Rapport 2 : Données

# Objectifs

L’objectif principal de ce projet est la prédiction des prix de vente des immobiliers de la région de Boston selon leur caractéristiques et leurs historiques. Et ce dans le but de déterminer une méthode l’évaluation de valeur des immobiliers plus performante.

Cette étude comprend plus précisément trois objectifs :

1. Démystifier les données afin de déterminer les attributs qui influence le plus la valeur d’immobilier.
2. Concevoir deux algorithmes de prédiction en utilisant les méthodes d’apprentissage automatique les plus efficaces pour ce sujet.
3. Établir une étude comparative des résultats des deux algorithmes. Et représenter des concussions sur les prix d’immobilier à Boston.

# Revu de littérature

Dans le but d’améliorer l’évaluation des valeurs d’immobilier, des études ont été réalisées afin de trouver une méthode alternative de prédiction de prix d’immobilier autre que les méthodes conventionnelles telle que la méthode Hedonic.

Dans l’étude comparative des modèles de prédiction des prix d’immobilier rural et urbain en Turquie, Hasan S. (2008) a démontré que le modelé basé sur les réseaux de neurones artificiels est significativement plus performante la régression de Hedonic avec une différence d’erreur quadratique (MSE) estimée à 2,03.

Selon Bourassa S. C. et al. (2007) la méthode de régression Hedonic ne prend pas en considération l’effet de la localisation sue les prix. Xiaalong L. (2012) a démontré l’importance de prise en considération des effets temporaire et spatiale sur l’estimation des prix. Ce dernier a mentionné aussi que le méthode Hedonic ne prend pas en considération les effets spatio-temporels.

# Description du projet

Dans le cadre de ce projet, nous allons non seulement évaluer l’effet spatio-temporel sur les prix des immobiliers à Boston mais aussi déterminer les autres caractéristiques qui aussi. En plus, nous allons prédire les prix des maisons avec deux méthodes afin d’effectuer une étude comparative entre d’autre méthode d’apprentissage automatique et apprentissage profond.

# Métrique de réussite

Bien entendu, la première validation à faire est la soumission de nos valeurs prédites sur le site du Kaggle pour comparer le taux de précision obtenu par notre modèle par rapport aux vraies valeurs. Cependant, cette comparaison est par RMSE (root mean square error) afin de discriminer si les prédictions

# Méthodologie

Cette étude comprendra l’exploration et compréhension des données afin de les nettoyer puis l’estimation des prix des immobiliers avec deux méthodes de prédiction dans le but de les comparées.

Ainsi les trois principales méthodologies pour repend aux objectifs du projet sont les suivantes :

**La méthodologie pour l’objectif 1 :**

Pour nettoyer les données, nous allons tout d’abord les comprendre. Pour ce faire nous explorerons les attributs des immobiliers en étudiant les caractéristiques de leurs distributions. Par la suite, nous allons explorer les anomalies au niveau de données comme ceux de structure de données et les données manquantes. Puis, nous déterminerons les algorithmes de prétraitement à utiliser pour traiter les anomalies découvertes. Dans le cas de données manquantes, une recherche sera effectuée afin de trouver les données manquantes et de les combiner aux données initiales. Nous évaluerons leurs pertinences en fonction du taux de données complètes.

# <http://www.neural.cz/dataset-exploration-boston-house-pricing.html>

<https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57>

Exploration des data avec pandas :

Étape 1 :

1. Liste des attributs
2. Calcul de prix moyen
3. Générer la liste des valeurs que peut prendre les attributs
4. Traiter ou normaliser les attributs quantitatifs
5. Nuage de points ou visualisation
6. Étude de corrélation

Étape 2 :

1. Prix d’achat par rapport à l’an de construction
2. Prendre en considération la superficie des maisons
3. Classification des maisons selon leur critère : faire des intervalles de superficie
4. Nombre de chambre

Étape 3 :

1. Étude de l’influence de la crise économique autour de l’année 2009 : 2008 a 2010

# Bibliographie

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), 143-160.

Bourassa, S., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research*, *32*(2), 139-159.

Limsombunchai, V. (2004, June). House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference* (pp. 25-26).

Liu, X. (2013). Spatial and temporal dependence in house price prediction. *The Journal of Real Estate Finance and Economics*, *47*(2), 341-369.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.