Projet Prédiction Wellbeing à Shanghai

I) Préparation des données

Pour prédire le wellbeing on va utiliser les inputs suivantes :

- Restaurant
- Bicycle_park
- Bus
- Railway_station_entrance_exit
- Convenience Store
- Scenic_spot
- Sport
- Taxi
- Green Space
- Mobike

Avant d'utiliser ses entrées en machine learning on apporte quelques modifications.
En effet, on transforme les longitudes et latitudes en point géométrique. Puis on regarde à quel polygone appartient chaque point. Ce travail nous permet d'avoir par polygone le nombre de restaurants, de vélo, de taxi, etc.

Notre output: Fichier 'TARGET-communities-extract-wellbeing' :

On va utiliser ce tableau pour déterminer notre output Y: Wellbeing of shanghainese communities. En effet ce fichier correspond à un sondage réalisé auprès d'un échantillon d'habitants de Shanghai et il est donc le mieux adapté pour représenter notre output.

Dans un premier temps on apporte quelques modifications au fichier :

- On supprime les endroits où il n'y a pas de valeurs.
- Si on suit notre logique, plus la moyenne sera grande et plus le wellbeing sera important. Or dans notre fichier 'TARGET-communities-extract-wellbeing': smell(0-1 avec 1 worst) et noise(0-1 avec 1 worst). Donc des bruits importants et des mauvaises odeurs vont entraîner une grande moyenne et biaiser le résultat. Pour empêcher cela on change la notation pour smell et noise: 0 pour pire et 1 pour bon.
- Pour chaque ligne, on transforme la longitude et la latitude en point géométrique
- Pour avoir notre output on fait une moyenne de clean, smell et noise avec des coefficients pour chacun : [(clean*a) + (noise*b) + (smell*c)] / (a+b+c)
  On rajoute donc une colonne dans notre fichier où il y aura cette moyenne pour chaque point géométrique

On regarde pour chaque point géométrique à quel polygone il appartient. Ensuite on crée une colonne hapiness qui correspond à la moyenne des moyennes pour chaque polygone. Cette colonne représentera donc le wellbeing par district.
Pour la partie Machine learning je choisis de travailler avec la classification. Je découpe donc mon output en deux catégories : **Unhappy** et **Happy**.

Choix des coefficients a, b et c :

Le choix des coefficients a, b et c va permettre de donner plus d'importance à une entité plutôt qu'une autre. Pour choisir ces derniers il faut avoir une idée bien précise de ce qu'est le wellbeing. Selon moi smell et noise jouent un rôle plus important que clean dans le bien être des habitants. En effet un lieu sale peut être mieux supporté qu'un lieu bruyant et où il y a des mauvaises odeurs. Parmi les trois, celui qui a le plus grand impact dans le wellbeing est Smell car des mauvaises odeurs peuvent être très difficilement acceptées.

Après cette réflexion je décide de choisir l'output suivante :

Y= (10*Clean+20*Noise+50*Smell)/80

II) Partie Machine Learning

Après avoir fixé ma sortie je réalise des tests pour voir comment les modèles se comportent.

- Quand on prend **Output= Smell** on obtient les scores suivants :

	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Tous les inputs	94%	96%	94%	93%
Restaurant	96%	96%	94%	96%
Bicycle_park	94%	91%	94%	94%
Bus	98,50%	98,50%	98,50%	98,50%
Railway_station_entrance_exit	95%	94%	95%	95%
Convenience Store	95%	92%	95%	95%
Scenic_spot	92%	94%	92%	92%
Sport	94%	89%	92%	94%
Taxi	95%	91%	94%	95,50%
Green Space	95,50%	95,50%	95,50%	92,50%
Mobike	95,50%	94%	92,50%	95,50%

On constate que l'ensemble des inputs donnent une excellente prédiction de smell.

- Quand on prend **Output= Noise** on obtient les scores suivants :

	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Tous les inputs	79%	79%	80%	82%
Restaurant	87%	73%	85%	85%
Bicycle_park	83,50%	70%	83,58%	83,58%
Bus	81,00%	81,00%	80,60%	80,60%
Railway_station_entrance_exit	78%	79%	78%	78%
Convenience Store	83%	73%	82%	82%
Scenic_spot	61%	66%	63%	64%
Sport	82%	73%	80%	82%
Taxi	76%	64%	77%	77,00%
Green Space	70,00%	67,00%	70,00%	70,00%
Mobike	79,00%	77%	74,00%	74,00%

On obtient également des bons scores lorsqu'on prend uniquement output=noise. Cependant les inputs prédisent mieux Smell que Noise.

- Quand on prend **Output= Clean** on obtient les scores suivants :

	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Tous les inputs	77%	74%	70%	76%
Restaurant	68%	76%	72%	72%
Bicycle_park	67%	49%	61%	61%
Bus	68,00%	68,00%	68,00%	68,00%
Railway_station_entrance_exit	61%	58%	65%	65%
Convenience Store	71%	62%	68%	68%
Scenic_spot	64%	67%	67%	67%
Sport	68%	61%	68%	68%
Taxi	73%	62%	70%	73,00%
Green Space	70,00%	58,00%	70,00%	70,00%
Mobike	70,00%	67%	68,00%	68,00%

Clean est le moins bien prédit par les inputs.

- Différentes combinaisons de coefficient pour voir comment les modèles se comportent.

Wellbeing	Decision Trees	Random Forest	Adaboost	Gradient Boosting
[(clean*1) + (noise*1) + (smell*1)] / 3	72%	72%	73%	63%
[(clean*2) + (noise*3) + (smell*1)] / 6	**76%**	**75%**	**75%**	**84%**
[(clean*2) + (noise*10) + (smell*1)] / 13	60%	60%	60%	63%
[(clean*2) + (noise*20) + (smell*1)] / 23	58%	66%	58%	54%
[(clean*3) + (noise*2) + (smell*1)] / 6	66%	66%	61%	70%
[(clean*30) + (noise*2) + (smell*1)] / 33	66%	63%	64%	70%
[(clean*2) + (noise*1) + (smell*3)] / 6	**67%**	**72%**	**78%**	**72%**
[(clean*1) + (noise*2) + (smell*3)] / 6	63%	66%	67%	72%
[(clean*2) + (noise*1) + (smell*10)] / 13	64%	64%	67%	69%
[(clean*1) + (noise*2) + (smell*10)] / 13	67%	67%	67%	64%
[(clean*1) + (noise*20) + (smell*20)] / 41	**72%**	**70%**	**76%**	**72%**
[(clean*1) + (noise*20) + (smell*40)] / 61	**76%**	**78%**	**76%**	**78%**
[(clean*1) + (noise*20) + (smell*60)] / 81	**82%**	**75%**	**85%**	**78%**
[(clean*1) + (noise*20) + (smell*100)] / 121	**73%**	**60%**	**76%**	**70%**
[(clean*20) + (noise*1) + (smell*20)] / 41	**63%**	**70%**	**73%**	**72%**
[(clean*1) + (noise*3) + (smell*2)] / 6	**69%**	**70%**	**66%**	**66%**
[(clean*3) + (noise*1) + (smell*2)] / 6	**76%**	**73%**	**78%**	**76%**
[(clean*10) + (noise*1) + (smell*1)] / 12	67%	73%	68%	66%
[(clean*1) + (noise*10) + (smell*1)] / 12	63%	64%	54%	72%
[(clean*1) + (noise*1) + (smell*10)] / 12	**76%**	**79%**	**84%**	**79%**
[(clean*1) + (noise*1) + (smell*20)] / 22	**73%**	**75%**	**72%**	**75%**
[(clean*1) + (noise*1) + (smell*100)] / 102	**78%**	**70%**	**67%**	**76%**
[(clean*1) + (noise*10) + (smell*10)] / 21	72%	66%	67%	67%

Le choix des coefficients va influer la qualité de notre prédiction. On remarque que lorsqu'on fixe les coefficients de clean et noise et qu'on augmente celui de smell on obtient une meilleur prédiction. On voit également que lorsqu'on augmente les coefficients de noise et smell en même temps au détriment de clean on obtient des bonnes prédictions.

- On regarde l'influence de chaque inputs

[(clean*2)+(noise*3)+(smell*1)] / 6	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Restaurant	67%	61%	66%	66%
Bicycle_park	67%	57%	67%	69%
Bus	**76%**	**72%**	**76%**	**76%**
Railway_station_entrance_exit	63%	61%	61%	60%
Convenience Store	69%	67%	63%	69%
Scenic_spot	54%	58%	58%	63%
Sport	**72%**	**69%**	**72%**	**72%**
Taxi	**72%**	**63%**	**72%**	**72%**
Green Space	67%	61%	67%	66%
Mobike	66%	69%	63%	66%
[(clean*3)+(noise*2)+(smell*1)] / 6	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Restaurant	**79%**	**76%**	**79%**	**79%**
Bicycle_park	69%	58%	64%	64%
Bus	**75%**	**73%**	**75%**	**75%**
Railway_station_entrance_exit	60%	60%	55%	60%
Convenience Store	**78%**	**66%**	**76%**	**76%**
Scenic_spot	64%	55%	67%	64%
Sport	**70%**	**63%**	**70%**	**70%**
Taxi	**78%**	**73%**	**78%**	**78%**
Green Space	**72%**	**66%**	**69%**	**72%**
Mobike	**76%**	**73%**	**72%**	**73%**
[(clean*1)+(noise*20)+(smell*20)] / 41	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Restaurant	**76%**	**73%**	**79%**	**79%**
Bicycle_park	67%	64%	61%	64%
Bus	**73%**	**73%**	**73%**	**73%**
Railway_station_entrance_exit	64%	63%	63%	63%
Convenience Store	**73%**	**76%**	**72%**	**72%**
Scenic_spot	67%	63%	61%	69%
Sport	**70%**	**63%**	**70%**	**70%**
Taxi	**72%**	**70%**	**73%**	**73%**
Green Space	**76%**	**69%**	**73%**	**75%**
Mobike	69%	75%	67%	67%
[(clean*1)+(noise*20)+(smell*40)] / 61	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Restaurant	66%	64%	67%	67%
Bicycle_park	61%	60%	57%	58%
Bus	**75%**	**75%**	**75%**	**75%**
Railway_station_entrance_exit	54%	55%	54%	53%
Convenience Store	64%	66%	70%	69%
Scenic_spot	66%	61%	67%	67%
Sport	69%	61%	66%	69%
Taxi	**75%**	**68%**	**75%**	**75%**
Green Space	58%	61%	57%	57%
Mobike	**82%**	**63%**	**84%**	**76%**

Quand on prédit avec une seule entrée à la fois on remarque que certaines entrées ont plus d'impact sur le wellbeing que d'autres. En effet, Restaurant, Bus, Sport, Taxi, Green Space, Convenience Store et Mobike à eux seuls donnent une bonne prédiction de notre wellbeing. Par contre des inputs comme Railway_station_entrance_exit vont avoir moins d'influence sur la sortie.

Résultats obtenus avec notre output : **Y= (10*Clean+20*Noise+50*Smell)/80**

[(clean*10) + (noise*20) + (smell*50)] / 80	Decision Trees	Random Forest	Adaboost	Gradient Boosting
Tous les inputs	82%	86%	76%	82%
Restaurant	76%	70%	70%	76%
Bicycle_park	60%	61%	60%	63%
Bus	73%	73%	73%	72%
Railway_station_entrance_exit	60%	59%	58%	56%
Convenience Store	68%	67%	67%	68%
Scenic_spot	64%	62%	62%	68%
Sport	67%	`60%	67%	67%
Taxi	71%	65%	65%	65%
Green Space	64%	62%	64%	64%
Mobike	72%	65%	67%	70%

En prenant tous les inputs et avec le modèle Gradient Boosting on obtient un score de **86%.**

Axe d'amélioration :
Pour améliorer ce score on peut rajouter d'autre inputs en lien direct avec le wellbeing et en supprimer certain comme par exemple Bicycle Park. En effet, ce dernier n'a pas trop d'importance puisqu'à Shanghai lorsqu'on loue un vélo (Mobike par exemple) on peut le redéposer où on le souhaite.
De plus avant de commencer à prédire il faut bien nettoyer et préparer les données. Je ne pense pas avoir été optimale durant cette phase notamment avec la dataset Mobike. Ainsi pour avoir le nombre de vélo par zone, j'ai compté le nombre de vélos qui quittent cette dernière. Cependant les vélos vont se retrouver dans une autre zone par la suite.
Le plus dur a été de définir le output puisqu'il fallait déterminer le niveau d'importance des trois critères (Smell, Noise et Clean) dans le wellbeing. Ce choix reste subjectif puisque chacun a sa propre définition du wellbeing. J'ai fait le choix de donner plus d'importance à Smell et Noise. Si j'ai obtenu ce score c'est parce que les inputs prédisaient très bien Smell et Noise.