**Machine Learning Algorithms for House Price Prediction**

Các nghiên cứu về định giá bất động sản đã phát triển dựa trên những lý thuyết kinh tế cốt lõi. Lý thuyết giá trị thuộc tính cho rằng giá trị của một căn nhà được hình thành từ các đặc tính cấu thành như diện tích, vị trí, số phòng và tiện ích xung quanh. Mô hình giá hedonistic mở rộng ý tưởng này bằng cách mô hình hóa giá nhà như một hàm số của các đặc trưng định lượng và định tính. Tuy nhiên, khi dữ liệu trở nên phức tạp, các mô hình truyền thống (ví dụ: hồi quy tuyến tính) thường không đạt hiệu quả cao, từ đó thúc đẩy việc ứng dụng các thuật toán học máy hiện đại [1-5].

**Dataset and Feature Engineering**

Chất lượng dữ liệu và kỹ thuật xử lý đặc trưng là yếu tố then chốt trong việc xây dựng mô hình học máy. Một số bộ dữ liệu được sử dụng phổ biến gồm:

* **Ames Housing Dataset:** Chứa hơn 2900 mẫu với hơn 70 đặc trưng về căn nhà.
* **Melbourne Housing Dataset [5]:** Dữ liệu thực tế từ thành phố Melbourne, Australia.
* **Korean Real Estate Dataset [4]:** Dữ liệu lớn phục vụ định giá bất động sản tại Hàn Quốc.
* **Vietnam Housing Dataset [1]:** Dữ liệu thực địa tại Việt Nam với nhiều biến định tính và định lượng.
* **Slovenia & India Housing Datasets [2,3]:** Được sử dụng để đánh giá mô hình trên các thị trường khác nhau.

Các kỹ thuật xử lý dữ liệu bao gồm:

* **Xử lý giá trị thiếu:** Điền giá trị trung bình/median hoặc loại bỏ dữ liệu thiếu.
* **Encoding biến phân loại:** Sử dụng One-Hot Encoding hoặc Label Encoding.
* **Scaling và Normalization:** Chuẩn hóa dữ liệu để tránh thiên vị.
* **Feature Selection:** Lựa chọn các đặc trưng quan trọng dựa trên tương quan và tầm ảnh hưởng.

**References (Part 2)**

**[1]. Truong, Q., Nguyen, M., Dang, H., et al. (2020) Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433-442.  
[2]. Ghosalkar, N. N., Dhage, S. N. (2018) Real estate value prediction using linear regression. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 1-5.  
[3]. Čeh, M., Kilibarda, M., Lisec, A., et al. (2018) Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, 7(5), 168.  
[4]. Hong, J., Choi, H., Kim, W. (2020) A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*.  
[5]. Phan, T.D. (2018) Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In: *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*. IEEE, 35-42.**