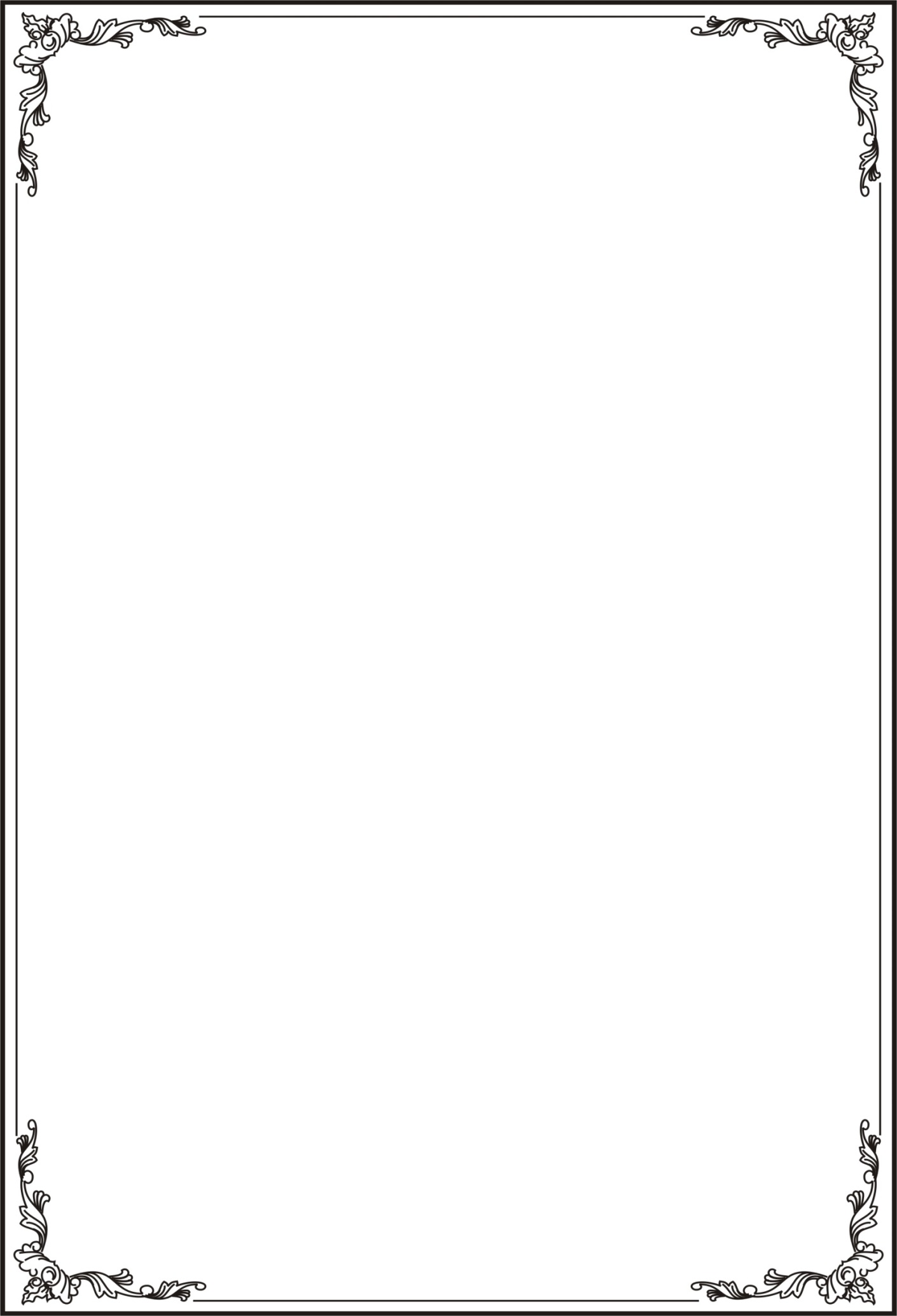
****

**PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

**----------------**



**BÁO CÁO MÔN CHUYÊN ĐỀ CNTT**

**Tên đề tài:**

**Phân tích dữ liệu và dự đoán giá chung cư tại**

**Tp. Hồ Chí Minh**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | ***ThS. Viên Thanh Nhã*** |
| **Sinh viên thực hiện:** | ***Trần Ngọc Hải - 2051067527*** |
|  | ***Nguyễn Như Ngọc - 2051067845*** |
| **Lớp:** | ***S23 – 62TH*** |

**TP.HCM, ngày tháng năm 2023**

Phụ lục

[Lời mở đầu 3](#_Toc146135770)

[Thuật Ngữ 4](#_Toc146135771)

[Bảng phân công công việc 7](#_Toc146135772)

[Chương I: XÁC ĐỊNH VÀ PHÂN TÍCH YÊU CẦU 8](#_Toc146135773)

[1. Giới thiệu chung 8](#_Toc146135774)

[2. Xác định yêu cầu 8](#_Toc146135775)

[3. Dự kiến kết quả đạt được 9](#_Toc146135776)

[4. Công cụ thực hiện đồ án 9](#_Toc146135777)

[Chương II: CƠ SỞ LÝ THUYẾT 9](#_Toc146135778)

[1. Tổng quan môn học 9](#_Toc146135779)

[2. Thuật toán Linear Regression 10](#_Toc146135780)

[3. Thuật toán Cây quyết định 10](#_Toc146135781)

[Chương III: MÔ HÌNH 10](#_Toc146135782)

[1. Mô tả các chức năng của tập dữ liệu 10](#_Toc146135783)

[2. Tiền xử lý dữ liệu (Data Preprocessing) 11](#_Toc146135784)

[3. Thuật toán hồi quy (Linear Regression) 14](#_Toc146135785)

[4. Thuật toán Cây quyết định (Decision Tree) 14](#_Toc146135786)

[Chương IV: Kết quả đạt được 14](#_Toc146135787)

[Chương V: DEMO - Hướng phát triển 14](#_Toc146135788)

# Lời mở đầu

Thành phố Hồ Chí Minh, là trung tâm kinh tế và tài chính hàng đầu của Việt Nam, đang trải qua một sự phát triển nhanh chóng trong lĩnh vực bất động sản, đặc biệt là thị trường chung cư. Với sự gia tăng đáng kể về dân số và nhu cầu về nhà ở, việc phân tích dữ liệu và dự đoán giá chung cư đã trở thành một yếu tố quan trọng để quản lý và định hình thị trường bất động sản đô thị.

Báo cáo môn học này tập trung vào việc áp dụng các phương pháp phân tích dữ liệu và dự đoán giá chung cư trong thành phố Hồ Chí Minh. Chúng em sử dụng các kỹ thuật, thuật toán trong Data Mining để xử lý và phân tích dữ liệu liên quan đến thị trường chung cư, từ đó giúp dự đoán và hiểu rõ hơn về xu hướng giá cả và yếu tố ảnh hưởng đến giá chung cư trong thành phố này.

Đặc biệt, nhóm em gửi lời cảm ơn sâu sắc và chân thành nhất đến Giảng viên Viên Thanh Nhã. Thầy đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báo giúp nhóm em hoàn thành tốt báo cáo đồ án của mình. Trong thời gian thực hiện đề tài, nhóm tụi đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới.

Xin chân thành cảm ơn các quý Thầy Cô !

# Thuật Ngữ

**Thuật ngữ công nghệ thông tin** là một trong những kiến thức cơ bản và quan trọng mà bất cứ ai đang hoạt động trong lĩnh vực này cũng đều nên biết. Việc nắm rõ cách thức sử dụng của chúng cũng vô cùng quan trọng. Dưới đây là một số liệt kê các thuật ngữ CNTT được dùng phổ biến tại Việt Nam hiện nay.

**Phần cứng (Hardware)** là từ dùng để gọi các bộ phận vật lý mà bạn có thể nhìn thấy, sờ được như màn hình, bàn phím, chuột, dây cáp, bộ vi xử lý CPU.

**Phần mềm (Software)** là thuật ngữ dùng để chỉ tập hợp các tập tin, dữ liệu hoặc câu lệnh có liên quan đến nhau nhằm thực hiện những nhiệm vụ, chức năng nhất định trên máy tính.

* Ứng dụng (Application) : Phần mềm được thiết kế để thực hiện các công cụ nhiệm vụ như xử lý văn bản, trình duyệt web hoặc trò chơi.
* Hệ điều hành (Operating System – OS) : Phần mềm quản lý và điều khiển tài nguyên của máy tính, cho phép các ứng dụng khác chạy trên nó.
* Mã nguồn mở (Open Source) : Phần mềm có mã nguồn mở được cộng đồng phát triển và sửa lỗi.

**Phát triển phần mềm (Software Development):** là quá trình tạo và phát triển các ứng dụng, hệ thống hoặc chương trình máy tính. Quá trình này bao gồm các công cụ thiết kế, viết mã nguồn, kiểm tra và phát triển phần mềm.

* Lập trình (Programming) : Tạo mã hóa để xây dựng ứng dụng và phần mềm hệ thống.
* Hợp nhất liên tục (Continuous Integration - CI) : Phương pháp tự động hóa công việc kiểm tra và tích hợp mã nguồn mới vào liên tục dự án phần mềm.
* SCRUM: Một phương pháp quản lý dự án phát triển phần mềm dựa trên quá trình làm việc theo các "sprint" (đợt phát triển ngắn hạn).
* API (Application Programming Interface): Giao diện cho phép các phần mềm khác tương tác với ứng dụng hoặc dịch vụ khác.

**Cookie**: Những tập tin mà một website lưu trên ổ cứng máy tính của người dùng khi họ truy cập trang web đó được gọi là Cookie.

**URL (Định vị tài nguyên thống nhất)**: Đây là thuật ngữ chỉ địa chỉ tham chiếu đến các tài nguyên web trên một mạng máy tính. URL hợp lệ chỉ dẫn đến duy nhất một tài nguyên như tài liệu CSS, trang HTML, tập tin PDF,… và có thể trỏ đến những tài nguyên đã bị xóa hoặc di chuyển.

**Server (Máy chủ)**: Đây là một hệ thống máy tính được thiết lập để hỗ trợ cung cấp dịch vụ mạng cho các máy tính khác và lưu trữ ở cùng một nơi.

**Mạng máy tính (Computer Networking):** Thuật ngữ này để chỉ một nhóm gồm tối thiểu 2 máy tính được kết nối với nhau để dùng chung các tài nguyên.

* IP (Internet Protocol): Giao thức quy định cách dữ liệu được định tuyến trên Internet.
* Mạng cục bộ (LAN – Local Area Network) : Mạng máy tính trong một khu vực nhỏ như một văn phòng hoặc nhà riêng.
* Mạng diện rộng (WAN - Wide Area Network) : Mạng máy tính kết nối các mạng LAN lớn hơn, thường xuyên trải qua các khu vực khác nhau.

**Bảo mật thông tin (Information Security)**:

* Tường lửa (Firewall): Hệ thống bảo mật giúp kiểm soát và giám sát lưu lượng mạng.
* Xác thực hai yếu tố (Two-Factor Authentication - 2FA) : Quy trình bảo mật yêu cầu hai phương tiện xác thực khác nhau để truy cập tài khoản.
* Sự tấn công mạng (Cyberattack): Hành động xâm nhập hoặc tấn công vào hệ thống máy tính hoặc mạng để gây hại.

**Trí tuệ nhân tạo (Artificial Intelligence - AI):**

* Học máy (Machine Learning) : Một loại trí tuệ nhân tạo cho phép máy tính học hỏi từ dữ liệu và cải thiện hiệu suất theo thời gian.
* Học sâu (Deep Learning) : Một nhánh của máy học sử dụng mạng nơ-ron sâu để xử lý phức tạp dữ liệu.

**Lưu trữ dữ liệu (Data Storage)** :

* Cơ sở dữ liệu (Cơ sở dữ liệu) : Hệ thống để lưu trữ và truy xuất dữ liệu.
* Đám mây (Cloud) : Lưu trữ và xử lý dữ liệu trên máy chủ từ xa được quản lý bởi các nhà cung cấp dịch vụ đám mây.

**Internet vạn vật ( Internet of Things - IoT)** :

* Kết nối vạn vật (Connected Devices) : Thiết bị không có khả năng thực hiện chức năng của nó nhưng vẫn kết nối và gửi dữ liệu qua internet.

**Phân tích dữ liệu (Data Analysis)**:

* Khai phá dữ liệu (Data Mining) : Quá trình tìm hiểu ẩn thông tin trong dữ liệu lớn.
* Trực quan hóa dữ liệu (Data Visualization) : Biểu đồ và đồ họa trực quan để hiển thị dữ liệu một cách trực quan.

**Trình lập trình ngôn ngữ (Programming Languages)**:

* Python : Lập trình ngôn ngữ phổ biến để phát triển ứng dụng, học máy và khoa học dữ liệu.
* Java : Sử dụng rộng rãi cho web ứng dụng và phát triển ứng dụng đa nền tảng.

**Công nghệ đám mây (Cloud Technology)**:

* SaaS (Software as a Service) : Mô hình phân phối phần mềm dựa trên đám mây, người dùng truy cập phần mềm qua internet thay vì cài đặt nó trên máy tính cá nhân.

Hãy nhớ rằng lĩnh vực Công nghệ Thông tin rất đa dạng và luôn tiến triển, vì vậy cơ hội học hỏi và phát triển kiến ​​thức là vô hạn.

# Bảng phân công công việc

|  |  |
| --- | --- |
| **Thành viên** | **Công việc** |
| Trần Ngọc Hải | * Tìm kiếm dataset và thu thập dữ liệu từ các trang mạng nhà ở và chung cư * Viết báo cáo |
| Nguyễn Như Ngọc | * Clean data * Viết báo cáo |

# Chương I: XÁC ĐỊNH VÀ PHÂN TÍCH YÊU CẦU

## Giới thiệu chung

Dự đoán giá chung cư tại Thành phố Hồ Chí Minh hoặc bất kỳ thị trường bất động sản nào là một nhiệm vụ phức tạp và khó khăn. Giá chung cư được ảnh hưởng bởi nhiều yếu tố, bao gồm vị trí, diện tích, chất lượng xây dựng, tiện ích, tình trạng thị trường và các yếu tố kinh tế và xã hội khác.

Lưu ý rằng dự đoán giá chung cư là một nhiệm vụ phức tạp và có thể liên quan đến rủi ro. Không có cách dự đoán chính xác 100% giá bất động sản trong tương lai. Thông tin từ nhiều nguồn và sự tư duy logic cùng với phân tích dữ liệu là cách tốt nhất để có cái nhìn tổng quan về tình hình thị trường và đưa ra dự đoán tương đối. Đối với đầu tư bất động sản, luôn lưu ý rằng nên tham khảo ý kiến chuyên gia và cân nhắc rủi ro trước khi ra quyết định.

## Xác định yêu cầu

Để dự đoán giá chung cư một cách chính xác, bạn cần xem xét các phương pháp và nguồn thông tin sau:

* Phân tích dữ liệu lịch sử: Nghiên cứu dữ liệu giá bất động sản chung cư trong quá khứ để hiểu xu hướng và biến động giá. Sử dụng các dịch vụ hoặc trang web cung cấp thông tin về giá như những giao dịch bất động sản đã diễn ra trong khu vực bạn quan tâm.
* Phân tích thị trường: Nắm bắt thông tin về tình trạng thị trường bất động sản tại Thành phố Hồ Chí Minh. Điều này bao gồm tình hình cung và cầu, tình trạng khu vực cụ thể, và các dự án bất động sản sắp tới.
* Ước tính từ các chuyên gia: Tìm hiểu ý kiến của các chuyên gia bất động sản về tình hình thị trường và dự đoán giá trong tương lai.
* Sử dụng mô hình dự đoán: Các mô hình dự đoán dựa trên dữ liệu có thể giúp bạn ước tính giá chung cư dựa trên các yếu tố quan trọng như vị trí, diện tích, và tiện ích. Mô hình hồi quy tuyến tính hoặc mô hình học máy như Random Forest hoặc Gradient Boosting có thể được sử dụng cho mục đích này.Tìm hiểu về các yếu tố tác động: Hiểu các yếu tố tác động đối với giá chung cư như hạ tầng, kế hoạch phát triển của khu vực, các dự án quy hoạch, và tình hình kinh tế tổng thể.

## Dự kiến kết quả đạt được

* Phân tích dữ liệu và dự đoán giá chung cư trong thành phố Hồ Chí Minh
* Chạy được thuật toán Linear Regression và nêu ra được kết luận.
* Chạy được thuật toán Cây quyết định và nêu ra được kết luận.

## Công cụ thực hiện đồ án

* Nhóm tụi em dùng **Colaboratory** hay còn gọi là **Google Colab** để thực hiện đồ án.
* Google Colab là một sản phẩm từ Google Research, nó cho phép chạy các dòng code python thông qua trình duyệt, đặc biệt phù hợp với Data analysis, machine learning và giáo dục.

# Chương II: CƠ SỞ LÝ THUYẾT

## Tổng quan môn học

Khoa học dữ liệu (Data Science) là một lĩnh vực nghiên cứu và ứng dụng dữ liệu để hiểu, phân tích, dự đoán và tạo ra giá trị từ thông tin chứa trong dữ liệu. Khoa học dữ liệu sử dụng kết hợp các phương pháp từ nhiều lĩnh vực như thống kê, toán học, khoa học máy tính, và nguồn dữ liệu thực tế để trích xuất thông tin hữu ích và kiến thức từ dữ liệu.

Khai phá dữ liệu (Data Mining) là quá trình tự động hoặc bán tự động tìm kiếm thông tin có giá trị từ tập dữ liệu lớn. Mục tiêu chính của khai phá dữ liệu là khám phá các mẫu, quy luật, sự kết nối, và thông tin ẩn trong dữ liệu mà có thể không dễ thấy bằng cách thường xuyên.

Tóm lại, khoa học dữ liệu là một lĩnh vực tổng hợp hơn, bao gồm cả việc khai phá dữ liệu, trong khi khai phá dữ liệu tập trung chủ yếu vào việc khám phá dữ liệu một cách tự động để tìm ra thông tin thú vị và quy luật trong dữ liệu.

## Thuật toán Linear Regression

Hồi quy tuyến tính (Linear Regression) là một trong những thuật toán máy học đơn giản và mạnh mẽ nhất được sử dụng để dự đoán một biến phụ thuộc liên tục dựa trên một hoặc nhiều biến độc lập. Thuật toán này xây dựng một mô hình tuyến tính để mô phỏng mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc.

Linear Regression là một công cụ quan trọng trong phân tích dữ liệu và dự đoán trong nhiều lĩnh vực, bao gồm kinh doanh, khoa học xã hội, y học, và nhiều lĩnh vực khác.

## Thuật toán Cây quyết định

Cây quyết định (Decision Tree) là một mô hình học máy phân loại và dự đoán được sử dụng trong nhiều ứng dụng khác nhau. Nó là một biểu đồ cây có cấu trúc, mô phỏng quá trình ra quyết định hoặc dự đoán bằng cách sử dụng luật điều kiện và giá trị của các thuộc tính.

Một cây quyết định được xây dựng từ dữ liệu huấn luyện bằng cách phân chia tập dữ liệu thành các phần con (subsets) dựa trên các thuộc tính. Quá trình này tiếp tục cho đến khi mỗi phần con chỉ chứa các điểm dữ liệu thuộc cùng một lớp hoặc có giá trị dự đoán gần nhau.

# Chương III: MÔ HÌNH

## Mô tả các chức năng của tập dữ liệu

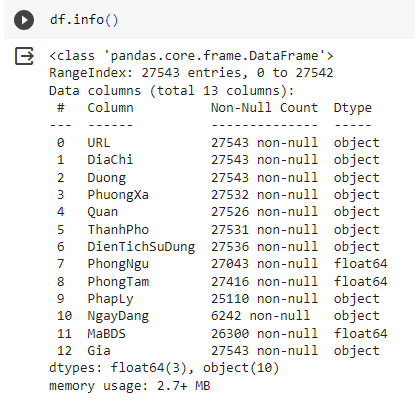
* DiaChi: địa chỉ của chung cư, ở thành phố Hồ Chí Minh
* DienTich: diện tích thực ở (sử dụng) trên sổ hồng, đơn vị: triệu/m2.
* PhongNgu: số lượng phòng ngủ.
* TenPhanKhu: căn hộ đó thuộc block nào trong khu chung cư ấý. Vì 1 khu chung cư có rất nhiều block, các block ở vị trí khác nhau sẽ có giá khác nhau.
* SoTang: căn hộ nằm ở tầng thứ mấy.
* PhongTam: số lượng nhà vệ sinh.
* Loại: chung cư hay nhà ở xã hội.
* GiayTo: giấy tờ pháp lý của căn hộ, có đang tranh chấp hay không, có hợp pháp hay không.
* MaCanHo: mã căn hộ (giống như số nhà).
* TinhTrangNoiThat: căn hộ đã có nội thất hay chưa(sofa, lò vi sóng, máy lạnh,...).
* HuongCuaChinh: hướng cửa chính của căn hộ.
* HuongBanCong: hướng ban công của căn hộ.
* DacDiem: Đặc điểm căn hộ ( căn trong góc, hay căn chính giữa,...).
* Gia: giá bán của căn hộ.

## Tiền xử lý dữ liệu (Data Preprocessing)

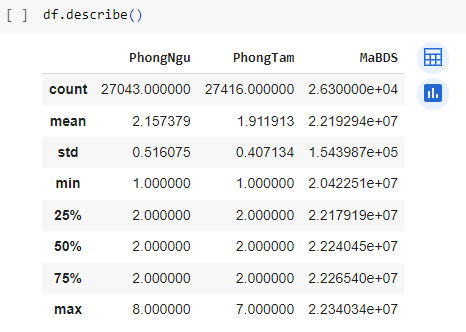
* QUY TRÌNH KHAI THÁC DỮ LIỆU
* Bộ dữ liệu thô chưa qua xử lý



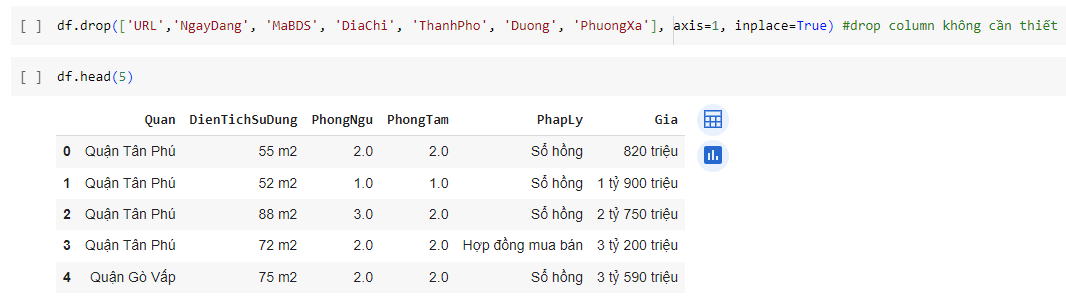
* Tổng quan về dữ liệu



* Mô tả dữ liệu số

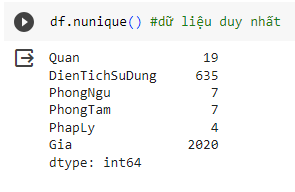


* Tiến hành drop đi các cột không cần thiết để xây model trong bộ dữ liệu và không ảnh hưởng đến quá trình dự đoán thuật toán

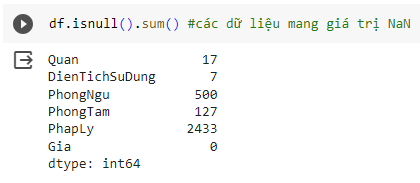


Sau khi drop xong chngs em chỉ còn giữ lại các cột cần thiết trong việc xây dựng model như : Quan , DienTichSuDung, PhongNgu, PhongTam, PhapLy, Gia

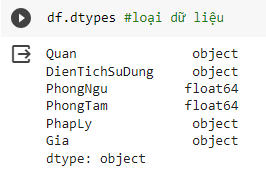
* Các giá trị duy nhất của từng cột



* Kiểm tra các dữ liệu mang giá trị NaN



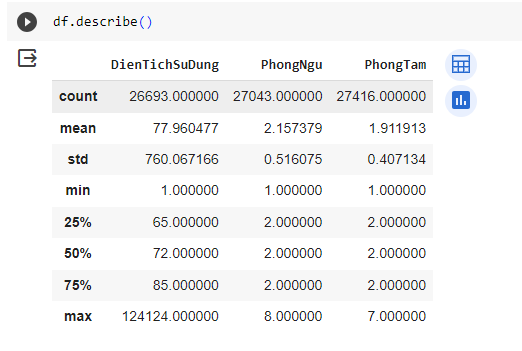
* Kiểu dữ liệu ở từng cột



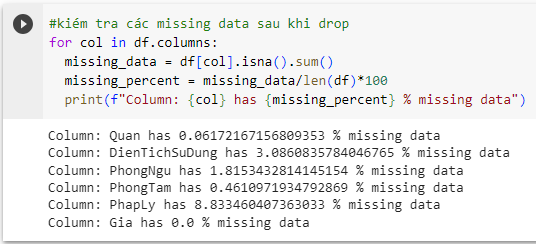
* BIẾN ĐỔI DỮ LIỆU
* Kiểu dữ liệu dạng String gồm có: Quan, DienTichSuDung, PhapLy, Gia
* Kiểu dữ liệu dạng số gồm có: PhongNgu, PhongTam
* Sau đó, chúng em tiếp tục chuyển đổi dữ liệu của cột “DienTichSuDung” đang là kiểu String sang dạng Numeric bằng code bên dưới để xóa bỏ đi đơn vị “m2” bên trong cột chỉ còn hiện số ở dạng float. Mục đích để chúng em làm vậy là vì muốn tất cả các cột bên dưới chỉ còn hiện những con số ở dang int64 hoặc float64 để thuận tiện hơn trong việc xây model.



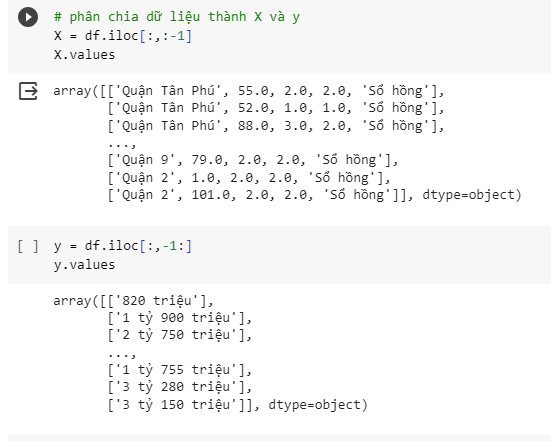
* Bộ mô tả dữ liệu dạng số sau khi chuyển đổi



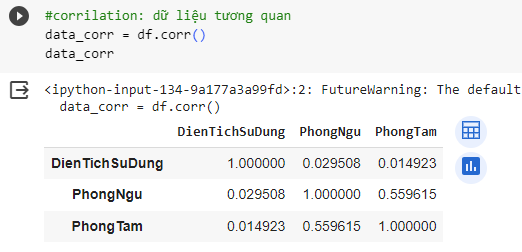
* Kiểm tra các giá trị bị missing data

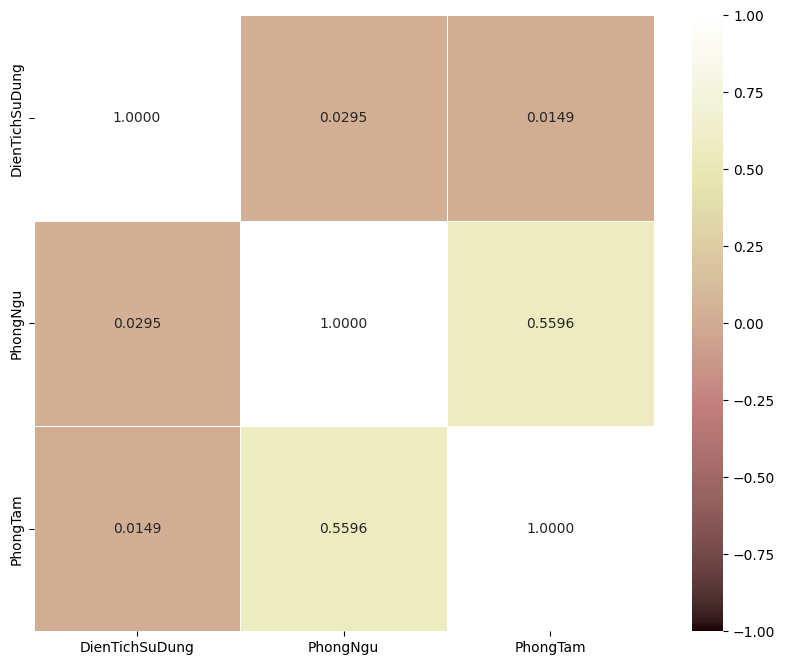


* Phân chia vùng dữ liệu thành X và y
* X gồm có các cột: Quan, DienTichSuDung, PhongNgu,PhongTam, PhapLy
* y gồm có các cột: Gia



* Tiến hành so sánh các dữ liệu tương quan với nhau





Dựa vào bảng vẽ trên ta có thể thấy được sự tương quan giữa các cột với nhau như:

* Cột DienTichSuDung có sự tương quan thấp với cột PhongNgu (có độ tương quan 0.0295) và PhongTam (có độ tương quan là 0.0149)
* Ở cột PhongNgu có sự tương quan tương đối cao với cột PhongTam với độ tương quan là 0.5596.
* Tích hợp dữ liệu (nếu có)

## Thuật toán hồi quy (Linear Regression)

## Thuật toán Cây quyết định (Decision Tree)

# Chương IV: Kết quả đạt được

# Chương V: DEMO - Hướng phát triển