**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KỸ THUẬT PHẦN MỀM**

**PHAN QUỐC PHONG - 1612498**

**PHẠM NGỌC SƠN - 1612565**

**ĐỊNH GIÁ BẤT ĐỘNG SẢN**

**VÀ TRỰC QUAN HÓA DỮ LIỆU**

**BẤT ĐỘNG SẢN Ở VIỆT NAM**

**KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**

**TP. HỒ CHÍ MINH, 2020**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KỸ THUẬT PHẦN MỀM**

**PHAN QUỐC PHONG - 1612498**

**PHẠM NGỌC SƠN - 1612565**

**ĐỊNH GIÁ BẤT ĐỘNG SẢN**

**VÀ TRỰC QUAN HÓA DỮ LIỆU**

**BẤT ĐỘNG SẢN Ở VIỆT NAM**

**KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**

**GIÁO VIÊN HƯỚNG DẪN**

**ThS. PHẠM HOÀNG HẢI**

**ThS. NGUYỄN HUY KHÁNH**

**NIÊN KHÓA 2016-2020**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Tp.HCM, ngày ... tháng.…. năm …...

Giáo viên hướng dẫn

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN**

Tp.HCM, ngày …. tháng …. năm ….

Giáo viên phản biện

**LỜI CẢM ƠN**

Trong quá trình thực hiện khóa luận tốt nghiệp, chúng em đã có cơ hội được tìm hiểu những kiến thức mới, vận dụng những kiến thức đã học cũng như rút ra cho riêng mình những kinh nghiệm và bài học quý giá.

Chúng em xin chân thành cảm ơn các thầy cô trong Khoa Công Nghệ Thông Tin, trường Đại Học Khoa Học Tự Nhiên, Tp.HCM đã tạo điều kiện tốt nhất cho chúng em thực hiện khóa luận tốt nghiệp này.

Chúng em xin gửi lời tri ân sâu sắc tới thầy Nguyễn Huy Khánh và thầy Phạm Hoàng Hải. Hai thầy là những người đã trực tiếp hướng dẫn, tận tình giải đáp thắc mắc và đưa ra những định hướng giúp chúng em trong quá trình thực hiện khóa luận tốt nghiệp.

Chúng em cũng xin gửi lời cảm ơn sâu sắc đến quý Thầy Cô đã tận tình giảng dạy, trang bị cho chúng em những kiến thức nền tảng vững chắc và quý báu trong những năm học vừa qua.

Chúng em xin gửi lòng biết ơn sâu sắc đến ba, mẹ, các anh chị và bạn bè đã giúp đỡ và động viên tinh thần lẫn vật chất để vượt qua những lúc khó khăn trong suốt quãng thời gian học tập và nghiên cứu.

Mặc dù chúng em đã cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép, nhưng chắc chắn sẽ không tránh khỏi những thiếu sót, kính mong sự cảm thông và tận tình chỉ bảo của quý Thầy Cô và các bạn.

Nhóm sinh viên thực hiện

1612498 – Phan Quốc Phong

1612565 – Phạm Ngọc Sơn

**ĐỀ CƯƠNG CHI TIẾT**

|  |  |
| --- | --- |
| **Tên đề tài**: | |
| **Giáo viên hướng dẫn:** | |
| **Thời gian thực hiện:** | |
| **Sinh viên thực hiện:** | |
| **Loại đề tài:** | |
| **Nội dung đề tài**: | |
| **Kế hoạch thực hiện:** | |
| **Xác nhận của GVHD** | **Ngày … tháng … nằm 2020**  **SV thực hiện** |

Tóm tắt

Ở Việt Nam, bất động sản đang là một lĩnh vực thu hút rất nhiều người tham gia. Thị trường này đem lại rất nhiều cơ hội rất để người tham gia có được nguồn thu nhập rất cao, nhưng cũng đem lại không ít rủi ro khiến nhiều người nhiều thua lỗ hoặc phá sản. Đề tài ra đời với mục tiêu xây dựng một mô hình định giá bất động sản để cung cấp thêm một kênh thông tin tham khảo có giá trị cho ngươi dùng. Mục tiêu thứ hai chính là trực quan hóa các dữ liệu liên quan thành các thông tin tiện ích để người dùng có thể phân tích được những lợi ích khi sở hữu bất động sản này. Nhóm thực hiện luận văn đã áp dụng các phương pháp học máy để xây dựng mô hình định giá và xây dựng một ứng dụng web để trực quan hóa các thông tin tiện ích. Kết quả trước mắt đã thành công xây dựng mô hình định giá cho nhà riêng [[1]](#footnote-1) và nhà mặt phố tại thành phố Hồ Chí Minh. Đồng thời cũng hoàn thiên ứng dụng web hiển thị các thông tin tiện ích có tính tương tác cao.

**Mục lục**

[Chương 1. Giới thiệu 1](#_Toc46160345)

[**1.1 Đặt vấn đề** 1](#_Toc46160346)

[1.1.1 Sức hút của đầu tư bất động sản 1](#_Toc46160347)

[1.1.2 Định giá bất động sản ở Việt Nam 1](#_Toc46160348)

[1.1.3 Những yếu tố ảnh hưởng giá của bất động sản 3](#_Toc46160349)

[**1.2 Mục tiêu** 4](#_Toc46160350)

[1.2.1 Xây dựng mô hình định giá 4](#_Toc46160351)

[1.2.2 Trực quan hóa dữ liệu bất động sản 5](#_Toc46160352)

[**1.3 Cách tiếp cận** 5](#_Toc46160353)

[1.3.1 Áp dụng trí tuệ nhân tạo trong định giá bất động sản 5](#_Toc46160354)

[1.3.2 Xây dựng ứng dụng web để trực quan hóa dữ liệu 6](#_Toc46160355)

[**1.4 Đóng góp** 6](#_Toc46160356)

[Chương 2. Khảo sát hiện trạng 8](#_Toc46160357)

[**2.1 Xây dựng mô hình giá** 8](#_Toc46160358)

[2.1.1 Cơ sở lý thuyết 8](#_Toc46160359)

[2.1.2 Các công trình liên quan 11](#_Toc46160360)

[**2.2 Ứng dụng trực quan hóa dữ liệu** 12](#_Toc46160361)

[2.2.1 Các ứng dụng tương tự: 12](#_Toc46160362)

[2.2.2 Nhận xét và đánh giá: 13](#_Toc46160363)

[2.2.3 Các framework và thư viện sẽ sử dụng: 14](#_Toc46160364)

[Chương 3. Phương pháp định giá 16](#_Toc46160365)

[**3.1 Dữ liệu bất động sản** 16](#_Toc46160366)

[3.1.1 Nguồn gốc và cấu trúc của dữ liệu 16](#_Toc46160367)

[3.1.2 Các kết luận quan trọng 19](#_Toc46160368)

[3.1.3 Tiền xử lý dữ liệu 25](#_Toc46160369)

[3.1.4 Tỉ lệ huấn luyện và kiểm tra 29](#_Toc46160370)

[**3.2 Độ đo để đánh giá** 30](#_Toc46160371)

[3.2.1 MSE (Mean squared error) 30](#_Toc46160372)

[3.2.2 RMSE (Root mean squared error) 30](#_Toc46160373)

[3.2.3 MAE (Mean Absolute error) 30](#_Toc46160374)

[3.2.4 R2 và R2 hiệu chỉnh 30](#_Toc46160375)

[**3.3 Mô hình học máy** 32](#_Toc46160376)

[3.3.1 Ngôn ngữ, Công cụ, môi trường 32](#_Toc46160377)

[3.3.2 Linear Regression 32](#_Toc46160378)

[3.3.3 Random Forest 34](#_Toc46160379)

[3.3.4 Mạng Nơ ron nhân tạo 35](#_Toc46160380)

[**3.4 Tối ưu mô hình học máy** 39](#_Toc46160381)

[3.4.1 Chọn lọc dữ liệu theo khoảng giá 39](#_Toc46160382)

[3.4.2 Hiệu chỉnh các thông số của mạng kỹ nơ ron 40](#_Toc46160383)

[Chương 4. Ứng dụng trực quan hóa dữ liệu 41](#_Toc46160384)

[**4.1 Xác định các chức năng cần thiết** 41](#_Toc46160385)

[4.1.1 Tìm hiểu nhu cầu thực tế 41](#_Toc46160386)

[4.1.2 Khảo sát từ các sản phẩm tương tự 42](#_Toc46160387)

[4.1.3 Phân tích cơ sở dữ liệu 42](#_Toc46160388)

[4.1.3 Các chức năng có thể xây dựng 44](#_Toc46160389)

[**4.2 Kiến trúc hệ thống** 45](#_Toc46160390)

[**4.3 Vai trò của các thành phần** 47](#_Toc46160391)

[4.3.1 Cở sở dữ liệu 47](#_Toc46160392)

[4.3.2 Api Server 47](#_Toc46160393)

[4.3.3 Machine Learning Server 47](#_Toc46160394)

[4.3.4 Web Appilcation 47](#_Toc46160395)

[**4.3 Giao tiếp giữa các thành phần** 48](#_Toc46160396)

[Chương 5. Kết quả 49](#_Toc46160397)

[**5.1 Kết quả xây dựng mô hình định giá** 49](#_Toc46160398)

[5.1.1 Sau khi lọc giữ liệu 49](#_Toc46160399)

[5.1.2 Hiệu chỉnh các thông số mạng nơ ron 49](#_Toc46160400)

[5.1.3 Kết quả cuối cùng 52](#_Toc46160401)

[5.1.2 Thảo luận 53](#_Toc46160402)

[**5.2 Kết quả xây dựng ứng dụng trực quan hóa dữ liệu** 56](#_Toc46160403)

[5.2.1 Danh sách các chức năng 56](#_Toc46160404)

[5.2.2 Minh họa các chức năng 57](#_Toc46160405)

[Chương 6. Kết luận 58](#_Toc46160406)

[**6.1 Mục tiêu** 58](#_Toc46160407)

[**6.2 Phương pháp** 58](#_Toc46160408)

[6.2.1 Phương pháp xây dựng mô hình định giá 58](#_Toc46160409)

[6.2.2 Phương pháp xây dựng ứng dụng trực quan hóa 59](#_Toc46160410)

[**6.3 Kết quả đạt được** 59](#_Toc46160411)

[**6.4 Những điểm hạn chế** 60](#_Toc46160412)

[**6.5 Hướng phát triển** 60](#_Toc46160413)

[Chương 7. Danh mục tài liệu tham khảo 62](#_Toc46160414)

[Phụ lục các chức năng của ứng dụng I](#_Toc46160415)

**Danh mục bảng**

[Bảng 1: Thông tin các căn nhà 9](#_Toc46131789)

[Bảng 2: Thông tin căn nhà mới 9](#_Toc46131790)

[Bảng 3: Một số trường quan trọng của một tin rao 17](#_Toc46131791)

[Bảng 4: Các trường được lựa chọn cho mô hình bán nhà mặt phố và bán nhà riêng 24](#_Toc46131792)

[Bảng 5: Các trường được lựa chọn của mô hình bán đất và bán đất nền 25](#_Toc46131793)

[Bảng 6: Kết quả thử nghiệm thuật toán Linear Regression đối với trường hợp bán nhà riêng 33](#_Toc46131794)

[Bảng 7: Thông số tốt nhất của thuật toán Random Forest cho mô hình bán nhà riêng 34](#_Toc46131795)

[Bảng 8: Kết quả thử nghiệm thuật toán Random Forest cho mô hình bán nhà riêng 35](#_Toc46131796)

[Bảng 9: Tham số ban đầu cho mạng nơ ron 36](#_Toc46131797)

[Bảng 10: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán nhà riêng 36](#_Toc46131798)

[Bảng 11: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán nhà mặt phố 37](#_Toc46131799)

[Bảng 12: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán đất 37](#_Toc46131800)

[Bảng 13: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán đất nền 37](#_Toc46131801)

[Bảng 14: Kết quả mô hình bán nhà riêng sau khi lọc dữ liệu 49](#_Toc46131802)

[Bảng 15: Kết quả mô hình bán nhà mặt phố sau khi lọc dữ liệu 49](#_Toc46131803)

[Bảng 16: Các thông số hiệu chỉnh và kết quả tương ứng của mô hình bán nhà mặt phố 51](#_Toc46131804)

[Bảng 17: Các thông số hiệu chỉnh và kết quả tương ứng của mô hình bán nhà riêng 52](#_Toc46131805)

[Bảng 18: Kết quả cuối cùng của mô hình bán nhà mặt phố 53](#_Toc46131806)

[Bảng 19: Kết quả cuối cùng của mô hình bán nhà riêng 53](#_Toc46131807)

**Danh mục hình ảnh**

[Hình 1: Cấu trúc mạng nơ ron đơn giản [1] 10](#_Toc46161367)

[Hình 2: Số lượng tin rao phân bố theo tỉnh thành năm 2018 - 2019 21](#_Toc46161368)

[Hình 3: Sự phân bố các tin rao theo loại giao dịch năm 2018-2019 22](#_Toc46161369)

[Hình 4: Ouliers trong tập dữ liệu bất động sản năm 2018 – 2019 26](#_Toc46161370)

[Hình 5: Đồ thị phân phối các thuộc tính Lat (vĩ độ), Lng (kinh độ) của các tin rao bán nhà mặt phố trong thời gian 2018-2019 29](#_Toc46161371)

[Hình 6: Biểu đồ phân tán giữa diện tích (Area) và giá (Price) 33](#_Toc46161372)

[Hình 7: Một góc biểu đồ phân phối tần suất theo giá của tập dữ liệu bán đất năm 2018-2019 39](#_Toc46161373)

[Hình 8: Kiến trúc toàn hệ thống 45](#_Toc46161374)

[Hình 9: Màn hình chức năng đăng ký tài khoản 64](#_Toc46161375)

[Hình 10: Màn hình chức năng đăng nhập 65](#_Toc46161376)

[Hình 11: Màn hình chọn loại bất động sản 66](#_Toc46161377)

[Hình 12: Màn hình chọn mục đích định giá 66](#_Toc46161378)

[Hình 13: Màn hình tìm kiếm thông qua vị trí và thể hiện thông tin quy hoạch bất động sản 67](#_Toc46161379)

[Hình 14: Màn hình nhập thông tin bất động sản 67](#_Toc46161380)

[Hình 15: Màn hình hiển thị thông tn cơ bản 68](#_Toc46161381)

[Hình 16: Màn hình kết quả định giá 69](#_Toc46161382)

[Hình 17: Bảnh điều hướng 69](#_Toc46161383)

[Hình 18: Màn hình chức năng tìm đường đi 70](#_Toc46161384)

[Hình 19: Chức năng lịch sử giá của khu vực 71](#_Toc46161385)

[Hình 20: Biểu đồ cơ cấu thị trường bất động sản trong khu vực 72](#_Toc46161386)

[Hình 21: Biểu đồ máu thể hiện giá trong khu vực 73](#_Toc46161387)

[Hình 22: Màn hình thông tin khu vực xung quanh 74](#_Toc46161388)

[Hình 23: Màn hình danh sách bất động sản tương tự 75](#_Toc46161389)

[Hình 24: Màn hình nhập liệu khi tìm kiếm bất động sản 75](#_Toc46161390)

[Hình 25: Màn hình kết quả tìm kiếm 76](#_Toc46161391)

[Hình 26: Các lựa chọn trên bản đồ 77](#_Toc46161392)

[Hình 27: Hiển thị các kết quả tìm kiếm và ranh giới khu vực 77](#_Toc46161393)

[Hình 28: Hiển thị kết quả tìm kiếm trên bản đồ màu 78](#_Toc46161394)

**Chương 1. Giới thiệu**

## **1.1 Đặt vấn đề**

### **1.1.1 Sức hút của đầu tư bất động sản**

Trong văn hoá của người Việt Nam, có một tư tưởng đặc trưng đã được truyền qua nhiều thế hệ đó là “an cư lạc nghiệp”. Một người chưa có được nơi ăn chốn ở, thường sẽ bị đánh giá là vẫn còn bấp bênh, chưa ổn định. Tuy rằng tư tưởng này có phần không phù hợp với xã hội ngày nay, ta không thể phủ nhận rằng đối với rất nhiều người, có được một căn nhà, mảnh đất cho bản thân là một điều tất yếu phải có trong cuộc sống.

Bên cạnh đó, bất động sản còn được sử dụng như một phương tiện để đầu tư đem lại lợi nhuận thông qua các hoạt động kinh doanh bất động sản như: bán, cho thuê, môi giới, tư vấn, … Kênh đầu tư này luôn đem lại cơ hội để người tham gia có được thu nhập cao: mua rẻ bán đắt, cho thuê nhà ở, cho thuê mặt bằng, cải tạo để tăng giá trị tài sản. Người sử dụng không phải mất nhiều chi phí để vận hành so với các loại tài sản khác. Hơn nữa, bất động sản có thể được sử dụng để làm tài sản thế chấp trong khi vẫn được giữ và sử dụng tài sản đó.

### **1.1.2 Định giá bất động sản ở Việt Nam**

Tại Việt Nam, có nhiều phương pháp để những người quan tâm có thể định giá một bất động sản. Có thể kể đến một số phương pháp sau đây.

* Thẩm định giá qua bảng giá đất địa phương

Giá trị bất động sản còn được ước lượng dựa trên bảng giá đất do cơ quan hành chính tỉnh hoặc thành phố đó ban hành. Thông thường, bảng giá được ban hành sẽ quy định cụ thể cách phân chia khu vục, vị trí, cách tính đối với từng loại đất. Tuy nhiên trong những năm gần đây, con sốt đất diễn ra tại nhiều địa phương khiến cho giá đất trên thực tế tăng ảo lên rất nhiều lần tùy thuộc vào khu vực.

* Thẩm định giá bất động sản dựa vào sự giúp đỡ của môi giới:

Kinh nghiệm của các nhà môi giới bất động sản là một điều không thể bỏ qua khi ta muốn thẩm định giá bất động sản của mình. Một người môi giới chuyên nghiệp và đủ tốt sẽ cho ta thông tin chính xác trong thời gian nhanh nhất. Thực tế trong nhiều năm gần đây, rất nhiều người không muốn thuê môi giới trong việc mua bán nhà đất bởi không thực sự tin tưởng bên thứ ba này. Tuy nhiên, không thể phủ nhận rằng rất nhiều giao dịch mua bán nhà đất thành công là nhờ dịch vụ này.

* Thẩm định giá qua giá bất động sản tương tự:

Ta có thể đóng vai một người có nhu cầu mua nhà, đất, đi khảo sát các căn nhà, đất đăng biển rao bán tại khu vực của mình. Tuy nhiên việc này sẽ tốn rất nhiều thời gian và công sức. Với sự phát triển của internet, người dân có thể tra cứu thông tin về giá nhà đất xung quanh khu vực của mình thông qua các sàn giao dịch bất động sản, hoặc kênh thông tin bất động sản khác nhau.

* Thông qua ngân hàng để định giá:

Cách thức đem lại sự yên tâm cho người sở hữu vì ngân hàng khá chặt chẽ trong việc thẩm định giá bất động sản để cho vay. Tuy nhiên cách thức này thường mất nhiều thời gian bởi phải thực hiện theo một quy trình nhiều bước của riêng ngân hàng đó. Người dùng có thể vụt mất cơ hội khi cần đưa ra quyết định trong thời gian ngắn.

Có thể thấy rằng có rất nhiều phương pháp để ta định giá một bất động sản. Nhưng một điều mà tất cả người tham gia vào lĩnh vực này đều nhất trí là không bao giờ được phụ thuộc vào một cách tính hoặc một nguồn thông tin nào cả mà phải kết hợp tất cả thông tin liên quan, các kết quả định giá khác nhau để đi đến một kết luận cuối cùng.

Những phương pháp kể trên có thể xem là những phương pháp truyền thống, đã được áp dụng lâu dài và đem đến những hiệu quả không tệ. Nhưng ngoài những phương pháp truyền thống này, người tham gia có thể có phương pháp nào khác hay không? Một phương pháp nào đó nhanh chóng hơn, có thể tin cậy được. Một phương pháp nào đó mang đậm hơi thở hiện đại của cuộc cách mạng công nghiệp 4.0.

### **1.1.3 Những yếu tố ảnh hưởng giá của bất động sản**

Có được kết quả định giá một bất động sản chưa phải là điểm kết thúc. Bất kỳ ai cũng cần làm rõ những câu hỏi: Căn nhà / mảnh đất này có thực sự tương xứng với số tiền bỏ ra hay không? Điều gì làm cho nó có giá đắt/ rẻ như vậy?

Để nhận định chính xác giá trị của một bất động sản cần xem xét đến rất nhiều yếu tố khác nhau. Quan hệ cung cầu trên thị trường tất nhiên là một yếu tố then chốt. Nhưng ngoài yếu tố đó, vẫn còn nhiều yếu tố ta phải xem xét đến: tự nhiên, kinh tế xã hội, …

Vị trí: có thể xem là yếu tố ảnh hưởng nhất. Những bất động sản nằm tại trung tâm đô thị, cụm công nghiệp, địa phương du lịch trọng điểm, … sẽ có giá trị cao hơn những bất động sản cùng loại nằm ở các vùng ven trung tâm. Ở phạm vi hẹp hơn, những bất động sản nằm tại giao lộ các con đường, trên các trục đường giao thông sẽ có giá trị lớn hơn những bất động sản nằm trong hẻm.

Kích thước, diện tích: Tất nhiên diện tích lớn hơn sẽ thường có giá cao hơn. Kích thước mảnh đất, căn nhà nào vuông vắn, mặt tiền lớn sẽ có ưu thế hơn.

Khả năng mang lại thu nhập: mức thu nhập hay giá trị lợi nhuận hàng năm mang lại có ảnh hưởng quan trọng đến giá trị của bất động sản đó. Những căn nhà nằm trong khu vực có giá nhà tăng tăng trưởng so với các năm trước sẽ đáng để sở hữu hơn với một nhà đầu tư. Điều này cũng giống như một mảnh đất bạn mua và bán được với giá tốt, sau đó người kế tiếp lại bán với giá tốt và cứ vậy giá mảnh đất đó tăng lên.

Ngoài các yếu tố đó ra thì an ninh khu vực, bệnh viện, trường học cũng rất được chú trọng khi một người xem xét đầu tự hoặc sinh sống tại một khu vực nào đó. Và còn rất rất nhiều các yếu tố khác có thể xem xét đến.

Những yếu tố kể trên cũng là những yếu tố để một người sở hữu bất động sản đánh giá số tiền mình bỏ ra có xứng đáng hay không. Hơn nữa, người dùng còn có xu hướng so sánh giá cả giữa các khu vực để lựa chọn ra những mục tiêu khả dĩ nhất.

Có thể nói những thông tin đáng thể người quan tâm phân tích là rất nhiều. Và những thông tin đó là kết quả quả một quá trình phân tích, tính toán dựa trên các thông tin thô sơ chứ không hề có. Việc trình bày các thông tin này so cho trực quan, rõ ràng, dễ hiểu, dễ truyền đạt là một thách thức không hề nhỏ.

## **1.2 Mục tiêu**

### **1.2.1 Xây dựng mô hình định giá**

Trong những năm gần đây, cụm từ “cách mạng công nghiệp 4.0” được nhắc đi nhắc lại rất nhiều tại Việt Nam. Dưới sự phát triển bùng nổ của thời đại Internet, công nghiệp 4.0 trong thời đại này đã tạo ra những phát minh vượt bậc. Những phát minh đó, ta không thể không nói đến trí tuệ nhân tạo.

Trí tuệ nhân tạo sẽ được ứng dụng thế nào trong vấn đề định giá bất động sản. Từ các thông tin về các hoạt động trong thị trường bất động sản, liệu ta có thể dạy cho máy tính học được các quy tắc trong đó và đưa ra các dự đoán về giá cả hay không? Từ đó cung cấp thêm cho người dùng một nguồn tham khảo về giá đáng tin cậy. Điều này sẽ đem lại rất nhiều lợi ích cho những người tham gia vào thị trường bất động sản. Họ có thể tiết kiệm được thời gian và công sức cũng như có được một cơ sở về giá để bắt đầu tham khảo và kết hợp với các nguồn thông tin khác.

Đây cũng chính là mục tiêu đầu tiên mà nhóm thực hiện luận văn muốn hướng tới.

### **1.2.2 Trực quan hóa dữ liệu bất động sản**

Để có thể truyền đạt những thông tin tiện ích về bất động sản cho người dùng, ta cần trực quan hóa nó thành các biểu đồ, đồ thị, bảng biểu, … Con người thường nhạy cảm với các hình ảnh hơn là các đoạn văn bản khô khan. Việc đưa ra các biểu đồ, đồ thị sẽ giúp người dùng nhanh chóng nắm bắt được thông tin, dễ dàng so sánh giữa các đối tượng, phát hiện ra các điểm bất thường, xu hướng biến động của đối tượng đang xem xét,… Lợi ích của việc trực quan hóa dữ liệu là không thể phủ nhận.

Với tính chất của đề tài, đòi hỏi nhóm phải việc kết hợp kết quả của mô hình định giá và các biểu đồ, đồ thị thông tin tiện ích. Qua quá trình phần tích, nhóm kết luận rằng việc xây dựng một ứng dụng Web để thể hiện các thông tin trên là một đòi hỏi tất yếu.

## **1.3 Cách tiếp cận**

### **1.3.1 Áp dụng trí tuệ nhân tạo trong định giá bất động sản**

Nhóm thực hiện luận văn đã tiến hành tìm hiểu về trí tuệ nhân tạo, học máy, học sâu để có những hiểu biết cơ bản về đối tượng nghiên cứu, các bài toán mà những phương pháp này có thể giải quyết. Đồng thời phân tích xem bản chất của bài toán định giá bất động sản dựa trên các thông tin về đặc điểm là gì. Nhóm kết luận rằng có thể dùng các phương pháp học máy để giải quyết bài toán đặt ra.

Sau khi có được kết luận quan trọng đó, nhóm tìm hiểu cách xử lý dữ liệu thu thập được về thị trường bất động sản để có được tập dữ liệu phù hợp. Đồng thời, nhóm phân tích các điều kiện cần và đủ có thể sử dụng các kỹ thuật học máy đối với vấn đề đặt ra thông qua cách các bài toán tương tự được giải quyết. Qua quá trình phần tích các phương pháp Linear Regression (hồi quy tuyến tính), Random Forest (rừng ngẫu nhiên), artificial neural network (mạng nơ ron nhân tạo, nhóm kết luận rằng giải pháp sử dụng mạng nơ-ron nhân tạo là phù hợp nhất với các tập cơ sở dữ liệu đã có.

### **1.3.2 Xây dựng ứng dụng web để trực quan hóa dữ liệu**

Trước tiên nhóm tìm hiểu nhu cầu thực tế của những người tham gia vào thị trường bất động sản. Sau đó nhóm tiến hành khảo sát các sản phẩm tương tự trên thị trường. Nhóm chọn ra hai sản phẩm nổi bật trong nhóm các sản phẩm tương tự là Property value và Zillow để tiến hành phân tích ưu nhược điểm. Sau khi phân tích ưu nhược điểm, nhóm phác thảo ra được các chức năng phù hợp với thị trường Việt Nam và có thể xây dựng được. Đồng thời, tìm cách bổ sung các thiếu sót của Property value và Zillow trên để sản phẩm hoàn thiện hơn. Sau khi phác thảo các chức năng cần thiết, nhóm tiến hành xây dựng ứng dụng web bằng các công nghệ sau: Front end sử dụng React, Back end sử dụng NodeJS + Express, cơ sở dữ liệu là MongoDB.

Song song với quá trình xây dựng ứng dụng còn là quá trình tìm kiếm các nguồn dữ liệu để có thể bổ sung thêm các chức năng trực quan hóa dữ liệu cũng như xem xét các chức năng được lên kế hoạch ban đầu có còn hợp lý để xây dựng hay không. Trải qua một quá tình dài xây dựng ứng dụng, đã có những chức năng được hoàn thành, những chức năng bị lược bỏ và cũng có những chức năng được bổ sung.

## **1.4 Đóng góp**

Xây dựng được hai mô hình định giá dựa trên kỹ thuật mạng nơ ron nhân tạo với sai số chấp nhận được:

* Định giá bán nhà mặt phố tại thành phố Hồ Chí Minh
* Định giá bán nhà riêng (không phải nhà mặt phố) tại thành phố Hồ Chí Minh

Xây dựng được ứng dụng trực quan hóa dữ liệu liên quan để truyền đạt các thông tin hữu ích ích:

* Vị trí trên bản đồ quy hoạch thành phố
* Các công trình công cộng: trường học, bệnh viện, nhà hàng, …
* Biểu đồ tăng trưởng giá trong khu vực.
* Bản đồ nhiệt giá trung bình trong thành phố.
* So sánh tăng trưởng giá của các địa phương khác nhau.
* Cảnh quan xung quanh căn nhà.

**Chương 2. Khảo sát hiện trạng**

## **2.1 Xây dựng mô hình giá**

### **2.1.1 Cơ sở lý thuyết**

**2.1.1.1 Học máy**

Học máy là quá trình giúp cho máy tính có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.

Dựa trên phương thức học, thông thường ta có thể phân loại các thuật toán trong học máy thành 2 nhóm:

Học có giám sát:  thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là (*data, label*), tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Học không giám sát: chúng ta không biết được *outcome* hay *nhãn* mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

Ngoài hai loại trên, trong một số tài liệu còn đề cập tới 2 nhóm khác là Học bán giám sát và học củng cố.

Mục tiêu cần đạt được chính là định giá một bất động sản có dữ liệu đầu vào là các đặc trưng của bất động sản đó như: chiều rộng, diện tích, vị trí địa lý, … Đây chính là đặc điểm của bài toán hồi quy (regression) trong học máy.

Từ tập dữ liệu gồm các thông tin về đặc điểm bất động sản và giá cả tương ứng, chúng ta tạo ra một chương trình có khả năng ước lượng giá cả một căn hộ dựa trên các thông tin mà ta cung cấp. Giả sử rằng ta có một tập dữ liệu gồm thông tin về một số căn hộ mặt tiền ở quận Thủ Đức, thành phố Hồ Chí Minh như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Diện tích (m2) | Số tầng | Mặt tiền (m) | Giá (tỷ) |
| 42 | 2 | 4 | 3,5 |
| 35 | 1 | 4 | 2,6 |
| 32 | 2 | 8 | 5 |

**Bảng 1: Thông tin các căn nhà**

Sử dụng dữ liệu này, chúng ta cần tạo ra một chương trình có thể ước lượng giá của bất kỳ căn hộ nào khác trong khu vực Thủ Đức: ví dụ:

|  |  |  |  |
| --- | --- | --- | --- |
| Diện tích (m2) | Số tầng | Mặt tiền (m) | Giá (tỷ) |
| 35 | 2 | 5 | ??? |

**Bảng 2: Thông tin căn nhà mới**

Để xây dựng chương trình này, ta áp dụng thuật toán học máy vào tập dữ liệu này. Quá trình này sẽ cố gắng tìm ra những phép toán cần thiết để tính toán ra kết quả cuối cùng. Thuật toán thường được sử dụng cho bài toán này là Linear Regression, Random Forest, …

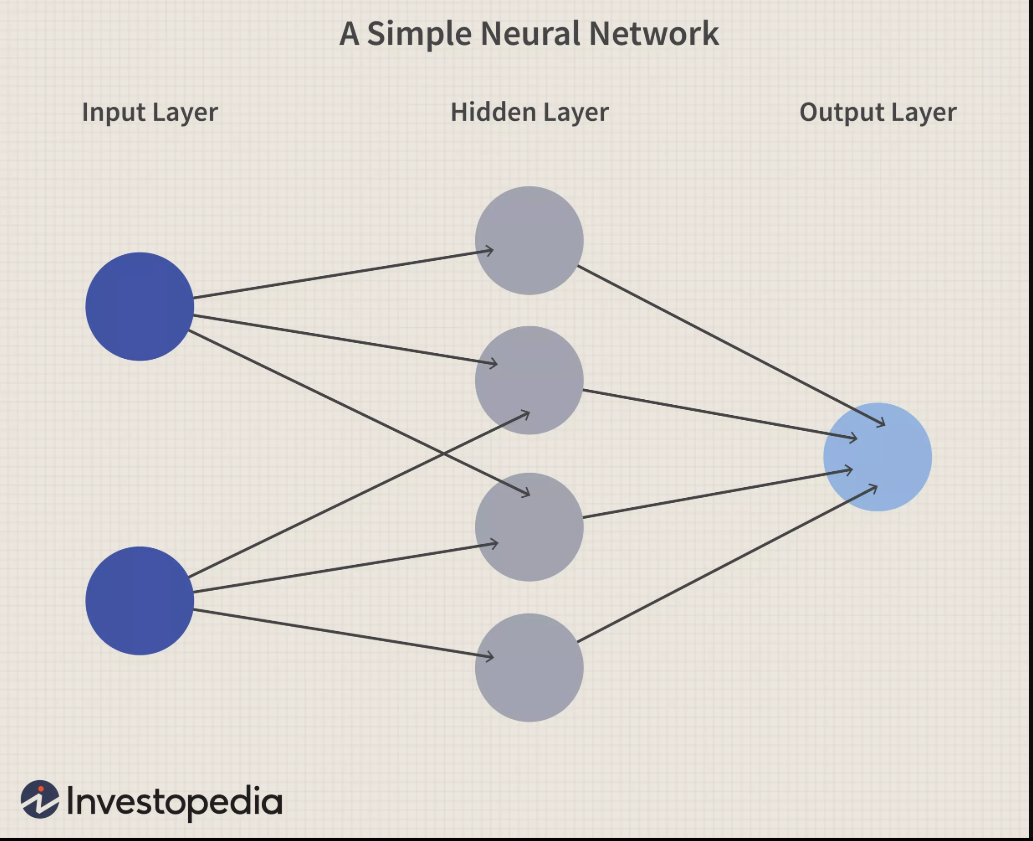
**2.1.1.2 Học sâu**

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, học máy đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là học sâu ( deep Learning ).

Học sâu là một tập hợp con của Machine Learning, có khả năng khác biệt ở một số khía cạnh quan trọng so với Machine Learning truyền thống, cho phép máy tính giải quyết một loạt các vấn đề phức tạp không thể giải quyết được.

Một ví dụ về một nhiệm vụ Machine Learning đơn giản có thể dự đoán doanh số bán kem sẽ thay đổi như thế nào dựa trên nhiệt độ ngoài trời (Machine learning regression masterclass in python). Việc đưa ra dự đoán chỉ sử dụng một vài tính năng dữ liệu theo cách này là tương đối đơn giản và có thể được thực hiện bằng cách sử dụng kỹ thuật Linear Regression với độ dốc giảm dần.

Học sâu lấy cảm hứng từ mạng nơ ron thần kinh của con người. Có 3 loại layer chính một mạng nơ ron là: Input layer, Các hidden layer, Output layer



**Hình 1: Cấu trúc mạng nơ ron đơn giản** [1]

Input layer nhận các dữ liệu đầu vào. Trong trường hợp của chúng ta là các thông tin về bất động sản như: diện tích, số phòng, chiều rộng, vị trí địa lý, …. Input layer sẽ đưa các đầu vào này vào hidden layer thứ nhất.

Các hidden layer thực hiện các phép tính toán cho các đầu vào.

Output layer trả về dữ liệu đầu ra, trường hợp của ta sẽ là đưa ra dự đoán về giá trị của bất động sản.

Mỗi một kết nối giữa nơ ron được liên kết với một trọng số (weight). Trọng số này chỉ rõ ra tầm quan trọng của giá trị đầu vào. Trọng số khởi tạo được chọn ngẫu nhiên. Mỗi một nơ ron sẽ có một Hàm kích hoạt ([Activation Function](https://en.wikipedia.org/wiki/Activation_function)) là một hàm phi tuyến. Những hàm phi tuyến giúp hệ thống giải quyết những quan hệ phi tuyến phức tạp giữa các đặc điểm của dữ liệu và kết quả đầu ra của mô hình.

### **2.1.2 Các công trình liên quan**

Từ lâu, bài toán định giá nhà dựa trên tập dữ liệu về giao dịch các giao dịch bất động sản trong quá khứ đã được sử dụng là bài tập các khóa học về học máy. Trên các nền tảng học trực tuyến Coursera và Udemy đều có các khóa học có các bài tập dự đoán giá nhà sử dụng các thuật toán regression hoặc mạng nơ ron nhân tạo [2] [3] [4]

Đồng thời cũng đã có nhiều luận văn trên thế giới nghiên cứu về việc ứng dụng các phương pháp học máy vào giải quyết bài toán định giá nhà [5] [6] [7].

Ngoài ra, kaggle – một công đồng trực tuyến nổi bật dành cho các nhà khoa học dữ liệu, cũng đã tổ chức nhiều cuộc thi có nội dung về dự đoán giá nhà bằng các phương pháp học máy. Lại phải nói thêm, ứng dụng web “Property value” [8] mà nhóm khảo sát cho mục tiêu xây dựng ứng dụng trực quan hóa cũng có chức năng định giá một bất động sản (ứng dụng hoạt động tại Úc). Chức năng này được xây dựng dựa trên các kỹ thuật học máy. Có thể thấy rằng, lĩnh vực này đã nghiên cứu và ứng dụng trên thực tế cũng như trong giảng dạy khá nhiều.

Tuy nhiên, có một cần chú ý đến là thị trường bất động sản Việt Nam có sự khác biệt các thị trường bất động sản nước ngoài. Hơn nũa, nguồn gốc và đặc điểm của cở sở dữ liệu mà nhóm sử dụng cũng có sự khác biệt với các công trình, ứng dụng trên. Những sự khác biệt này có sự ảnh hưởng lớn đến mục đích và kết quả của nhóm thực hiện luận văn. Các nội dung này sẽ được trình bày chi tiết ở chương 3 và chương 4.

## **2.2 Ứng dụng trực quan hóa dữ liệu**

### **2.2.1 Các ứng dụng tương tự:**

**2.2.1.1 Property value** [8]

Property value là sản phẩm của CoreLogic (trước đây là RP Data). CoreLogic là nhà cung cấp thông tin bất động sản, phân tích và dịch vụ quản lý rủi ro liên quan đến bất động sản lớn nhất tại Úc và New Zealand. Property value cung cấp dữ liệu để cho khách hàng có cái nhìn sâu sắc hơn đối với một bất động sản để người dùng có thể đưa ra các quyết định sáng suốt hơn tùy vào mục đích.

Ưu điểm:

* Giao diện người dùng thân thiện.
* Thể hiện đầy đủ các thông tin cơ bản cần có của một bất động sản
* Trực quan hóa các thông tin về trường học đầy đủ, đa dạng.
* Cho phép so sánh với các căn hộ khác theo nhiều tiêu chí: thuê, bán, đã bán.
* Các thông tin về khu vực đa dạng: thu nhập theo hộ gia đình, tình trạng nhà ở, độ tuổi và giới tính người trong khu vực, …

Nhược điểm:

* Thiếu các thông tin về công trình công cộng khác như công viên, bệnh viện, …
* Thiếu các thông tin về thị trường bất động sản trong khu vực
* Thiếu các biểu đồ so sánh giá cả với các khu vực xung quanh.

**2.2.1.2 Zillow**

Zillow [9] là thị trường hàng đầu về bán và cho thuê bất động sản với phương châm trao cho khách hàng dữ liệu và thông tin về nhà và khu vực xung quanh đó. Zillow phục vụ toàn bộ vòng đời của việc sở hữu và sinh sống trong một ngôi nhà: mua, bán, cho thuê, tu sửa và nhiều hơn nữa.

Ưu điểm:

* Giao diện người dùng thân thiện.
* Thể hiện đầy đủ các thông tin cơ bản cần có của một bất động sản
* Cung cấp các thông tin về thuế và lịch sử đóng thuế cũng như lịch sử giá
* Cung cấp thông tin về chi phí hàng tháng cho việc sinh sống
* Cung cấp giá nếu khách hàng muốn cho thuê.
* Có liệt kê trường học và một số thông tin liên quan như: khoảng cách, đánh giá (rating), liên kết đến trang chuyên đánh giá trường học.

Nhược điểm:

* Thiếu các thông tin về công trình công cộng khác như công viên, bệnh viện, …
* Không có sự so sánh về giá cả với khu vực xung quanh.
* Thiếu các thông tin về thị trường bất động sản trong khu vực

### **2.2.2 Nhận xét và đánh giá:**

Việc khảo sát 2 ứng dụng web trên giúp nhóm có cái nhìn sơ bộ về các chức năng cơ bản, những điểm mạnh cũng như hạn chế của các ứng dụng web đó, từ đó góp phần định hướng cho việc xây dựng các chức năng cần thiết cho đề tài.

Ngoài ra, nhóm cũng phân tích các thông tin có được từ tập dữ liệu thu thập được (sẽ được trình bày rõ ở chương 3) để xem xét có thể khắc phục một vài thiếu sót của các ứng dụng trên hay không.

Từ những gì đã phân tích, nhóm sẽ xây dựng các chức năng để trực quan hóa dữ liệu như sau:

* Vị trí trên bản đồ quy hoạch thành phố
* Các công trình công cộng: trường học, bệnh viện, nhà hàng, …
* Biểu đồ tăng trưởng giá trong khu vực.
* Bản đồ nhiệt giá trung bình trong thành phố.
* So sánh tăng trưởng giá của các địa phương khác nhau.
* Cảnh quan xung quanh căn nhà.

### **2.2.3 Các framework và thư viện sẽ sử dụng:**

**2.2.3.1 NodeJs và Express:**

Node.js [10] là một Javascript run-time environment mã nguồn mở và đa nền tảng, được xây dựng dựa trên engine V8 của Chrome. Node.js cho phép người sử dụng có thể sử dụng Javascript ở môi trường bên ngoài trình duyệt, từ đó có thể xây dựng ứng dụng web ở phía server. Ưu điểm của Node.js là tốc độ xử lí nhanh, xử lí được nhiều request đồng thời với cơ chế bất đồng bộ, khả năng mở rộng ứng dụng tốt hơn. Bên cạnh đó, hệ thống các thư viện hỗ trợ npm và cộng đồng sử dụng Node.js trên thế giới là rất lớn.

Express [11] là một framework xây dựng ứng dụng web sử dụng Node.js. Express cung cấp một bộ khung tương đối hoàn chỉnh để người dùng có thể phát triển ứng dụng web của mình một cách nhanh chóng và hiệu quả.

**2.2.3.2 Flask và Flask-RESTful**

Flask [12] là một microframework dùng để xây dựng ứng dụng web bằng ngôn ngữ Python. Ưu điểm của Flask là sự gọn nhẹ, đơn giản và nhanh chóng trong quá trình xây dựng ứng dụng web.

Flask RESTful [13] là một phần mở rộng cho Flask mà thêm hỗ trợ cho việc nhanh chóng xây dựng các API REST.

**2.2.3.3 Các thư viện hỗ trợ khác**

React: thư viện Javascript hỗ trợ trong việc xây dựng giao diện người dùng, được duy trì bởi Facebook. Thư viện hỗ trợ xây dựng ứng dụng single page theo hướng gồm nhiều component tách biệt, dễ quản lý, mở rộng và tái sử dụng.

Scikit-learn: thư viện Python dành cho máy học. Thư viện hỗ trợ việc xây dựng, huấn luyện các mô hình máy học với nhiều giải thuật đa dạng đồng thời hỗ trợ trong các công việc như đánh giá hiệu quả của mô hình theo các thông số, tiền xử lý dữ liệu, …

Pandas: thư viện Python chuyên biệt cho việc xử lý và phân tích dữ liệu. Thư viện cung cấp các công cụ đọc và ghi dữ liệu theo nhiều định dạng, hỗ trợ các dạng cấu trúc dữ liệu và hàm giúp cho việc xử lý các bộ dữ liệu nhanh chóng và hiệu quả hơn

Tensorflow: là một thư viện phần mềm mã nguồn mở hỗ trợ mạnh mẽ các phép toán học để tính toán trong machine learning và deep learning.

**Chương 3. Phương pháp định giá**

## **3.1 Dữ liệu bất động sản**

### **3.1.1 Nguồn gốc và cấu trúc của dữ liệu**

Nhóm thực hiện luận văn sử dụng dữ liệu gồm các tin rao bán hoặc thuê bất động sản trên thị trường bất động sản của Việt Nam từ năm 2015 đến 9 tháng đầu năm 2019 với số lượng tin rao xấp xỉ 9.4 triệu tin từ nhiều đối tượng đăng khác nhau (môi giới, chủ sở hữu). Từ phần này trở về sau, khi nói tới năm 2019, hàm ý nhóm muốn nhắc tới là 9 tháng đầu năm 2019.

Các trường quan trọng của một tin rao như sau:

|  |  |
| --- | --- |
| **Tên trường** | **Ý nghĩa** |
| CateId | Id của loại giao dịch: bán đất, bán nhà riêng, … |
| CityCode | Mã tỉnh thành |
| DistricId | Mã quận |
| CityName | Tên tỉnh thành |
| DistrictName | Tên quận |
| WardId | mã phường |
| StreetId | Mã đường |
| Price | Giá trị của bất động sản |
| Area | Diện tích của bất động sản |
| Rooms | Số phòng |
| Direction | Hướng |
| Floors | Số tầng |
| RoadIn | Độ dài đường vào nhà |
| Toilet | Số nhà vệ sinh |
| FrontEnd | Chiều dài mặt tiền |
| Lat | Vĩ độ của bất động sản trên bản đồ |
| Lng | Kinh độ của bất động sản trên bản đồ |

**Bảng 3: Một số trường quan trọng của một tin rao**

Ngoài ra còn một số các thuộc tính khác như: tên, số điện thoại liên lạc của người đăng, các thuộc tính phục vụ riêng cho truy vấn của ứng dụng web đó, …

Như đã trình bày ở chương hai, thị trường bất động sản Việt Nam có sự khác biệt các thị trường bất động sản nước ngoài và nguồn gốc, đặc điểm của cở sở dữ liệu mà nhóm sử dụng cũng có sự khác biệt với các công trình, ứng dụng đã khảo sát. Những sự khác biệt này có sự ảnh hưởng lớn đến mục đích và kết quả của nhóm thực hiện luận văn. Sau đây, ta đi vào làm rõ sự khác biệt này.

**3.1.1.2 Sự khác biệt của cơ sở dữ liệu**

Trước hết phải nói tới dữ liệu được sử dụng làm bài tập xây dựng mô hình định giá bất động sản. Các tập dữ liệu này thường xuất phát từ kaggle. Nói cách khác, các bài tập này, và các cuộc thi tại kaggle thường sẽ có chung nguồn dữ liệu. Các dữ liệu này được những người đóng góp thu thập từ các giao dịch bất động sản tại một địa phương nhất định. Đối với ứng dụng Property value, ứng dụng này sử dụng cơ sở dữ liệu của CoreLogic, một tập đoàn chuyên cung cấp các thông tin tài chính, tài sản và tiêu dùng. Những dữ liệu này là sự phản chiếu của các giao dịch thật được ghi nhận, nói cách khác, các dữ liệu này tương đối chính xác và có tính khách quan cao. Ngoài ra, những người đóng góp hoặc các chuyên gia phân tích đã xử lý các vấn đề của một tập dữ liệu thu thập như trùng lặp, ngoại lai, sai số, thiếu dữ liệu, … Các dữ liệu này đã được làm sạch để có thể huấn luyện.

Cơ sở dữ liệu mà nhóm sử dụng được tạo ra từ các tin rao của người đăng. Không có sự ràng buộc về tính chính xác của thông tin này. Nói cách khác, các thông tin này mang đậm tính chủ quan của người đăng và đặc biệt có độ chính xác không hề cao. Việc người dùng nhập liệu các thông tin này cũng tìm ẩn nhiều sai sót bên trong dữ liệu: thiếu giá trị các trường dữ liệu, nhập sai giá trị, trùng lặp tin đăng, ... Để có thể huấn luyện, cần phải xử lý kỹ dữ liệu này.

Tới đây, một câu hỏi được đặt ra là với một cơ sở dữ liệu “xấu” như vậy, liệu có thể xây dựng mô hình định giá bất động sản hay không. Để trả lời vấn đề này, ta hãy phân tích qua thị trường bất động sản Việt Nam qua một vài góc nhìn.

**3.1.1.3 Giá bất động sản và cơn sốt đất ảo**

Khái niệm cơn sốt đất ảo chỉ sự gia tăng giá đất trên diện rộng với mức tăng đột biến trong thời gian ngắn nhưng nhu cầu sử dụng đất không có thật. Người mua đất chủ yếu là giới đầu cơ, không mua để ở, cũng không xây dựng để kinh doanh mà thường bỏ hoang và mua không quan tâm đến giá trị thực tế. Cơn sốt đất ảo giá đất được hình thành do tin đồn hoặc thông tin không rõ ràng, chưa có cơ sở chắc chắn. Mặt bằng giá đất liên tục tăng do tâm lý đám đông, dẫn đến xuất hiện nhiều nhu cầu ảo, khiến giá đất tăng thiếu cơ sở thực tế hoặc bị kỳ vọng thái quá [[2]](#footnote-2).

Cơn sốt đất thông thường chỉ diễn ra cục bộ tại một vài vị trí có thay đổi quy hoạch, biến chuyển hạ tầng, nơi có các dự án cầu, đường, công trình tầm cỡ được xây dựng đúng tiến độ. Sự phát triển mạnh của hệ thống hạ tầng đô thị đã kích thích làm tăng mặt bằng giá bất động sản. Những nơi mà người dân tin rằng sẽ phát triển kinh tế xã hội tương lai, việc tăng giá lên nhiều lần là điều hợp lý mà thôi. Nếu không mua vào điểm còn rẻ, sau này sẽ không thể mua được. Có thể nói giá bất động sản của Việt Nam trong nhiều nằm gần đây không phụ thuộc vào giá trị của chính bất động sản đó mà phụ thuộc vào quan điểm của những người tham gia.

Qua những phân tích nhỏ, ta có thể thấy rằng giá bất động sản Việt Nam mang đậm tính chủ quan của người rao bán. Việc một giao dịch bất động sản đã hoàn thành được ghi nhận chưa đủ để nói lên đây là một thông tin chính xác khi có thể trong vài tuần, giá cả nếu rao bán sẽ cao gấp 2 đến 3 lần. Điều này đem đến một quyết định quan trọng ảnh hưởng đến mục tiêu của xây dựng mô hình định giá bất động sản mà nhóm muốn thực hiện

**3.1.1.4 Kết luận về tính khả dĩ của cơ sở dữ liệu**

Cơ sở dữ liệu mà là các tin rao trên thị trường bất động sản của Việt Nam. Nói cách khác, bản thân nó phản ánh quan điểm chủ quan của người tham gia vào thị trường bất động sản. Như ta đã phân tích, quan điểm chủ quan này lại là yếu tố tác động lớn đến giá cả của bất động sản. Tóm lại, ta có thể sử dụng cơ sở dữ liệu này để huấn luyện nên mô hình định giá bất động sản tại Việt Nam và trực quan hóa các thông tin tiện ích có được từ cơ sở dữ liệu này.

Việc xây dựng mô hình định giá dựa trên dữ liệu quá lâu ở trong quá khứ sẽ không phù hợp với mục tiêu mà nhóm thực hiện luận văn muốn giải quyết: cung cấp thêm một kênh thông tin tham khảo giá cả bất động sản cho người dùng. Từ những nguyên nhân đó, nhóm quyết định sử dụng các tin rao từ 2018 đến 2019.

### **3.1.2 Các kết luận quan trọng**

Sau khi tìm hiểu về đặc thù của thị trường bất động sản Việt Nam, nhóm tiến hành kết hợp các thông tin đó và phân tích kỹ cơ sở dữ liệu. Các kết luận quan trọng được rút ra như sau:

**3.1.2.1 Phân loại mô hình theo giao dịch bất động sản**

Dữ liệu có sự phân loại theo loại giao dịch trên thị trường bất động sản: Bán nhà mặt phố; bán nhà riêng; bán nhà biệt thự, liền kề; cho thuê chung cư; Cho thuê nhà riêng; Bán đất; Bán đất nền, …

Tuy rằng cấu trúc dữ liệu của mỗi tin rao đều giống nhau, mỗi tin rao thuộc một loại giao dịch này về bản chất sẽ có sự khác biệt so với tin rao thuộc loại giao dịch khác. Ví dự như đối với một tin rao thuôc loại giao dịch báo nhà mặt phố sẽ có các thông tin đáng để xem xét như số tầng, số phòng, số nhà vệ sinh. Tuy nhiên, một tin rao thuộc loại giao dịch bán đất sẽ không có các thông tin này.

Có thể thấy rằng nếu chỉ sử dụng các dữ liệu thuộc riêng một loại giao dịch để xây dựng mô hình định giá riêng cho loại giao dịch đó thì sẽ đáng tin cậy hơn.

Từ những nguyên nhân trên, nhóm quyết định sẽ xây dựng các mô hình định giá theo từng loại giao dịch khác nhau.

**3.1.2.2 Tập trung vào thành phố Hồ Chí Minh**

Các tin rao về bất động sản đến từ nhiều tỉnh thành trên cả nước. Tuy nhiên, số lượng các tin rao chủ yếu tập trung ở một số tỉnh thành lớn: Hồ Chí Minh, Hà Nội.

**Hình 2: Số lượng tin rao phân bố theo tỉnh thành năm 2018 - 2019**

Nếu ta sử dụng dữ liệu này để xây dựng mô hình định giá thì máy tính sẽ chỉ học được dữ liệu từ một số tỉnh thành nhất định. Hơn nữa, do đặc điểm về kinh tế, dân cư, địa lý, … giữa các tỉnh thành là khác nhau, mà các yếu tố này có quan hệ mật thiết với giá cả bất động sản. Nếu ta sử dụng mô hình định giá được xây dựng từ dữ liệu chỉ học được từ tỉnh thành này, áp dụng lên tỉnh thành khác sẽ không lại kết quả đúng đắn. Điều này không đem lại ý nghĩa thực tế nào.

Ngoài ra, trong phần trực quan hóa dữ liệu bất động sản. Nhóm cần sử dụng đến thông tin quy hoạch của một địa phương để người dùng có những trải nghiệm tốt nhất. Qua sự tìm kiếm của nhóm, thành phố Hồ Chí Minh đã xây dựng ứng dụng web Thông tin quy hoạch thành phố do Sở quy hoạch – kiến trúc phát triển và quản lý [14]. Các thông tin quy hoạch của một đối tượng (thửa đất) mà ứng dụng cung cấp được trực quan hóa trên bản đồ và cho phép người dùng tương tác bản đồ để thay đổi đối tượng cần tra cứu. Các thông tin này là kết quả của lời gọi API đến máy chủ của hê thống. Điều này đem lại nhiều thuận lợi cho nhóm thực hiện khi muốn thể hiện các thông tin quy hoạch về bất động sản mà người dùng đang muốn xem xét.

Từ những nguyên nhân khách quan trên, nhóm quyết định đưa đối tượng cần xây dưng mô hình định giá và trực quan hóa dữ liệu trở về bất động sản tại thành phố Hồ Chí Minh. Điều này không làm mất đi tính tổng quát của mục tiêu mà nhóm hướng tới vì những gì được thực hiện với các bất động sản tại thành phố Hồ Chí Minh đều có thể áp dụng tại các địa phương khác.

**3.1.2.3 Chọn lựa các loại giao dịch bất động sản**

Mặc dù cơ sở dữ liệu chứa nhiều tin rao khác nhau nhưng sự phân bố theo các loại giao dịch bất động sản lại không đều. Chủ yếu các tin rao nằm trong các loại giao dịch sau: bán nhà riêng, bán nhà mặt phố, bán đất và bán đất nền.

**Hình 3: Sự phân bố các tin rao theo loại giao dịch năm 2018-2019**

Các loại giao dịch còn lại bao gồm: bán trang trại, khu nghỉ dưỡng; bán căn hộ chung cư, bán nhà biệt thự liền kề; bán kho, nhà xưởng, bán các loại bất động sản khác; cho thuê căn hộ trung cư; cho thuê nhà trọ; cho thuê văn phòng; cho thuê nhà riêng; cho thuê nhà mặt phố; … Số lượng các tin rao của các loại giao dịch này ít hơn nhiều so với 4 loại giao dịch kể trên.

Ngoài ra, ta sẽ không thể chỉ dựa vào tập dữ liệu thô này mà huấn luyện mà cần phải tiền xử lý để loại trừ các sai số. Phần tiền xử lý sẽ được trình bày chi tiết ở mục sau. Việc xử lý này sẽ làm giảm số lương tin rao cần thiết xuống.

Từ những điều đã trình bày, nhóm quyết định sẽ xây dựng mô hình định giá đối với các loại giao dịch bán nhà riêng, bán nhà mặt phố, bán đất và bán đất nền. Tuy nhiên, vì nhiều yếu tố khách quan, kết quả nhóm đạt được chỉ là hai mô hình định giá là bán nhà riêng và bán nhà mặt phố. Nguyên nhân cụ thể sẽ được trình bày trong tại phần kết quả.

Từ các phần sau, khi nói tới cơ sở dữ liệu, ta có thể ngầm hiểu là các dữ liệu thuộc 4 loại giao dịch như trên.

**3.1.2.4 Các thuộc tính được lựa chọn**

Nhóm phân tích các thuộc tính của một tin rao vào quan sát thấy được các kết quả sau:

Trường đầu tiên cần nói tới là Directions tức hướng của bất động sản. Qua tìm hiểu, trường này có 9 giá trị lựa chọn: Đông, Tây, Nam, Bắc, Đông – Nam, Đông – Bắc, Tây – Nam, Tây – Bắc và Không xác định. Tuy nhiên, khi vẽ biểu đồ phân phối các giá cả theo trường Directions, Không xác định chiếm từ 85% đến 90 %. Nhóm kết luận thuộc tính này không thích hợp để vào huấn luyện.

Trường hợp thứ hai mà nhóm muốn đề cập không chỉ đến từ quá trình phân tích mà còn đến từ quá trình chạy thực nghiệm một số thuật toán sẽ được trình bày sau. Để mô tả các yếu tố về vị trí, dữ liệu sử dụng các trường: DistrictId, WardId, StreetId, Lat, Lng. Tuy nhiên từ kết quả huấn luyện, nhóm thấy rằng việc có các trường này hay không hay lại không làm ảnh hưởng đến kết quả quá nhiều. Sai lệch giữa việc có và không có các trường này là không đáng kể. Đôi khi, trong một số trường hợp các sai số RMSE và MAE còn tăng nhẹ (các sai số này sẽ được trình bày tại mục 3.2). Từ đó nhóm quyết định không sử dụng đến các trường thông tin này.

Qua quá trình phân tích trêm, nhóm quyết định lựa chọn các thuộc tính sau:

|  |  |
| --- | --- |
| **Bán nhà riêng và Bán nhà mặt phố** | |
| **Tên trường** | **Ý nghĩa** |
| Price | Giá trị của bất động sản |
| Area | Diện tích của bất động sản |
| Rooms | Số phòng |
| Floors | Số tầng |
| RoadIn | Độ dài đường vào nhà |
| Toilet | Số nhà vệ sinh |
| FrontEnd | Chiều dài mặt tiền |
| Lat | Vĩ độ của bất động sản trên bản đồ |
| Lng | Kinh độ của bất động sản trên bản đồ |

**Bảng 4: Các trường được lựa chọn cho mô hình bán nhà mặt phố và bán nhà riêng**

|  |  |
| --- | --- |
| **Bán đất – bán đất nền** | |
| **Tên trường** | **Ý nghĩa** |
| Price | Giá trị của bất động sản |
| Area | Diện tích của bất động sản |
| RoadIn | Độ dài đường vào nhà |
| FrontEnd | Chiều dài mặt tiền |
| Lat | Vĩ độ của bất động sản trên bản đồ |
| Lng | Kinh độ của bất động sản trên bản đồ |

**Bảng 5: Các trường được lựa chọn của mô hình bán đất và bán đất nền**

### **3.1.3 Tiền xử lý dữ liệu**

Tiền xử lý dữ liệu là một bước rất quan trọng trong việc giải quyết bất kỳ vấn đề nào trong lĩnh vực Học Máy. Hầu hết các bộ dữ liệu được sử dụng trong các vấn đề liên quan đến Học Máy cần được xử lý, làm sạch và biến đổi trước khi một thuật toán Học Máy có thể được huấn luyện trên những bộ dữ liệu này.

**3.1.3.1 Số hóa dữ liệu**

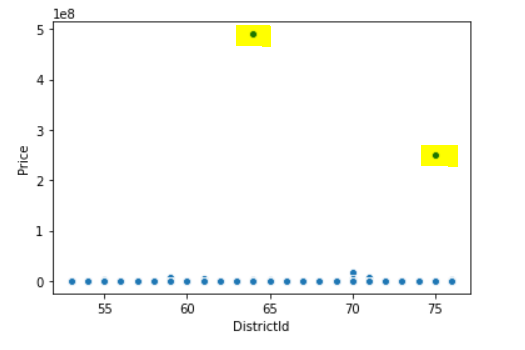
Dữ liệu đầu vào cần thiết của một bài toán trong học máy phải là dữ liệu mà máy tính có thể xử lý được. Trong quá trình phân tích cơ sở dữ liệu, nhóm quan sát thấy 2 đặc trưng quan trọng của một bất động sản trong cơ sở dữ liệu thu thập là được là giá cả, diện tích đều ở dạng chuỗi. Điều này bắt buộc nhóm phải chuyển đổi các giá trị chuỗi này thành giá trị số tương ứng.

Thuộc tính giá cả của cơ sở dữ liệu có các dạng đơn vị là: triệu, tỉ, tỷ, trăm, triệu/m2 và các dạng viết hoa chữ cái đầu tiên. Nhóm chuyển giá cả thành các số có đơn vị chung là triệu Việt Nam Đồng (VND), từ giờ sẽ gọi tắt là đồng.

Thuộc tính diện tích có chuẩn chung là “x m2” với x là diện tích bất động sản. Nhóm số hóa giá trị này thành x, đơn vị là m2.

**3.1.3.2 Các ngoại lai (outliers)**

Các ngoại lai là những điểm có giá trị khác xa phần còn lại của dữ liệu. Có hiểu đơn giản rằng, các ngoại lai này có khoảng cách xa bất thường với các điểm còn loại của tập dữ liệu.



**Hình 4: Ouliers trong tập dữ liệu bất động sản năm 2018 – 2019**

Nguyên nhân gây ra điều này trong tập dữ liệu đang được xem xét tới từ việc người dùng sai sót khi nhập liệu. Rất nhiều trường hợp các tin rao có giá nhà là hàng trăm tỷ đồng chỉ với diện tích chưa tới 50 m2 và không có đặc trưng gì đặc biệt khác.

Có nhiều cách để xử lý trường hợp này như chọn lọc dữ liệu theo giá trị Z-score hoặc theo khoảng tứ phân vị (IQR). Nhóm thực hiện xóa dữ liệu ngoại lai theo phương pháp 1.5 IQR [[3]](#footnote-3):

* Tính khoảng phân vị thứ nhất Q1
* Tính khoảng phân vị thứ ba Q3
* Tính IQR = Q3 – Q1
* Chăn trên upper\_bound = Q3 + 1.5\*IQR
* Chăn dưới lower\_bound = Q1 – 1.5\*IQR

**3.1.3.3 Trùng lặp dữ liệu**

Việc trùng lặp dữ liệu xảy ra xuất phát từ nhiều nguyên nhân. Có thể kể đến việc người đăng tin đăng nhiều tin rao về một bất động sản vào nhiều thời điểm khác nhau. Mỗi tin rao đều có một khoảng thời gian có hiệu lực. Khi hết khoảng thời gian này, nếu không có người liên hệ, người đăng tin sẽ có xu hướng đăng lại tin rao này.

Ngoài ra còn có một trường hợp có thể coi là trùng lặp dữ liệu cần phải xử lý. Cùng một bất động sản, nhưng sẽ có giá khác nhau ở các tin rao khác nhau. Nguyên nhân của việc này liên hệ mật thiết đến ý kiến chủ quan của người đăng tin. Có thể đến việc một người đăng tin bán nhà với giá x tỷ đồng, nhưng vì nhiều nguyên nhân xảy, người đăng cần bán gấp căn nhà đó nên giá có thể bị hạ xuống 500 triệu đồng. Ngược lại, nhiều trường hợp giá căn nhà sẽ được tăng thêm. Nguyên nhân có thể kể đến như căn nhà được rao bởi nhiều người môi giới khác nhau hoặc ý kiến chủ quan của chủ nhân căn nhà. Nếu nhìn từ nhiều góc độ, các tin rao này đều có thể coi là chính xác. Giá trị của một bất động sản không cố định mà sẽ biến động theo thời gian. Thêm vào đó, với đặc thù của thị trường bất động sản Việt Nam, giá trị một bất động sản liên hệ mật thiết tới niềm tim của người tham gia thị trường này. Vì vậy, nhóm quyết định sẽ coi các trường hợp này là dữ liệu trùng lặp.

Việc xử lý dữ liệu trùng lặp trong python tương đối dễ dàng với thư viện pandas. Thư viện này cung cấp phương thức để xóa dữ liệu trùng lặp và cho phép ta chọn giữa việc xóa tất cả, giữ lại phần tử đầu tiên hoặc cuối cùng. Nhóm quyết định giữ lại phần tử đầu tiên.

**3.1.3.4 Khuyết thiếu dữ liệu**

Theo như nhóm quan sát, trong cơ sở dữ liệu có nhiều tin rao thiếu thông tin ở một số thuộc tính. Điều này không khó lý giải vì dữ liệu được tạo ra từ người dùng. Việc người dùng bỏ sót hoặc sai sót khi nhập liệu là điều có thể xảy ra.

Thông thường có hai cách xử lý vấn đề này:

* Thay thế giá trị bị thiếu bằng một giá trị khác như trung bình, trung vị, 0.
* Loại bỏ giá trị bị thiếu này đi: bỏ đi cột hoặc bỏ đi dòng.

Để sử dụng phương pháp nào phụ thuộc vào bản chất của vấn đề ta đang xem xét. Với các tin rao khuyết thiếu các thông tin quan trọng như diện tích, giá cả; ta không nên thay thế các giá trị này vì sẽ tạo ra một thông tin sai hoàn toàn. Nhóm quyết định sẽ xóa đi các dòng dữ liệu khuyết thiếu giá trị quan trọng.

**3.1.3.5 Co giãn dữ liệu**

Thông thường, các thuật toán học máy sẽ không hoạt động tốt nếu các thuộc tính dữ liệu có độ giãn cách quá khác nhau. Đây chính là trường hợp mà nhóm gặp phải: diện tích phần lớn biến thiên từ 30 đến 120 m2 trong khi giá cả thì biến động từ khoảng 500 đến 25000 triệu đồng tùy theo loại giao dịch.

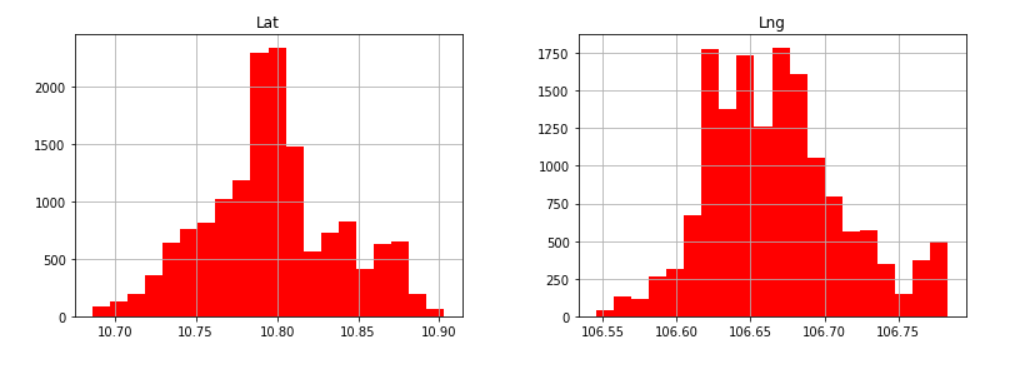
Có hai cách phổ biến để xử lý trường hợp này: min-max scaling và standarlization.

Đối với *min-max scaling* (nhiều người gọi nó là *normalization*) thì khá đơn giản: Các giá trị được điều chỉnh sao cho dải giá trị của nó nằm gọn trong đoạn [0;1]. Chúng ta làm điều này bằng cách trừ các giá trị cho min, sau đó chia giá trị vừa tính được cho (max - min). Thư viện Scikit-Learn cung cấp lớp MinMaxScaler cho mục đích này. Nó có một hyper-parameter feature\_range để bạn thay đổi dải giá trị mặc định từ [0;1] sang một dải khác nếu muốn.

Với *standardization* thì lại khá khác biệt. Nó làm cho dữ liệu của chúng ta có được cái gọi là phân phối chuẩn tắc - standard normal distribution. Trước tiên nó trừ các giá trị cho trung bình cộng của thuộc tính. Sau đó nó chia giá trị vừa tính được cho phương sai (bình phương độ lệch chuẩn) để thuộc tính cuối cùng có phương sai đơn vị. Tất nhiên thư viện Scikit-Learn cũng cung cấp lớp StandardScaler cho mục đích này.

Normalization thường tốt hơn khi dữ liệu của chúng ta không có đồ thị phân phối theo hình chuông; ngoài ra cách co giãn dữ liệu này phù hợp với mạng nơ ron nhân tạo hơn là standarlization [15].

Kết quả vẽ một số đồ thị phân phối giữa cho thấy không hề có sự phân phối dữ liệu theo hình chuông.



**Hình 5: Đồ thị phân phối các thuộc tính Lat (vĩ độ), Lng (kinh độ) của các tin rao bán nhà mặt phố trong thời gian 2018-2019**

Từ những nguyên nhân trên, nhóm quyết định xử dụng min-max scaling để co giãn dữ liệu.

### **3.1.4 Tỉ lệ huấn luyện và kiểm tra**

Đối với một tập dữ liệu dành cho một mô hình, nhóm sử dụng 75 % số lượng dữ liệu cho việc huấn luyện với cách lấy ngẫu nhiên, phần còn lại sẽ dành cho việc kiểm tra. Việc này được thực hiện thông qua đối tượng *train\_test\_split* của thư viện sk-learn.

## **3.2 Độ đo để đánh giá**

Trong luận văn này, nhóm chọn 4 thông số để đánh giá hiệu quả của mô hình.

### **3.2.1 MSE (Mean squared error)**

**MSE =**

Trong đó là gì giá trị thực tế của đối tượng i, là giá trị dự báo đối tượng i, n là số quan sát.

MSE (sai số bình phương trung bình) là trung bình của các bình phương sai số, tức là sự khác biệt giữa các ước lượng và những gì dự báo được. Chỉ số này càng nhỏ thì độ chính xác của mô hình xây dựng được càng cao.

### **3.2.2 RMSE (Root mean squared error)**

**RMSE =**

RMSE là căn bậc hai của MSE. Vì vậy, RMSE có cùng đơn vị với các giá trị được dự đoán. RMSE cho biết mức độ phân tán các giá trị dự đoán so với các giá trị thực.

Sai số này càng nhỏ thì độ chính xác của mô hình xây dựng được càng cao.

### **3.2.3 MAE (Mean Absolute error)**

**MAE =**

Từ công thức tính có thể thấy MAE (sai số tuyệt đối trung bình) là trung bình của các sai số. Tương tự như RMSE và MSE, chỉ số này càng nhỏ thì mô hình có độ chính xác càng cao. MAE có cùng đơn vị với giá dự đoán

### **3.2.4 R2 và R2 hiệu chỉnh**

**R2 = 1 – ESS / TSS**

RSS: tổng sai số bình phương do hồi quy

TSS: tổng sai số bình phương.

R2 (R bình phương) xuất phát từ ý tưởng xem toàn bộ biến thiên quan sát được của các biến phụ thuộc chia thành 2 phần: phần biến thiên do hồi quy (regression) (ESS) và phần biến thiên do phần dư.

Giá trị R bình phương dao động từ 0 đến 1. R bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp với bộ dữ liệu dùng để huấn luyện [16].

Ví dụ: Giá trị R bình phương của một mô hình dự đoán là 0.7, tức là 70% sự biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập; 30% còn lại do sai số trong quá trình thu thập dữ liệu, hoặc do các biến độc lập khác có thể giải thích sự thay đổi nhưng chưa được đưa vào mô hình).

R bình phương có hạn chế là càng đưa nhiều biến vào mô hình, dù các biến đó chưa giải thích được sự thay đổi của biến phụ thuộc thì giá trị R bình phương vẫn sẽ tăng. Lý do là càng đưa thêm biến vào thì sẽ khiến phần dư giảm xuống. Để ngăn chặn tình trạng này, R bình phương hiệu chỉnh được sử dụng.

**R2­adj =**

Việc thêm vào một biến dẫn đến tăng R bình phương nhưng cũng làm giảm đi một bậc tự do, vì chúng ta đang ước lượng thêm một tham số nữa. R2 hiệu chỉnh là một phép đo độ thích hợp tốt hơn bởi vì nó cho phép đánh đổi giữa việc tăng R2 và giảm bậc tự do [16].

Đối với bài toán regression thì R2 hiệu chỉnh rơi vào khoảng trên 50% là có thể chấp được [16].

Thư viện sklean đã cung cấp các phương thức tính MSE, MAE, R bình phương; các chỉ số RMSE có thể tự tính được bằng MSE và R phương

## **3.3 Mô hình học máy**

Nhóm thực hiện luận văn đã tham khảo khóa học “*Machine learning regression masterclass in python*” [2] trên website học trực tuyến Udemy, đặc biệt là ở các nội dung định giá nhà bằng thuật toán Linear Regression (chương 6) và Mạng Nơ ron nhân tạo (chương 8). Ngoài ra, nhóm còn tham khảo một số blog về áp dụng thuật toán rừng ngẫu nhiên (Random Forest) trong định giá nhà được tác giả sử dụng khi tham gia cuộc thi trên kaggle [17]. Trong suốt quá trình làm luận văn, nhóm đã có sự thay đổi giữa việc chọn lựa các kỹ thuật học máy.

### **3.3.1 Ngôn ngữ, Công cụ, môi trường**

Nhóm chọn python phiên bản 6.8 64bit là ngôn ngữ lập trình cho việc xây dựng mô hình định giá.

Môi trường huấn luyện Google Colab [18] là một dịch vụ miễn phí của Google nhằm hỗ trợ nghiên cứu và học tập về AI. Colab cung cấp môi trường Code như Jupyer Notebook và có thể sử dụng GPU và TPU miễn phí.

Sử dụng các công cụ đã được trình bày là tensorflow, sklearn, pandas, …

### **3.3.2 Linear Regression**

Thuật toán mà nhóm lựa chọn đầu tiên để huấn luyện chính là Linear Regression. Đây là một thuật toán đơn giãn và thường được chọn làm thuật toán bắt đầu cho những người tìm hiểu về học máy. Nhóm bắt đầu với thuật việc xây dựng mô hình định giá nhà riêng.

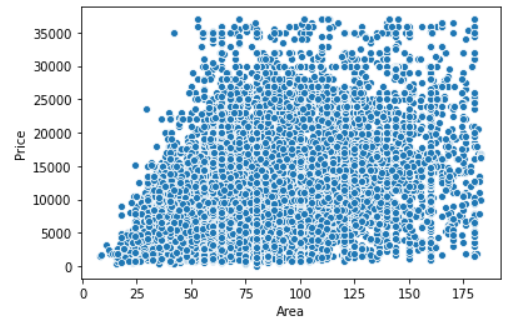
Tập dữ liệu dùng để xây dựng mô hình này sau khi trải qua quá trình tiền xử lý dữ liệu có kích thước là: **53207** dòng. Tuy nhiên, kết quả đánh giá của mô hình này rất xấu.

|  |  |
| --- | --- |
| Số lượng dòng trong tập test: **13302**  Giá trung bình trên tập test: **3995.019** (triệu đồng) | **RMSE: 1785.842** |
| **MSE: 3189230.815** |
| **MAE: 1365.419** |
| **R2 hiệu chỉnh: 0.38** |

**Bảng 6: Kết quả thử nghiệm thuật toán Linear Regression đối với trường hợp bán nhà riêng**

Để cải thiện kết quả, nhóm không chọn hết tập dữ liệu mà tiến hành chọn lọc ra các tập dữ liệu theo khoảng giá để huấn luyện và phân tích kết quả. Các kết quả này đều không hề có sự cải thiện đáng chú ý khi chỉ số R2 hiệu chỉnh luôn giao động trong khoảng 35 -45 %, các sai số RMSE và MAE luôn cao so với giá trung bình trên tập test (xấp xỉ 30 – 45 %).

Để tìm hiểu nguyên nhân, nhóm sử dụng biểu đồ phân tán scatterplotsđể biểu thị mối tương quan giữa một vài cặp biến (các thuộc tính). Kết quả cho thấy, các cặp biến đều có mối quan hệ phi tuyến tính. Ví dụ sau đây là giữa diện tích và giá cả.



**Hình 6: Biểu đồ phân tán giữa diện tích (Area) và giá (Price)**

Từ biểu đồ, ta không thể rút ra được mối quan hệ giữa các cặp biến này là gì. Thông thường diện tích và giá cả sẽ có mối quan hệ tuyến tính: diện tích càng lớn thì giá bất động sản càng lớn. Tuy nhiên ngoài diện tích còn rất nhiều yếu tố chi phối đến giá cả đặc biệt là vị trí địa lý (cặp tọa kinh độ, vĩ độ). Tình trạng này xất hiện tương tự ở các cặp thuộc tính khác. Và tương tự đối với các loại giao dịch khác.

Trải qua quá trình tìm hiểu nguyên nhân, nhóm rút ra được kết luận thuật toán Linear Regression không thích hợp với tập dữ liệu huấn luyện.

### **3.3.3 Random Forest**

Sau khi thất bại với thuật toán Linear Regression, nhóm quyết định sử dụng một thuật toán về cây quyết định là Random Forest.

Random Forest là một phương pháp học có giám sát nên có thể xử lý được các bài toán về classification và regression. Các các điều kiện tương tự trường hợp Linear Regression, nhóm tiến hành xây dựng mô hình định giá bằng phương pháp Random Forest.

Lúc này, nhóm xử dụng tới thư viện fastai và parfit, cho phép máy tính tìm ra bộ thông số đầu vào cho ra giá trị tốt nhất. Bảng sau là kết quả của chương trình tìm ra bộ thông số tốt từ các khoảng đầu vào của từng thông số.

|  |  |  |
| --- | --- | --- |
| **Tên thông số** | **Khoảng giá trị cho thông số** | **Giá trị tốt nhất** |
| min\_samples\_leaf | **[1,3,5,10,15,25]** | 1 |
| max\_features | **['sqrt', 'log2', 0.4, 0.5, 0.6, 0.7]** | 0.7 |
| n\_estimators | **[5,10,20,40,60]** | 40 |

**Bảng 7: Thông số tốt nhất của thuật toán Random Forest cho mô hình bán nhà riêng**

Kết quả đạt của mô hình tương đối khả quan.

|  |  |
| --- | --- |
| Số lượng dòng trong tập test: **13302**  Giá trung bình trên tập test: **3995.019** (triệu đồng) | **RMSE: 858.902** |
| **MSE: 737712.67** |
| **MAE: 557.5** |
| **R2 hiệu chỉnh: 0.856** |

**Bảng 8: Kết quả thử nghiệm thuật toán Random Forest cho mô hình bán nhà riêng**

Tuy nhiên theo như nhóm tìm hiểu thì phương pháp này lại có hạn chế cần phải xem xét. Phương pháp này không hoạt động tốt khi vượt ngoài phạm vi được huấn luyện. Nói cách khác, phạm vi dự đoán mà phương pháp thể đưa ra bị ràng buộc bởi các giá trị cao nhất và thấp nhất trong dữ liệu huấn luyện [[4]](#footnote-4) [[5]](#footnote-5).

Trải qua quá trình phân tích và trao đổi với giáo viên hướng dẫn, nhóm quyết định không sử dụng phương pháp này.

### **3.3.4 Mạng Nơ ron nhân tạo**

Sau khi bỏ qua phương pháp Random Forest, nhóm đến với một phương pháp khá mới đó là sử dụng mạng nơ ron nhân tạo (học sâu). Phương pháp này có thể giải quyết cả bài toán classification và regression. Đặc biệt, mạng nơ ron nhân tạo cho phép ta giải quyết các vấn đề phi tuyến gặp phải ở trường hợp đầu tiên.

**3.3.4.1 Thử nghiệm**

Tham khảo từ khóa học “Machine learning regression masterclass in python” trên Udemy, nhóm bắt đầu với các tham số cơ bản sau:

|  |  |
| --- | --- |
| Input layer | **50** nodes, Activation function: **relu** |
| Hidden layer 1 | **50** nodes, Activation function: **relu** |
| Output layer | **1** node, Activation function: **linear** |
| Optimizer | **Adam** |
| Loss | **Mean\_squared\_error** |
| Epochs | **100** |
| Batch\_size | **50** |
| Validation\_split | **0.2** |

**Bảng 9: Tham số ban đầu cho mạng nơ ron**

Sử dụng tập dữ liệu tương tự các trường hợp Linear Regression và Random Forest để huấn luyện và kiểm tra. Kết quả khi kiểm tra trên tập test của bán nhà riêng cho thấy kết quả tương đối khả quan

|  |  |
| --- | --- |
| **Bán nhà riêng** | |
| Số lượng dòng trong tập test: **13302**  Giá trung bình trên tập test: **3995.019** (triệu đồng) | **RMSE: 970.114** |
| **MSE: 941121.716** |
| **MAE: 690** |
| **R2 hiệu chỉnh: 0.817** |

**Bảng 10: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán nhà riêng**

Sau khi phân tích và trao đổi với giáo viên hướng dẫn, nhóm quyết định chọn mạng nơ ron nhân tạo làm giải pháp giải quyết vấn đề đặt ra trong việc xây dựng mô hình định giá.

**3.3.4.2 Xây dựng các mô hình định giá**

Với các thông số đầu vào tương tự như trường hợp bán nhà mặt phố, tiến hành thực nghiệm với trường hợp **bán nhà mặt phố**.

Kết quả đối với trường hợp này như sau:

|  |  |
| --- | --- |
| **Bán nhà mặt phố** | |
| Tổng số dòng trong tập dữ liệu: **16663**  Số lượng dòng trong tập test: **4166**  Giá trung bình trên tập test: **10814.5** (triệu đồng) | **RMSE: 3553.663** |
| **MSE: 12628521.773** |
| **MAE: 2418.905** |
| **R2 hiệu chỉnh: 0.743** |

**Bảng 11: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán nhà mặt phố**

Có thể thấy kết quả đối với trường hợp bán nhà mặt phố và bán nhà riêng là tương đối khả quan. Tuy nhiên đối với các trường hợp bán đất và bán đất nền, kết quả khi tiến hành kiểm tra trên tập test không khả quan.

|  |  |
| --- | --- |
| **Bán đất** | |
| Tổng số dòng trong tập dữ liệu: **65381**  Số lượng dòng trong tập test: **16346**  Giá trung bình trên tập test: **1979** (triệu đồng) | **RMSE: 916.917** |
| **MSE: 840737.499** |
| **MAE: 632** |
| **R2 hiệu chỉnh:0.554** |

**Bảng 12: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán đất**

|  |  |
| --- | --- |
| **Bán đất nền** | |
| Tổng số dòng trong tập dữ liệu: **24591**  Số lượng dòng trong tập test: **6148**  Giá trung bình trên tập test: **2288** (triệu đồng) | **RMSE: 998.775** |
| **MSE: 997552** |
| **MAE: 657.292** |
| **R2 hiệu chỉnh: 0.667** |

**Bảng 13: Kết quả thử nghiệm mạng Nơ ron cho mô hình bán đất nền**

Trong quá trình phân tích nguyên nhân của các kết quả này, nhóm đã phát hiện ra một số vấn đề cần xử lý ở dữ liệu như việc giá trung bình trên tập test của cả 2 trường hợp trên thấp so với thực tế (sẽ được trình bày chi tiết tại mục 3.4).

Nhóm đã tiến hành xử lý vấn đề đó và tiến hành thử nghiệm chạy với các đầu vào khác nhau của thuật toán nhưng kết quả của mô hình dự đoán giá đất hay đất nền đều không được cải tiến đáng kể. Điều này bắt buộc nhóm phải xem xét bản chất vấn đề của việc định giá đất và đất nền ngoài thực tế.

**3.3.4.3 Các vấn đề đối với mô hình định giá đất, đất nền**

Xuyên suốt quá trình tìm hiểu về thị trường bất động sản Việt Nam, nhóm nhận thấy rằng, đất là đối tượng có sự biến động giá lớn hơn rất nhiều trong so với nhà cửa. Như đã trình bày, nguyên nhân của điều này là quan điểm chủ quan của người tham gia thị trường bất động sản. Tập cơ sở dữ liệu nói chung hay các trường hợp về đất đang được đề cập được tạo ra từ người đăng tin rao nên có thể suy ra nó sự biến động giá lớn.

Đồng thời, qua quá trình phân tích cấu trúc cơ sở dữ liệu của các trường hợp bán đất, bán đất nền; so sánh với các trường hợp về nhà, nhóm rút ra được những điều sau:

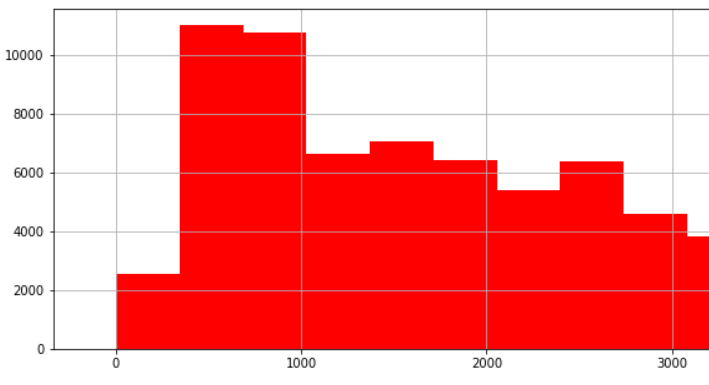
* Các trường hợp giao dịch về nhà có nhiều trường thông tin hơn cho máy học: độ dài mặt tiền, số phòng, số tầng, …
* Các trường khác biệt kể trên lại mang thông tin quan trọng về đặc điểm của căn nhà khi định giá ngoài thực tế.
* Các trường thông tin hiện tại chưa đủ mô tả các yếu tố làm cho giá đất biến động: vị trí trung tâm đô thị, gần các khu quy hoạch lớn, hình dáng mảnh đất, sự phát triển của kinh tế trong khu vực, …

Từ những gì đã phân tích và dưới sự hướng dẫn của giáo viên hướng dẫn, nhóm cho rằng dữ liệu hiện tại chưa đáp ứng được yêu cầu để có thể xây dựng mô hình định giá đất, đất nền. Mục tiêu trở thành cải thiện các kết quả của mô hình định giá nhà mặt phố và nhà riêng.

## **3.4 Tối ưu mô hình học máy**

### **3.4.1 Chọn lọc dữ liệu theo khoảng giá**

Quá trình phân tích những vấn đề về đất và đất nền đã giúp nhóm phát hiện ra các dữ liệu này tồn tại sai sót. Đó là sai sót của người dùng khi tạo sai loại tin giao dịch.



**Hình 7: Một góc biểu đồ phân phối tần suất theo giá của tập dữ liệu bán đất năm 2018-2019**

Nhóm phát hiện ra có sự phân bố không hề ít các tin rao có giá từ xấp xỉ 1 cho đến dưới 100. Xét dưới góc độ địa phương, ta đang nghiên cứu các mô hình tại thành phố Hồ Chí Minh, thành phố có nền kinh tế phát triển nhất cả nước, việc tồn tại giá bán một căn nhà bất kể tại vị trí nào với giá như trên đều không hợp lý. Nguyên nhân của việc này là một số tin rao về thuê nhà, thuê chung cư, thuê kho bãi, đã được tạo với phân loại là bán nhà mặt phố. Vấn đề tương tự cũng xuất hiện đối với trường hợp khác.

Sau vẽ sơ đồ phân phối của theo biến giá cả, nhóm quyết định chọn lọc lại các tin rao với chặn dưới của giá là 500 triệu đồng. Sau khi xử lọc, tiến hành xử lý trùng dữ liệu, kết quả số lượng các dòng trong cả 2 tập dữ liệu của bán nhà mặt phố, bán nhà riêng đều tăng lên. Có thể thấy, khi không lọc dữ liệu với chặn dưới, một số dữ liệu đã bị xóa nhầm trong thao tác xử lý dữ liệu trùng lặp

### **3.4.2 Hiệu chỉnh các thông số của mạng kỹ nơ ron**

Tại giai đoạn này, nhóm tiến hành thử và sai các thông số kỹ thuật để tìm ra bộ thông số phù hợp nhất.

Các yếu tố sau đây đã được điều chỉnh để tìm kiếm kết quả tốt nhất:

* Số hidden layer trong mạng nơ ron
* Số nơ ron trong mỗi lớp
* activation function
* Optimizer của mạng,
* Loss function
* Thêm các lớp Dropout để chống over-fitting
* Số lượng epoch
* Kích thước batch

Kết quả của các thay đổi sẽ được trình bày ở chương 5.

**Chương 4. Ứng dụng trực quan hóa dữ liệu**

## **4.1 Xác định các chức năng cần thiết**

### **4.1.1 Tìm hiểu nhu cầu thực tế**

Để xây dựng tính năng của ứng dụng, trước tin nhóm tìm hiểu về nhu cầu cần biết thông tin về một bất động sản của con người [[6]](#footnote-6). Vì mục tiêu thứ hai của luận văn là có thể trực quan hóa các dữ liệu hữu ích cho người dùng khi sở hữu một bất động sản nên ta sẽ không xét tới các yếu tổ của bản thân bất động sản như là diện tích, mặt tiền … Người tham gia vào thị trường bất động sản dù là để đầu tư, để sinh sống hay để kinh doanh, … đều quan tâm tới các vấn đề như cơ sở hạ tầng của khu vực, các công trình công cộng, mật độ dân cư, … Không khó để lý giải điều này. Các nơi có mật độ dân cư cao sẽ là các khu vực có kinh tế phát triển. Mọi người thường có xu hướng tập trung ở những vùng có kinh tế phát triển bởi các khu vực này đem lại cơ hội việc làm cao, thu nhập cao hơn các nơi khác. Tương tự, ở các khu vực này sẽ có nhiều các công trình bệnh viện, trường học, trung tâm mua sắm, … Những nơi có kinh tế phát triển thì cơ sở hạ tầng sẽ được chú trọng đầu tư để thu hút đầu tư, giao thông đi lại thuận tiện. Ngoài các yếu tố kinh tế, các yếu tố về xã hội cũng được quan tâm nhiều. Các khu vực có chất lượng y tế, giáo dục, trình độ dân trí, trật tự trị an cũng sẽ được mọi người quan tâm. Một người muốn định cư lâu dài chắc hẳn đều sẽ quan tâm nơi mình ở có an toàn hay không, có gần trường học, bệnh viện hay không, …

Đối với những người đầu tư, họ còn quan tâm tới yếu tố tăng trường về giá trong khu vực cũng như các khu vực xung quanh. Để đầu tư, ta cần cân bằng giữa nhiều yếu tố mới có thể đem lại lợi nhuận. Hãy thử xem xét vấn đề sau: một người quan tâm tới 2 căn nhà có giá trị tương tự nhau, 1 ở quận A, 1 ở quận B. Căn nhà quận A có gần trung tâm thành phố hơn căn nhà quận B. Tuy nhiên, trong 2 năm trở lại đây, khu vực quận A không có sự tăng trưởng lớn về giá nhà. Khu vực nhà B thì ngược lại, trong vòng 2 năm, giá nhà tăng tương đối cao nhờ các dự án mới được xây dựng. Trong trường hợp này, rõ ràng đầu tư vào căn nhà B sẽ có xác suất thu lợi cao hơn. Đôi khi, người yếu tố tăng trương, các nhà đầu tư còn quan tâm tới sự sôi nổi của thị trường. Một khu vực có nhiều tin tức giao dịch sẽ khu vực được quan tâm nhiều, thông thường là các khu vực có đang nóng lên và đáng để đầu tư.

### **4.1.2 Khảo sát từ các sản phẩm tương tự**

Sau khi tiến hành tìm hiểu nhu cầu từ thực tế, nhóm chọn ra hai ứng dụng web là “Property value” và “Zillow” có các chức năng đáp ứng được các nhu cầu trên. Hai ứng dụng này đều cung cấp nhiều chức năng trực quan hóa dữ liệu hữu ích giúp người dùng dễ dàng nắm bắt thông tin. Về trường học, Property value vẽ các marker đại diện cho các trường học lên bản đồ và cung cấp các thông tin chi tiết như các lớp học, số học sinh,… Property value còn đặc biệt cung cấp các thông tin khu vực dưới dạng các biểu đồ tròn: thu nhập theo hộ gia đình, tình trạng nhà ở, độ tuổi và giới tính, … Cho phép so sánh căn nhà với các căn nhà đã bán, được rao bán và cho thuê trong khu vực. Đối với thông tin trường học, Zillow thì liệt kê danh sách, khối đào tạo, đánh giá (rating) và khoảng cách từ nhà tới các trường. Đặc biệt, Zillow cung cấp thông tin chi phí cần thiết trong một tháng để sinh sống tại căn nhà, lịch sử mức thuế và giá trị cho nếu để cho thuê.

### **4.1.3 Phân tích cơ sở dữ liệu**

Sau khi khảo sát từ các sản phẩm tương tự, nhóm đi đến phân tích cơ sở dữ liệu để phác thảo các chức năng có thể xây dựng. Cơ sở dữ liệu không có các trường thông tin về các công trình bệnh viện hay trường học. Tuy nhiên, ta có thể sử dụng dịch vụ của Google để tìm kiếm danh sách các bệnh viện, trường học mở rộng ra là nhà hàng, khu mua sắm, … trong một phạm vi nhất định so với căn nhà. Sau khi đã có thông tin này, việc còn lại là vẽ chúng lên bản đồ. Cũng từ quá trình tìm hiểu này, nhóm quan sát thấy có thể cho người dùng quan sát căn nhà (Street view) và tìm đường đi tới một địa điểm nào bằng dich vụ của google map. Cơ sở dữ liệu gồm các tin rao từ năm 2015 đến năm 2019, ta có thể dùng các tin rao này để trực quan hóa cơ cấu thị trường trong khu vực dưới dạng biểu đồ tròn. Cũng từ các tin rao này, ta tính được tăng trưởng giá trung bình theo khu vực qua các năm, giá trung bình theo m2 của các khu vực. Thông qua quá trình tìm hiểu về trực quan hóa dữ liệu [19], nhóm cho rằng các yếu tố này nên được vẽ thành các loại biểu đồ khác nhau. Giá tăng trưởng qua các năm sẽ phù hợp với dạng biểu đồ đường vì sẽ giúp người nhìn thấy được sự biến động giá. Giá trung bình theo m2 phù hợp để vẽ thành bản đồ nhiệt, bởi ta cần so sánh giữa các khu vực với nhau.

Để thể hiện căn nhà lên bản đồ, ta chỉ cần thông tin về tọa độ nếu sử dụng google map. Tuy nhiên nhóm nhận thấy rằng như vậy chỉ có thể vẽ marker lên bản đồ, không có các thông tin quan trong về quy hoạch. Bởi đất đai trong thành phố đều được quy hoạch cho các cơ quan liên quan quản lý. Nếu có thể có được các thông tin này, ta sẽ có nhiều thông tin hơn về vị trí của một căn nhà như giới hạn của thửa đất mà căn nhà được xây dựng trên đó, số thửa, số tờ, diện tích thửa đất. Như chương 3 đã trình bày trong lý do chuyển mục tiêu trở về thành phố Hồ Chí Minh, thành phố Hồ Chí Minh đã xây dựng ứng dụng web Thông tin quy hoạch thành phố do Sở quy hoạch – kiến trúc phát triển và quản lý [14]. Điều này giúp cho việc thể hiện các thông tin về bất động sản chi tiết hơn. Cơ sở dữ liệu cũng không có bất kỳ thuộc tính nào liên quan đến đánh giá khu vực. Nhóm quyết định xây dựng một chức năng cho phép người dùng đánh giá và xem các đánh giá từ người khác.

Để tăng trải nghiệm của người dùng, nhóm thấy rằng cần phát triển một số tính năng để người dùng thuận tiện hơn khi sử dụng ứng dụng. Đó là lưu danh sách các bất động sản đã định để dễ dàng xem lại danh sách mình đã định giá. Ngoài ra còn là xây dựng tính năng tìm kiếm bất động sản từ cơ sở dữ liệu trong trường hợp người dùng không muốn định giá nhà mà muốn tìm kiếm trong cơ sở dữ liệu và phân tích giá của những kết quả tìm kiếm qua các thông tin được trực quan.

### **4.1.3 Các chức năng có thể xây dựng**

Trải qua quá tìm hiểu thực tế, khảo sát ứng dụng tương tự và phân tích cơ sở dữ liệu. Nhóm tiếp tuc tìm hiểu về cách sử dụng các loại biểu đồ đường, biểu đồ tròn, bản đồ, bản đồ nhiệt, … trên React, là thư viện được chọn để phát triển ứng dụng Web).

Sau khi xong quá trình tìm hiểu, nhóm xác định được các chức năng sau là phù hợp để xây dựng:

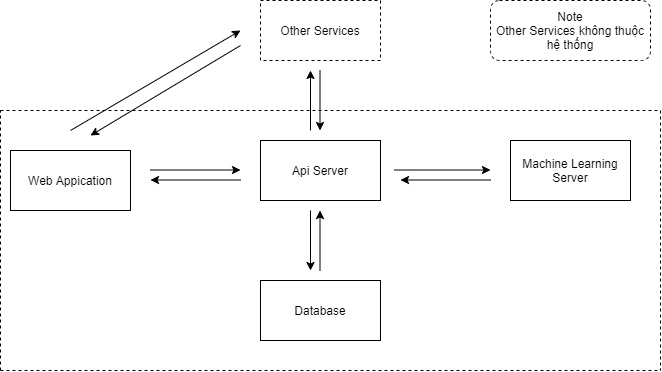
* Hiển thị thông tin quy hoạch
* Xem giá bất động sản
* Xem các công trình công cộng trong khu vực
* Xem cảnh quan (street view) của căn nhà.
* Xem và thêm nhận xét về một địa điểm.
* Xem cơ cấu thị trường bất động sản.
* Xen danh sách các bất động sản tương tự.
* Xem bản đồ nhiệt về giá trung bình trên m2.
* Xem lịch sử giá trung bình qua các năm.
* Tìm kiếm theo khu vực quận, huyện, loại bất động sản, diện tích, mức giá, thông tin cơ bản bất động sản và hiển thị danh sách lên bản đồ
* Xem thời gian, đường đi từ bất động sản tới địa điểm khác.
* Lưu và xem danh sách bất động sản đã định giá.

## **4.2 Kiến trúc hệ thống**

Hệ thống được xây dựng theo kiến trúc microservice. Mỗi thành phần trong hệ thống sẽ thực hiện một nhiệm vụ riêng biệt với nhau, không phụ thuộc lẫn nhau. Việc các thành phần độc lập với nhau như vậy sẽ giúp cho việc phát triển mỗi thành phần trong hệ thống trở nên đơn giản hơn.

Việc các thành phần được tách rời nằm còn giúp ta có thể phát triển với nhiều ngôn ngữ và công nghệ khách nhau để tận dụng những công nghệ phù hợp cho nhiệm vụ mà thành phần đó đảm nhiệm.

Khi xảy ra lỗi, ta dễ dàng xác định được nơi xảy ra lỗi và nguyên nhân gây ra lỗi để có thể nhanh chóng khắc phục. Ngoài ra, một lợi ích cực kỳ to lớn của kiến trúc này chính là việc bảo trì và nâng cấp hệ thống trở nên đơn giản hơn.



**Hình 8: Kiến trúc toàn hệ thống**

Luồng làm việc của hệ thống như sau:

1. Khi người dùng ở phía UI (ứng dụng web) sử dụng các chức năng của hệ thống, các lời gọi Api đến Api Server được kích hoạt
2. Api Server nhận được các yêu cầu, kiểm tra và xử lý các yêu cầu này

2a. Đối với các yêu cầu định giá bất động sản, Api server sẽ tạo yêu cầu rồi gửi tới Machine Learning Server

Tại Machine Learning Server, quá trình thực thi diễn ra như sau:

* Tạo dữ liệu đầu vào từ nội dung của yêu cầu được gửi tới
* Tải mô hình tương ứng với đối tượng
* Tiến hành định giá bất động sản
* Trả về kết quả định giá cho Api Server

Api Server trả về kết quả cho ứng dụng Web.

2b. Đối với các yêu cầu tìm kiếm bất động sản, Api Server sẽ tiến hành truy vấn cơ sở dữ liệu với điều kiện tương ứng và trả về kết quả cho ứng dụng Web.

2c. Đối với các yêu cầu cần sử dụng dịch vụ của google, Api Server sẽ tạo các yêu cầu rồi gửi tới Server của Google, sau khi có kết quả sẽ tính toán các giá trị nếu cần thiết rồi sau đó gửi lại cho ứng dụng Web.

1. Một số chức năng tại ứng dụng Web không tạo các yêu cầu với Api Server mà sẽ gửi tới Server của Google.
2. Ứng dụng Web sau khi nhận các kết quả từ Google và Api Server sẽ tiến hành trực quan hóa dữ liệu tùy theo chức năng

* Sử dụng bản đồ của google và tạo các marker cần thiết.
* Sử dụng google streetview.
* Vẽ bản đồ nhiệt.
* Biểu đồ tăng trưởng.
* …

## **4.3 Vai trò của các thành phần**

### **4.3.1 Cở sở dữ liệu**

Đảm điểm chức năng lưu trữ các dữ liệu phục vụ cho các yêu cầu từ Api Server:

* Lưu trữ các tin rao bất động sản.
* Lưu trữ thông tin quy hoạch thành phố Hồ Chí Minh.
* Lưu trữ lịch sử giá trung bình theo phường, xã

Cở sở dữ liệu đã được tọa index để tăng tốc các truy vấn

### **4.3.2 Api Server**

Api Server là thành phần xử lý các yêu cầu của người dùng từ phía ứng dụng web. Api Server thực hiện các chức năng chính sau.

* Kết nối và đưa thông tin đến ứng dụng web.
* Thực hiện các truy vấn tới dịch vụ của google.
* Thực hiện các truy vấn thông tin quy hoạch thành phố Hồ Chí Minh
* Thực hiện các truy vấn dữ liệu bất động sản cho tìm kiếm.
* Thực hiện các yêu cầu về xác thực và phân quyền.
* Trung gian trong việc thực hiện định giá nhà giữa ứng dụng Web và Python server

### **4.3.3 Machine Learning Server**

Đảm nhận nhiệm vụ thực thi định giá theo yêu cầu được gửi từ Api Server. Machine Learning Server thực hiện các chức năng sau:

* Tạo dữ liệu đầu vào cho mô hình
* Dựa vào loại mô hình yêu cầu, tiến hành co giãn dữ liệu
* Tải mô hình theo yêu cầu và tiến hành dự đoán giá
* Đảo nghịch co giãn dữ liệu với kết quả định giá.
* Trả về kết quả cho Api Server

### **4.3.4 Web Appilcation**

Đảm nhận nhiệm vụ trực quan hóa dữ liệu liên quan thành các thông tin tiện ích tạo điều kiện thuận lợi cho người dùng nắm bắt các thông tin. Các chức năng của ứng dụng Web sẽ được trình bày chi tiết hơn tại chương 6 kết quả.

## **4.3 Giao tiếp giữa các thành phần**

Các thành phần trong hệ thống giao tiếp với nhau thông qua HTTP API. Các Server nhận yêu cầu được thiết kế theo tiểu chuẩn RESTful API để tiện cho việc quản lý tài nguyên. Dữ liệu trả về cho các lời gọi API này đều ở dạng JSON.

Đối với các lời gọi API tới Api Server, thành phần này sẽ xác thực API bằng token trong header “Authorization” trong yêu cầu gửi lên. Việc xác thực các API nhằm đảm bảo tính bảo mật của các thành phần trong hệ thống, ngăn ngừa nguy cơ bị tấn công hoặc rò rỉ dữ liệu.

Việc sử dụng RESTFul API mang tới một số ưu điểm sau:

* Giúp cho ứng dụng rõ ràng hơn.
* Dữ liệu được trả về với nhiều định dạng khác nhau như: xml, html, json….
* Code đơn giản và ngắn gọn.
* Quản lý tài nguyên tốt hơn.

**Chương 5. Kết quả**

## **5.1 Kết quả xây dựng mô hình định giá**

### **5.1.1 Sau khi lọc giữ liệu**

|  |  |
| --- | --- |
| **Bán nhà riêng** | |
| Tổng số dòng trong tập dữ liệu: **58026**  Số lượng dòng trong tập test: **14507**  Giá trung bình trên tập test: **3957.280** (triệu đồng) | **RMSE: 968.903** |
| **MSE: 938772.661** |
| **MAE: 681.144** |
| **R2 hiệu chỉnh: 0.812** |

**Bảng 14: Kết quả mô hình bán nhà riêng sau khi lọc dữ liệu**

Kết quả cho thấy hiệu quả có cải thiện có sự cải thiện nhẹ nhưng không đáng kể so với trường hợp ban đầu.

|  |  |
| --- | --- |
| **Bán nhà mặt phố** | |
| Tổng số dòng trong tập dữ liệu: **16897**  Số lượng dòng trong tập test: **4225**  Giá trung bình trên tập test: **10003** (triệu đồng) | **RMSE: 2932.878** |
| **MSE: 8601772.773** |
| **MAE: 2095.499** |
| **R2 hiệu chỉnh: 0.741** |

**Bảng 15: Kết quả mô hình bán nhà mặt phố sau khi lọc dữ liệu**

Trường hợp này là lại cho thấy sự hiệu quả khi sai số giảm đi đáng kể so với trước:

* RMSE giảm từ 35553.663 xuống 2932.878
* MAE giảm từ 2418.905 xuống 2095.499

### **5.1.2 Hiệu chỉnh các thông số mạng nơ ron**

Nhóm đã tiến hành thử và sai rất nhiều lần để tìm ra đáp án tốt nhất. Nhóm sẽ liệt kê các trường hợp tốt nhất trong số các lần tiến hành thí nghiệm.

Để đánh giá hiệu quả, các thông số về tập dữ liệu được giữa nguyên. Mặc định output layer có 1 node và activation là ‘linear’. Epoch là 100, batch size là 50

Sau đây là các kết quả của việc thay đổi thông số trong mạng nơ ron.

|  |  |
| --- | --- |
| **Bán nhà mặt phố** | |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ | **RMSE = 2894.805**  **MSE = 8379893.311**  **MAE = 2070.07**  **R2 hiệu chỉnh = 0.748** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Hidden layer 2: 128 nodes, activation ‘relu’ | **RMSE = 2933.169**  **MSE = 603477.518**  **MAE = 2089.916**  **R2 hiệu chỉnh = 0.741** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Dropout layer 1: rate = 0.2 * Hidden layer 2: 128 nodes, activation ‘relu’ | **RMSE = 2859.86**  **MSE = 8178796.713**  **MAE = 2036.8**  **R2 hiệu chỉnh = 0.754** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Dropout layer 1: rate = 0.2 * Hidden layer 2: 128 nodes, activation ‘relu’ * Dropout layer 2: rate = 0.2 * Hidden layer 3: 128 nodes, activation ‘relu’ | **RMSE = 2832.296**  **MSE = 8021898.323**  **MAE = 2028.422**  **R2 hiệu chỉnh = 0.758** |

**Bảng 16: Các thông số hiệu chỉnh và kết quả tương ứng của mô hình bán nhà mặt phố**

|  |  |
| --- | --- |
| **Bán nhà riêng** | |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ | **RMSE = 941.178**  **MSE = 885815.49**  **MAE = 656.07**  **R2 hiệu chỉnh = 0.822** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Hidden layer 2: 128 nodes, activation ‘relu’ | **RMSE = 958.115**  **MSE = 917985.221**  **MAE = 650**  **R2 hiệu chỉnh = 0.816** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Dropout layer 1: rate = 0.2 * Hidden layer 2: 128 nodes, activation ‘relu’ | **RMSE = 929.922**  **MSE = 864754.223**  **MAE = 650.3**  **R2 hiệu chỉnh = 0.826** |
| * Input layer: 32 nodes, activation ‘relu’ * Hidden layer 1: 128 nodes, activation ‘relu’ * Dropout layer 1: rate = 0.2 * Hidden layer 2: 128 nodes, activation ‘relu’ * Dropout layer 2: rate = 0.2 * Hidden layer 3: 128 nodes, activation ‘relu’ | **RMSE = 922.862**  **MSE = 851673.584**  **MAE = 629.23**  **R2 hiệu chỉnh = 0.829** |

**Bảng 17: Các thông số hiệu chỉnh và kết quả tương ứng của mô hình bán nhà riêng**

Ở các hai trường hợp, các thử nghiệm đều có sự cải thiện so với trước nhưng không đáng kể khi sai số giảm rất ít.

### **5.1.3 Kết quả cuối cùng**

Sau khi trải qua quá trình thử và sai, nhóm chọn ra trường hợp tốt nhất để làm kết quả cuối cùng. Hai bảng sau là tổng hợp kết quả của từng mô hình nhóm xây dựng.

|  |  |
| --- | --- |
| **Bán nhà mặt phố** | |
| Tổng số dòng trong tập dữ liệu: **16897**  Số lượng dòng trong tập test: 4225  Khoảng giá: **[615; 25000]** (triệu đồng)  Giá trung bình trên tập test: **10003** (triệu đồng) | **RMSE = 2832.296** |
| **MSE = 8021898.323** |
| **MAE = 2028.422** |
| **R2 hiệu chỉnh = 0.758** |

**Bảng 18: Kết quả cuối cùng của mô hình bán nhà mặt phố**

|  |  |
| --- | --- |
| **Bán nhà riêng** | |
| Tổng số dòng trong tập dữ liệu: **58026**  Số lượng dòng trong tập test: **14507**  Khoảng giá: **[500; 12500]** (triệu đồng)  Giá trung bình trên tập test: **3957.28** (triệu đồng) | **RMSE = 922.862** |
| **MSE = 851673.584** |
| **MAE = 629.23** |
| **R2 hiệu chỉnh = 0.829** |

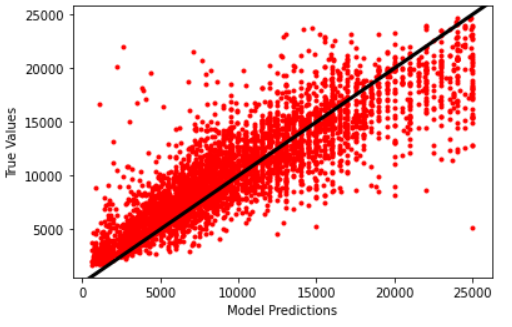
**Bảng 19: Kết quả cuối cùng của mô hình bán nhà riêng**

### **5.1.2 Thảo luận**

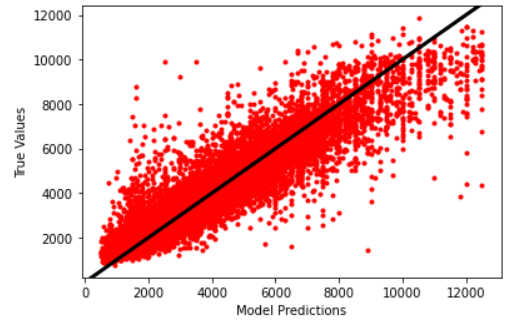
Nhìn chung, ta có thể thấy sai số ở cả 2 mô hình tương đối cao:

* Bán nhà mặt phố với giá trung bình gần 10 tỷ đồng có trung bình sai số tuyệt đối (MAE) là gần 2.3 tỷ.
* Bán nhà riêng với giá trung bình gần 4 tỷ có trung bình sai số tuyệt đối là gần 629 triệu.

Trước khi đi vào phân tích kết quả, ta hãy thử vẽ đồ thị Oxy sau. Trục tung là các giá trị thực trong tập dữ liệu test, trục hoành là các giá trị định giá. Từ giả thiết này, ta thấy rằng giá trị định giá chính xác sẽ nằm trên đường thẳng y = x.



**Hình Bán nhà mặt phố - đồ thị so sánh giá trị định giá và giá trị thực**



**Hình Bán nhà mặt riêng - đồ thị so sánh giá trị định giá và giá trị thực**

Có thể thấy, ở cả hai mô hình, các kết quả đều có xu hướng hội tự theo đường thẳng y = x, tức là gần giá trị tương ứng. Tuy nhiên có thể thấy vẫn có nhiều điểm cách quá xa với vị trí tương ứng trên đường thẳng y = x. Những trường hợp này chính là các sai số nghiêm trọng, làm cho giá trị MSE tăng cao vì MSE sẽ bình phương sai số lên. Vì giá trị RMSE là căn bậc hai của MSE, nên giá trị RMSE sẽ cao hơn giá MAE tương ứng.

Câu hỏi đặt ra là liệu đây có phải là một kết quả chấp nhận được?

Đối với các vấn đề ứng dụng trí tuệ nhân tạo để giải quyết, việc đánh giá các sai số cần đòi hỏi ta phải xem xét các tính chất của vấn đề và mục tiêu ta hướng tới khi xây dựng mô hình định giá. Ví dụ như việc dự đoán giá cổ phiếu, một mô hình có độ chính xác 70% có thể xem là rất tốt bởi tính biến động và khó đoán của thị trường đầu tư này là thách thức không nhỏ. Trong trường hợp khác, một mô hình dự đoán bệnh có độ chính xác 70% lại được xem là cực kỳ xấu.

Trở lại vấn đề ta đang giải quyết, như những gì đã phân tích ở mục 3.1 chương 3, cơ sở dữ liệu mà là các tin rao trên thị trường bất động sản của Việt Nam. Bản thân nó phản ánh quan điểm chủ quan của người tham gia vào thị trường bất động sản. Một người có giá một căn nhà định giá theo ngân hàng là 3 tỷ, sẽ có xu hướng tăng giá khi đăng tin rao với giá 4 hoặc 5 tỷ, đôi khi là tăng gấp lên 6 nếu đất trong khu vực đó đang sốt lên. Tương tự nếu cần phải bán gấp căn nhà, chủ sở hữu hoàn toàn có thể hạ giá nhà xuống để có thể thanh khoản nhanh chóng. Từ điều này, có thể thấy sự sai số lớn trong kết quả của hai mô hình là điều có thể hiểu được.

Cuối cùng, mục tiêu đầu tiên của luận văn là xây dựng mô hình định giá nhằm cung cấp cho người dùng thêm một kênh thông tin tham khảo, không phải là một nguồn tin duy nhất mà người dùng nên phụ thuộc. Trong tất cả các lĩnh vực đầu tư, người tham gia không nên phụ thuộc vào một nguồn thông tin duy nhất mà phải kết hợp tất cả các thông tin lại với nhau để đi đến một kết luận cuối cùng.

Từ những gì đã phân tích, nhóm kết luận rằng kết quả của 2 mô hình định giá bán nhà mặt phố và bán nhà riêng tại thành phố Hồ Chí Minh là kết quả chấp nhận được.

## **5.2 Kết quả xây dựng ứng dụng trực quan hóa dữ liệu**

### **5.2.1 Danh sách các chức năng**

Dưới đây là danh sách các tính năng mà nhóm đã hoàn thành:

* Hiển thị thông tin quy hoạch bất động sản,
* Xem giá bất động sản.
* Xem các công trình công cộng xung quanh bất động sản
* Xem các trường học cấp một, cấp hai, cấp ba, đại học xung quanh bất động sản
* Xem cảnh quan đường phố gần khu vực bất động sản.
* Xem và thêm lời nhận xét của các người dùng về khu vực bất động sản
* Xem cơ cấu thị trường bất động sản dưới dạng biểu đồ tròn
* Xem danh sách bất động sản tương tự.
* Xem bản đồ nhiệt về giá theo các tiêu chí khác nhau: giá trung bình trên m2, tăng trưởng, thị trường rao tin.
* Xem lịch sử giá trung bình của khu vực bất động sản qua các năm.
* Xem thời gian, đường đi từ bất động sản tới nơi khác.
* Thêm một địa điểm mới của người dùng vào danh sách các địa điểm quan trọng.
* Tìm kiếm bất động sản theo khu vực quận, huyện, loại bất động sản, diện tích, mức giá, thông tin cơ bản bất động sản và hiển thị danh sách lên bản đồ
* Hiển thị thống kê về khu vực bất động sản.
* Chia sẻ bất động sản lên trang facebook cá nhân.
* Lưu bất động sản vào danh sách ưu thích.
* Xem danh sách bất động sản đã định giá.
* Xem danh sách bất động sản đã lưu.

### **5.2.2 Minh họa các chức năng**

Để thuận tiện cho việc trình bày luận văn, nhóm sẽ trình bày các màn hình mình họa cho chức năng tại phần phụ lục.

**Chương 6. Kết luận**

## **6.1 Mục tiêu**

Ứng dụng trí tuệ nhân tạo vào vấn đề định giá bất động sản dựa trên các thông tin về các hoạt động trong thị trường bất động sản. Từ đó cung cấp thêm cho người dùng một nguồn tham khảo về giá đáng tin cậy bên cạnh các nguồn tham khảo truyền thống khác. Điều này sẽ đem lại rất nhiều lợi ích cho những người tham gia vào thị trường bất động sản khi họ có thể tiết kiệm được thời gian và công sức cũng như có được một cơ sở về giá để bắt đầu tham khảo.

Nếu như mục tiêu đàu tiên trả lời cho câu hỏi giá là bao nhiêu? thì mục tiêu thứ hai sẽ trả lời cho câu hỏi tại sao lại có giá đó. Muốn làm được điều đó ta phải xây dựng một ứng dụng trực quan hóa dữ liệu về bất động sản để truyền đạt những thông tin tiện ích có liên quan cho người dùng. Ta cần trực quan hóa nó thành các biểu đồ, đồ thị, bảng biểu, …. Việc đưa ra các biểu đồ, đồ thị sẽ giúp người dùng nhanh chóng nắm bắt được thông tin, dễ dàng so sánh giữa các đối tượng, phát hiện ra các điểm bất thường, xu hướng biến động của đối tượng đang xem xét, …

## **6.2 Phương pháp**

### **6.2.1 Phương pháp xây dựng mô hình định giá**

Để xây dựng mô hình định giá, trước tiên nhóm tìm hiểu các kiến thức cơ bản về lĩnh vực này để có một cái nhìn tổng quan về vấn đề mình cần giải quyết. Sau đó, nhóm phân tích cơ sở dữ liệu để xác định xem liệu có thể sử dụng cơ sở dữ liệu này hay không. Song song với quá trình đó là việc xác định phạm vi và đối tượng có thể xử lý dựa trên các điều kiện đang có. Nhóm kết luận có thể sử dụng cơ sở dữ liệu này và vấn đề cần giải quyết là định giá nhà mặt phố, nhà riêng tại thành phố Hồ Chí Minh. Tuy nhiên cơ sở dữ liệu còn có rất nhiều sai số cần được làm sạch. Nhóm tiến hành làm sạch dữ liệu để có thể huấn luyện. Qua quá trình tham khảo tài liệu về các vấn đề tương tự, nhóm thực nghiệm các phương pháp xử lý khác nhau bao gồm Linear Regression, Random Forest, mạng nơ ron nhân tạo. Kết quả cho thấy phương pháp mạng nơ ron nhân tạo là phương pháp phù hợp để xây dựng mô hình định giá. Cuối cùng, nhóm tiến hành các phương pháp để tối ưu mô hình: lọc giữ liệu để loại các sai số, thử nghiệm các thông số khác.

### **6.2.2 Phương pháp xây dựng ứng dụng trực quan hóa**

Để xây dựng ứng dụng trực quan hóa dữ liệu, nhóm trước tiên tìm hiểu về các thông tin hữu ích mà một người muốn biết trong thực tế. Sau đó tiến hành khảo sát các ứng dụng tương tự để có cái nhìn sơ bộ về các chức năng cũng như giao diện cần xây dựng. Tiếp theo là quá trình phân tích cơ sở dữ liệu để xác định các chức năng có thể xây dựng. Song song đó là quá trình tìm kiếm các nguồn thông tin khác để có thể khắc phục những thiếu sót trong cơ sở dữ liệu. Sau khi xác định được các chức năng, nhóm xác định kiến trúc của hệ thống cần xây dựng. Cuối cùng là quá trình cài đặt các chức năng của hệ thống.

## **6.3 Kết quả đạt được**

Trong khóa luận này, nhóm đã thực hiện việc tìm hiểu về thị trường bất động sản, xử lý dữ liệu, huấn luyện mô hình trí tuệ nhân tạo, xây dựng hệ thống định giá và trực quan hóa dữ liệu gồm nhiều thành phần, hoàn thiện được nhiều chức năng giúp cho việc trực quan hóa dữ liệu tốt hơn.

Cụ thể những kết quả đạt được như sau:

* Tìm hiểu về định giá bất động sản và thị trường bất động sản tại Việt Nam.
* Tìm hiểu việc phân tích và xử lý dữ liệu trước khi huấn luyện.
* Áp dụng mạng nơ ron nhân tạo trong việc xây dựng mô hình định giá: bán nhà riêng và bán nhà mặt phố tại thành phố Hồ Chí Minh.
* Tìm hiểu và ứng dụng các kiến thức trong xây dựng Server sử dụng Express với Nodejs và Flask với Python.
* Tìm hiểu về thư viện React để xây dựng ứng dụng UI có tính tương tác cao.
* Xây dựng hệ thống microservice gồm 4 thành phần: Database, Api Server, Machine Learning Server và ứng dụng Web

## **6.4 Những điểm hạn chế**

Trong quá trình thực hiện luận văn, với thời gian thực hiện có hạn và nhiều yếu tố khách quan khác, hệ thống vẫn còn một số hạn chế:

* Chưa thể xây dựng đầy đủ các mô hình định giá cho các loại giao dịch trên thị trường
* Mô hình định giá có sai số tương đối cao, chưa tối ưu.
* Chưa thể mở rộng định giá và trực quan hóa tới các tỉnh thành khác.
* Chưa thể tự động hóa việc xây dựng mô hình.
* Còn rất nhiều thông tin có thể trực quan hóa cho người dùng: thời tiết, mật độ dân cư, lịch sử giá, …

## **6.5 Hướng phát triển**

Quan những kết quả đạt được và hạn chế nêu trên, nhóm thực hiện luận văn cho rằng hướng phát triển của hệ thống trong tương lai bao gồm:

* Hoàn thiện ứng dụng web, hỗ trợ thêm nhiều chức năng trực quan hóa dữ liệu hơn như thời tiết, mật độ phương tiện giao thông, lịch sử giá, thu nhập bình quân, mật độ dân cư, trị an xã hội, …
* Mở rộng phạm vi ra các tỉnh thành khác trong nước.
* Hỗ trợ ứng dụng di động.
* Thu thập dữ liệu có tính chính xác cao, ghi nhận từ các giao dịch ngoài đời thật.
* Xử lý dữ liệu tốt hơn để có thể huấn luyện ra một mô hình định giá tốt hơn.
* Thực hiện kết hợp nhiều mô hình máy học để cho ra kết quả tốt hơn.

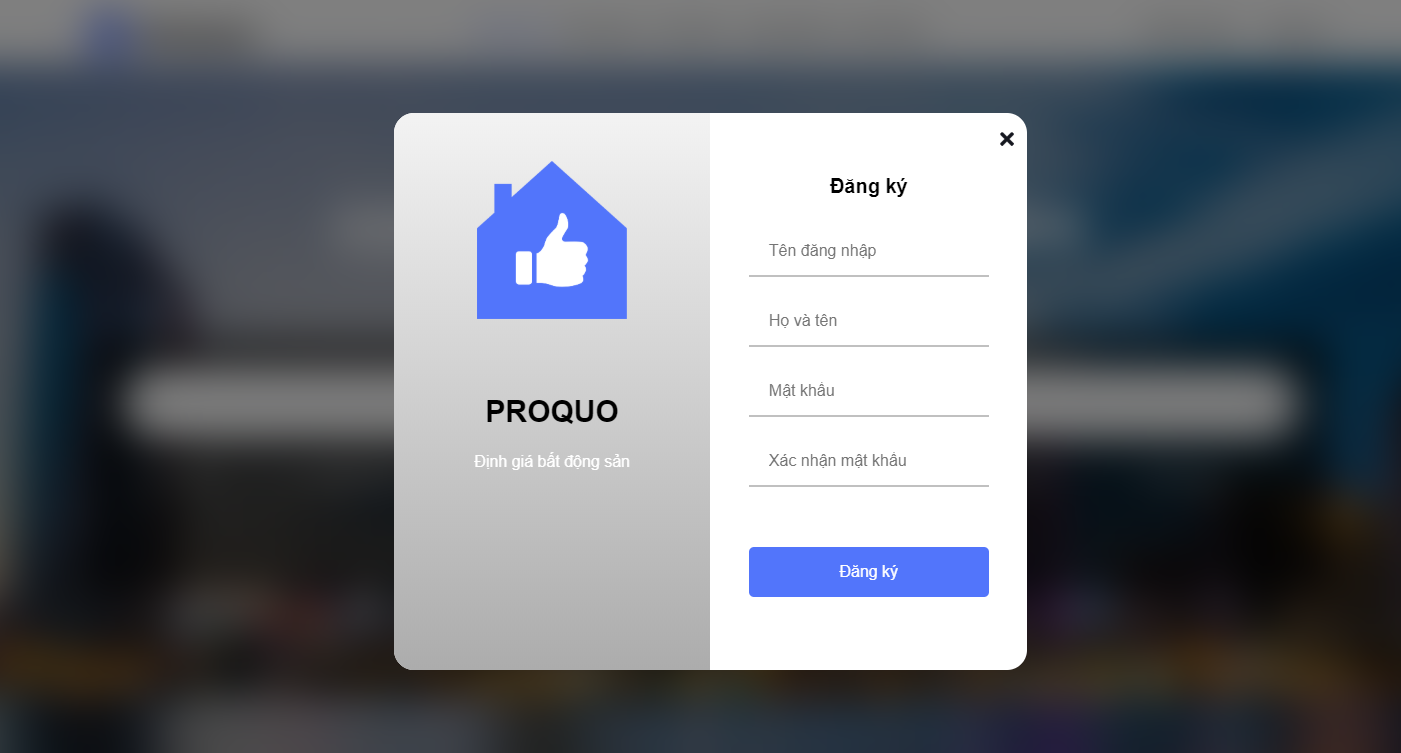
# Chương 7. Danh mục tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | "Neural Network," [Online]. Available: https://www.investopedia.com/terms/n/neuralnetwork.asp. |
| [2] | "Machine Learning Regression Masterclass in Python," [Online]. Available: https://www.udemy.com/course/machine-learning-regression-masterclass-in-python/. |
| [3] | "Keras Deep Learning in Python," [Online]. Available: https://www.udemy.com/course/keras-deep-learning-in-python/. |
| [4] | "Predicting house prices with regression using TensorFlow," [Online]. Available: https://www.coursera.org/projects/tensorflow-beginner-predicting-house-prices-regression. |
| [5] | B. K. D. A. Afonso, L. C. Melo, W. Dihanster and S. Sousa, "Housing Prices Prediction with a Deep Learning and Random Forest Ensemble". |
| [6] | A. Ng, "Machine Learning for a London Housing Price Prediction Mobile Application". |
| [7] | J. Oxenstierna, "Predicting house prices using Ensemble Learning with Cluster Aggregations". |
| [8] | "Property Value," [Online]. Available: https://www.propertyvalue.com.au/. |
| [9] | "Zillow," [Online]. Available: https://www.zillow.com/. |
| [10] | "Node.js," [Online]. Available: https://nodejs.org/en/. |
| [11] | "Express," [Online]. Available: https://expressjs.com/. |
| [12] | "Flask," [Online]. Available: https://flask.palletsprojects.com/en/1.1.x/. |
| [13] | "Flask RESTful," [Online]. Available: https://flask-restful.readthedocs.io/en/latest/. |
| [14] | "Thông tin quy hoạch thành phố Hồ Chí Minh," [Online]. Available: https://thongtinquyhoach.hochiminhcity.gov.vn/. |
| [15] | "How, When, and Why Should You Normalize / Standardize / Rescale Your Data?," [Online]. Available: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff. |
| [16] | "Hệ số R bình phương, R bình phương hiệu chỉnh," [Online]. Available: http://phantichspss.com/r-binh-phuong-r-binh-phuong-hieu-chinh-cong-thuc-y-nghia-cach-tinh-thu-cong-va-cach-tinh-bang-spss.html. |
| [17] | "My First Kaggle Competition," [Online]. Available: https://towardsdatascience.com/my-first-kaggle-competition-using-random-forests-to-predict-housing-prices-76efee28d42f. |
| [18] | "Welcome To Colaboratory," [Online]. Available: https://colab.research.google.com/. |
| [19] | "Tổng quan về trực quan hóa dữ liệu," [Online]. Available: https://bigdatauni.com/vi/tin-tuc/tong-quan-ve-data-visualization-truc-quan-hoa-du-lieu.html. |

**Phụ lục các chức năng của ứng dụng**

**1. Đăng ký tài khoản**

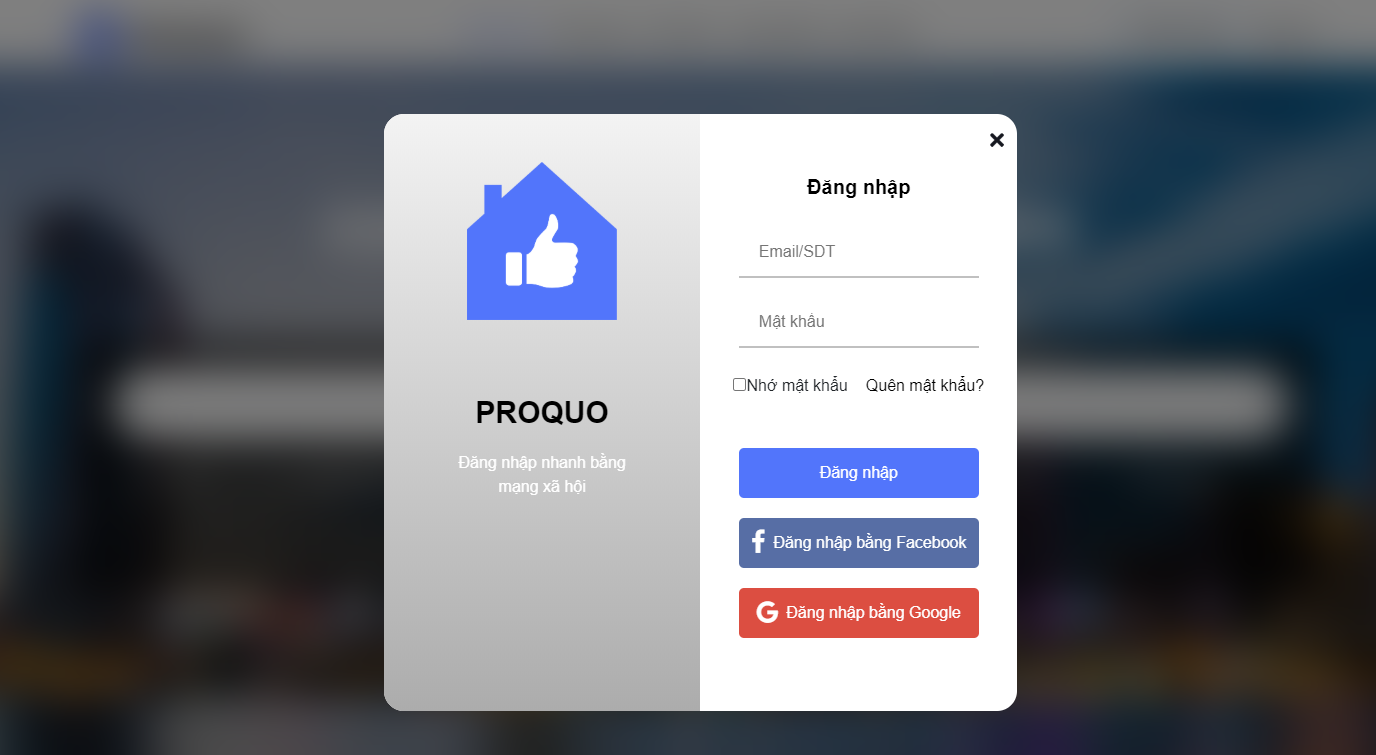
Ứng dụng cho phép người dùng tạo tài khoản với các thông tin cơ bản bao gồm tên đăng nhập, tên người dùng, mật khẩu. Mỗi tài khoản được mặc định với số lần định giá là 5 lần.



**Hình 9: Màn hình chức năng đăng ký tài khoản**

**2. Đăng nhập**

Ứng dụng giúp người dùng đăng nhập vào website ứng dụng với tài khoản đã tạo. Ứng dụng còn hỗ trợ người dùng sử dụng các mạng xã hội để đăng nhập.



**Hình 10: Màn hình chức năng đăng nhập**

**3. Chức năng định giá**

Sau khi đăng nhập tài khoản người dùng, người dùng bắt đầu nhập các thông tin của bất động sản theo các bước thông tin.

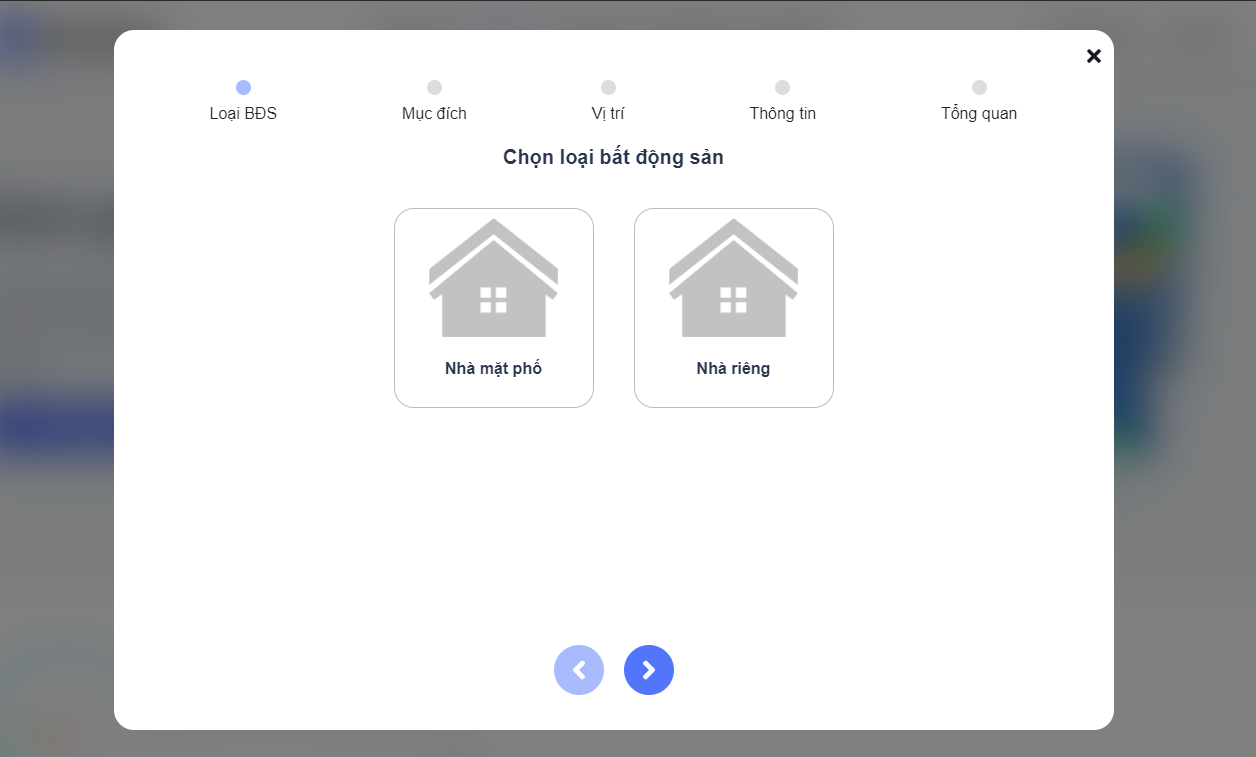
Bước 1: Chọn loại bất động sản cần định giá.

Bước 2: Chọn mục đính động sản trong việc định giá bao gồm đầu tư và định cư. Dựa vào mục đích website đưa ra các thông tin phù hợp nhất với người đùng

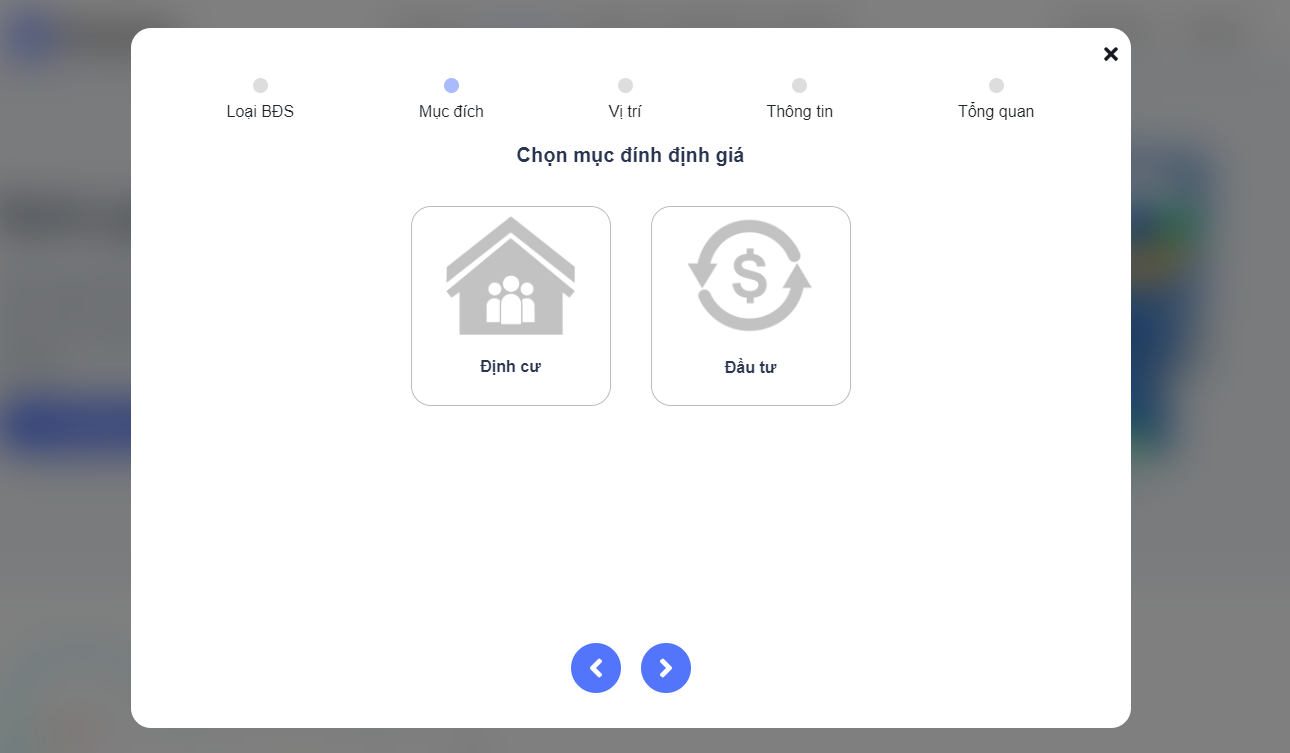
Bước 3: Nhập vị trí bất động sản và trực quan hóa quy hoạch nếu có.

Bước 4: Nhập các thông tin bất động sản

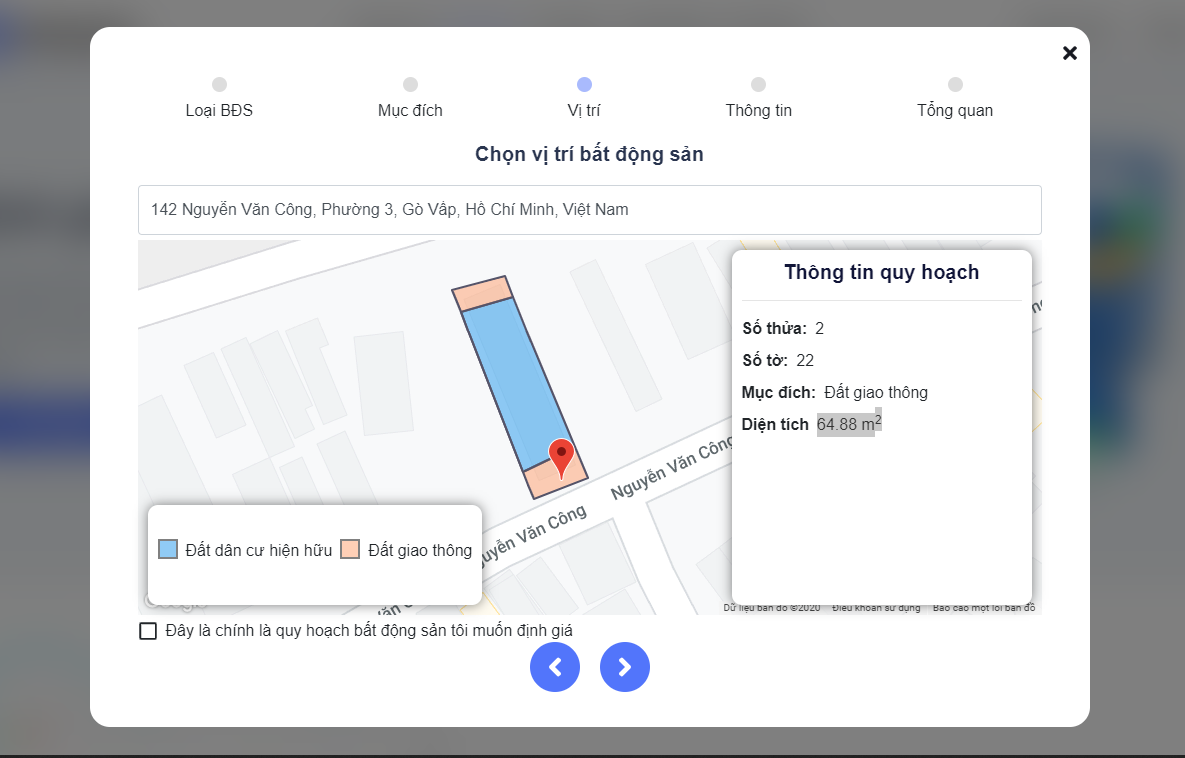
Bước 5: Thể hiện thông tin cơ bản của bất động sản. Người dùng nhấn “định giá” để địch giá



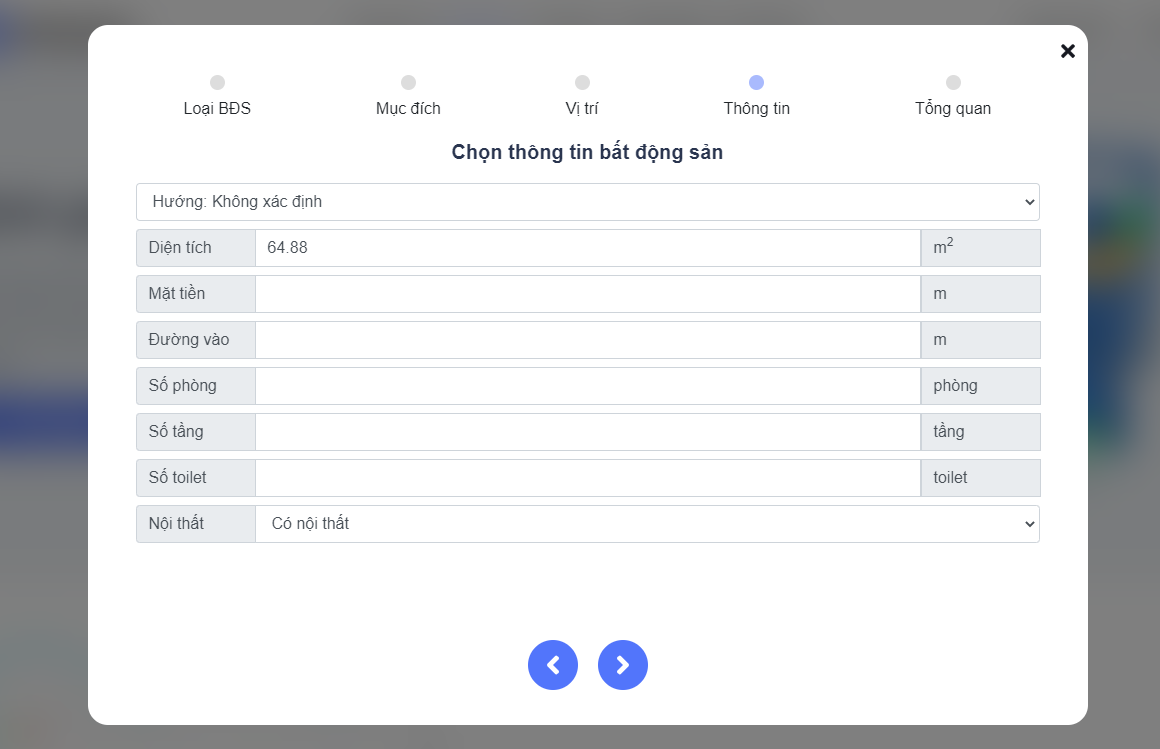
**Hình 11: Màn hình chọn loại bất động sản**



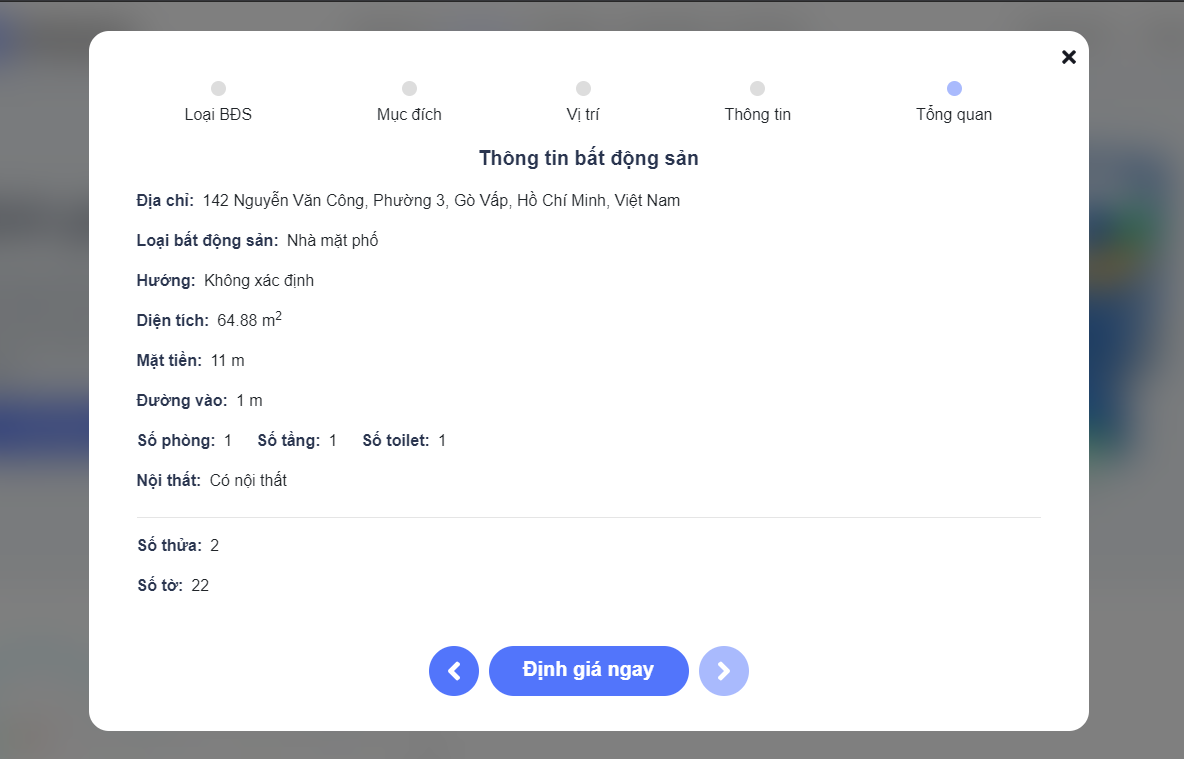
**Hình 12: Màn hình chọn mục đích định giá**



**Hình 13: Màn hình tìm kiếm thông qua vị trí và thể hiện thông tin quy hoạch bất động sản**



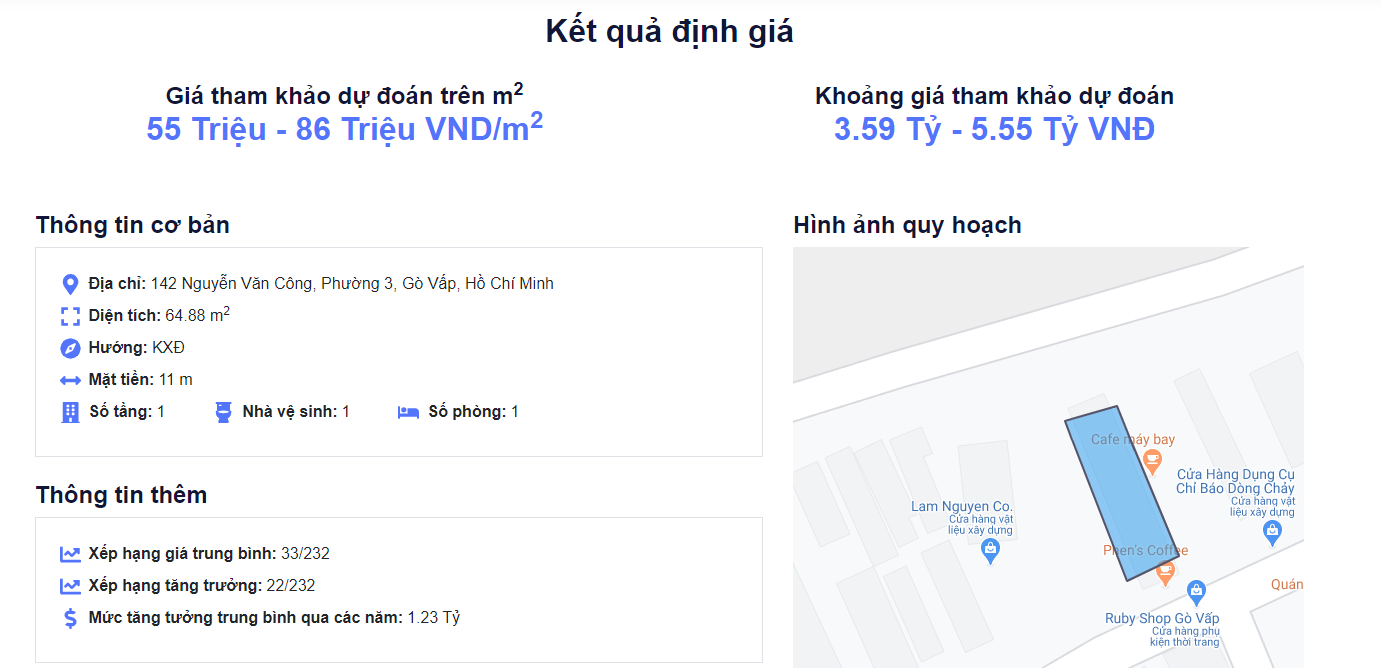
**Hình 14: Màn hình nhập thông tin bất động sản**



**Hình 15: Màn hình hiển thị thông tn cơ bản**

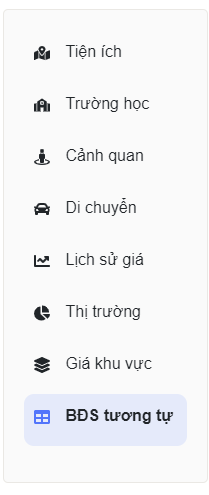
Sau khi người dùng xác nhận thông tin và thực hiện định giá. Website sẽ thể hiện các kết quả định giá bao gồm các thông tin:

* Khoảng giá dự đoán.
* Khoảng giá trung bình theo m2.
* Các thông tin cơ bản người dùng đã nhập.
* Một số thông tin hữu ích thêm.
* Hình ảnh quy hoạch của bất động sản.



**Hình 16: Màn hình kết quả định giá**

Bảng điều hướng gồm các phím tắt giúp người dùng nhanh chóng đi tới thông tin cần thiết.

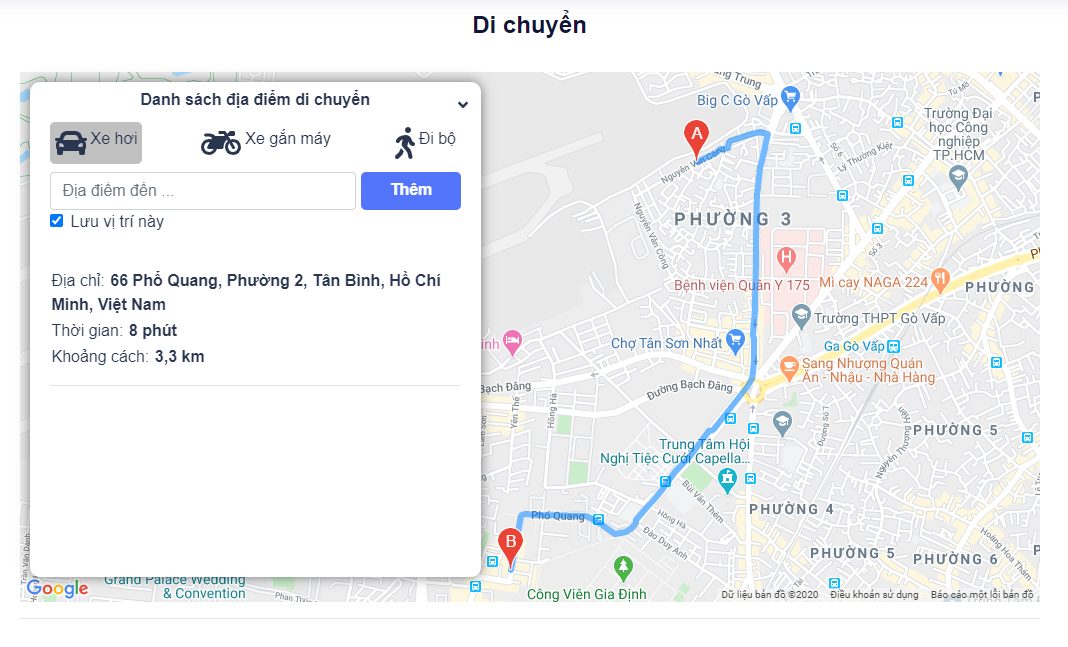


**Hình 17: Bảnh điều hướng**

**4 Di chuyển**

Nhằm mục đích giúp người dùng hiện thị thời gian và khoảng cách đến các vị trí quan trọng như công ty, trường học…

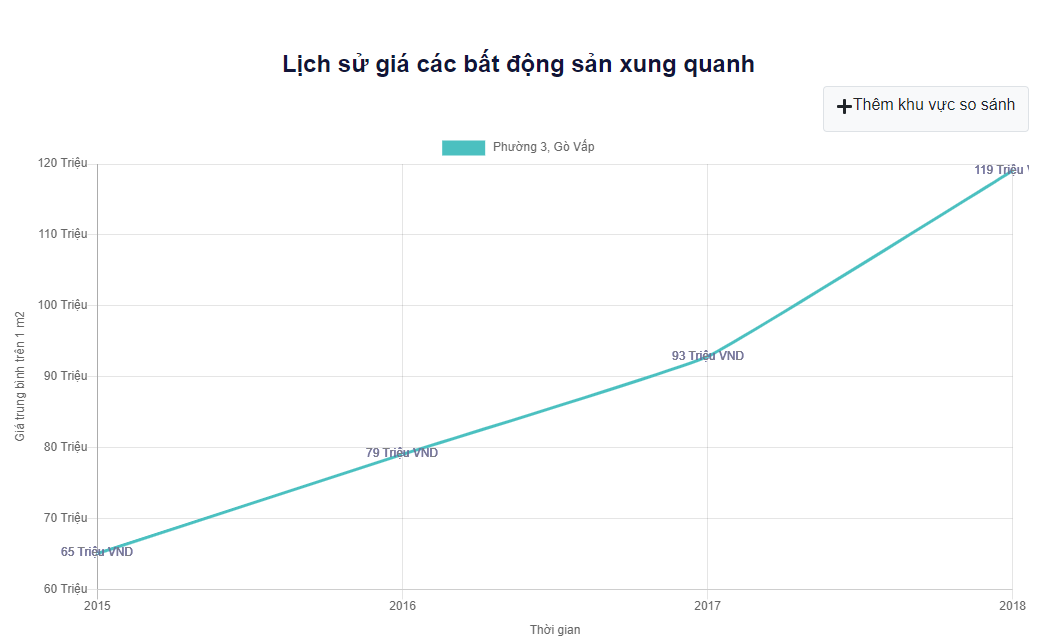
Sau khi đăng nhập người dùng có thể thêm các vị trí quan trọng với người dùng.



**Hình 18: Màn hình chức năng tìm đường đi**

**5 Lịch sử giá khu vực bất động sản.**

Với dữ liệu thông kê được, Website hiển thị thông tin lịch sử giá trung bình theo m2 qua các năm. Người dùng có thể thêm một khu vực bất động sản khác để so sách với khu vực của mình.



**Hình 19: Chức năng lịch sử giá của khu vực**

**6 Cơ cấu thị trường**

Với dữ liệu tin tức bất động sản, Website cung cấp cho người dùng thông tin về khu vực đó một năm số lượng các loại bất động sản được rao. Người dùng có thể chọn năm để hiện thị lịch sử cơ cấu thị trường,



**Hình 20: Biểu đồ cơ cấu thị trường bất động sản trong khu vực**

**7 Bản đồ giá khu vực**

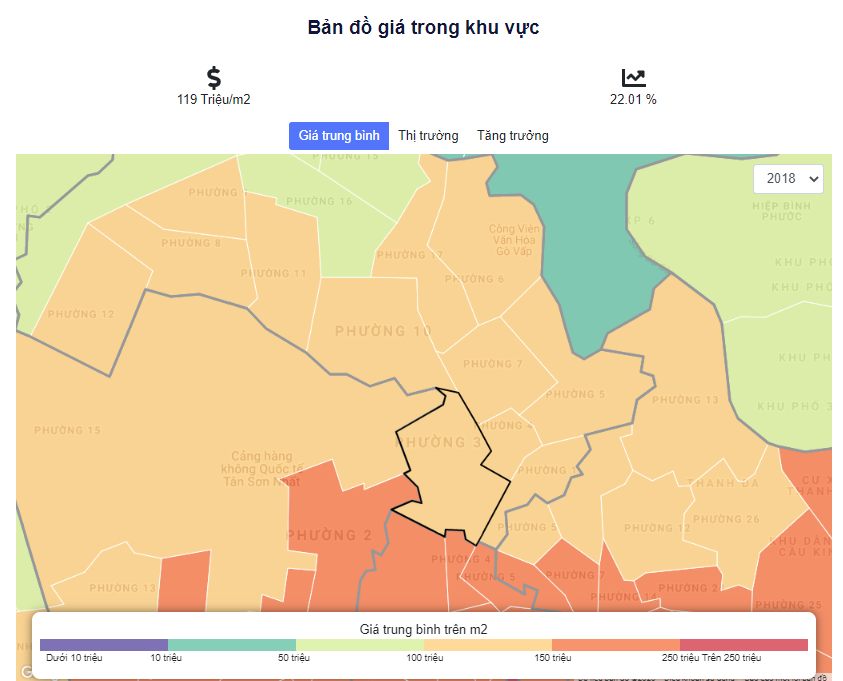
Thể hiện một các trực quan toàn bộ khu vực thành phố Hồ Chí Minh giúp người dùng nhanh chóng hình dung được mức giá ở các khu vực khác với cùng loại bất động sản của người dùng

Website cung cấp 3 loại thông tin bao gồm

+ Giá trung bình trên m2

+ Cơ cấu thị trường

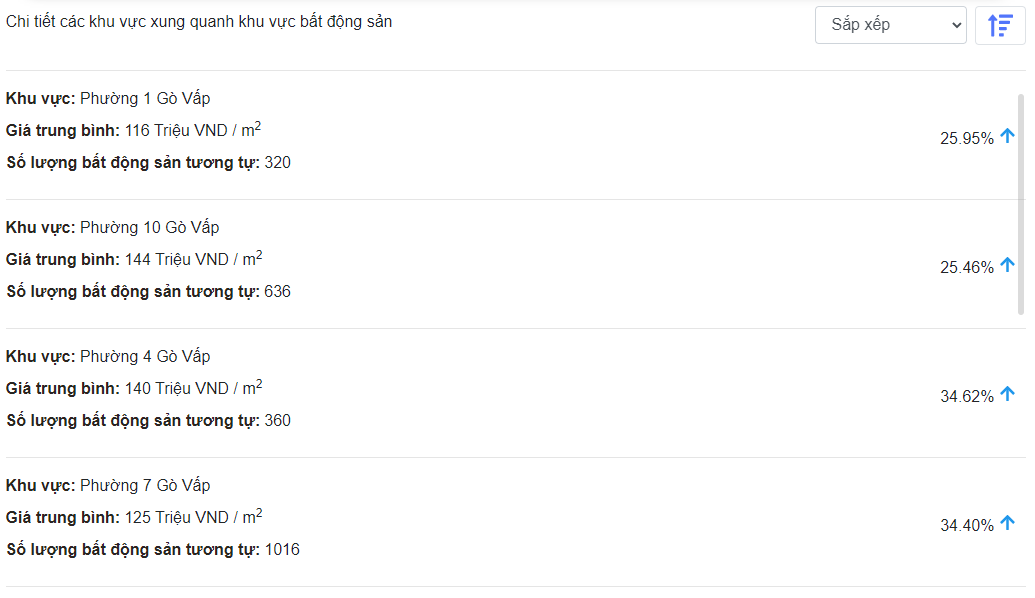
+ Tăng trưởng theo phần trăm so với năm trước



**Hình 21: Biểu đồ máu thể hiện giá trong khu vực**

**8 Thông tin các khu vực xung quanh bất động sản người dùng định giá**

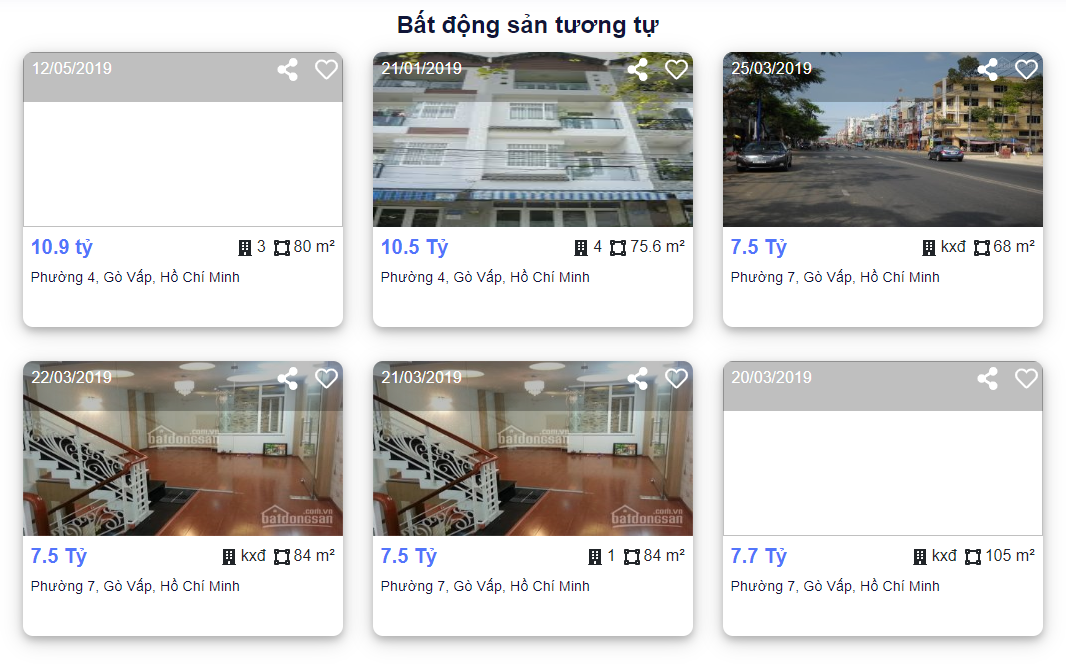
Giúp người dùng có thông tin của các khu vực liền kề với khu vực của người dùng.



**Hình 22: Màn hình thông tin khu vực xung quanh**

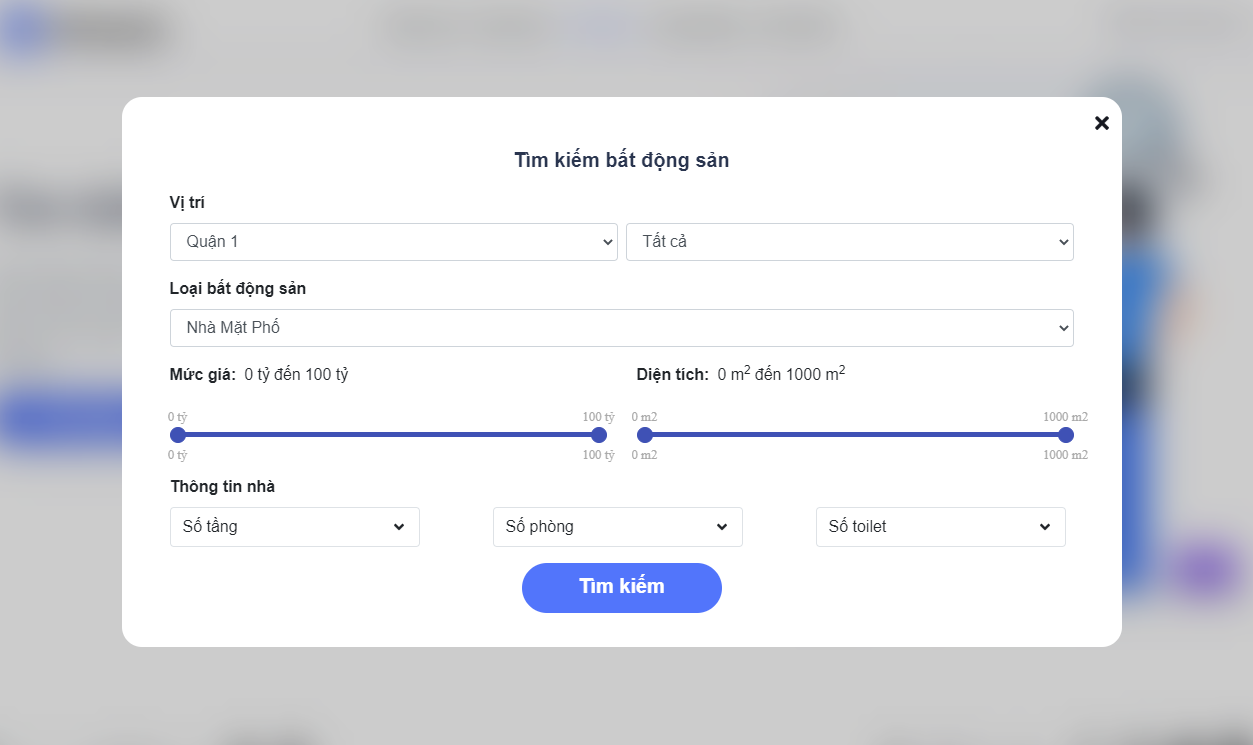
**9 Các bất động sản tương tự**

Website cung cấp danh sách các bất động sản thực tế có cùng các tiêu trí với bất động sản người dùng định giá. Giúp cho người dùng biết được những giá thực tế đang được rao bán cũng như tìm kiếm được bất động sản tốt hơn.

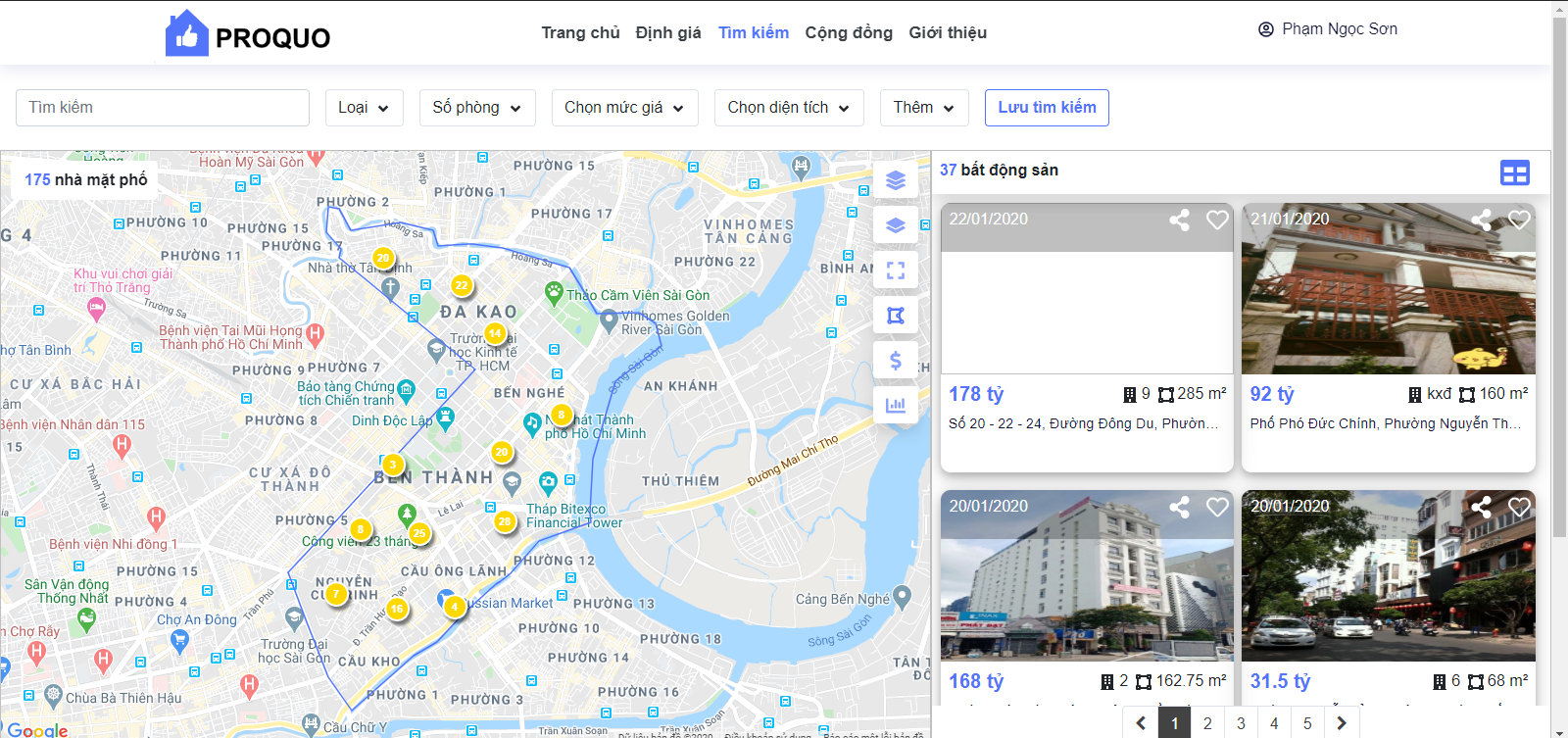


**Hình 23: Màn hình danh sách bất động sản tương tự**

**10 Tìm kiếm bất động sản**



**Hình 24: Màn hình nhập liệu khi tìm kiếm bất động sản**



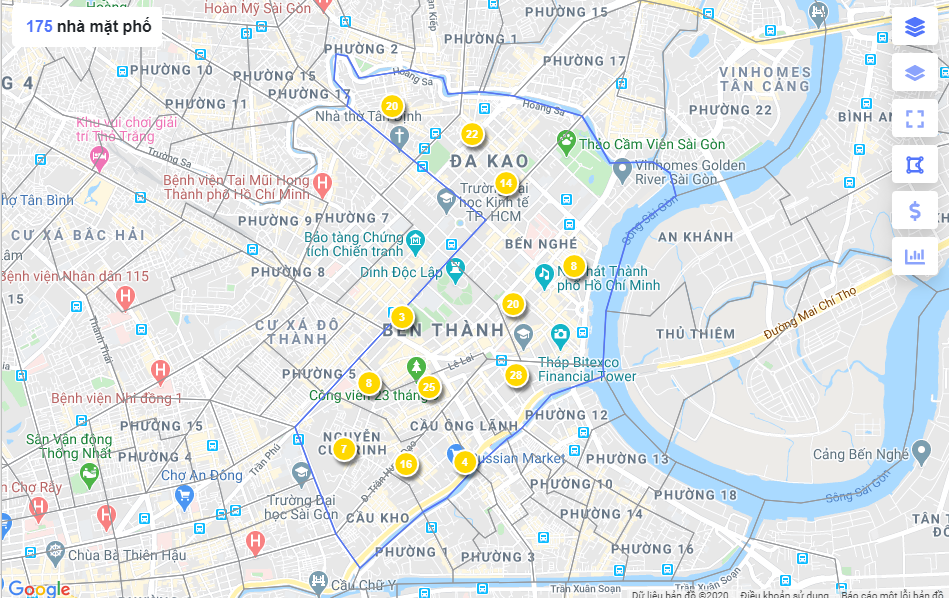
**Hình 25: Màn hình kết quả tìm kiếm**

Các chức năng trên bản bồ

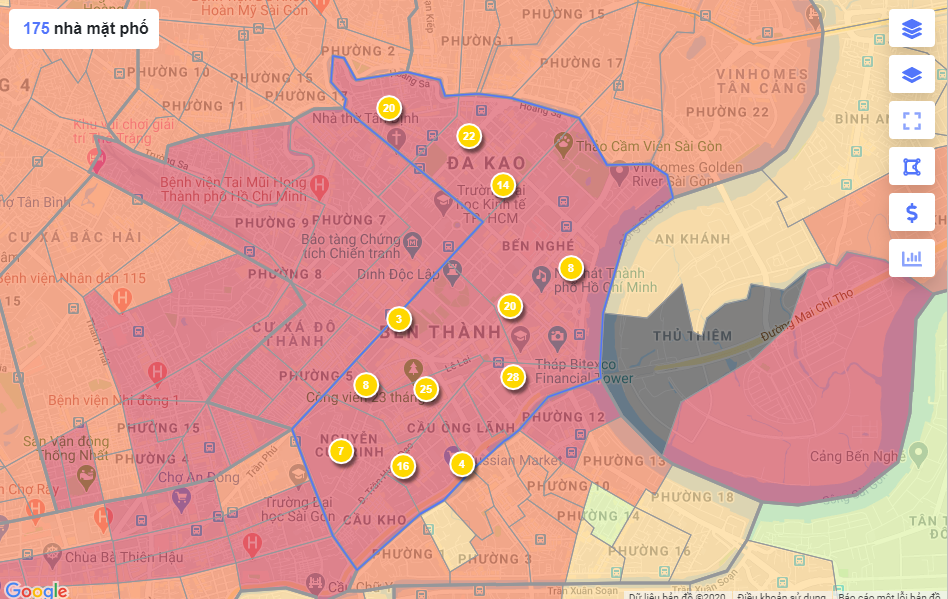
* Hiển thị ranh giới quận
* Hiển thị ranh giới huyện
* Hiển thị bản đồ màu giá khu vực
* Thống kê khu vực đang tìm kiếm

****

**Hình 26: Các lựa chọn trên bản đồ**



**Hình 27: Hiển thị các kết quả tìm kiếm và ranh giới khu vực**



**Hình 28: Hiển thị kết quả tìm kiếm trên bản đồ màu**

1. Các căn nhà không phải nhà mặt phố hay biệt thự. [↑](#footnote-ref-1)
2. <https://www.diaochoangkhang.vn/truyen-thong/kien-thuc-bat-dong-san/con-sot-dat-ao-dau-hieu-va-nhung-he-luy-can-luu-y> [↑](#footnote-ref-2)
3. <https://medium.com/@prashant.nair2050/hands-on-outlier-detection-and-treatment-in-python-using-1-5-iqr-rule-f9ff1961a414> [↑](#footnote-ref-3)
4. <https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a-random-forest-algorithm> [↑](#footnote-ref-4)
5. <https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f> [↑](#footnote-ref-5)
6. <https://dandautu.vn/kinh-nghiem/yeu-to-anh-huong-gia-tri-bat-dong-san> [↑](#footnote-ref-6)