**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ CHO THUÊ CĂN HỘ KHU VỰC THỦ ĐỨC**

|  |  |  |  |
| --- | --- | --- | --- |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Trần Tuyết Nhi | 21520393 | KHMT |
| 2 | Nguyễn Trọng Ân | 21520548 | KHMT |
| 3 | Hoàng Kim Ngọc Anh | 21520560 | KHMT |
| 4 | Trần Khánh Duy | 21520787 | KHMT |

**TP. HỒ CHÍ MINH – 12/2022**

# GIỚI THIỆU

Đề tài "Xây dựng mô hình dự đoán giá thuê căn hộ chung cư ở Thành phố Thủ Đức" là một nghiên cứu nhằm xây dựng mô hình dự đoán giá thuê căn hộ chung cư trong khu vực Thành phố Thủ Đức. Mục tiêu của đề tài là cung cấp một công cụ hoặc giải pháp cho việc ước lượng giá thuê căn hộ chung cư dựa trên các yếu tố như diện tích, số phòng ngủ, số tầng, tên dự án, tên phường,…

Để thực hiện đề tài này, nhóm sử dụng các công cụ lập trình là Google Colab, Jupyter Notebook, ngôn ngữ lập trình là Python với các framework như: pandas, numpy, seaborn, matplotlib, scipy, sklearn. Các phương pháp và thuật toán được áp dụng cho đề tài này là thuật toán hồi quy tuyến tính. Đánh giá hiệu suất của mô hình bằng thang đo đánh giá hệ số xác định (R-squared), sử dụng kiểm chứng chéo K-fold để chọn ra mô hình tốt nhất.

Kết quả đạt được từ đề tài này là mô hình dự đoán giá cho thuê căn hộ ở khu vực Thủ Đức với các biến đã biết, người dùng có thể nhập từ 1 đến 13 biến, không cần phải cung cấp đủ thông tin 13 biến. Kết quả đánh giá mô hình trên tập dữ liệu test chứa missing data là 0.8133. Mô hình trên được nhóm xây dựng dựa trên việc xây dựng và đánh giá kết quả của 24 mô hình gốc (13 mô hình đơn biến và 11 mô hình đa biến) dự đoán giá thuê căn hộ chung cư ở Thành phố Thủ Đức.

Bộ dữ liệu phân tích tự thu thập tại trang web nhatot.com[1] ghi vào tài liệu tham khảo với giới hạn khu vực tại Thành phố Thủ Đức, Thành phố Hồ Chí Minh.

# MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu phân tích được thu thập thủ công, gồm có 674 dòng và 14 cột. Trong đó có 13 cột là biến độc lập và 1 cột là biến phụ thuộc. Bộ dữ liệu phân tích tự thu thập tại trang nhatot.com[1] ghi vào tài liệu tham khảo.

Ý nghĩa của các biến trong bộ dữ liệu được trình bày ở bảng 1:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên biến** | **Ý nghĩa** |
| 1 | Diện tích | Kích thước căn hộ cho thuê tính theo đơn vị m2 |
| 2 | Số phòng ngủ | Số phòng ngủ của căn hộ cho thuê |
| 3 | Số tầng | Căn hộ cho thuê ở tầng nào của tòa nhà |
| 4 | Tên dự án | Tên tòa nhà nơi có căn hộ cho thuê |
| 5 | Tên phường | Tên phường nơi có căn hộ cho thuê |
| 6 | Tình trạng nội thất | Tình trạng nội thất sẵn có của căn hộ cho thuê |
| 7 | Số phòng vệ sinh | Số phòng vệ sinh của căn hộ cho thuê |
| 8 | Loại hình căn hộ | Loại hình nhà ở của căn hộ cho thuê |
| 9 | Ban công | Căn hộ cho thuê có ban công hay không |
| 10 | Khu mua sắm | Căn hộ cho thuê có gần chợ hoặc siêu thị hay không |
| 11 | Trường học | Căn hộ cho thuê có gần trường học (cấp 1, cấp 2, cấp 3. đại học, cao đẳng) hay không |
| 12 | Bệnh viện | Căn hộ cho thuê có gần bệnh viện hay không |
| 13 | Số ảnh | Số ảnh chụp căn hộ cho thuê được người đăng tin cung cấp |
| 14 | Giá cho thuê | Giá cho thuê của căn hộ được tính theo đơn vị VND |
| Các biến Khu mua sắm, Trường học, Bệnh viện được tính là gần khi xét trong phạm  vi bán kính 2km thông qua Google Map | | |

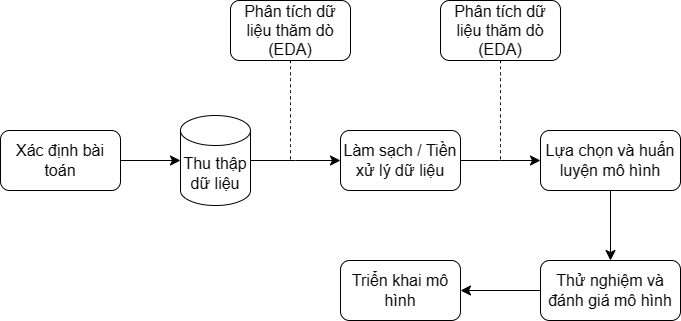
*Bảng 1: Bảng ý nghĩa các biến của bộ dữ liệu*

Bộ dữ liệu có 6 biến kiểu số và 8 biến kiểu phân loại. Kiểu của các cột, các giá trị và số lượng missing data của các biến được trình bày ở bảng 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên biến** | **Dtype** | **Các giá trị** | **Số lượng dữ liệu bị thiếu** |
| Diện tích | int64 | Các giá trị nằm trong khoảng từ 20-280m2 | 0 |
| Số phòng ngủ | int64 | Các giá trị nằm trong khoảng từ 1-5 phòng | 0 |
| Số tầng | float64 | Các giá trị nằm trong khoảng từ tầng 1-26 | 581 |
| Tên dự án | object | Có 95 dự án khác nhau | 176 |
| Tên phường | object | Có 31 tên phường | 0 |
| Tình trạng nội thất | object | Có 3 giá trị:  ● Đầy đủ: Căn hộ cho thuê được trang bị đầy đủ nội thất  ● Một phần: Căn hộ cho thuê được trang bị nội thất cơ bản  ● Chưa có: Căn hộ cho thuê chưa có nội thất | 0 |
| Số phòng vệ sinh | int64 | Các giá trị nằm trong khoảng từ 1-4 ảnh | 0 |
| Loại hình căn hộ | object | Có 5 giá trị:  ● Chung cư  ● Căn hộ dịch vụ, mini  ● Duplex  ● Penthouse  ● Officetel | 0 |
| Ban công | bool | Có 2 giá trị:  ● True: nếu căn hộ có ban công  ● False: nếu căn hộ không có ban công | 0 |
| Khu mua sắm | bool | Có 2 giá trị:  ● True: nếu căn hộ gần các địa điểm được xét  ● False: nếu căn hộ xa các địa điểm được xét  (Được tính trong phạm vi bán kính 2km) | 0 |
| Trường học | bool | 0 |
| Bệnh viện | bool | 0 |
| Số ảnh | int64 | Các giá trị nằm trong khoảng từ 2-14 ảnh | 0 |
| Giá cho thuê | int64 | Có giá trị trong khoảng từ 1100000-108000000 | 0 |

*Bảng 2: Bảng các giá trị của các biến trong bộ dữ liệu*

# PHƯƠNG PHÁP PHÂN TÍCH



*Hình 1. Quy trình Phân tích dữ liệu*

## 3.1. Xác định bài toán

Xây dựng mô hình dự đoán giá cho thuê căn hộ khu vực Thủ Đức

## 3.2. Thu thập dữ liệu

Bộ dữ liệu phân tích được nhóm thu thập thủ công từ trang web nhatot.com[1]

## 3.3. Làm sạch / Tiền xử lý dữ liệu

### 3.3.1. Xử lý giá trị bị khuyết (missing data)

Theo thống kê, có 2 cột thuộc tính có missing data, đó là cột “Số tầng” và cột “Tên dự án”. Đối với thuộc tính ‘Số tầng’, có 581 dòng dữ liệu bị khuyết, nhóm thực hiện điền khuyết bằng phương pháp điền trung vị (median) nhằm giữ nguyên tính chất của cột thuộc tính. Cột thuộc tính ‘Tên dự án’, có 176 dòng dữ liệu bị khuyết, thể hiện các căn hộ không thuộc các dự án lớn có tên, tiến hành điền khuyết một giá trị “None” mô tả dữ liệu căn hộ không có tên dự án. Sau bước xử lý giá trị bị khuyết, 100% giá trị bị khuyết được xử lý.

### Xử lý giá trị ngoại lệ (outliers)

Cột thuộc tính “Loại hình căn hộ” gồm 4 giá trị là: Chung Cư, Căn Hộ Dịch Vụ, Mini, Duplex và Penthouse. Để giới hạn đề tài nhóm chỉ xem xét và xây dựng mô hình với 2 loại hình căn hộ là Chung Cư và Căn Hộ Dịch Vụ. Các loại hình căn hộ như Duplex và Penthouse không phổ biến và có giá cho thuê chênh lệch khá nhiều gây nhiễu cho bộ dữ liệu nên nhóm quyết định loại bỏ. Nhóm đã loại bỏ 100% outliers của cột thuộc tính.

Quan sát boxplot phân bố giá thuê nhà theo từng phường của cột “Tên phường”, ta thấy tồn tại khá nhiều outlier. Những outlier này là các giá trị thiểu số và nhóm cần dự đoán giá cho thuê nhà của các căn hộ phổ biến trong khu vực. Vì vậy nhóm sẽ loại bỏ bớt outlier theo từng phường. Nhóm đã loại bỏ 99.5% outliers của cột thuộc tính.

Tương tự với việc xét giá thuê nhà ở từng phường, xét cột “Tên dự án”, tại cụ thể từng dự án, ta thấy vẫn tồn tại các outlier. Vì vậy nhóm loại bỏ các outlier theo từng dự án để giúp kết quả phân tích chính xác hơn. Nhóm đã loại bỏ 8.5% outliers của cột thuộc tính.

Sau bước loại bỏ giá trị ngoại lệ, nhóm đã loại bỏ tất cả 92 giá trị ngoại lệ. Bộ dữ liệu còn 582 dòng.

### 3.3.3. Chuẩn hóa dữ liệu

Đối với các biến số, sử dụng StandardScaler và Polynomial Features để chuẩn hóa dữ liệu về cùng tỉ lệ và tìm ra bậc tốt nhất cho mô hình hồi quy với từng thuộc tính.

Đối với các biến phân loại, sử dụng OneHotEncoder và PolynomialFeatures chuẩn hoá dữ liệu về dạng số và tìm ra bậc tốt nhất cho mô hình hồi quy với từng thuộc tính.

## 3.4. Lựa chọn và huấn luyện mô hình

Sau khi làm sạch dữ liệu qua các bước xử lý missing data và outlier, nhóm đã thực hiện xây dựng và chọn lọc mô hình tốt nhất để dự đoán giá cho thuê căn hộ khi người dùng lần lượt nhập vào thông tin từ 1 biến đến 13 biến. Với mỗi mô hình đơn biến nhóm thực hiện dò tự động degree trong khoảng (1,4) cho từng biến và chọn mô hình được kiểm định chéo với k=5 cho kết quả kiểm định trung bình cao nhất. Với mô hình đa biến, nhóm phân ra 2 loại biến: biến số và biến phân loại. Với từng loại biến, nhóm thực hiện dò tự động degree trong khoảng (1,4) và chọn mô hình được kiểm định chéo với k=5 cho kết quả kiểm định trung bình cao nhất. Kết quả mô hình tốt nhất với degree là dạng tuple (a,b) trong đó a là chỉ số degree cho biến số và b là chỉ số degree cho biến phân loại.

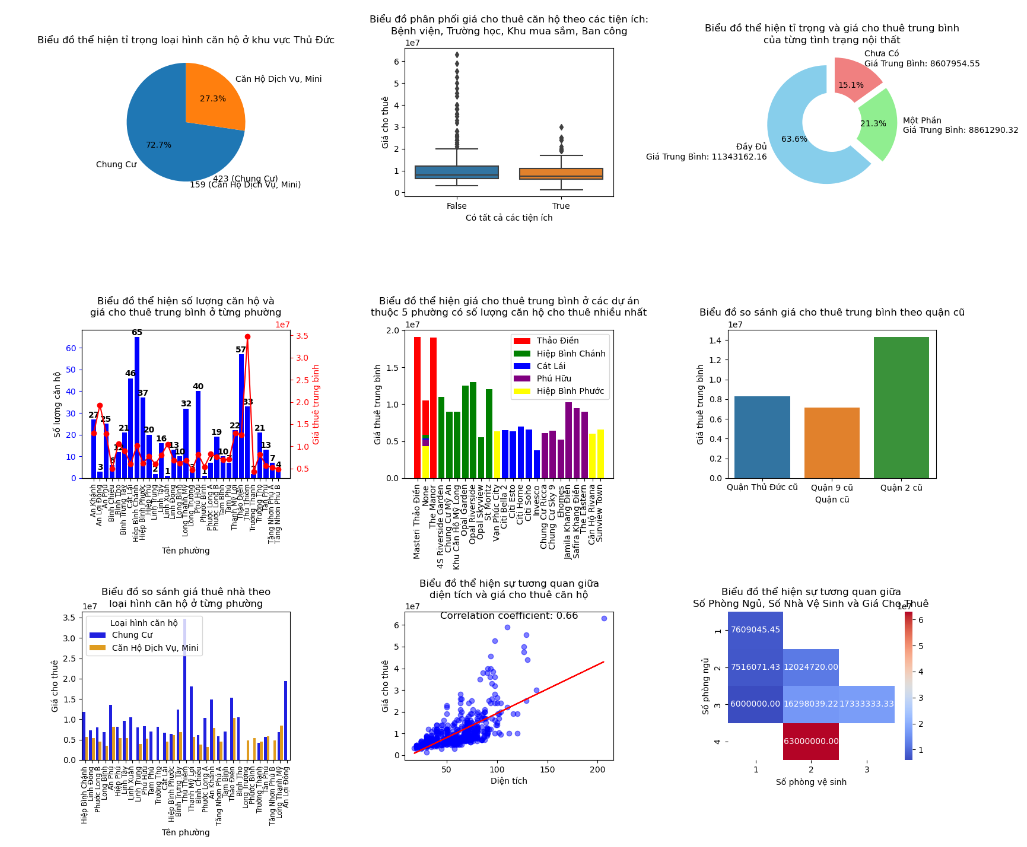
## 3.5. Thử nghiệm và đánh giá mô hình

Sau khi thử nghiệm trên bộ dữ liệu, mô hình được đánh giá dựa trên R2 score và K-fold Validation với k = 5.

## 3.6. Triển khai mô hình

Nhóm đã triển khai đc 100% dự định về mô hình

# PHÂN TÍCH THĂM DÒ/SƠ BỘ



*Hình 2: Trực quan hoá dữ liệu*

Các biểu đồ được vẽ trên cùng 1 trang bằng python. Phân tích chi tiết các biểu đồ như sau:

Biểu đồ [0,0] thể hiện tỷ trọng các loại hình căn hộ ở Thủ Đức cho ta biết khu vực Thủ Đức sẽ bao gồm 2 loại hình căn hộ là Chung Cư và Căn Hộ Dịch Vụ, Mini. Trong đó chiếm đa số là loại hình căn hộ Chung Cư (72.7%).

Biểu đồ [0,1] thể hiện sự phân phối giá cho thuê căn hộ theo các tiện ích, các căn hộ có tất cả 4 tiện ích Bệnh viện, Trường học, Khu mua sắm và Ban công có giá trị True, các căn hộ có 1, 2 hoặc 3 trong 4 tiện ích có giá trị False. Ta thấy giá cho thuê của các căn hộ đầy đủ diện tích có độ phân bố ngắn hơn (1 triệu 3 - 30 triệu) so với các căn hộ không đầy đủ tiện ích (3 triệu – 63 triệu). Giá cho thuê tối thiểu của các căn hộ đầy đủ tiện nghi cũng thấp hơn so với các căn hộ không đầy đủ tiện nghi.

Biểu đồ [0,2] thể hiện tỷ trọng và giá cho thuê trung bình của từng tình trạng nội thất, ta thấy các căn hộ đầy đủ nội thất chiếm tỉ trọng cao nhất hơn 60% (63.6%). Các căn hộ được trang bị nội thất đầy đủ chiếm tỉ trọng thấp nhất 15.1%. Giá cho thuê trung bình của từng loại tình trạng giảm dần theo tỉ trọng. Không có sự khác biệt quá lớn giữa giá cho thuê trung bình của căn hộ chưa có nội thất và căn hộ trang bị một phần nội thất.

Biểu đồ [1,0] thể hiện số lượng và giá cho thuê trung bình ở từng phường cho ta biết phường Hiệp Bình Chánh có số lượng căn hộ cho thuê nhiều nhất (65), và các phường Linh Xuân, Phước Bình có số lượng căn hộ cho thuê ít nhất (1). Bên cạnh đó, phường có giá cho thuê trung bình cao nhất là phường Thủ Thiêm (khoảng 34 triệu 7), tiếp theo sau là phường An Lợi Đông (khoảng 19 triệu 3). Phường có giá thuê trung bình thấp nhất là phường Trường Thạnh với 2 căn hộ và giá trung bình là 4 triệu 350. Tổng cộng số lượng căn hộ cho thuê ở khu vực Thủ Đức - Tổng số lượng đối tượng của bộ dữ liệu là 582 dòng.

Biểu đồ [1,1] thể hiện giá cho thuê trung bình của các dự án thuộc 5 phường có số lượng căn hộ cho thuê nhiều nhất cho thấy giá cho thuê trung bình của các dự án bị ảnh hưởng theo phường, các dự án thuộc cùng một phường có giá cho thuê không chênh lệch nhau quá nhiều. Giá cho thuê trung bình của 2 dự án The Manor và Masteri Thảo Điền là cao nhất, hơn 19 triệu.

Biểu đồ [1,2] so sánh giá cho thuê trung bình theo Quận cũ, nhóm thực hiện chia giỏ các tên phường theo địa chỉ quận cũ trước khi sát nhập thành Thành phố Thủ Đức. Kết quả phân tích biểu đồ cho thấy giá cho thuê trung bình của các phường thuộc Quận 2 cũ cao hơn 6 triệu so với các phường thuộc Quận Thủ Đức cũ và Quận 9 cũ, cho thấy yếu tố vị trí địa lý ảnh hưởng mạnh đến giá cho thuê căn hộ.

Biểu đồ [2,0] so sánh giá thuê nhà theo loại hình căn hộ ở từng phường cho thấy hầu hết giá thuê nhà của loại hình căn hộ chung cư cao hơn hẳn so với giá thuê nhà của loại hình căn hộ dịch vụ, mini ở từng phường. Bảy phường Linh Xuân, Tam Phú, Trường Thọ, Thủ Thiêm, Tam Bình, Bình Thọ và An Lợi Đông chỉ có thông tin căn hộ cho thuê là loại hình chung cư. Ba phường Long Trường, Phước Bình, Tăng Nhơn Phú B chỉ có thông tin cho thuê căn hộ dịch vụ và giá cho thuê của các phường này không cao, vào khoảng 6 triệu.

Biểu đồ [2,1] thể hiện sự tương quan giữa diện tích và giá cho thuê căn hộ cho thấy biến diện tích ảnh hưởng mạnh theo chiều thuận đến giá cho thuê căn hộ. Đa số các căn hộ cho thuê có diện tích dưới 100m2.

Biểu đồ [2,2] thể hiện sự tương quan giữa Số Phòng Ngủ, Số Nhà Vệ Sinh và Giá Cho Thuê, các căn hộ có cùng số lượng phòng ngủ, giá thuê nhà tăng theo số lượng phòng vệ sinh. Có sự tăng mạnh về giá thuê nhà khi tăng số lượng phòng vệ sinh từ 1 lên 2 phòng. Xét các căn hộ có 1 phòng vệ sinh, căn hộ có số lượng phòng ngủ càng tăng thì giá thuê nhà lại càng giảm. Căn hộ có 2 phòng vệ sinh có giá thuê nhà tương quan thuận với số phòng ngủ. Biểu đồ còn cho ta thấy giá thuê nhà trung bình của căn hộ 3 phòng ngủ, 1 phòng vệ sinh là thấp nhất (6 triệu), căn hộ 4 phòng ngủ và 2 phòng vệ sinh có giá thuê trung bình cao nhất (63 triệu).

# KẾT QUẢ PHÂN TÍCH

Nhóm đã xây dựng được 24 mô hình dự đoán giá cho thuê, trong đó có mô hình dự đoán đơn biến và 11 mô hình dự đoán đa biến. Các kết quả đánh giá mô hình như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Loại mô hình** | **Tên biến sử dụng** | **Best degree** | **Score (R2) on train set** | **Score (R2) on test set** |
| Mô hình đơn biến | Diện tích | 2 | 0.5188 | 0.4472 |
| Số phòng ngủ | 2 | 0.1891 | 0.0418 |
| Số tầng | 3 | 0.022 | 0.0848 |
| Tên dự án | 2 | 0.8298 | 0.7977 |
| Tên phường | 1 | 0.7051 | 0.7173 |
| Tình trạng nội thất | 3 | 0.0248 | 0.0251 |
| Số phòng vệ sinh | 1 | 0.1351 | 0.0853 |
| Loại hình căn hộ | 3 | 0.1085 | 0.0281 |
| Ban công | 3 | 0.0234 | 0.026 |
| Khu mua sắm | 1 | 0.015 | 0.0059 |
| Trường học | 2 | 0.1644 | -0.0113 |
| Bệnh viện | 2 | 0.0186 | 0.0155 |
| Số ảnh | 1 | 0.0324 | 0.0842 |
| Mô hình 2 biến | Diện tích, Tên dự án | Biến số: 1  Biến phân loại: 1 | 0.9026 | 0.8492 |
| Mô hình 3 biến | Diện tích, Số phòng ngủ, Tên dự án | Biến số: 1  Biến phân loại: 2 | 0.9030 | 0.8464 |
| Mô hình 4 biến | Diện tích, Số phòng ngủ, Tên dự án, Tình trạng nội thất | Biến số: 1  Biến phân loại: 3 | 0.9329 | 0.8558 |
| Mô hình 5 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất | Biến số: 1  Biến phân loại: 3 | 0.9332 | 0.8559 |
| Mô hình 6 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường | Biến số: 1  Biến phân loại: 3 | 0.9470 | 0.8667 |
| Mô hình 7 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ | Biến số: 1  Biến phân loại: 3 | 0.9493 | 0.8644 |
| Mô hình 8 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh | Biến số: 1  Biến phân loại: 3 | 0.9494 | 0.8634 |
| Mô hình 9 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh, Số tầng | Biến số: 1  Biến phân loại: 3 | 0.9494 | 0.8628 |
| Mô hình 10 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh, Số tầng, Bệnh viện | Biến số: 1  Biến phân loại: 1 | 0.9245 | 0.8608 |
| Mô hình 11 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh, Số tầng, Bệnh viện, Trường học | Biến số:1  Biến phân loại: 1 | 0.9246 | 0.8606 |
| Mô hình 12 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh, Số tầng, Bệnh viện, Trường học,Ban công | Biến số: 1  Biến phân loại: 1 | 0.9248 | 0.8593 |
| Mô hình 13 biến | Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án,Tình trạng nội thất, Tên phường, Loại hình căn hộ, Số ảnh, Số tầng,Khu mua sắm, Bệnh viện, Trường học,Ban công | Biến số: 1  Biến phân loại: 1 | 0.9249 | 0.8574 |

*Bảng 3: Kết quả xây dựng mô hình dự đoán giá thuê căn hộ tại khu vực Thủ Đức*

Với mô hình đơn biến, biến độc lập cho mô hình dự đoán giá thuê chung cư tốt nhất là Tên dự án với score trên tập train là 0.83 và tập test là 0.79. Với mô hình đa biến, mô hình 6 biến bao gồm: Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án, Tình trạng nội thất, Tên phường cho kết quả dự đoán tốt nhất với score tập train là 0.95 và tập test là 0.87. Các mô hình đa biến còn lại kết quả score tập test ở vào khoảng 0.84 - 0.86.

# CHỈNH SỬA SAU BÁO CÁO

Sau khi xây dựng và đánh giá kết quả 24 mô hình với số lượng biến đầu vào khác nhau, nhóm quyết định chọn mô hình 6 biến: Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án, Tình trạng nội thất, Tên phường với bậc đa thức của biến số là 1 và biến phân loại là 3 làm mô hình dự đoán giá cho thuê căn hộ tại khu vực Thủ Đức.

Trong trường hợp người dùng nhập đầy đủ 13 biến, mô hình sẽ xử lý các thông tin liên quan đến 6 biến được nói đến ở trên để dự đoán ra kết quả cuối cùng. Các cột thông tin khác sẽ bị bỏ qua, do đó là các thông tin gây nhiễu làm ảnh hưởng xấu đến kết quả dự đoán. Kết quả đánh giá mô hình trên tập train với score là 0.9470 và tập test là 0.8667.

Đối với trường hợp người dùng nhập không đầy đủ các biến, nếu các thông tin bị thiếu rơi vào các cột thuộc tính dùng để huấn luyện mô hình thì nhóm áp dụng biện pháp dùng đối số mặc định để điền vào các giá trị bị trống. Các giá trị dùng làm đối số mặc định tương ứng với các cột như bảng sau:

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Giá trị dùng làm đối số mặc định** |
| Diện tích | 60 |
| Số phòng ngủ | 2 |
| Số phòng vệ sinh | 2 |
| Tên dự án | None |
| Tình trạng nội thất | Đầy đủ |
| Tên phường | Hiệp Bình Chánh |
| Các giá trị thay thế được chọn theo các giá trị xuất hiện nhiều nhất của thuộc tính (mode) | |

*Bảng 4: Bảng các giá trị đối số mặc định*

Kết quả đánh giá mô hình trên tập test gồm 20 dòng ngẫu nhiên lấy từ bộ dữ liệu với 30 missing data được khởi tạo ngẫu nhiên ở các cột thuộc tính có score là 0.8133.

# KẾT LUẬN

Sau khi thu thập thủ công từ trang web nhatot.com[1] nhóm đã xây dựng được bộ dữ liệu gồm 674 dòng dữ liệu với 14 biến trong đó có 13 biến độc lập và 1 biến phụ thuộc cho đề tài xây dựng mô hình dự đoán giá thuê căn hộ tại khu vực Thủ Đức. Thực hiện xử lý missing data và outlier, nhóm được bộ dữ liệu ổn định hơn gồm 582 dòng. Xây dựng mô hình với kỹ thuật pipeline, hiệu suất của mô hình bằng thang đo đánh giá hệ số xác định (R-squared), thực hiện dò tự động để tìm degree tốt nhất cho các biến và sử dụng kiểm chứng chéo K-fold để chọn ra mô hình tốt nhất.

Kết quả đạt được từ đề tài này là mô hình dự đoán giá cho thuê căn hộ ở khu vực Thủ Đức với các biến đã biết (có thể nhập từ 1 đến 13 biến, không cần phải đủ 13 biến), kết quả đánh giá mô hình trên tập dữ liệu test chứa missing data là 0.8133. Mô hình trên được nhóm xây dựng dựa trên việc xây dựng và đánh giá kết quả của 24 mô hình gốc (13 mô hình đơn biến và 11 mô hình đa biến) dự đoán giá thuê căn hộ chung cư ở Thành phố Thủ Đức. Trong 13 mô hình đơn biến, mô hình sử dụng biến "Tên dự án" đã cho ra kết quả tốt nhất với R-squared (hệ số xác định) trên tập test là 0.79. Điều này có nghĩa là biến "Tên dự án" có mối quan hệ mạnh mẽ với giá thuê căn hộ chung cư, và việc sử dụng chỉ một biến này đã đủ để dự đoán giá thuê một cách chính xác. Trong 11 mô hình đa biến còn lại, mô hình sử dụng 6 biến bao gồm: Diện tích, Số phòng ngủ, Số phòng vệ sinh, Tên dự án, Tình trạng nội thất, Tên phường đã cho kết quả tốt nhất với R-squared trên tập test là 0.8667. Các mô hình khác cũng cho kết quả khá gần với R-squared trên tập test trong khoảng từ 0.84 đến 0.86.

**TÀI LIỆU THAM KHẢO**

1. Nhà tốt. Link: <https://www.nhatot.com> (7/11/2023)
2. Mastering Data Analytics. Link: <https://mastering-da.com/bieu-do-dung-trong-data-visualization> (12/12/2023)
3. Daniel Nelson, David Landup, Jovana Ninković, Data Visualization in Python, 2020
4. Assure Shift. Link: <https://www.assureshift.in/blog/factors-affecting-rental-prices> (30/10/2023)
5. Data.word. Link: <https://data.world/bob-wakefield/house-price-data> (30/10/2023)
6. Scikit learn. Link: https://scikit learn.org/stable/supervised\_learning.html#supervised-learning (10/12/2023)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Nguyễn Trọng Ân | Thu thập dữ liệu, phân tích thăm dò, làm sạch, tiền xử lý dữ liệu, viết báo cáo |
| 2 | Trần Tuyết Nhi | Thu thập dữ liệu, phân tích thăm dò, huấn luyện và đánh giá mô hình, viết báo cáo. |
| 3 | Hoàng Kim Ngọc Anh | Thu thập dữ liệu, phân tích thăm dò, làm sạch, tiền xử lý dữ liệu, viết báo cáo |
| 4 | Trần Khánh Duy | Thu thập dữ liệu, phân tích thăm dò, huấn luyện và đánh giá mô hình, viết báo cáo. |