus l'HDI est haut ;
- La population totale n'a pas d'influence sur l'HDI ;
- Lorsque le taux de fertilité (2000-2007) est bas, l'HDI est haut ;
- Plus le taux de mortalité infantile est bas, plus l'HDI est haut ;
- La présence ou non du HIV n'a que peu d'influence sur l'HDI ;
- Le taux de chômage n'a pas beaucoup d'influence sur l'HDI ;
- Plus le coefficient d'inégalité humaine est bas, plus l'HDI est haut ;

CONSTRUCTION DES MODELES

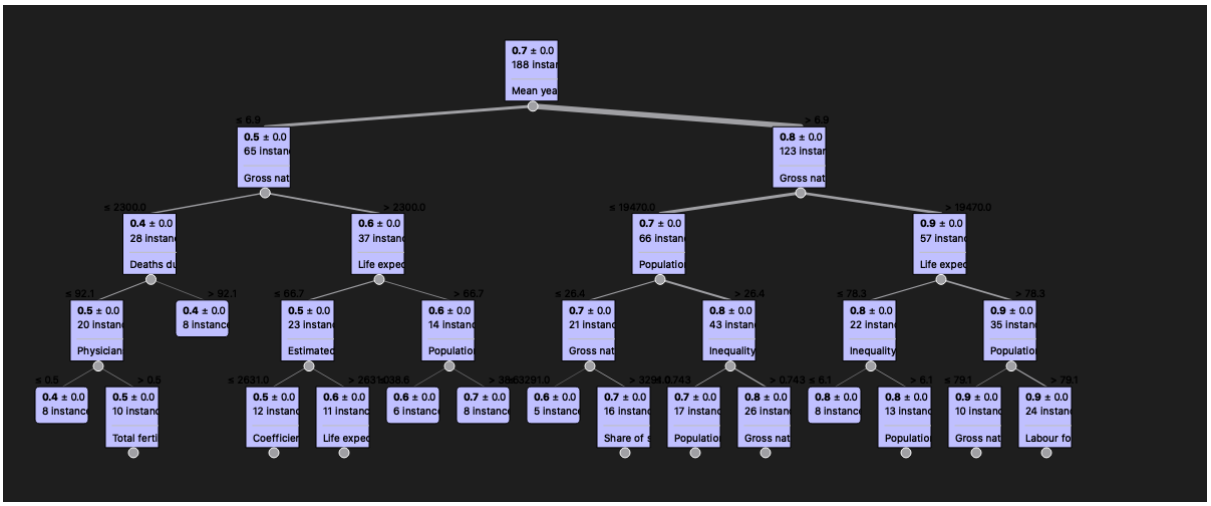
Le premier modèle utilisera un arbre de décision dont le nombre minium d'instance par feuille sera de 5, les ensembles de moins de 5 ne seront pas divisé et la profondeur maximale sera de 6.

Le deuxième modèle utilisera la régression via un kNN. Celui-ci se basera sur l'utilisation de 5 voisins.

Caractériser les prédictions

CARACTERISATION DE L'ARBRE

Voici l'arbre de décision donnée par *orange*.

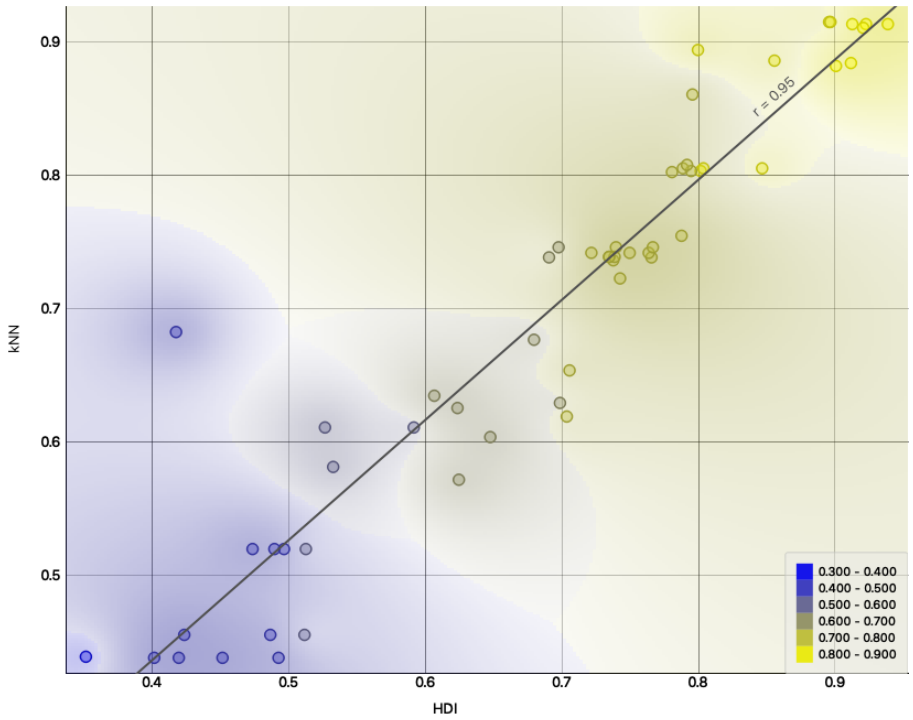


On peut remarquer qu'il utilise des caractéristiques importantes pour le modèle HDI comme le nombre d'années d'étude moyen, l'espérance de vie, l'âge médian de la population ou encore les revenus de la population.

Les choix de caractéristiques pour créer l'arbre me semble bon pour pouvoir prédire l'HDI d'un pays.

CARACTERISTIQUES DU KNN

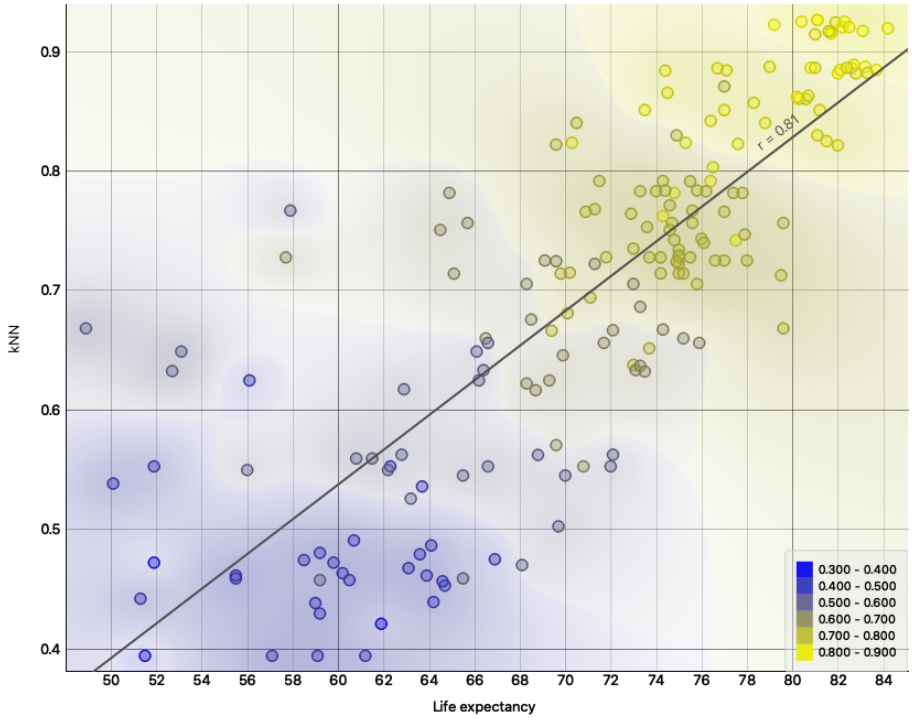
Voici le *scatter plot* du kNN pour l'HDI.



On peut remarquer que les différents points restent assez proche de la droite lorsque l'HDO est compris entre 0.7 et 1 même si certains éléments s'en écartent un peu.

Lorsque l'HDI est inférieur à 0.7, les valeurs s'écartent plus de la droite du kNN.

Pour d'autres caractéristiques, dont la durée de vie, le kNN n'est plus aussi précis. On peut remarquer sur le graphe si dessous une plus grande dispersion des points.



Cela signifie qu'une seule caractéristique (dans ce cas, l'espérance de vie) ne peut pas être prise à part pour évaluer l'HDI avec le kNN.

UTILISATION DES MODELES

UTILISATION DE L'ARBRE DE DECISION

L'arbre de décision peut être utilisé pour pouvoir prédire l'HDI d'un pays suivant les caractéristiques que l'arbre utilise. Il sera alors possible d'avoir une approximation de l'HDI du pays, ou du moins, un ordre de grandeur de celui-ci.

Par exemple, en prenant le cas de la Namibie, en suivant l'arbre avec les caractéristiques du pays, nous arrivons à un HDI de ± 0.6 pour un HDI réelle de 0.640.

UTILISATION DU KNN

L'interprétation des prédictions du kNN n'est pas facile pour les prédictions. Sur l'HDI, les éléments sont assez proches de la droite mais sur les caractéristiques séparées, il y a de grande dispersion des valeurs.