1-BOUCLE PBL

Tout afficher Tout réduire

PRÉREQUIS

Objectifs d'apprentissage

Algorithmique

Environnement python

SUJET

Énoncé

EXERCICE : ON SÉPARE LES APPARTEMENTS

Contexte et Objectif :

L'agence immobilière Loft's Craft, basée en Californie, est reconnue pour sa qualité de service et son excellent bouche-à-oreille, attirant à la fois des clients particuliers et professionnels. Afin d'optimiser ses propositions de biens immobiliers et de réduire le nombre de visites nécessaires, y compris virtuelles, l'agence envisage de recourir à l'intelligence artificielle.

Solution Envisagée :

L'intelligence artificielle offre de nombreuses possibilités pour améliorer les services de Loft's Craft. Une étude est en cours pour identifier les meilleures applications de l'IA adaptées à leurs besoins. En attendant, un jeu de données issu du recensement de 1990, correspondant à la zone géographique couverte par l'agence, a été récupéré pour être exploité.

Description du Jeu de Données :

Le jeu de données comprend des informations sur les maisons et leurs habitants, regroupées par secteur de recensement, une zone homogène en termes de démographie et d'économie. Voici les principales colonnes :

- **longitude** : Mesure de la longitude centroïde d'un secteur de recensement. Plus la valeur est élevée, plus il se trouve à l'ouest (varie de -124,3 à -114,3).
- **latitude** : Mesure de la latitude centroïde d'un secteur de recensement. Plus la valeur est élevée, plus il se trouve au nord (varie de 32,5 à 42,5).
- **housing_median_age** : Âge médian des logements dans le secteur (varie de 1 à 52 ans).
- **total_rooms** : Nombre total de pièces dans un secteur (varie de 2 à 37,937).
- **total_bedrooms** : Nombre total de chambres dans un secteur (varie de 1 à 6,445, certaines valeurs peuvent être manquantes).
- **population** : Nombre total de résidents dans un secteur (varie de 3 à 35,682).
- **households** : Nombre total de foyers dans un secteur (varie de 1 à 6,082).
- **median_income** : Revenu médian des foyers dans un secteur (varie de 0,5 à 15 dizaines de milliers de dollars).
- **median_house_value** : Valeur médiane des logements dans un secteur (varie de 14,999 à 500,001 dollars).
- **ocean_proximity** : Proximité par rapport à l'océan, avec des valeurs comme INLAND, <1 OCEAN, NEAR OCEAN, NEAR BAY, ISLAND.

Préparation des Données :

Dans cette boucle, nous allons préparer notre jeu de données pour l'analyse et l'apprentissage automatique. Cela inclut l'importation des données, l'identification et la gestion des valeurs manquantes, et la création de visualisations pour une meilleure compréhension des données. Nous suivrons le processus suivant :

1. Importation des données : Tout d'abord, nous allons importer les bibliothèques nécessaires et configurer notre environnement pour une meilleure lisibilité des données et des graphiques.

2. Chargement et exploration des données : Nous commençons par charger les données du fichier CSV extrait d'une archive .tgz. Ensuite, nous explorons la structure du jeu de données pour obtenir un aperçu des colonnes, des types de données et des valeurs manquantes.

3. Compréhension des données : dans cette phase nous produirons un résumé du jeu de données pour comprendre la distribution des variables et identifier les valeurs manquantes. Les méthodes `info()` et `describe()` de pandas sont utilisées pour fournir des résumés concis et des statistiques descriptives.

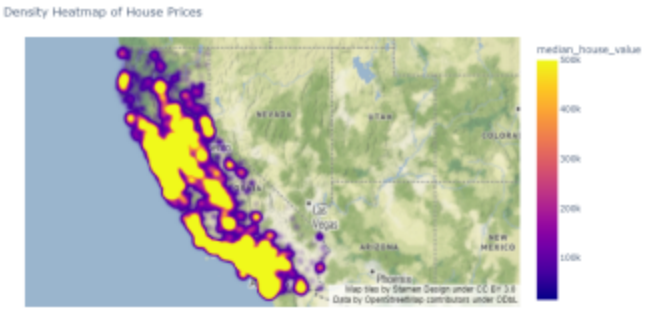
4. Gestion des valeurs manquantes : Pour simplifier, nous remplissons les valeurs manquantes dans les colonnes numériques avec la valeur médiane de chaque colonne. Pour les colonnes catégorielles, nous utilisons la valeur la plus fréquente (mode) pour l'imputation.

5. Visualisation des données : Nous utilisons diverses visualisations pour analyser la distribution des données et les relations entre les variables. Cela inclut des histogrammes pour les caractéristiques numériques, des diagrammes en barres pour les caractéristiques catégorielles, des cartes de chaleur pour les corrélations et des cartes géographiques pour la visualisation spatiale des prix des maisons.

Cette préparation des données nous permet de transformer notre jeu de données brut en un format propre et prêt pour les analyses ultérieures et les modèles d'apprentissage automatique.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.00	20640.00	20640.00	20640.00	20433.00	20640.00	20640.00	20640.00	20640.00
mean	-119.57	35.63	28.64	2635.76	537.87	1425.48	499.54	3.87	206855.82
std	2.00	2.14	12.59	2181.62	421.39	1132.46	382.33	1.90	115395.62
min	-124.35	32.54	1.00	2.00	1.00	3.00	1.00	0.50	14999.00
25%	-121.80	33.93	18.00	1447.75	296.00	787.00	280.00	2.56	119600.00
50%	-118.49	34.26	29.00	2127.00	435.00	1166.00	409.00	3.53	179700.00
75%	-118.01	37.71	37.00	3148.00	647.00	1725.00	605.00	4.74	264725.00
max	-114.31	41.95	52.00	39320.00	6445.00	35682.00	6082.00	15.00	500001.00

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```



Les résultats semblent prometteurs, les points rouges semblent correspondre à des zones où l'immobilier est plus cher : au Nord avec les environs de San Francisco et la côte sud avec notamment Los Angeles...

On lui a conseillé d'utiliser l'algorithme de K-means qui devrait donner de meilleurs résultats en préparant au mieux son jeu de données de manière à exploiter le potentiel de l'ensemble des données fournies.

Cet algorithme serait capable de séparer les données de la base en k parties (ici k=2) parties distinctes sans aucun paramètres préalables ! Uniquement par une analyse intelligente des données entre elles ! C'est génial !

Ressources pour les étudiants

Définition et éthique

CNIL - Comment permettre à l'Homme de garder la main ?

Préparation des données

Conférence DataWrangling : Conférence Mohamed-Amin BENATIA - Déc. 2020

BasicsStatisticsForDataScience [pdf] : Support de la conférence

Conseils d'IBM

Standardisation et Normalisation

Algorithmes de classification automatique (non supervisés)

Apprentissage supervisé et non-supervisé

Classification Ascendante Hiérarchique et en plus poussé ce lien (pour approfondir)

Présentation K-means

Possibilités de K-means

Techniques de l'ingénieur : IA

Machine learning et données

IA et éthique

Objectivité de l'IA

IA et transformation numérique

IA et données en entreprise

IA et avenir

IA pour les ventes

Ouvrages de référence

Python Data Science Handbook, Jake VanderPlas : https://jakevdp.github.io/PythonDataScienceHandbook/ (version complète disponible en ligne)

Pour aller plus loin

Artificial Intelligence with Pythons de Joshi, Prateek : https://univ.scholarvox.com/catalog/book/docid/88842652?searchterm=artificial%20intelligence