|  |  |
| --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**ĐỒ ÁN CƠ SỞ**

**thu thập và phân tích giá chung cư trên trang web:batdongsan.com.vn**

Ngành: **Công Nghệ Thông Tin**

Chuyên ngành: **Khoa Học Dữ Liệu**

Giảng viên hướng dẫn: Lê Nhật Tùng

Sinh viên thực hiện:

Trần Văn Thiện MSSV:22864000 Lớp:22DKHA1  
Trần Minh Thông MSSV: 22864000 Lớp:22DKHA1

Trần Tuấn Kiệt MSSV:2286400013 Lớp: 22DKHA1

TP. Hồ Chí Minh, ngày 31 tháng 10 năm 2024

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

**KHOA CÔNG NGHỆ THÔNG TIN**

**HƯỚNG DẪN TRÌNH BÀY ĐỒ ÁN CƠ SỞ**

1. **BỐ CỤC**

Bố cục đồ án được minh họa qua mục lục:

|  |
| --- |
| **MỤC LỤC**  Trang phụ bìa  Lời cam đoan  Mục lục 1  Danh mục các ký hiệu, các chữ viết tắt 2  Danh mục các bảng 3  Danh mục các hình vẽ, đồ thị 4  Chương 1. TỔNG QUAN 6  1.1 … 6  1.1.1 ... 6  1.1.2… 7  1.2 …. 8  Chương 2. … 9  2.1 … 9  2.1.1… 9  2.1.2… 10  2.2 …  …  Chương … - KẾT LUẬN VÀ KIẾN NGHỊ xx  TÀI LIỆU THAM KHẢO xx  PHỤ LỤC (Nếu có) xx |

**1. Giới thiệu đề tài**

**1.1 Bối cảnh và lý do chọn đề tài**  
Trong những năm gần đây, tốc độ phát triển của các đô thị lớn ở Việt Nam, đặc biệt là TP.HCM, đã khiến giá nhà đất tăng mạnh. Điều này đặt ra thách thức cho các nhà đầu tư, nhà phát triển bất động sản và cả người mua nhà trong việc định giá, đầu tư và lựa chọn bất động sản phù hợp. Các yếu tố như vị trí, tiện ích, và tình hình kinh tế cũng ảnh hưởng lớn đến quyết định đầu tư.

**1.2 Mục đích của đề tài**  
Đề tài "Phân tích giá nhà" nhằm đưa ra một phương pháp khoa học để thu thập và phân tích dữ liệu về thị trường bất động sản, đặc biệt là phân khúc chung cư tại TP.HCM. Nghiên cứu này không chỉ hỗ trợ các nhà đầu tư xác định tiềm năng của từng khu vực mà còn giúp người mua có cái nhìn tổng quan về giá cả và điều kiện mua nhà.

**1.3 Lợi ích cho nhà đầu tư và người dùng cuối**  
Các nhà đầu tư sẽ có thể đưa ra các quyết định chính xác hơn dựa trên dữ liệu cụ thể và các phân tích về xu hướng giá cả. Người mua nhà cũng có thể sử dụng hệ thống để tra cứu các chung cư phù hợp với điều kiện tài chính và yêu cầu cá nhân của mình.

**2. Nhiệm vụ của đề tài**

Đề tài sẽ thực hiện một loạt các nhiệm vụ sau để hoàn thành mục tiêu đã đề ra:

**2.1 Thu thập dữ liệu**  
Áp dụng các kỹ thuật thu thập dữ liệu từ các nguồn đáng tin cậy, trong đó trang web Batdongsan.vn là một lựa chọn chủ yếu do chứa nhiều thông tin liên quan đến bất động sản ở Việt Nam. Quá trình này bao gồm việc xây dựng các công cụ thu thập (crawler) bằng các thư viện như Selenium để tự động hóa việc lấy dữ liệu từ trang web.

**2.2 Xây dựng cơ sở dữ liệu**  
Dữ liệu thu thập sẽ được tổ chức và lưu trữ trong MongoDB, một hệ quản trị cơ sở dữ liệu NoSQL mạnh mẽ, linh hoạt và phù hợp cho việc lưu trữ dữ liệu bán cấu trúc như thông tin bất động sản. Cơ sở dữ liệu này cần được tối ưu hóa để đảm bảo khả năng truy xuất nhanh và hiệu quả cho quá trình phân tích dữ liệu.

**2.3 Phân tích dữ liệu và trực quan hóa**  
Dữ liệu được xử lý và phân tích để đưa ra các biểu đồ trực quan giúp nhà đầu tư và người mua dễ dàng tiếp cận thông tin. Các thư viện như Matplotlib, Seaborn có thể được sử dụng để xây dựng các biểu đồ này, cung cấp thông tin về xu hướng giá cả theo khu vực, so sánh các tiện ích, và xác định các "điểm nóng" về giá trị bất động sản.

**3. Tính cấp thiết và ý nghĩa khoa học của đề tài**

**3.1 Tính cấp thiết của dự án**  
TP.HCM là thành phố lớn nhất Việt Nam với mật độ dân số cao và nhu cầu về nhà ở ngày càng tăng, đặc biệt là trong bối cảnh người dân đổ về đô thị để tìm kiếm cơ hội việc làm và cải thiện chất lượng cuộc sống. Với mức giá nhà đang tăng nhanh, nhiều người mua nhà gặp khó khăn trong việc tìm kiếm bất động sản phù hợp. Dự án này giúp người mua tiếp cận thông tin minh bạch, hỗ trợ họ trong việc lựa chọn đúng đắn và phù hợp với tài chính.

**3.2 Ý nghĩa khoa học**  
Đề tài không chỉ mang lại lợi ích thực tiễn mà còn có ý nghĩa khoa học khi ứng dụng các phương pháp xử lý dữ liệu hiện đại trong lĩnh vực bất động sản. Việc thu thập và phân tích dữ liệu một cách có hệ thống giúp tạo ra một bộ dữ liệu quan trọng cho các nghiên cứu tiếp theo và góp phần xây dựng nền tảng khoa học dữ liệu bất động sản ở Việt Nam.

**4. Mục tiêu**

**4.1 Mục tiêu chung**  
Xây dựng một cơ sở dữ liệu chi tiết và cập nhật về các đặc điểm và giá cả của các chung cư trên thị trường TP.HCM. Phân tích dữ liệu và phát triển hệ thống trực quan để cung cấp thông tin hữu ích cho người dùng.

**4.2 Mục tiêu cụ thể**

* **Thu thập dữ liệu**: Từ trang web Batdongsan.vn, thu thập thông tin chi tiết về giá cả, tiện ích, diện tích, vị trí của các chung cư.
* **Thiết kế cơ sở dữ liệu**: Lựa chọn MongoDB làm nền tảng để lưu trữ dữ liệu và tối ưu hóa cho khả năng truy xuất dữ liệu nhanh chóng.
* **Phân tích dữ liệu**: Sử dụng công cụ trực quan hóa dữ liệu để tạo ra các biểu đồ và đồ thị về xu hướng giá nhà, so sánh các khu vực, và phát hiện các khu vực tiềm năng.
* **Xây dựng hệ thống truy vấn**: Phát triển các chức năng tìm kiếm để người dùng có thể lọc và xem chi tiết các chung cư phù hợp với nhu cầu của họ.

**5. Đối tượng và phạm vi nghiên cứu**

**5.1 Đối tượng nghiên cứu**  
Đối tượng của nghiên cứu này là các dữ liệu liên quan đến chung cư tại TP.HCM, bao gồm các thông tin như giá bán, diện tích, tiện ích, và vị trí. Ngoài ra, đề tài cũng xem xét thị hiếu và nhu cầu của người mua nhà để từ đó tạo ra các báo cáo phân tích phù hợp.

**5.2 Phạm vi nghiên cứu**  
Phạm vi của đề tài là thị trường bất động sản chung cư tại TP.HCM. Đề tài sẽ tập trung vào việc thu thập dữ liệu từ các khu vực trung tâm, các quận có tốc độ phát triển cao, cũng như các khu vực đang nổi lên trong thời gian gần đây. Những dữ liệu này sẽ được giới hạn ở TP.HCM để đảm bảo tính chính xác và tập trung của nghiên cứu.

**6. Phương pháp nghiên cứu**

**6.1 Phương pháp thu thập dữ liệu**  
Dữ liệu sẽ được thu thập từ các trang web như Batdongsan.vn bằng cách sử dụng kỹ thuật web scraping. Công cụ Selenium sẽ được sử dụng để tự động hóa quá trình thu thập dữ liệu và lưu trữ chúng trong cơ sở dữ liệu MongoDB. Để đảm bảo tính hợp pháp và tuân thủ quy định, việc thu thập dữ liệu sẽ thực hiện trong giới hạn tần suất cho phép.

**6.2 Phương pháp phân tích dữ liệu**  
Sau khi thu thập, dữ liệu sẽ được xử lý và phân tích bằng các công cụ như Pandas và Scikit-learn. Các kỹ thuật như phân tích hồi quy sẽ được áp dụng để hiểu mối quan hệ giữa giá cả và các đặc điểm của chung cư. Ngoài ra, các công cụ trực quan hóa như Matplotlib và Seaborn sẽ được sử dụng để tạo các biểu đồ trực quan giúp minh họa xu hướng giá cả và các "điểm nóng" về bất động sản.

**6.3 Phương pháp trực quan hóa dữ liệu**  
Để giúp người dùng dễ dàng tiếp cận và hiểu thông tin, đề tài sẽ sử dụng các biểu đồ và bản đồ để trình bày dữ liệu. Các biểu đồ giá nhà theo thời gian, bản đồ các khu vực tiềm năng và biểu đồ so sánh giữa các chung cư sẽ được sử dụng để người dùng có thể nắm bắt thông tin một cách trực quan.

**7. Tổng quan lý thuyết**

7.1 Khái niệm và phân loại bất động sản Bất động sản là một loại tài sản bao gồm đất đai và những tài sản gắn liền với đất, ví dụ như nhà cửa, các công trình xây dựng. Bất động sản có thể chia thành nhiều loại như đất ở, đất thương mại, đất công nghiệp, và các loại hình bất động sản hỗn hợp khác. Trong phạm vi nghiên cứu này, chúng ta sẽ tập trung vào phân khúc chung cư - một dạng bất động sản nhà ở phổ biến trong các đô thị lớn như TP.HCM.

7.2 Các yếu tố ảnh hưởng đến giá nhà Giá trị bất động sản nói chung và giá nhà nói riêng phụ thuộc vào nhiều yếu tố, bao gồm vị trí, cơ sở hạ tầng, tiện ích xung quanh, tình hình kinh tế, và chính sách của nhà nước. Những yếu tố này có tác động khác nhau ở mỗi khu vực. Trong TP.HCM, vị trí địa lý có thể coi là yếu tố quyết định lớn nhất đối với giá nhà. Ngoài ra, việc tiếp cận với các khu vực tiện ích như trường học, bệnh viện, trung tâm mua sắm, và giao thông công cộng cũng ảnh hưởng đáng kể đến giá trị của các căn hộ.

7.3 Ứng dụng khoa học dữ liệu trong bất động sản Khoa học dữ liệu hiện nay được sử dụng rộng rãi trong lĩnh vực bất động sản để phân tích và dự đoán giá trị bất động sản. Các phương pháp khoa học dữ liệu, như phân tích hồi quy, phân tích cụm, và các mô hình học máy khác, có khả năng cung cấp các dự đoán chính xác hơn về xu hướng giá nhà, từ đó hỗ trợ các nhà đầu tư và người mua đưa ra quyết định thông minh hơn. Việc áp dụng các mô hình dữ liệu hiện đại vào nghiên cứu giá nhà sẽ là một bước tiến quan trọng trong ngành bất động sản ở Việt Nam.

**8. Tình hình thị trường bất động sản tại TP.HCM**

8.1 Tăng trưởng dân số và nhu cầu nhà ở TP.HCM là thành phố đông dân nhất Việt Nam và có tốc độ gia tăng dân số nhanh chóng. Điều này làm tăng nhu cầu về nhà ở, đặc biệt là ở các khu vực trung tâm thành phố và các quận đang phát triển nhanh. Theo thống kê, nhu cầu về nhà ở tại TP.HCM tăng đều đặn qua các năm, trong đó phần lớn là nhu cầu về căn hộ chung cư - loại hình phù hợp với người dân đô thị do tiện lợi và chi phí hợp lý.

8.2 Các khu vực "nóng" của thị trường Các quận trung tâm như Quận 1, Quận 3, Bình Thạnh, và Quận 2 thường có giá nhà đất cao do có hệ thống cơ sở hạ tầng phát triển, tiện ích đa dạng và gần các trung tâm kinh tế. Bên cạnh đó, một số quận ven như Quận 9, Thủ Đức (nay là thành phố Thủ Đức), và Quận 7 cũng đang nổi lên với nhiều dự án bất động sản mới, tạo ra nhiều cơ hội cho các nhà đầu tư.

8.3 Thách thức của thị trường bất động sản TP.HCM Thị trường bất động sản TP.HCM không chỉ mang lại nhiều cơ hội mà còn tiềm ẩn nhiều thách thức. Việc giá nhà đất tăng cao, hạ tầng giao thông chưa đáp ứng kịp nhu cầu, và tình trạng "bong bóng" bất động sản là những yếu tố đáng lo ngại. Đối với người mua nhà, việc tìm kiếm một căn hộ có giá hợp lý và đáp ứng nhu cầu trở thành một vấn đề nan giải. Đề tài này sẽ giúp cung cấp thông tin và dự đoán xu hướng thị trường để người dùng có thể đưa ra quyết định tốt hơn.

**9. Chiến lược phân tích dữ liệu**

9.1 Phân tích mô tả (Descriptive Analysis) Trong giai đoạn đầu, dữ liệu thu thập được sẽ được phân tích mô tả để cung cấp cái nhìn tổng quan về thị trường chung cư TP.HCM. Phân tích mô tả bao gồm việc thống kê về giá bán trung bình, diện tích trung bình, số lượng phòng ngủ, nhà vệ sinh, và tiện ích của các chung cư. Các thông tin này sẽ được trình bày qua các biểu đồ, bảng số liệu để người đọc dễ dàng hình dung.

9.2 Phân tích khám phá (Exploratory Data Analysis - EDA) Phân tích khám phá sẽ giúp tìm ra mối quan hệ giữa các yếu tố ảnh hưởng đến giá nhà. Ví dụ, chúng ta sẽ xem xét các yếu tố như vị trí địa lý, tiện ích xung quanh, và diện tích ảnh hưởng đến giá bán của các chung cư như thế nào. Việc phân tích khám phá giúp xác định các "điểm nóng" về giá trị bất động sản và có thể làm cơ sở cho việc xây dựng mô hình dự đoán.

9.3 Phân tích dự báo (Predictive Analysis) Để đưa ra các dự đoán về giá nhà trong tương lai, chúng ta sẽ sử dụng các mô hình học máy như hồi quy tuyến tính (Linear Regression), cây quyết định (Decision Tree) và các thuật toán khác phù hợp. Phân tích dự báo sẽ cho phép xác định các yếu tố quan trọng nhất ảnh hưởng đến giá nhà và đưa ra các dự báo giá cho các khu vực khác nhau.

**10. Công cụ và nền tảng hỗ trợ**

10.1 Các công cụ lập trình và phân tích dữ liệu Python là ngôn ngữ lập trình chính được sử dụng trong nghiên cứu này, nhờ vào các thư viện mạnh mẽ dành cho khoa học dữ liệu như Pandas, Scikit-learn, Matplotlib và Seaborn. Những thư viện này cho phép xử lý và phân tích dữ liệu một cách linh hoạt, đồng thời tạo ra các biểu đồ và đồ thị để trực quan hóa dữ liệu.

10.2 Cơ sở dữ liệu MongoDB MongoDB là một lựa chọn phù hợp cho việc lưu trữ dữ liệu bán cấu trúc như dữ liệu bất động sản. Nó cho phép dễ dàng mở rộng và hỗ trợ truy vấn dữ liệu nhanh chóng, đáp ứng nhu cầu xử lý lượng dữ liệu lớn trong nghiên cứu này.

10.3 Môi trường chạy và phát triển Ngoài Python và MongoDB, chúng ta sẽ sử dụng Jupyter Notebook để phát triển và chạy các tập lệnh phân tích. Jupyter Notebook là môi trường phù hợp cho các dự án khoa học dữ liệu nhờ khả năng hiển thị kết quả tức thì và giúp kiểm tra, đánh giá mã một cách linh hoạt.

**11. Các khó khăn và giải pháp**

11.1 Khó khăn trong quá trình thu thập dữ liệu Do các trang web bất động sản thường có cơ chế chặn bot, việc thu thập dữ liệu sẽ phải tuân thủ các quy định của trang web và đảm bảo không vượt quá tần suất cho phép. Giải pháp là thực hiện thu thập dữ liệu một cách hợp lý, sử dụng proxy và các khoảng thời gian nghỉ giữa các lần truy vấn để tránh bị chặn.

11.2 Xử lý dữ liệu không đồng nhất Dữ liệu từ các nguồn có thể không đồng nhất về định dạng và chất lượng. Do đó, việc làm sạch và tiền xử lý dữ liệu là rất quan trọng. Các kỹ thuật như loại bỏ dữ liệu bị thiếu, chuẩn hóa dữ liệu và xử lý các giá trị bất thường sẽ được áp dụng để đảm bảo dữ liệu có chất lượng cao cho phân tích.

11.3 Độ chính xác của mô hình dự đoán Việc dự đoán giá nhà không dễ dàng vì có nhiều yếu tố tác động và một số yếu tố có thể biến đổi theo thời gian. Để cải thiện độ chính xác, chúng ta sẽ thử nghiệm với nhiều mô hình khác nhau và sử dụng phương pháp điều chỉnh siêu tham số (hyperparameter tuning) để tìm ra mô hình phù hợp nhất.

**12. Kết quả dự kiến và ứng dụng**

12.1 Kết quả dự kiến Dự án sẽ cung cấp một cơ sở dữ liệu chi tiết về thị trường chung cư tại TP.HCM, cùng với hệ thống trực quan hóa dữ liệu để người dùng có thể dễ dàng xem các thông tin về giá cả, diện tích, tiện ích và các yếu tố khác. Ngoài ra, các mô hình dự đoán sẽ giúp nhà đầu tư và người mua nắm bắt được xu hướng giá trong tương lai và có quyết định phù hợp.

12.2 Ứng dụng thực tiễn Các nhà đầu tư có thể sử dụng các kết quả phân tích và dự đoán để xác định các khu vực tiềm năng, tối ưu hóa chiến lược đầu tư. Người mua nhà có thể dễ dàng tìm kiếm và so sánh các chung cư để đưa ra quyết định phù hợp với nhu cầu cá nhân. Dự án này cũng sẽ là một nền tảng tham khảo cho các nhà nghiên cứu và phát triển công nghệ trong lĩnh vực bất động sản ở Việt Nam.

**13. Đóng góp của đề tài**

13.1 Đóng góp thực tiễn Đề tài này mang lại lợi ích thực tiễn cho cả nhà đầu tư và người mua nhà, giúp họ ra quyết định chính xác và tiết kiệm thời gian trong quá trình tìm kiếm bất động sản. Đây là một công cụ hữu ích, minh bạch giúp tối ưu hóa quá trình mua bán và đầu tư bất động sản.

13.2 Đóng góp khoa học Đề tài cũng có ý nghĩa khoa học khi áp dụng các phương pháp xử lý dữ liệu và học máy hiện đại vào phân tích thị trường bất động sản. Nó góp phần xây dựng cơ sở dữ liệu lớn cho các nghiên cứu tiếp theo và tạo điều kiện để phát triển các công nghệ tiên tiến trong lĩnh vực này.

**14. Kết luận và Kỳ vọng của đề tài**

Dự án này hy vọng sẽ cung cấp một cơ sở dữ liệu chi tiết, hữu ích cho các nhà đầu tư và người mua nhà, cũng như xây dựng một hệ thống truy vấn mạnh mẽ để người dùng có thể dễ dàng tìm kiếm thông tin về các chung cư phù hợp. Các kết quả phân tích không chỉ hỗ trợ các quyết định đầu tư mà còn giúp người mua tiếp cận thị trường bất động sản một cách dễ dàng hơn.

Ngoài ra, đề tài này kỳ vọng sẽ trở thành một nguồn dữ liệu tham khảo quan trọng cho các nghiên cứu tiếp theo về bất động sản ở Việt Nam. Việc áp dụng các phương pháp khoa học dữ liệu cũng tạo cơ hội để phát triển các công cụ hỗ trợ và phân tích thị trường trong tương lai.