Choix des données

Différents nombres de features

État d'Iowa: 89 features
King County:18 features
État de Californie: 8 features

Choix des algorithmes

- Regression Lineaire
- Random Forest
- Bagging Regressor avec arbres de décision

Pré-traitement des données

Iowa dataset

- Encodage des features catégoriques : avoir des chiffres et pas du texte
- Remplacer valeurs *NA* par valeur moyenne de chaque attribut de l'ensemble d'entraînement

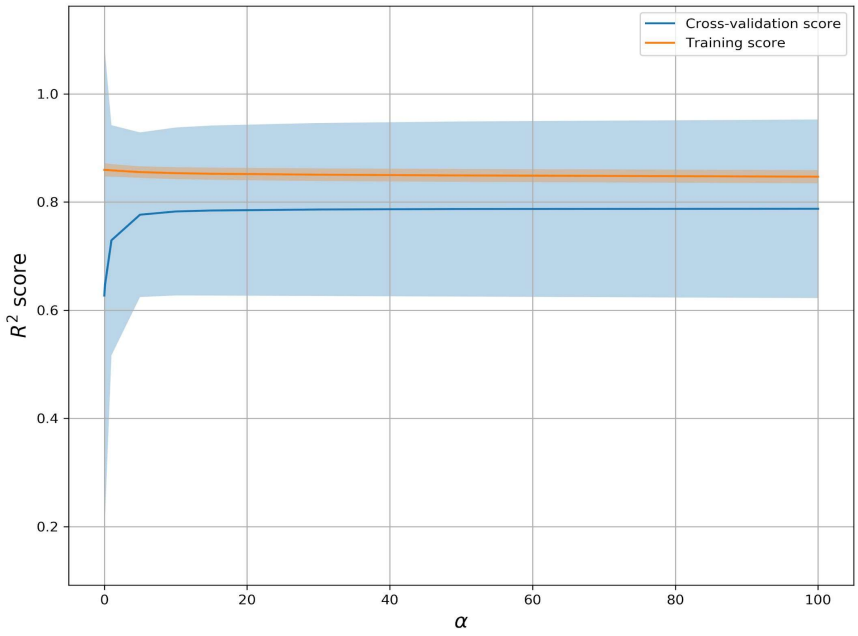
Pré-traitement des données

California dataset

- Effacer la colonne "ocean proximity" afin de n'avoir que des données numériques
- Remplacer les valeurs manquantes par la médiane de la colonne considérée
- # chambres par maison= # chambres dans le district / # maisons dans le district
- # salles de séjour par maison = # salles de séjour dans le district / # maisons dans le district

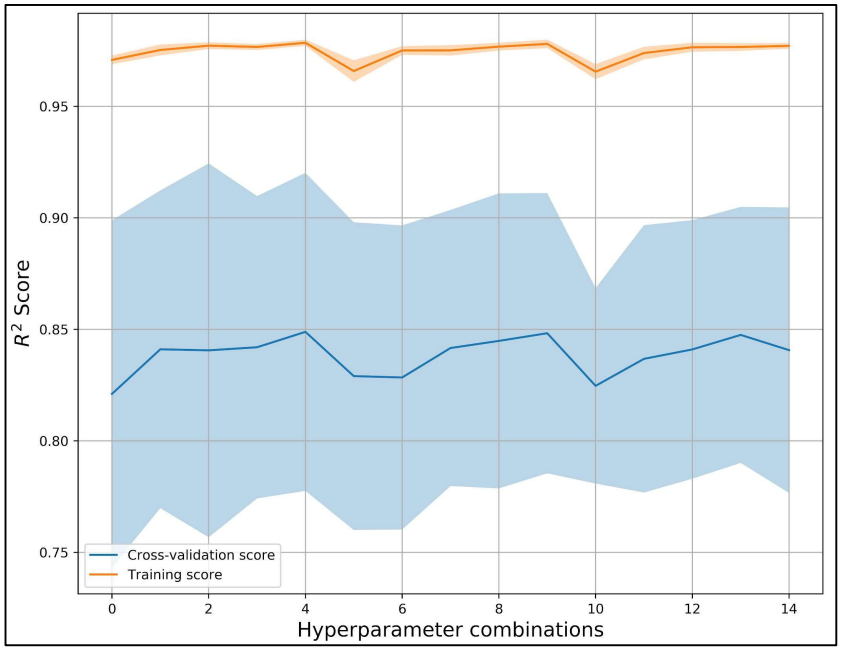
Résultats Iowa (1)

Ridge regression: courbes de validation



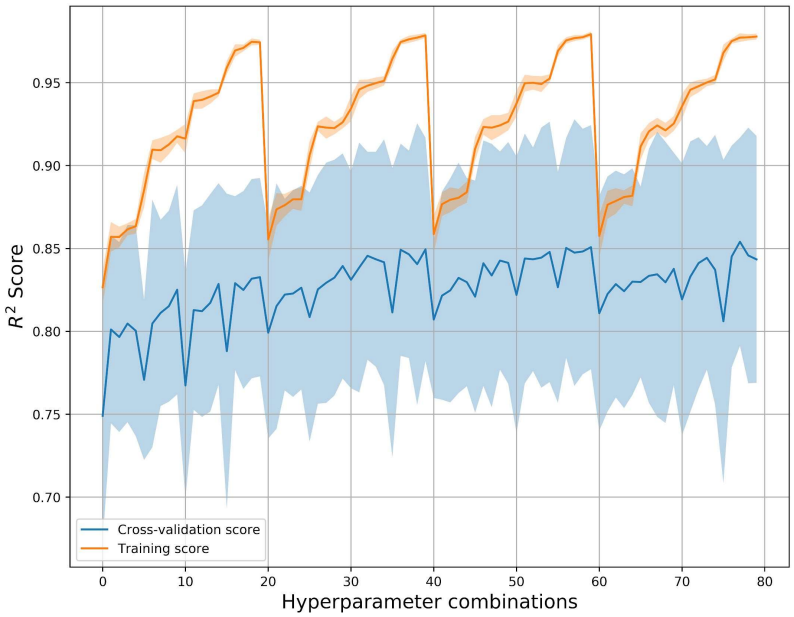
Résultats Iowa (2)

Random Forest : courbes de validation



Résultats Iowa (3)

Bagging regressor: courbes de validation



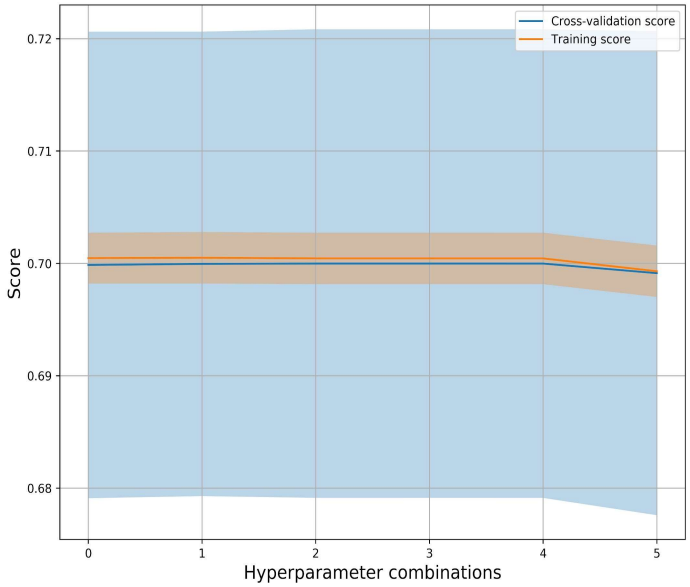
Résultats Iowa (4)

Iowa : conclusion des résultats

Modèle\score R^2	**Training**	**Testing**
Ridge	0.844	0.853
Random forest	0.979	0.896
Bagging	0.981	0.883

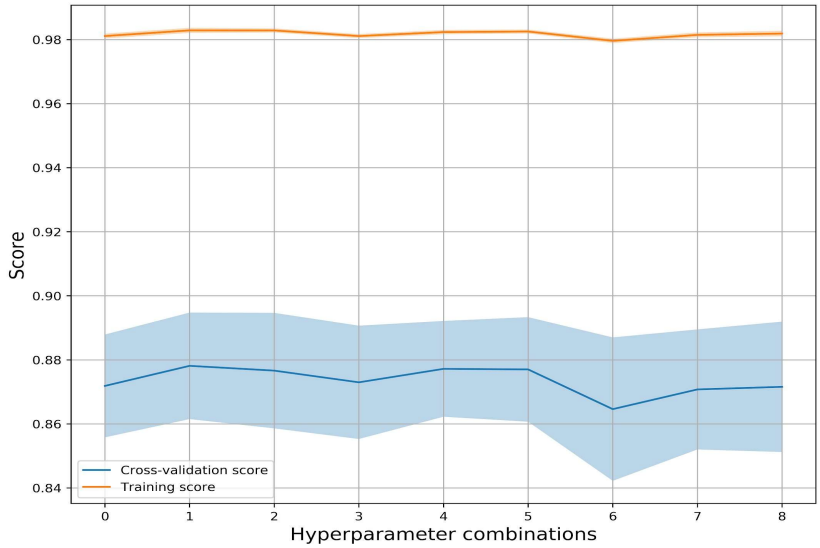
Résultats King County (1)

Ridge regression: courbes de validation



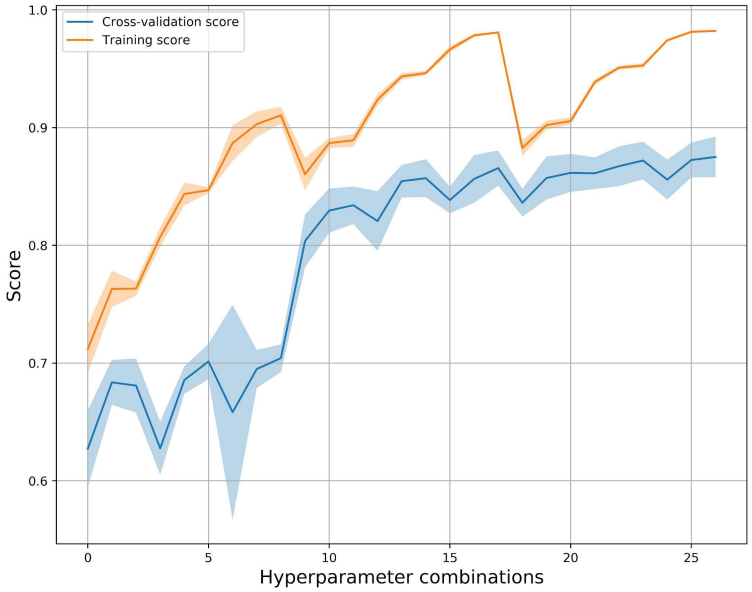
Résultats King County (2)

Random Forest : courbes de validation



Résultats King County (3)

Bagging regressor: courbes de validation



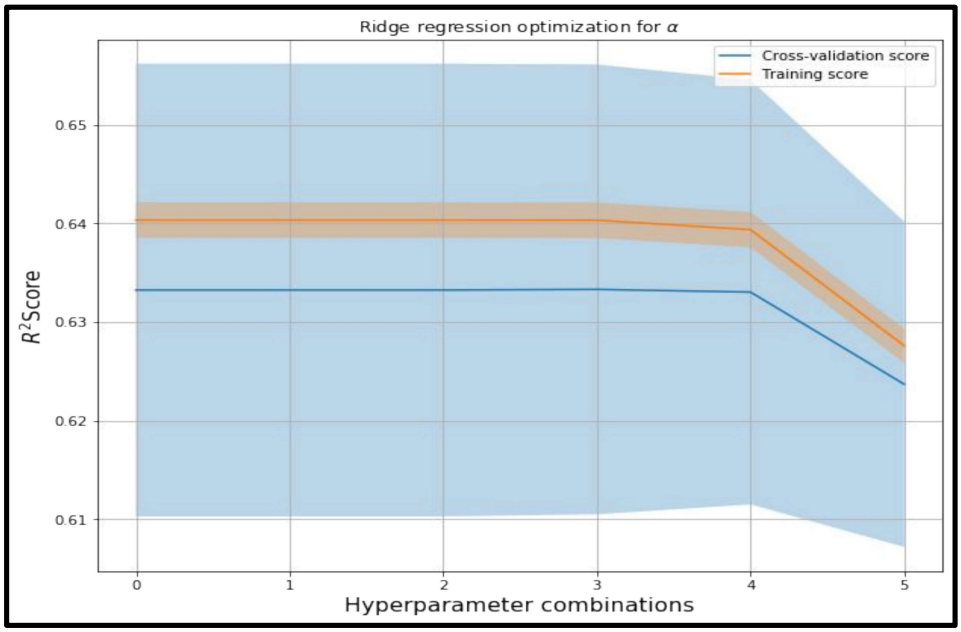
Résultats King County (4)

King County: conclusion des résultats

Modèle\score R^2	**Training**	**Testing**
Ridge	0.697	0.724
Random forest	0.982	0.880
Bagging	0.982	0.882

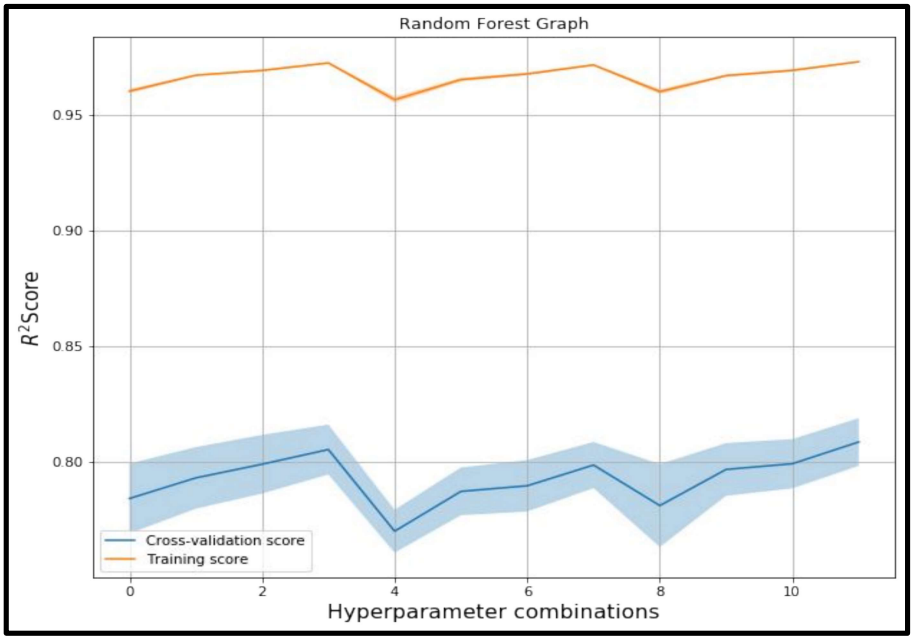
Résultats Californie (1)

Ridge regression: courbe validation



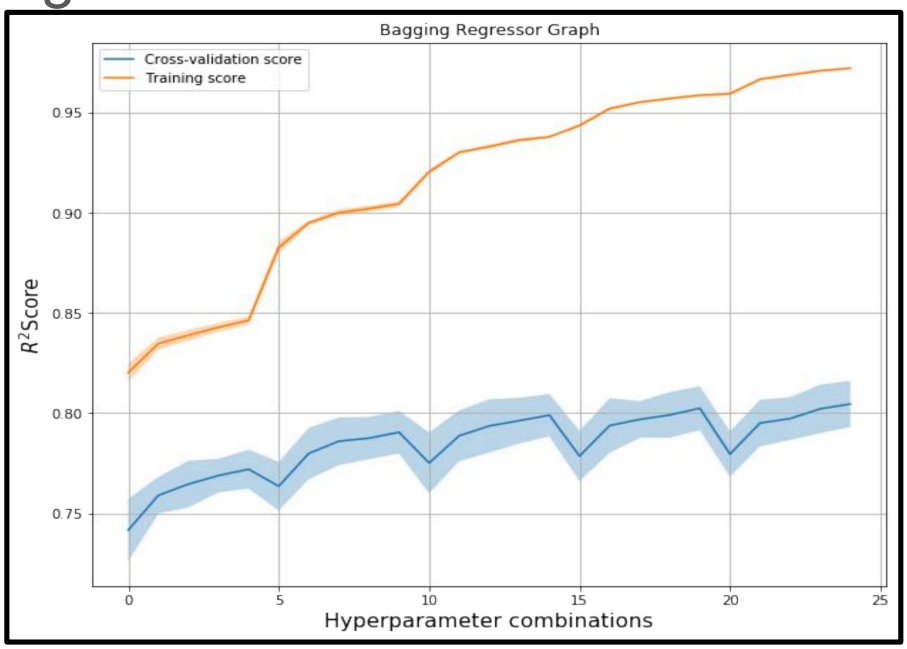
Résultats Californie (2)

Random forest : courbe validation



Résultats Californie (3)

Bagging regressor: courbe validation



Résultats Californie (4)

Californie: conclusion résultats

Modèle\score R^2	**Training**	**Testing**
Ridge	0.640	0.619
Random forest	0.961	0.805
Bagging	0.878	0.802

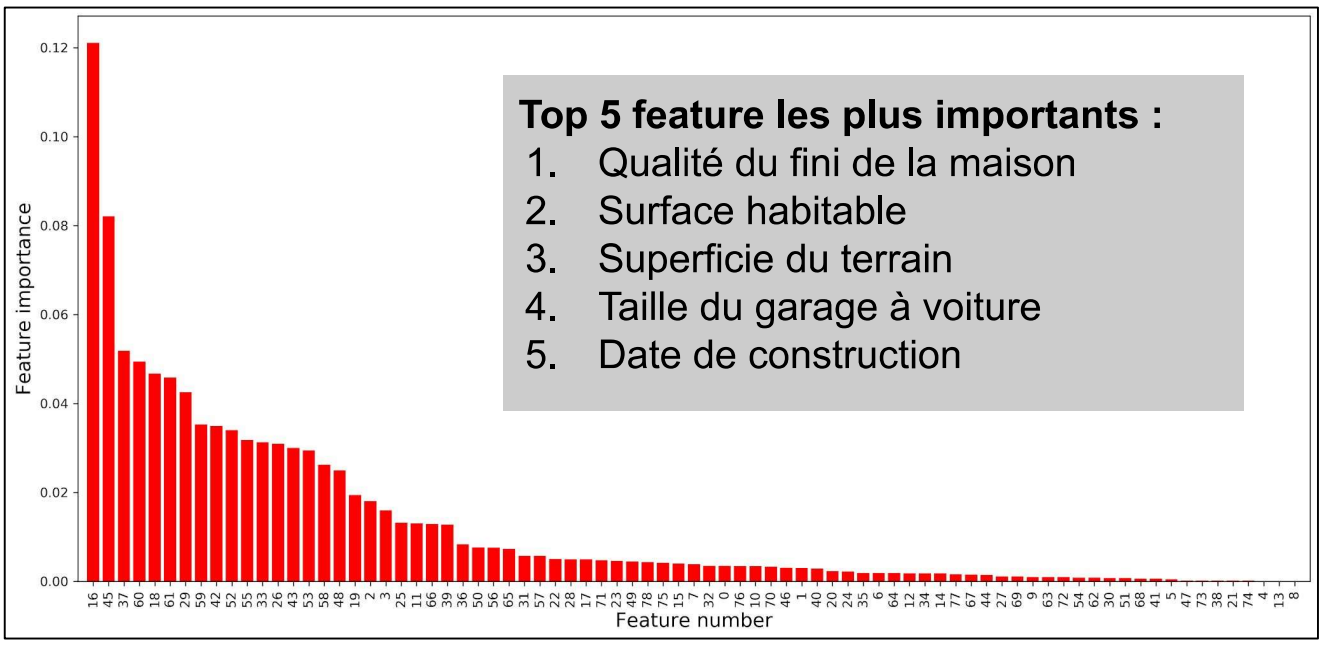
Comparaison des datasets

Quel dataset a eu la 1ere performance?

- Ridge regression : Iowa
- Bagging regression:Iowa
- Random Forest regression:King County et Iowa

Features les plus importants

- Iowa

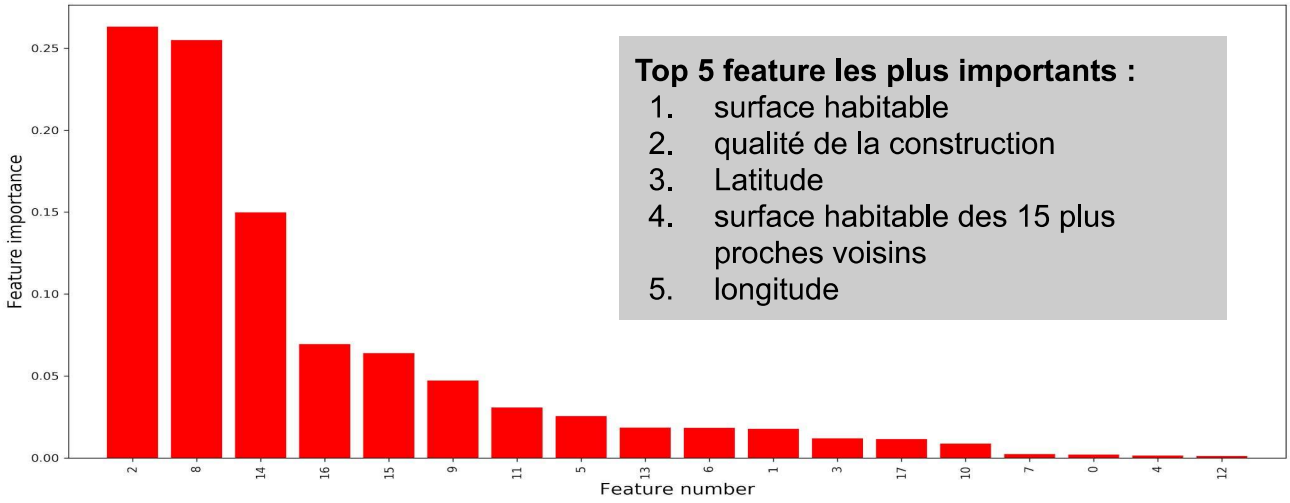


Top 5 feature les plus importants :

1. Qualité du fini de la maison
2. Surface habitable
3. Superficie du terrain
4. Taille du garage à voiture
5. Date de construction

Features les plus importants

- King County

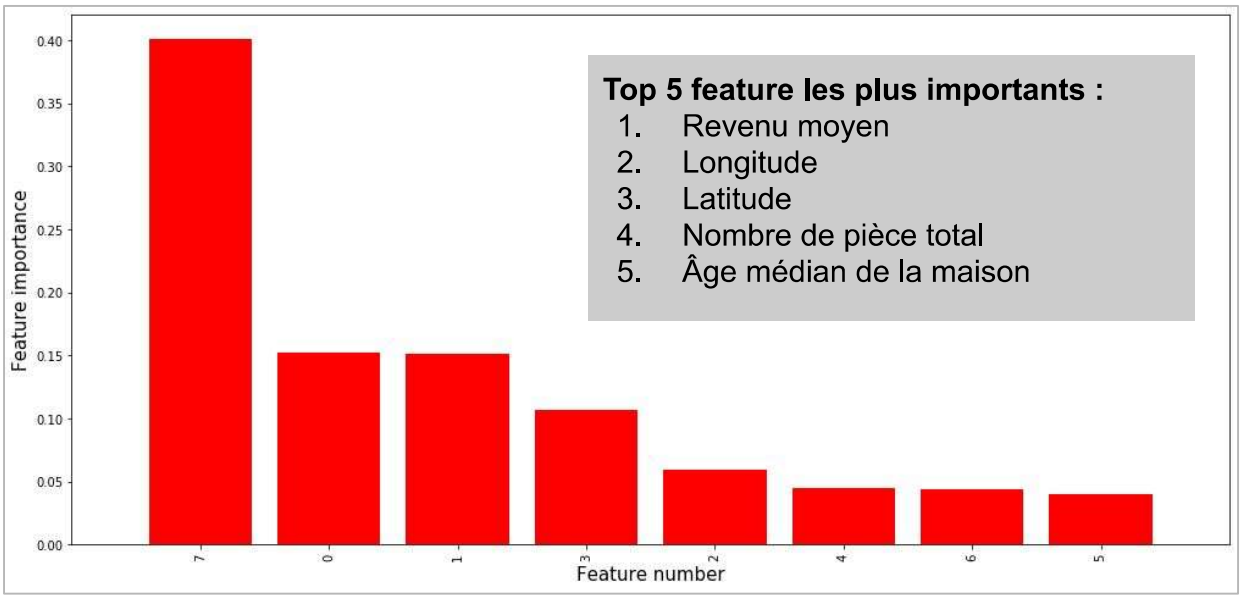


Top 5 feature les plus importants :

1. surface habitable
2. qualité de la construction
3. Latitude
4. surface habitable des 15 plus proches voisins
5. longitude

Features les plus importants

- Californie



Top 5 feature les plus importants :

1. Revenu moyen
2. Longitude
3. Latitude
4. Nombre de pièce total
5. Âge médian de la maison

Conclusion des features

- Ce qu'on en retire? Suggestions pour quelqu'un qui estime des maisons.

Les features qui reviennent dans le top des 5 des meilleurs features

-qualité de la construction
-latitude
-longitude
-surface habitable
-ancienneté de la maison

Remerciements

Université de Montréal

Mila