ĐẠI HỌC BÁCH KHOA HÀ NỘI

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

A red and white flag with a yellow and red design

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN AI**

**Chủ đề: Dự đoán giá nhà đất Hà Nội**

|  |  |
| --- | --- |
| Giảng viên hướng dẫn:  Nhóm 12  Sinh viên tham gia: | **Trần Thế Hùng**  **Phạm Công Hào 20215045**  **Đặng Thái Tuấn 20210907**  **Phạm Đức Lưu 20215084** |
|  |  |

***Hà Nội, 6/2024***

1. Giới thiệu bài toán

* Bài toán dự đoán giá nhà đất là một trong những bài toán quan trọng trong lĩnh vực bất độ sản. Mục đích của bài toán là xây dựng một mô hình dự đoán giá nhà đất dự trên các yếu tố đầu vào như vị trí, diện tích, số phòng, số tầng và nhiều yếu tố khác. Với sự phát triển không ngừng và nhu cầu nhà ở tăng cao, việc dự đoán chihs xác giá nhà đất sẽ giúp các nhà đầu tư, người mua nhà và các bên liên quan đưa ra các quyết định chính xác và hiệu quả hơn.

1. Phân công công việc

|  |  |
| --- | --- |
| **Công việc** | **Tham gia** |
| Craw data | Đặng Thái Tuấn,  Phạm Đức Lưu |
| Xử lý dữ liệu | Phạm Công Hào |
| Xây dựng mô hình | Phạm Công Hào,  Đặng Thái Tuấn,  Phạm Đức Lưu |
| Xây dựng giao diện | Phạm Đức Lưu,  Đặng Thái Tuấn |

1. Chi tiết công việc
2. Craw data
3. Xử lý dữ liệu
4. Missing values (xử lý các bản ghi bị thiếu dữ liệu)
   * Sau khi kiểm tra lượng dữ liệu bị thiếu trong mỗi bản ghi là 3% < 10% => thực hiện sinh dữ liệu cho các feature bị thiếu. Cụ thể:
     + Diện tích: gán giá trị bằng giá trị trung bình diện tích của các căn nhà trong bộ dữ liệu.
     + Số tầng: gán giá trị bằng giá trị xuất hiện nhiều lần nhất trong Feature “Số tầng” của bộ dữ liệu.
     + Số phòng ngủ, WC, Thang máy: Xử lý tương tự như “Số tầng”.
     + Mặt tiền: việc gán giá trị bị thiếu cho Feature này tùy thuộc vào vị trí của căn nhà trong bộ dữ liệu:
       - “Nhà hẻm, ngõ” => bằng 0.
       - “Nhà mặt tiền, phố” => bằng 5.
       - “Biệt Thự liền kề” => bằng giá trị trung bình cộng Mặt tiền của các căn nhà có vị trí “Biệt thự liền kề”.
     + Vị trí, Quận, Pháp lý: gán các giá trị bằng giá trị xuất hiện nhiều nhất trong mỗi Feature tương ứng.
5. Loại bỏ các Outliers (xử lý các bản ghi có dữ liệu khác nhiều so với đa số)
   * Sử dụng phương pháp Interquartile Range (IQR) tính giới hạn trên và giới hạn dưới sau đó loại bỏ các bản ghi có giá trị nằm ngoài khoảng từ giới hạn dưới đến giới hạn trên.
   * Ghi chú: bản ghi là các thông số của một căn nhà (diện tích, số tầng, vị trí, ......). Bộ dữ liệu được tạo thành từ các bản ghi.
6. Biến đổi dữ liệu dạng chữ

Do các Feature bên dưới đều là kiểu dữ liệu dạng Ordinal feature (có thứ tự) nên cần mã hóa phân cấp cho từng Feaure. Cụ thể:

* + Quận: gán bằng giá trị cho 1m2 đất của mỗi Quận. Cụ thể:
    - Quận Đống Đa: 192
    - Quận Hai Bà Trưng: 214
    - Quận Hà Đông: 144
    - ....
  + Vị trí: có các loại giá trị (“Nhà hẻm, ngõ”, “nan”, “Nhà đường nội bộ, Cổ Nhuế”, “Biệt thự liền kề”, “Nhà mặt tiền, phố”) sẽ được mã hóa lần lượt (0, 1, 1, 2, 3).

Giải thích giá trị “nan” do dữ liệu của Chung cư không có thuộc tính “Vị trí” => không xác định => gán giá trị “nan”.

* + Pháp lý: có các giá trị (“Sổ hồng, Sổ đỏ”) do chỉ có hai thuộc tính để tránh việc mô hình hiểu nhầm đây là dữ liệu kiểu Boolean sẽ mã hóa (0, 2) thay vì (0, 1).
  + Kiểu nhà: có hai giá trị (“Nhà”, Chung cư”) tương tự sẽ được mã hóa (0, 2).

1. Xây dựng mô hình

Giới thiệu: phương pháp Ensemble là một kỹ thuật trong học máy nhằm cải thiện độ chính xác và hiệu suất của mô hình bằng cách kết hợp nhiều mô hình dự đoán khác nhau thay vì chỉ dựa vào một mô hình duy nhất, phương pháp Ensemble tận dụng sức mạnh của nhiều mô hình để tạo ra một mô hình tổng hợp mạnh mẽ hơn.

Nhóm sử dụng phương pháp Ensemble để kết hợp hai mô hình dự đoán RandomForestRegressor và XGBoost

1. RandomForestRegressor
2. XGBoost
3. Xây dựng giao diện