TD9 et TD10

Travail noté à présenter le 28 juin

Analyse des données DVF

Le travail à réaliser, en trinôme, du 22 juin au 28 juin. Chaque trinôme présentera ses réalisations pendant 10 min à la dernière séance soit le 28 juin.
Le dataset utilisé représente les données DVF c'est-à-dire **Demandes de valeurs foncières.**
Il s'agit des transactions immobilières intervenues au cours des cinq dernières années sur le territoire métropolitain et les DOM-TOM, à l'exception de l'Alsace, de la Moselle et de Mayotte.
Les données contenues sont issues des actes notariés et des informations cadastrales.

Il s'agit des données OpenData:
https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/

Les équipes

Ce travail est à réaliser en équipe. Chaque équipe est constituée d'un **trinôme** issu du même groupe TD.
Les binômes sont acceptés uniquement par la contrainte du nombre d'étudiants dans un groupe de TD. Le travail en solo n'est pas accepté ni les équipes de 4 étudiants.
Nous ne pouvons pas faire d' exceptions, ni pour des équipes de 4 ni pour plus de binômes ou des solos.

Le contenu et les livrables

Il s'agit d'appliquer certains algorithmes de ML sur les DVF, pour y arriver il faut passer par une **phase de préparation de données** suivie d'une **analyse exploratoire des données**.

La préparation des données comporte généralement les tâches suivantes :

- Fusion des ensembles et/ou enregistrements de données
- Sélection d'un sous-ensemble de données
- Calcul de nouveaux attributs
- Tri des données en vue de la modélisation
- Suppression ou remplacement des blancs ou des valeurs manquantes
- Fractionnement en sous-ensembles d'apprentissage et de test

Vous pouvez toujours proposer d'autres traitements mais il faudra expliquer à chaque fois. Pour bien maîtriser cette phase, je vous invite à consulter le lien suivant : Data Preparation with pandas

Analyse exploratoire des données :

Pour bien maîtriser cette phase, je vous invite à consulter le lien suivant : Python for Data Science: Implementing Exploratory Data Analysis (EDA) and K-Means Clustering
Cette phase consiste à :
1. Identifier la variable cible qui doit être prédite.
2. Effectuer une analyse exploratoire des données (EDA, Exploratory Data Analysis) par une variété de graphiques et de statistiques en suivant les étapes suivantes :

- Faire un aperçu des variables en examinant leur type, leur distribution, leur plage de valeurs et leur signification.
- Examiner la relation entre les variables caractéristiques et la variable cible en utilisant des techniques graphiques telles que des histogrammes, des diagrammes en boîte et des graphiques de dispersion
- Construire une matrice de corrélation entre les variables
- Interpréter les résultats de l'EDA pour identifier les caractéristiques importantes qui influencent le prix du mètre carré et les relations significatives entre les variables

Nous serons attentifs sur:

- Une dizaine d'interprétations et visualisations pertinentes autour d'une année (en particulier 2022) prix moyen du mètre carré d'un département d'une ville, comparaison entre différents types, appart, maison, geomap, différence entre ville, départements, régions,...
- En bonus, quelques interprétations et visualisations pertinentes en comparaison avec au moins une autre année (en particulier avant covid)

La partie model ML :

Vous avez deux parties:
La première partie "**Apprentissage non supervisé**" : modelez les données selon un algorithme ML non supervisé, interprétez les clusters obtenus.

La deuxième partie Apprentissage supervisé : Il s'agit de travailler sur les prix des **appartements** (vous pouvez limiter votre travail à une région ou un département)
mise en place d'un modèle de prédiction des prix des appartements dans une région ou département. Une évaluation de votre modèle est indispensable. Vous avez bien évidemment le droit de combiner avec d'autres données externes.

Les livrables:

1- **Un notebook détaillé** (avec son exportation en html, permettant de voir tout le travail fourni sans être contraint à exécuter le notebook.
2- **Une présentation de 10 minutes pendant la séance TD10** décrivant votre travail, votre démarche et vos conclusions.
3- Une petite note à transmettre à votre prof avant la présentation contenant la description de la contribution de chaque membre du trinôme et le **ratio de contribution** (par exemple 33% pour chacun ou 40% pour Nom1, 35% pour Nom2, 25% pour Nom3...)

Notation

La notation est faite par votre prof **pendant la séance TD10**, sur la base de votre présentation et des livrables.
Elle prendra en compte les modules utilisés, le chargement de vos données, les efforts de nettoyage et d'interprétation, le nombre de visualisation, la pertinence des visualisations, les géomap, le merge avec d'autres données externes et bien évidemment vos model ML et évaluation des models.
Nous comptons sur vous pour avoir un travail authentique et ne pas avoir le même code ni les mêmes indicateurs ou visus. Partager certaines idées pourquoi pas, mais pas la démarche, et surtout pas le code. Ne tombez pas dans le piège du plagiat ou plagiat déguisé. jouez le jeu!

Evaluation par les paires

Nous envisageons de mettre en place une évaluation par les pairs. Les modalités sont à suivre.

Deadline et zone de dépôt sur moodle

Le devoir est à rendre au plus tard le 27 juin à 23h55.
A déposer sur moodle (une seule copie par trinôme) (d'autres modalités sont à suivre)
La présentation est à déposer en fin de la séance TD10.

Quelques liens utiles et autres données:

L'Institut national de la statistique et des études économiques (une mine de données si vous voulez combiner avec données démographiques, économique…etc)
https://www.insee.fr
https://www.insee.fr/fr/statistiques
Geojson pour afficher les départements et les régions (avec choropleth de plotly ou folium)
https://france-geojson.gregoiredavid.fr/

Communes de france - Base des codes postaux
https://www.data.gouv.fr/fr/datasets/communes-de-france-base-des-codes-postaux/
Départements de france
https://www.data.gouv.fr/fr/datasets/departements-de-france/
API Découpage Administratif - (API Geo)
https://api.gouv.fr/les-api/api-geo
Qu'est-ce que DVF ?
https://www.groupe-dvf.fr
https://www.groupe-dvf.fr/wp-content/uploads/2019/06/DVF-1-Qu-est-ce-que-DVF_2019-05-17.pdf