جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

Master : SIDI-TA

Module :DATA MINING

Pr: MOHAMMED SABIRI

la prédiction de prix immobiliers

Une Approche Data Mining

Réalisé par : Amal FATHI

2023/2024

Introduction :

La Data Mining, également connue sous le nom de fouille de données, est une discipline du domaine de l'informatique qui consiste à découvrir des motifs, des relations et des informations significatives à partir de grandes quantités de données. Elle utilise des techniques avancées d'analyse statistique, de modélisation prédictive et de machine learning pour extraire des connaissances utiles à partir de jeux de données complexes.

L'objectif principal de ce projet de Data Mining est de développer un modèle de prédiction de prix immobiliers. Cela implique l'exploration et l'analyse approfondies des données immobilières afin de comprendre les relations entre différentes caractéristiques des propriétés (telles que la superficie, le nombre de chambres à coucher, etc.) et leurs prix de vente respectifs.

Les principales étapes de ce projet comprennent la préparation des données, l'identification des caractéristiques pertinentes, la construction d'un modèle de régression linéaire pour prédire les prix immobiliers, et l'application de techniques de clustering pour identifier des groupes homogènes de propriétés partageant des caractéristiques similaires.

I. L’environnement de programmation :

1. Language de programmation :

Python est un langage de programmation interprété, de haut niveau, orienté objet, et dynamique. Il est réputé pour sa simplicité et sa lisibilité, ce qui en fait un choix populaire pour les débutants en programmation. Python prend en charge plusieurs paradigmes de programmation, notamment la programmation impérative, fonctionnelle et orientée objet. Il dispose également d'une vaste bibliothèque standard, ainsi que d'une communauté active qui contribue à de nombreux modules et frameworks.

2. Les outils :

Visual Studio Code (VS Code) :

est un éditeur de code source gratuit et open-source développé par Microsoft. Il est léger, extensible et prend en charge une variété de langages de programmation.

Anaconda est une distribution open-source des langages de programmation Python et R, principalement utilisée pour la science des données, le machine learning, et la programmation scientifique.

II. Le dérouLement d'une étude de data mining

1. Définition du problème :

Le problème de prédiction des prix immobiliers consiste à développer un modèle capable d'estimer le prix d'une propriété en fonction de ses caractéristiques. L'objectif est de créer un outil de prédiction précis qui peut aider les acteurs du marché immobilier à estimer la valeur d'une propriété sans avoir à recourir à une évaluation manuelle approfondie.

- **Entrées :**

Caractéristiques de la propriété : Des variables telles que la superficie de la propriété, le nombre de chambres à coucher, la localisation géographique, les équipements, etc.

- **Sortie :**

Prix de la propriété : La valeur estimée de la propriété en termes monétaires.

- **Objectif :**

Développer un modèle de prédiction robuste et fiable qui utilise des techniques de Data Mining et de machine learning pour apprendre à partir de données historiques et à fournir des estimations de prix précises pour de nouvelles propriétés.

2. Collecte de Données - Base de Données Housing :

La base de données Housing est un ensemble de données fréquemment utilisé dans le domaine de l'apprentissage automatique et de l'analyse prédictive. Elle contient des informations sur les prix immobiliers en fonction de diverses caractéristiques. Pour le projet de prédiction de prix immobiliers, nous avons choisi cette base de données en raison de sa pertinence et de sa disponibilité.

- **Caractéristiques Principales :**

  - ✓ **area :** Superficie de la propriété en mètres carrés.
  - ✓ **bedrooms :** Nombre de chambres à coucher.
  - ✓ **price** : Prix de la propriété en unités monétaires.

La base de donnée Housing :

	A	B	C	D	E
1	price,area,bedrooms,bathrooms,stories,mainroad,guestroom,b				
2	13300000,7420,4,2,3,yes,no,no,no,yes,2,yes,furnished				
3	12250000,8960,4,4,4,yes,no,no,no,yes,3,no,furnished				
4	12250000,9960,3,2,2,yes,no,yes,no,no,2,yes,semi-furnished				
5	12215000,7500,4,2,2,yes,no,yes,no,yes,3,yes,furnished				
6	11410000,7420,4,1,2,yes,yes,yes,no,yes,2,no,furnished				
7	10850000,7500,3,3,1,yes,no,yes,no,yes,2,yes,semi-furnished				
8	10150000,8580,4,3,4,yes,no,no,no,yes,2,yes,semi-furnished				
9	10150000,16200,5,3,2,yes,no,no,no,no,0,no,unfurnished				
10	9870000,8100,4,1,2,yes,yes,yes,no,yes,2,yes,furnished				
11	9800000,5750,3,2,4,yes,yes,no,no,yes,1,yes,unfurnished				
12	9800000,13200,3,1,2,yes,no,yes,no,yes,2,yes,furnished				
13	9681000,6000,4,3,2,yes,yes,yes,yes,no,2,no,semi-furnished				
14	9310000,6550,4,2,2,yes,no,no,no,yes,1,yes,semi-furnished				
15	9240000,3500,4,2,2,yes,no,no,yes,no,2,no,furnished				
16	9240000,7800,3,2,2,yes,no,no,no,no,0,yes,semi-furnished				
17	9100000,6000,4,1,2,yes,no,yes,no,no,2,no,semi-furnished				
18	9100000,6600,4,2,2,yes,yes,yes,no,yes,1,yes,unfurnished				

3. Nettoyage et transformation des données

Le processus de nettoyage et de transformation des données est essentiel pour garantir la qualité des informations utilisées dans notre modèle de prédiction de prix immobiliers. Parmi les étapes clés, la gestion des valeurs manquantes revêt une importance particulière.

Dans notre approche, nous avons opté pour une méthode de suppression des lignes contenant des valeurs manquantes. Cette méthode est appliquée aux lignes où au moins une des caractéristiques nécessaires pour la prédiction (par exemple, la superficie, le nombre de chambres à coucher ou le prix) comporte une valeur manquante.

La méthode utilisée :

```
def preprocess_data(self):
    # Supprimer les lignes avec des valeurs manquantes
    self.df = self.df.dropna()
```

4. Appliquer les techniques de fouille de données

Le projet de prédiction de prix immobiliers met en œuvre deux techniques de fouille de données clés pour analyser et extraire des informations significatives à partir des données immobilières.

Régression Linéaire :

Objectif : La régression linéaire est utilisée pour établir une relation linéaire entre les caractéristiques (surface et nombre de chambres) et le prix immobilier. Le modèle apprend à partir des données existantes et peut ensuite prédire les prix pour de nouvelles données.
Application : À travers la bibliothèque scikit-learn en Python, le modèle de régression linéaire est entraîné sur les données immobilières, et ses prédictions sont comparées aux valeurs réelles. Le modèle fournit une vision quantitative de l'influence de chaque caractéristique sur le prix.

```
# Entraîner le modèle de régression linéaire
model = LinearRegression()
model.fit(X, y)

# Prédire les prix pour les données existantes
self.df['predicted_price'] = model.predict(X)
```

K-Means Clustering :

Objectif : Le clustering est appliqué pour regrouper les données immobilières en clusters distincts en fonction de leurs caractéristiques. Dans ce cas, le K-Means clustering est utilisé pour identifier des groupes homogènes.
Application : Le modèle K-Means est employé pour regrouper les données en clusters en fonction des scores de sentiment prédits précédemment. Ces clusters représentent différents groupes d'opinions concernant les propriétés immobilières.

```python
def apply_clustering(self):
    # Sélectionner les caractéristiques pour le clustering
    X_cluster = self.df[['area', 'bedrooms']]

    # Appliquer l'algorithme de K-Means
    kmeans = KMeans(n_clusters=3, random_state=42)
    self.df['cluster'] = kmeans.fit_predict(X_cluster)

def predict_prices(self):
    if hasattr(self, 'df'):
        # Appliquer le prétraitement des données
        self.preprocess_data()

        # Appliquer le clustering
        self.apply_clustering()

        X = self.df[['area', 'bedrooms']]
        y = self.df['price']
```

5. Interprétation du modèle et établissement des conclusions :

L'interprétation du modèle de prédiction de prix immobiliers, basé sur une régression linéaire, peut se faire en examinant les coefficients des variables explicatives et en évaluant la qualité globale du modèle. Voici une interprétation générale des aspects clés :

Coefficients des Variables :

Surface (Area) : Un coefficient positif indique que, toutes choses égales par ailleurs, une augmentation de la surface est associée à une augmentation du prix immobilier prédit.

Nombre de chambres (Bedrooms) : Un coefficient positif indique que, toutes choses égales par ailleurs, un plus grand nombre de chambres est associé à une augmentation du prix immobilier prédit.

Intercept (Constante) :

Le terme intercept dans le modèle représente le prix de base, indépendamment des variables explicatives.

Qualité Globale du Modèle :

R^2 (Coefficient de détermination) : Il mesure la proportion de la variance totale du prix immobilier qui est expliquée par le modèle. Une valeur proche de 1 indique un bon ajustement.

Graphique de Prédictions :

Le graphique de dispersion des prédictions par rapport aux vraies valeurs permet d'évaluer visuellement la performance du modèle. Une dispersion étroite autour de la ligne de régression suggère une bonne adéquation.

6. Gérer la connaissance découverte

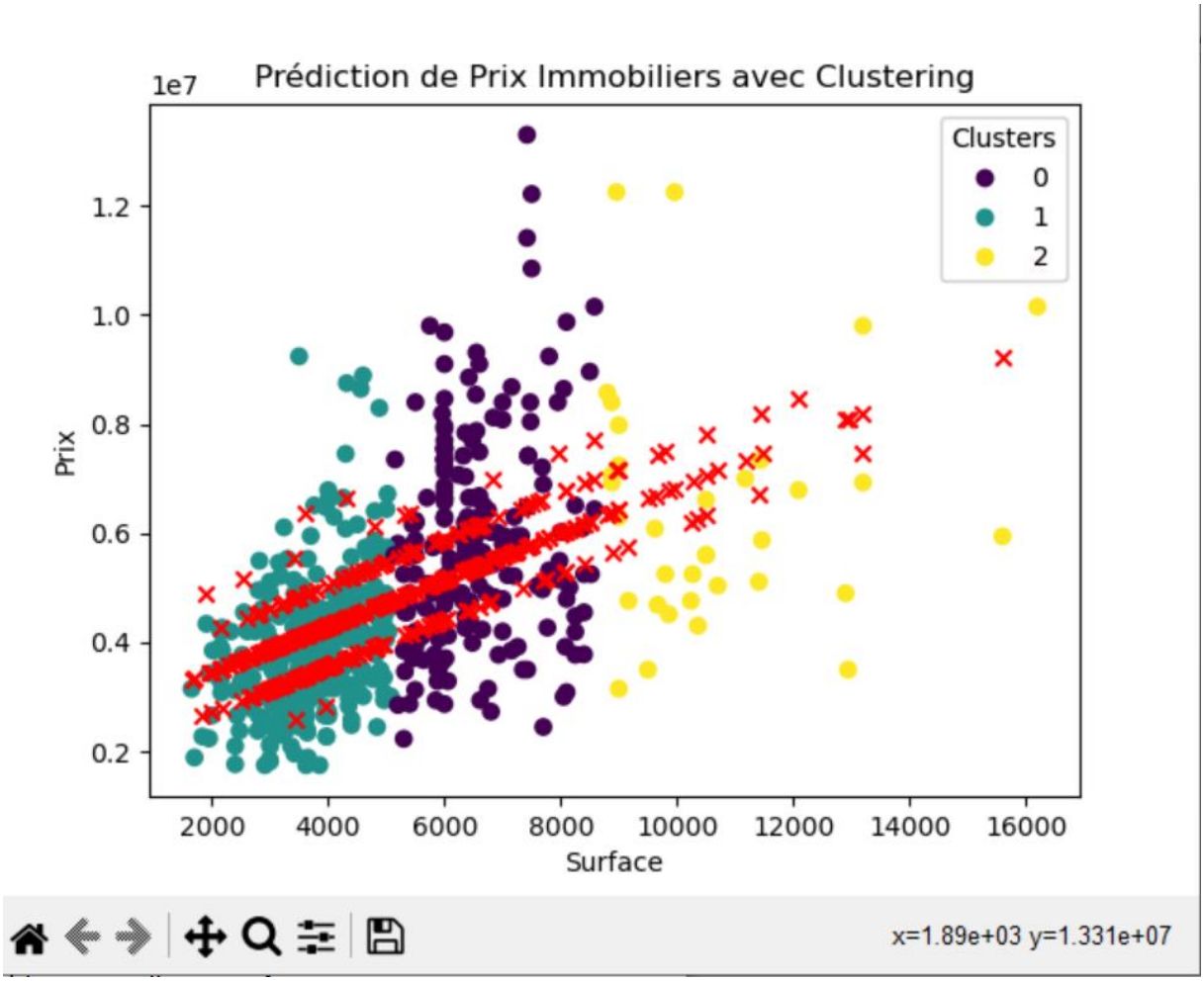
La gestion des connaissances découvertes dans le cadre d'un projet de fouille de données, tel que la prédiction de prix immobiliers, est cruciale pour tirer des enseignements pertinents et prendre des décisions éclairées.

Documentation des Découvertes :

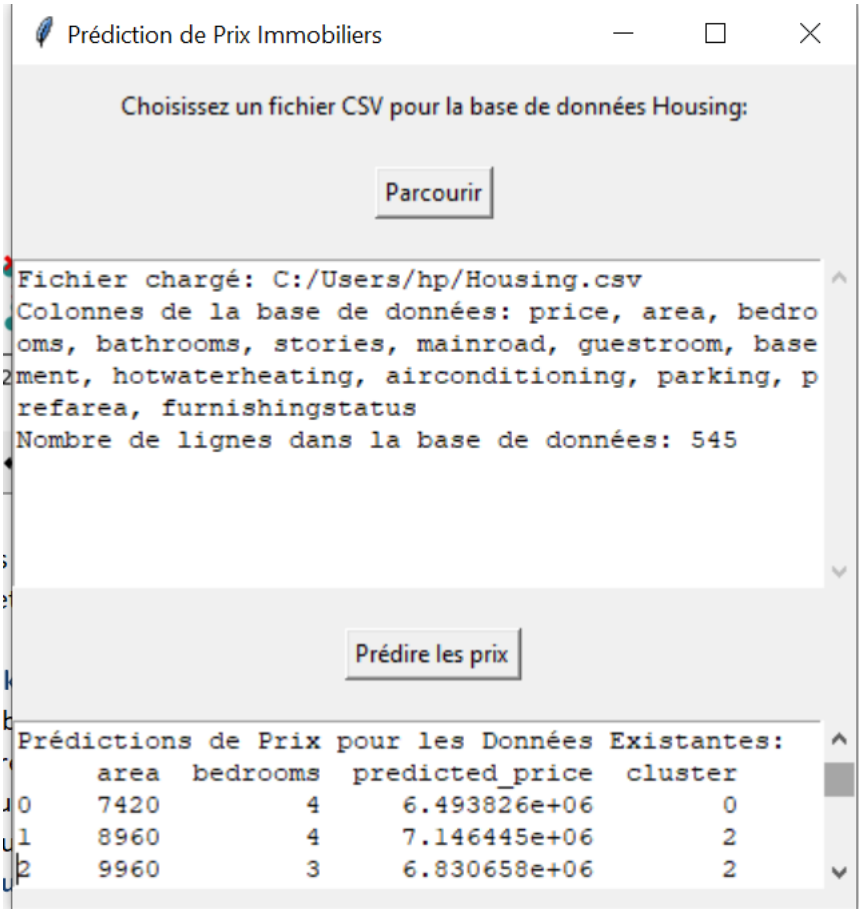
Documentez soigneusement chaque étape du processus de fouille de données, y compris les méthodes utilisées, les paramètres choisis, et les résultats obtenus.

Visualisation des Résultats :

Utilisez des graphiques et des visualisations pour représenter les résultats de manière claire et compréhensible.



Ajoutez des commentaires aux graphiques pour expliquer les tendances observées et les conclusions importantes.



Stockage Structuré des Données :

Utilisez des bases de données ou des structures de données organisées pour stocker les résultats, les prédictions et d'autres informations importantes.
Assurez-vous que les données sont facilement accessibles et peuvent être utilisées pour de futures analyses.

Évaluation Continue :

Mettez en place des mécanismes d'évaluation continue pour suivre les performances des modèles au fil du temps.
Réévaluez périodiquement les résultats à la lumière de nouvelles données pour maintenir la pertinence des modèles.

Conclusion :

Ce projet de prédiction des prix immobiliers a été réalisé dans le contexte de la science des données et de la fouille de données, en utilisant des techniques de modélisation prédictive. Voici quelques points clés tirés de ce projet :

Objectif du Projet : L'objectif principal de ce projet était de développer un modèle de prédiction des prix immobiliers en utilisant des techniques de régression linéaire. Ce modèle vise à estimer le prix d'une propriété en fonction de ses caractéristiques, telles que la surface et le nombre de chambres.

Collecte de Données : La collecte de données a été effectuée à partir de la base de données "Housing" qui contient des informations sur la surface, le nombre de chambres et le prix de différentes propriétés.

Nettoyage et Transformation des Données : Des étapes de nettoyage ont été effectuées pour éliminer les données manquantes et les valeurs aberrantes, assurant ainsi la qualité des données utilisées pour l'entraînement du modèle.

Modélisation : Un modèle de régression linéaire a été entraîné sur les caractéristiques sélectionnées (surface et nombre de chambres). Ce modèle a été utilisé pour prédire les prix immobiliers.

Visualisation : La visualisation a été utilisée pour présenter les résultats de manière compréhensible. Un graphique a été généré pour comparer les prédictions du modèle aux vraies valeurs.

Interprétation du Modèle : L'analyse des coefficients des variables (surface et nombre de chambres) a été effectuée pour comprendre leur impact sur le prix immobilier. Cette interprétation fournit des informations sur la contribution de chaque caractéristique à la prédiction des prix.

Explication des Résultats : Une explication a été fournie sous la forme d'un paragraphe expliquant le graphique généré, mettant en évidence l'alignement entre les prédictions du modèle et les vraies valeurs.