Machine Learning : Exploration et prétraitement des données

-TP1-

Objectifs :

À la suite d'une première étape de définition de la problématique et de génération des idées et des opportunités à exploiter, vient une étape cruciale dans une démarche machine learning: **la création et la préparation d'un jeu de données exploitable.**

- Elle consiste à définir et collecter les données d'intérêt, et à réaliser des premières analyses permettant d'identifier les ***corrections*** à faire (traitement des valeurs manquantes, des valeurs aberrantes, ...), et les ***agrégations*** possibles entre les différentes sources de données.
- Lors de cette étape, des ***premières analyses statistiques*** sont réalisées pour permettre au Data scientist d'affiner sa compréhension de la problématique et de mieux cerner les pistes à exploiter pour y répondre. Ces analyses exploratoires consistent notamment à étudier les distributions des différentes variables, les dépendances entre celle-ci, etc.
- Ces premières analyses vont également permettre d'identifier, si nécessaires, de nouvelles variables à créer (à partir des données disponibles ou à compléter par des données externes) ***=> Feature engineering.***

Dans ce TP, on va supposer que la problématique s'articule autour de l'analyse de la distribution des logements pour comprendre les tendances du marché et identifier les possibilités d'investissement.

1. Chargement des librairies et des données fournies

Pour réaliser ce TP, on aura besoin d'importer les librairies suivantes :

- La bibliothèque **numpy** (pour l'analyse numérique et statistique)
- La bibliothèque **pandas** (pour le traitement des DataFrame)
- La bibliothèque **matplotlib** (pour la visualisation)
- La bibliothèque **seaborn** (pour la visualisation)

Utiliser le code suivant pour charger et lire le contenue du fichier *Housing_dataset.csv* dans un dataframe (Housing).

```
import pandas as pd
import numpy as np
import matplotlib as plt
import seaborn as sns
----------------------------------------------------------------------
housing=pd.read_csv(Path+' / Houseing_dataset.csv')
housing.head(10)
```

2. Premières explorations des données

a. Afficher la dimensionnalité de votre jeu de données en utilisant la méthode ***shape***.
b. Donner une description détaillée des variables de votre jeu de données suite à l'exécution de la commande **housing.info().**
c. En faisant appel aux méthodes **isnull()** et **sum()**, vérifier s'il existe des valeurs manquantes. Si c'est le cas proposer une méthode pour le traitement de cette problématique
d. Vérifier s'il existe des données redondantes dans votre jeu de données en utilisant les méthodes **duplicated()** et **sum()**. Si c'est le cas procéder avec le suivant :
-Afficher les échantillons qui souffrent de cette redondance
-Calculer la fréquence de redondance de chaque échantillon. Que peut être la cause de cette redondance ? Et que peut-on conclure ?
- Corriger cette redondance et effectuer un test pour vérifier sa correction.
e. Supprimer les colonnes *'BROKERTITLE', 'MAIN_ADDRESS', 'FORMATTED_ADDRESS'.* A votre avis, pourquoi on a procédé par cette suppression ?
f. Classer les données par prix, type et adresse.

3. Analyse des variables

Pour la suite, nous réaliserons des analyses statistiques sur les différentes variables du jeu de données afin de raffiner la compréhension du contexte et des objectifs à atteindre.

/ Dans un premier temps, ***chaque variable sera étudiée de façon unitaire*** : cette analyse univariée permet de mieux comprendre les données, d'identifier des valeurs potentiellement aberrantes, et de tirer des premières conclusions sur la suite des travaux.
/ ***Des analyses multivariées*** permettront ensuite d'identifier de nouveaux enseignements métier (liens entre les variables), et de définir de nouvelles variables à créer pour enrichir le jeu de données.

a. Analyses univariées

- Calculer des statistiques descriptives (moyenne, écart type, ...) des variables ***Price***, ***Beds***, ***Bath***, et ***Propertysqrt,*** en utilisant la méthode ***describe().*** Peut-on générer des premières idées et connaissances (répartition des données, valeurs aberrantes, etc.) à ce stade de l'analyse?
- Ecrire une fonction qui permet d'afficher les statistiques, l'histogramme et le box plot d'une variable numérique étudiée de votre choix.
Hints : La fonction peut être définit comme suit `def stats(df, variable, bins, stat="count")`
Utiliser **seaborn** pour l'histogramme et le boxplot

- Afficher le bar plot de la variable catégorielle 'Type'. Quelle est la répartition des types de logement dans les observations de notre jeu de données.
- Résumer l'ensemble des enseignements que vous avez pu tirer à partir de cette première étape d'exploration des données. Quelles sont à votre avis les corrections à appliquer ?

b. Analyses multivariées

➢ ***Pour les variables numériques :***

- Calculer et visualiser (à l'aide d'une ***heatmap***) la matrice de corrélation entre les différentes variables de votre jeu de donnée. Qu'est ce vous constater ?
- Utiliser la méthode ***pairplot*** pour tracer les graphes de dispersion (scatter plot) entre chaque pair de variables numériques.

➢ ***Pour les variables catégorielles :***

- Utiliser la méthode de vectorisation par *One hot encoding* pour vectoriser les variables catégorielles et afficher un échantillon du résultat. Que remarquez-vous ?
- Utiliser la méthode de vectorisation par *One hot encoding* pour vectoriser uniquement la variable '*LOCALITY',* puis calculer et visualiser (à l'aide d'une heatmap) la matrice de corrélation entre les différentes variables de votre jeu de donnée. Qu'est ce vous constater ?
- Utiliser la méthode ***pairplot*** pour tracer les graphes de dispersion (scatter plot) entre chaque pair de variables

c. L'ingénierie des variables :

Pour appliquer la réduction de dimensionnalité sur notre jeu de données, décrit sur des variables numériques et catégorielles, on peut faire appel à des méthodes d'ACP adaptées aux données mixtes, comme **MCA (Multiple Correspondence Analysis)** ou **FAMD (Factor Analysis of Mixed Data)**, disponibles dans des bibliothèques comme ***prince.***

- Exécuter le code suivant et commenter le résultat. Comment peut-on améliorer les résultats de cette analyse ?

```
!pip install prince
import prince

famd = prince.FAMD(n_components=2)
df_famd = famd.fit_transform(housing)
plt.scatter(df_famd[0], df_famd[1])
plt.title("Projection FAMD")
plt.xlabel("Composante 1")
plt.ylabel("Composante 2")
plt.show()
```

4. Sauvegarde du jeu de données

Utiliser la méthode **to_csv** pour enregistrer votre jeu de données à la suite des modifications réalisées (sans la partie concernant le feature engineering).

```
df.to_csv("../data_housing_prepared.csv", index=False)
```