INSTITUT FRANCOPHONE INTERNATIONAL

ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐHQGHN VNU Since 1906

**FOUILLE DE DONNÉES ET DE RECHERCHE D'INFORMATIONS**

# RAPPORT FINAL

Option: SIM

IFI-Promotion :23

---

***APPRENTISSAGE***

---

Présenté par :
Abdoul-Fatao OUEDRAOGO
Timothee BOHINBO

Encadreur:
Mme. Nguyen Thi Minh Huyen
:

Table de Matieres

INTRODUCTION **3**
OBJECTIF 4
II. INFORMATION SUR LE JEU DE DONNÉE **4**
III. PRETRAITEMENT DES DONNEES **5**
IV. CHOIX DE LA MÉTHODE D'APPRENTISSAGE SUPERVISÉE **6**
a. Multiple linear regression 6
b. Test 7
V. COMPARAISON AVEC LA MÉTHODE SUPPORT VECTOR REGRESSION 8
a. Phase d'entraînement 8
b. Test 9
c. Analyse comparative des deux méthodes 9
CONCLUSION **10**
Référence: 10

## INTRODUCTION

L’apprentissage supervisé, dans le contexte de l’intelligence artificielle (IA) et de l’apprentissage automatique, est un système qui fournit à la fois les donnees en entree et les données attendues en sortie. Les donnees en entree et en sortie sont étiquetés en vue de leur classification, afin d'établir une base d’apprentissage pour le traitement ultérieur des données. Les systèmes d’apprentissage automatique supervisé alimentent les algorithmes d’apprentissage avec des quantités connues qui égayeront les futures décisions.

Dans notre cas d'espèce, nous l'appliquerons sur notre jeu de données traitant de la valorisation immobilière afin d’en dégager les résultats afin de les interpréter. Pour ce faire donc, nous allons reposer l’essentiel de notre travail sur cet algorithme qu’est la régression linéaire.

# I. OBJECTIF

Le problème de l'immobilier est d'actualité et met en exergue un bon nombre de détails. En effet, le prix d'un logement est conditionné par l'ˆage de la maison, la distance par rapport à la station MRT la plus proche, le nombre de dépanneurs dans le cercle vivant a pied, les coordonnées géographiques(latitude et longitude) et la date de transaction. Notre analyse vise donc l'objectif de pr´edire le plus proche possible de la r´ealit´e la valeur qu'aura un logement ayant réuni les critères relevés plus haut.

# II. INFORMATION SUR LE JEU DE DONNÉE

Le problème posé dans notre jeu de données est la prédiction de la valeur d'un bien immobilier compte tenu des facteurs cités plus haut. Notre jeu de donnée comporte 414 instances et 7 attributs et porte sur la valorisation immobilière. L'objectif est de déterminer le prix d'une maison à partir de certaines caractéristique telques l'âge de maison, la distance par rapport à la station MRT, et les coordonnées géographiques.

Les attributs présents dans notre jeu de données sont les suivants :

X1 = la date de la transaction

X2 = l'âge de la maison

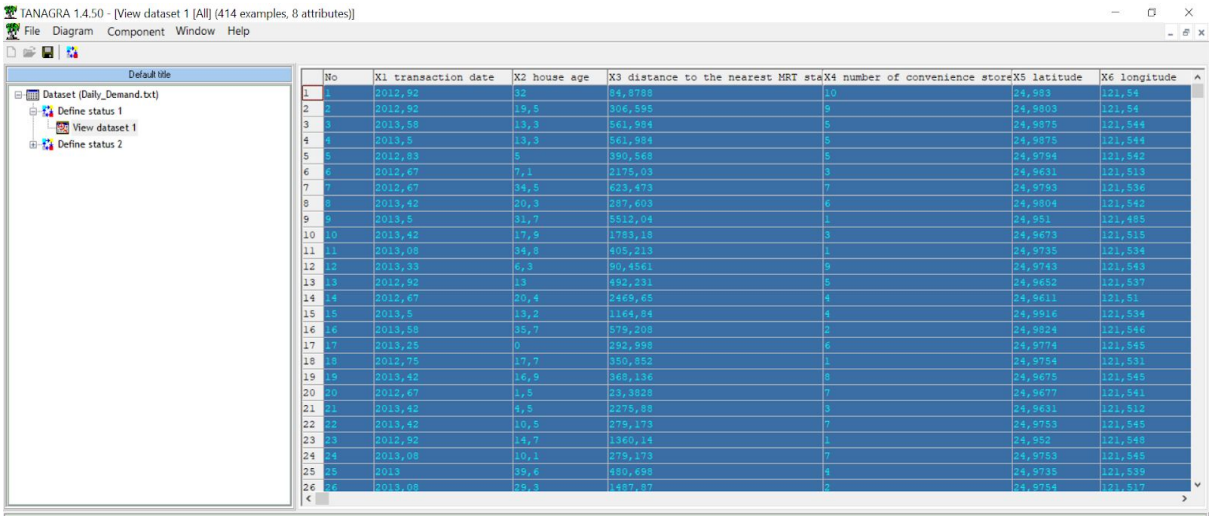
X3 = la distance par rapport à la station MRT la plus proche

X4 = le nombre de dépanneurs dans le cercle

X5 = la latitude

X6 = la longitude.

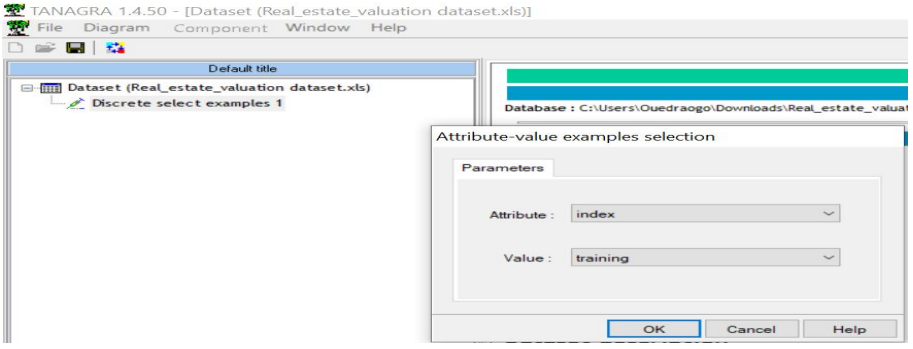
Y = prix du logement



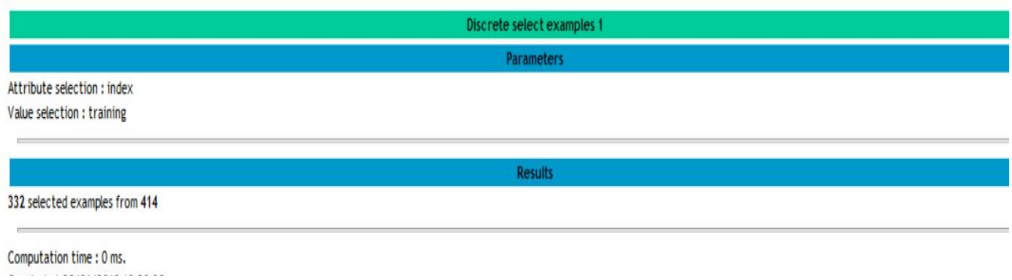

| | No | X1 transaction date | X2 house age | X3 distance to the nearest MRT sta | X4 number of convenience store | X5 latitude | X6 longitude |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2012,92 | 32 | 84,8788 | 10 | 24,983 | 121,54 |
| 2 | 2 | 2012,92 | 19,5 | 306,595 | 9 | 24,9803 | 121,54 |
| 3 | 3 | 2013,58 | 13,3 | 561,984 | 5 | 24,9875 | 121,544 |
| 4 | 4 | 2013,5 | 13,3 | 561,984 | 5 | 24,9875 | 121,544 |
| 5 | 5 | 2012,83 | 5 | 390,568 | 5 | 24,9794 | 121,542 |
| 6 | 6 | 2012,67 | 7,1 | 2175,03 | 3 | 24,9631 | 121,513 |
| 7 | 7 | 2012,67 | 34,5 | 623,473 | 7 | 24,9793 | 121,536 |
| 8 | 8 | 2013,42 | 20,3 | 287,603 | 6 | 24,9804 | 121,542 |
| 9 | 9 | 2013,5 | 31,7 | 5512,04 | 1 | 24,951 | 121,485 |
| 10 | 10 | 2013,42 | 17,9 | 1783,18 | 3 | 24,9673 | 121,515 |
| 11 | 11 | 2013,08 | 34,8 | 405,213 | 1 | 24,9735 | 121,534 |
| 12 | 12 | 2013,33 | 6,3 | 90,4561 | 9 | 24,9743 | 121,543 |
| 13 | 13 | 2012,92 | 13 | 492,231 | 5 | 24,9652 | 121,537 |
| 14 | 14 | 2012,67 | 20,4 | 2469,65 | 4 | 24,9611 | 121,51 |
| 15 | 15 | 2013,5 | 13,2 | 1164,84 | 4 | 24,9916 | 121,534 |
| 16 | 16 | 2013,58 | 35,7 | 579,208 | 2 | 24,9824 | 121,546 |
| 17 | 17 | 2013,25 | 0 | 292,998 | 6 | 24,9774 | 121,545 |
| 18 | 18 | 2012,75 | 17,7 | 350,852 | 1 | 24,9754 | 121,531 |
| 19 | 19 | 2013,42 | 16,9 | 368,136 | 8 | 24,9675 | 121,545 |
| 20 | 20 | 2012,67 | 1,5 | 23,3828 | 7 | 24,9677 | 121,541 |
| 21 | 21 | 2013,42 | 4,5 | 2275,88 | 3 | 24,9631 | 121,512 |
| 22 | 22 | 2013,42 | 10,5 | 279,173 | 7 | 24,9753 | 121,545 |
| 23 | 23 | 2012,92 | 14,7 | 1360,14 | 1 | 24,952 | 121,548 |
| 24 | 24 | 2013,08 | 10,1 | 279,173 | 7 | 24,9753 | 121,545 |
| 25 | 25 | 2013 | 39,6 | 480,698 | 4 | 24,9735 | 121,539 |
| 26 | 26 | 2013,08 | 29,3 | 1487,87 | 2 | 24,9754 | 121,517 |

## III. PRETRAITEMENT DES DONNEES

Nous voulons utiliser la colonne index pour la sous-division du jeu de données. Nous insérons les EXEMPLES DISCRETE SELECT(Onglet INSTANCE SELECTION) dans le diagramme. Nous définissons les paramètres suivants en cliquant sur le menu contextuel PARAMETERS.




Nous validons et nous cliquons sur le menu VIEW. Nous voyons que 332 exemples sont affectés à la phase d'apprentissage ; les autres (82 exemples) seront utilisés pour l'évaluation des modèles.




## IV. CHOIX DE LA MÉTHODE D'APPRENTISSAGE SUPERVISÉE

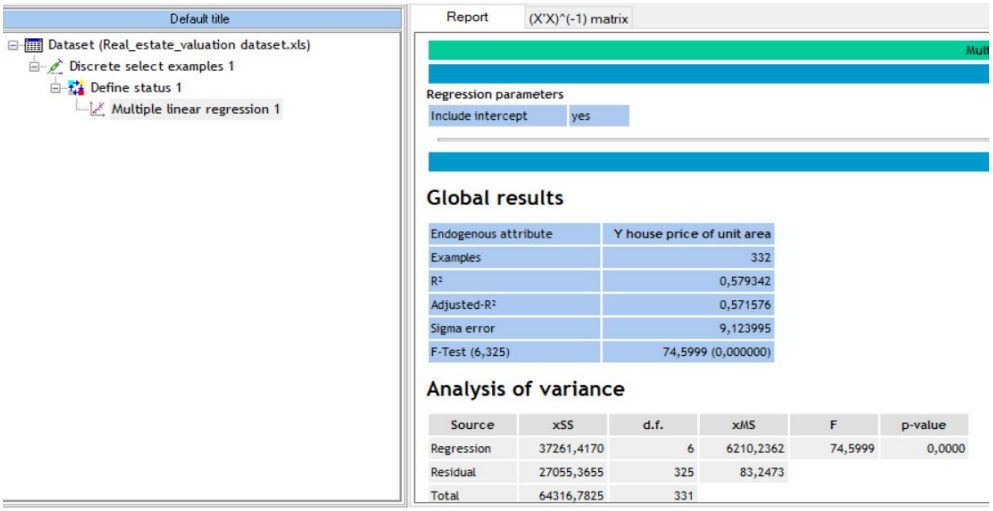
Dans notre cas nous allons utiliser la methode de regression lineaire multiple pour application à notre jeu de donnée. Nous avons choisi d'utiliser le modèle de régression linéaire car nos variables sont des variables continues en plus nos valeurs sont réelles.

Dans ce modèle de régression linéaire, on a plusieurs variables dont une qui est une variable explicative et les autres qui sont des variables expliquées.

NB: Nous avons considéré tous nos variables non négligeable donc elles seront tous utilisés pour pouvoir faire la prédiction des prix de maison.

### a. Multiple linear regression

Nous allons effectuer une régression linéaire multiple en intégrant toutes les variables explicatives (les variables exogènes). Nous introduisons le composant MULTIPLE LINEAR REGRESSION (onglet RÉGRESSION) dans le diagramme, a la suite de DEFINE STATUS Nous activons le menu VIEW pour 4 acceder aux resultats.

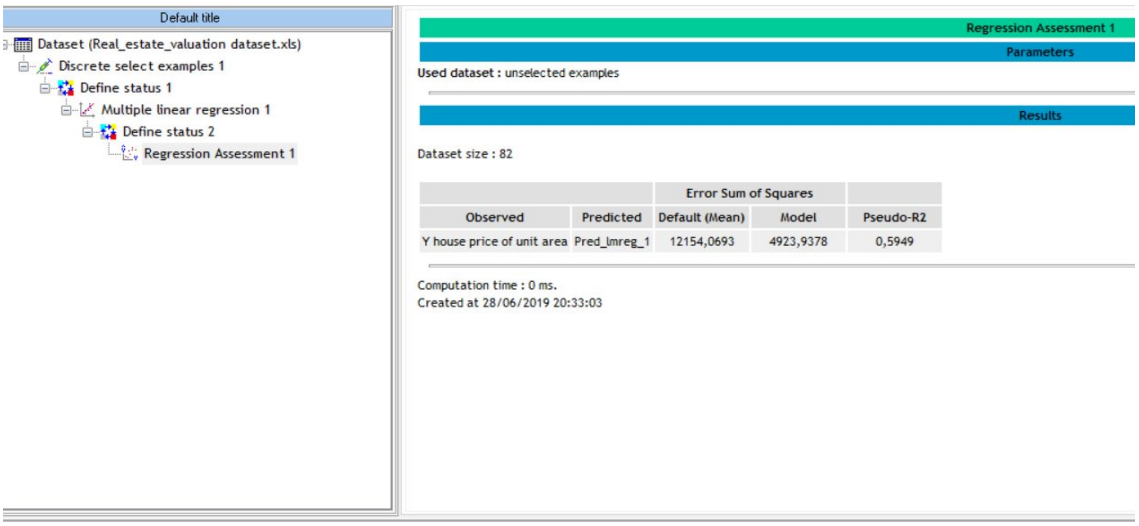



La régression semble excellente, le coefficient de détermination $R^2$ est égal à 0.582370 près de 58% de la variance du prix de la maison est expliquée par la régression. Nous sommes plutôt confiants par rapport aux résultats.

Mais arrêter l'analyse à ce stade serait une erreur. Le nombre de variables indépendantes est élevé dans relation avec le nombre d'exemples. Un overfitting peut survenir. Il serait judicieux d'évaluer la qualité de le modèle sur un jeu de données qui n'a pas contribué à sa construction. C'est le but de l'échantillon de test.

## b. Test

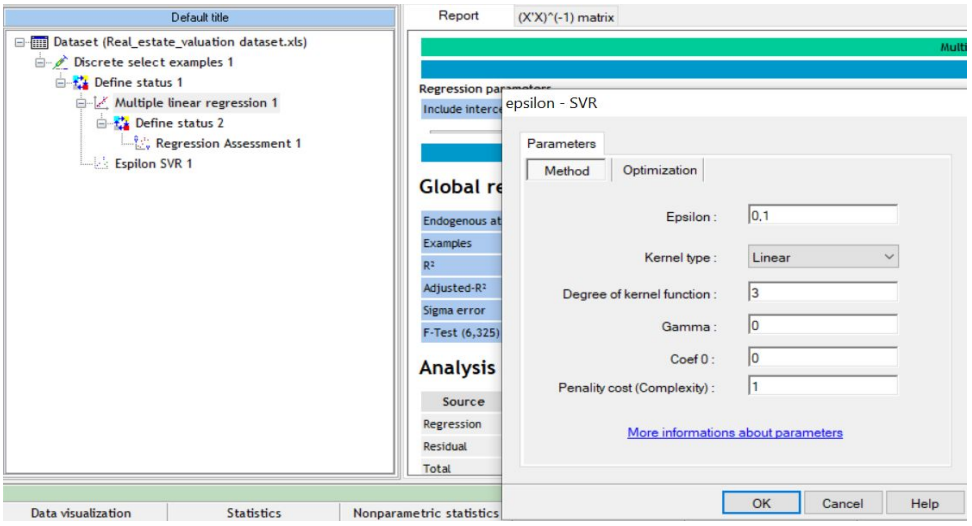
Nous voulons comparer les valeurs observées de l'attribut dépendant (ACTIVITY) avec les valeurs prédites du modèle sur l'échantillon test défini par SUBSET = TEST. Nous devons spécifier le type de variables avec le Composant DEFINE STATUS : nous définissons Y HOUSE PRICE comme target et PREDLMREG1, ajoute automatiquement parle composante de régression, comme input. Nous pouvons maintenant insérer le composant RÉGRESSION ASSESSMENT. Nous définissons les paramètres afin de rendre la comparaison sur les exemples non sélectionnés au d´ebut du diagramme, c'est-à-dire l'échantillon à tester.



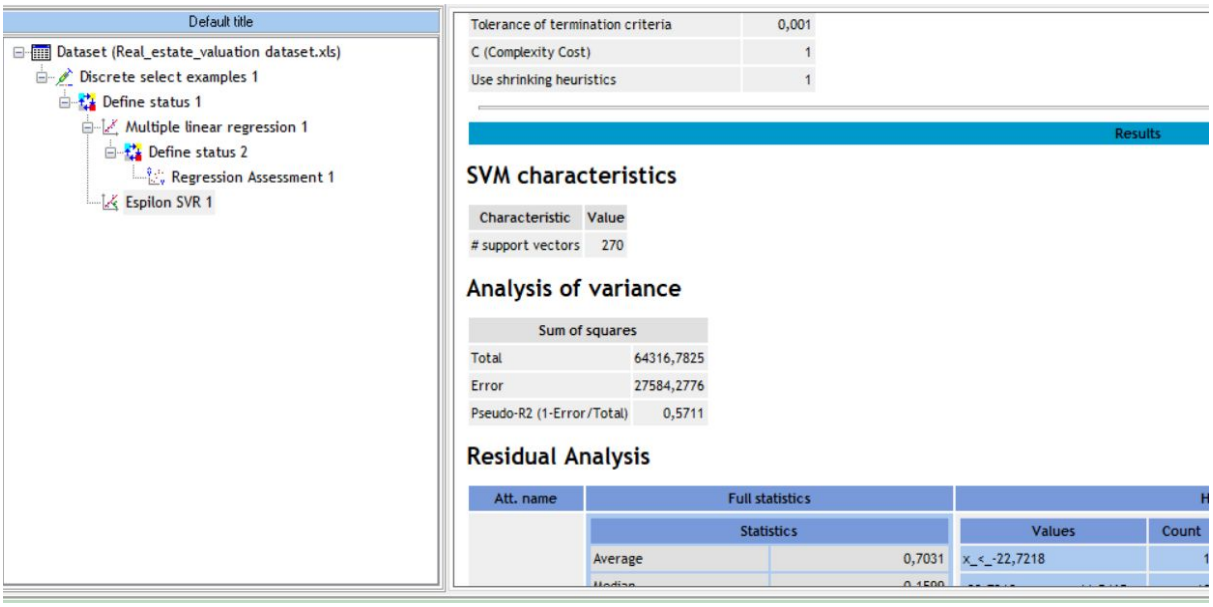

## V. COMPARAISON AVEC LA MÉTHODE SUPPORT VECTOR REGRESSION

### a. Phase d'entraînement

Nous voulons maintenant impl´ementer le composant Epsilon - SVR (onglet REGRESSION). On l'insère après le DEFINE STATUS 1 dans le diagramme. Nous cliquons sur le menu PARAMETRES. Les paramètres par défaut sont les suivants :



Nous ne modifions pas ces paramètres. Nous validons et nous cliquons sur le menu VIEW. On obtient:




Le nombre de vecteurs de support est 270. Le pseudo-R2 sur l'échantillon d'apprentissage est 0,5711. La régression semble très bon.

### b. Test

Nous voulons utiliser l'échantillon test pour évaluer la régression. Nous insérons dans le diagramme : le Composant DEFINE STATUS ( Y HOUSE PRICE comme target et PREDESV R1 comme input)et L'ÉVALUATION DE RÉGRESSION composant (nous utilisons l'ensemble de données c'est à dire l'échantillon de test ).

Le pseudo-R2 sur l'échantillon d'apprentissage est 0,5686. La régression semble très bon.

## C. Analyse comparative des deux méthodes

En se basant sur les résultats nous dirons que les deux sont très proches. Le coefficient de détermination $R^2$ est de 0.58 soit 58% avec la methode de regression lineaire et 0.57 soit 57%. Nous aurons pu obtenir des meilleurs si on avait pas considéré tous les variables, dans notre cas ici nous avons vu que le coefficient X6 (longitude) pouvait être négligé.

CONCLUSION

En somme, ces différents travaux nous ont permi de mieux comprendre les techniques d'analyse des donn´ees en utilisant les algorithmes d'apprentissage supervisés et non supervis´es avec l'outil Tanagra.Ces démarches nous ont servi à explorer toutes les techniques de manipulation des données.

**Référence:**


- https://eric.univ-lyon2.fr/~ricco/cours/cours/Regression_Lineaire_Multiple.pdf
- https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-regmult.pdf
- http://irma.math.unistra.fr/~fbertran/enseignement/Master1_FC_MCB/Cours5.pdf
- http://data-mining-tutorials.blogspot.com/
- http://www.sciences.ch/dwnldbl/informatique/DataMiningTanagra.pdf