**TRẦN TRỌNG CHÍNH (1821050679)**

PHÂN TÍCH GIÁ NHÀ

**BÀI TẬP LỚN**

**MÔN: TÊN MÔN HỌC – MÃ MÔN HỌC**

**Hà Nội, Năm 2022**

**BÀI TẬP LỚN**

**Môn học: Tên môn học**

**Hà Nội, Năm 2022**

**ĐỀ TÀI:**

**PHÂN TÍCH GIÁ NHÀ**

**SINH VIÊN THỰC HIỆN**

Trần Trọng chính (1821050679)

Lớp: Công nghệ Phần mềm K63A

*Ghi chú: Nếu các thành viên cùng lớp thì ghi 1 tên lớp.*

**GIẢNG VIÊN HƯỚNG DẪN**

**Hồ Thị Thảo Trang**

Bộ môn: Công nghệ phầm mềm

Thông tin chung

|  |  |
| --- | --- |
| Tên đề tài | Xây dựng, phân tích giá bất động sản |
| Họ và tên sinh viên: | Trần Trọng Chính |
| Điện thoại liên lạc: | 0329998172 |
| Email: | Chinhtran2k000@gmail.com |
| Lớp: | Công Nghệ Phần MềmA K63 |
| Hệ đào tạo: | Đại học chính quy |
| Đồ án tốt nghiệp được thực hiện tại: | Hà Nội |
| Thời gian làm ĐATN: | 2022 |

**Chương 1: Tổng quan**

**1.lý do chọn đề tài**

Trong thời đại ngày nay, công nghệ thông tin đang được phát triển mạnh mẽ, nhanh chóng và xâm nhập vào nhiều lĩnh vực khoa học, kỹ thuật cũng như trong cuộc sống. Nó trở thành công cụ đắc lực trong nhiều ngành nghề như giao thông, quân sự, y học....

Trước đây, khi máy tính chưa được ứng dụng rộng rãi, các công việc quản lý nhà hàng đều được làm thủ công nên rất mất thời gian và tốn kém về nhân lực cũng như tài chính. Ngày nay, với sự phát triển công nghệ thông tin mà máy tính đã được sử dụng rộng rãi trong các cơ quan, nhà máy, trường học... giúp cho công việc được tốt hơn. Việc sử dụng máy tính vào công tác quản lý nhà hàng là một yêu cầu cần thiết nhằm xóa bỏ những phương pháp quản lý lạc hậu, lỗi thời gây tốn kém về nhiều mặt. Vì vậy, chúng em đã xây dựng đề tài “phân tích giá nhà” với mong muốn giúp cho các hộ gia đình sẽ mua được giá nhà tối ưu nhất.

Chúng em chân thành cảm ơn sự hướng dẫn, chỉ bảo tận tình của cô giáo Hồ Thị Thảo Trang giảng viên trường Đại học Mỏ - Địa Chất. Cô đã cho chúng em nhiều kiến thức quan trọng và nhiều ý kiến đóng góp quý giá. Cảm ơn sự đóng góp ý kiến chân thành và quý báu của các Thầy Cô và những người quan tâm đã giúp đỡ chúng em để hoàn thành đề tài này.

Tuy chúng em đã hết sức cố gắng để có thể hoàn thiện đề tài một cách tốt nhất, nhưng do thời gian có hạn và sự hiểu biết về ngôn ngữ lập trình còn hạn chế, do còn chưa có nhiều kinh nghiệm nên chắc chắn đề tài sẽ không tránh khỏi những thiếu sót, chúng em rất mong nhận được những ý kiến đóng góp của được hoàn thiện hơn từ quý Thầy, Cô và những người quan tâm để đề tài

Chúng em xin chân thành cảm ơn!

**2. Mục tiêu của ĐATN**

Bất động sản là một thuật ngữ pháp luật, còn được gọi là địa ốc hoặc nhà đất (ở các nước Anh, Úc, Mỹ, Canada, Bahama…). Nó bao gồm đất đai và những gì liên quan chặt chẽ không thể tách rời với mảnh đất. Những thứ đó là nhà cửa, kiến trúc bên trên hoặc những gì bên dưới mặt đất như dầu mỏ, khoáng chất. Còn những thứ có thể tách rời khỏi mảnh đất thì không được xem là bất động sản (túp lều, nhà di động, nhà tạm, …). Ngoài ra, bất động sản tiếng Anh là [Real Estate](https://en.wikipedia.org/wiki/Real_estate)hoặc [Real Property](https://vi.wikipedia.org/wiki/B%E1%BA%A5t_%C4%91%E1%BB%99ng_s%E1%BA%A3n)

Bất động sản liên quan mật thiết đến tài chính và vốn, và phân biệt rõ ràng với từ động sản. Nếu sở hữu bất động sản, bạn có mọi quyền lợi đối với chính mảnh đất và những gì liên quan với nó. Bao gồm: sử dụng để kinh doanh hay cho thuê, xây nhà, bán lại, ký quỹ…

Thấy được tầm quan trọng của bất động sản đối với nền kinh tế xã hội nhất là đối với nền kinh tế Việt Nam đang trên đà phát triển, nhu cầu thiết yếu của mỗi hộ gia đình sẽ có nhu cầu mua những ngôi nhà cho gia đình nên em đã nghĩ tới việc xây dựng một hệ thống gợi ý giá nhà và như những căn nhà theo nhu cầu của người mua.

Cụ thể như sau:

Dự đoán giá chung cư khu vực thành phố Hồ Chí Minh với tập dữ liêụ 24.949 entries từ trang [Chợ tốt](https://nha.chotot.com/)

Xây dựng hệ thống recommend system cho hệ thống trên

**1.2. Khảo sát thực trạng**

Với vị thế là một nước đang phát triển thu hút mọi người như việt nam cũng nhưng nhu cầu mua cho mình một căn hộ để an cư lập nghiệp, ở việt nam xuất hiện hàng trăm nghìn các nhà, căn hộ hoặc chung cư để phục vụ nhu cầu đó. Tuy nhiên cùng với việc phân tích giá nhà các nhà, căn hộ được xây dựng lên đã hình thành một bài toán quản lý nhà hàng khó khắn. Việc phân tích giá nhà sẽ giải quyết được vấn đề đau đầu của người có nhu cầu mua hoặc thuê cho mình một căn nhà để sinh sống là giá thành theo nhưng yêu cầu mà chủ sở hữu muốn mua như một số thông tin cơ bản về căn nhà như diện tích, phòng ngủ, nhà vệ sinh, phòng khách ... Trên thực tế đang có rất người mua nhà thông qua mua giới mà không dự đoán được giá nhà mà theo yêu cầu mình muốn mua là ở khoảng tầm giá như thế nào nên rất nhiều hộ gia đình đã mua một căn nhà với giá đắt hơn so với giá thị trường hiện tại

**1.3. Đề xuất phương pháp giải quyết**

Dựa trên khảo sát thực tế việc hộ gia đình muốn mua một căn hộ thông qua mua giới là không tối ưu không bắt kịp với thời đại công nghiệp hoá hiện đại hoá hiện nay. Cho nên em sẽ làm một phân tích giá nhà mua nhà cho mọi người được thuận tiện, giảm sai sót, nhân lực… Và giúp được các chủ bất động sản tiếp cần được tệp khách hàng khổng nồ trên internet từ đó giúp doanh thu tăng cao giúp phát triển cũng như giúp đất nước ta ngày một lớp mạnh

**CHƯƠNG 2: MÔ TẢ BÀI TOÁN**

**2.1. Giới thiệu bài toán**

Bài toán: “Phân tích giá nhà”

Phần mềm phân tích giá nhà phục vụ cho nhu cầu mua nhà của các hộ giá đình có nhu cầu mua cho mình căn nhà để ở, họ sẽ biết được giá phải mua cho căn nhà mà mình muốn mua theo nhu cầu là khoảng trừng là bao nhiêu tiền dữ liệu được lấy từ chotot.vn

**2.2 Quy trình làm bài toán**

**2.2.1. chuẩn bị dư liệu cho bài toán**

Dữ liệu của bài toán được lấy từ website chotot.com và sử dụng thư viện selenium để craw dữ liệu

Graphical user interface, text, application, email

Description automatically generated

Hình 1: Giao diện chotot.com

Selenium: Selenium là công cụ tự động hóa trình duyệt web ban đầu được thiết kế để tự động hóa các ứng dụng web cho mục đích thử nghiệm (testing). Nó hiện được sử dụng cho nhiều ứng dụng khác như tự động hóa các tác vụ quản trị dựa trên web, tương tác với các nền tảng không cung cấp Api, cũng như để thu thập thông tin web (web crawling)

**Mô tả các biến dữ liệu**

**DiaChi**: địa chỉ của chung cư, ở thành phố Hồ Chí Minh

TinhTrangBDS: là chung cư này đã bàn giao chưa, hay vẫn còn đang trong quá trìnhg xây dựng.

**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.

Graphical user interface, application

Description automatically generated

Hình 2: thông tin data trên chotot.com

Text

Description automatically generated

Hình 3: html data chotot.com

Dữ liệu sau khi thu thập được bao gồm 24019 dòng và 15 trường:

A screenshot of a computer

Description automatically generated with medium confidence

→ được lưu [dataset.csv](https://github.com/HungTrinhIT/FinalProject-Datascience/blob/main/data/chotot/rawdata.csv)

Hình 4: data sau khi được craw

**2.2.2: sử lý dữ liệu (data minning)**

Vì khi craw data về nhiều trường bị missing lý do là người bán và người mua chưa nhập đủ thông tin các thông số đó nên em sẽ phải sử lý tất cả các dữ liệu missing để có được tệp dữ liệu chính xác nhất có thể để đưa vào mô hình trainnig.

Để sử lý dữ liệu cho data: em dùng thư viện pandas:

Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ, linh hoạt và mang hàm ý. Tên thư viện được bắt nguồn từ panel data (bảng dữ liệu). Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thời gian.

Mục tiêu của pandas là trở thành khối căn bản (building block) cấp cao cơ bản cho công việc thực tế, phân tích dữ liệu thế giới thực trong Python, và rộng hơn là trở thành công cụ thao tác / phân tích mã nguồn mở mạnh mẽ và linh hoạt nhất có sẵn trong bất kỳ loại ngôn ngữ lập trình nào.

Thông tin dữ liệu missing của dataset

Table

Description automatically generated

Hình 5: tỷ lệ dữ liệu missing

**2.2.2.1: trường giấy tờ**

trường giấy tờ: sẽ fill cùng trường usd để sử đưa ra kết quả tối ưu nhất:

Text

Description automatically generated

Hình 6: sử lý data trường giấy tờ

Dựa vào thông tin hình 6, em sẽ sử lý trường giấy tờ theo mốc giá nhà như sau:

Giá > 121.000 USD thì Giấy tờ = Đã có sổ

Giá <= 121.000 USD và Giá > 112.000 USD thì Giấy tờ = Đang chờ sổ

Giá <= 112.000 USD thì Giấy tờ = Giấy tờ khác

Sau khi sử lý xong trường giấy tờ kết quả thu được:

Graphical user interface, application

Description automatically generated

Hình 7: data trường giấy tờ sau khi được sử lý

**2.2.2.2: trường tình trang nội thất**

trường trường tình trạng nội thất: cũng sẽ fill cùng trường usd để sử đưa ra kết quả tối ưu nhất:

Text

Description automatically generated

Hình 8: sử lý data trường tình trạng nội thất

Dựa vào thông tin hình 8, em sẽ sử lý trường tình trạng nội thất theo mốc giá nhà như sau:

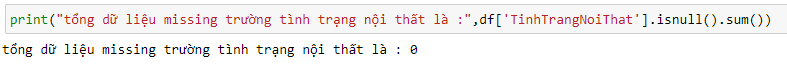
Giá > 170.000 USD thì Tình trạng nội thất = Bàn giao thô

Giá <= 170.000 USD và Giá> 130.000 USD thì Tình trạng nội thất = Nội thất cao cấp

Giá <= 130.000 và Giá >105.000 USD thì Tình trạng nội thất = Nội thất đầy đủ

Giá <= 105.000 USD thì Tình trạng nội thất = Hoàn thiện cơ bản

Sau khi sử lý xong trường Tình trạng nội thất kết quả thu được:



Hình 9: data trường tình trạng nội thất sau khi được sử lý

**2.2.2.3: trường số tầng**

trường số tầng: sẽ fill cùng trường usd để sử đưa ra kết quả tối ưu nhất:

Graphical user interface, table

Description automatically generated

Hình 10: sử lý data trường số tầng

Vì có rất nhiều tầng nên em sẽ gom vào giá trị trung bình của từng tầng x 1x

Dựa vào thông tin hình 10, em sẽ sử lý trường số tầng theo mốc giá nhà như sau:

Giá > 260.000 USD thì Số tầng = 45

Giá <= 260.000 USD và Giá> 170.000 USD thì Số tầng = 35

Giá <= 170.000 và Giá >120.000 USD thì Số tầng = 25

Giá <= 120.000 USD và Giá> 115.000 USD thì Số tầng = 5

Giá <= 115.000 USD thì Số tầng = 15

Sau khi sử lý xong số tầng kết quả thu được:

Graphical user interface, text, application

Description automatically generated

Hình 11: data trường số tầng sau khi được sử lý

**2.2.2.4: Trường hướng ban công**

trường hướng ban công: sẽ fill cùng trường usd để sử đưa ra kết quả tối ưu nhất:

Graphical user interface, application

Description automatically generated

Hình 12: sử lý data trường hướng ban công

Dựa vào thông tin hình 12, em sẽ sử lý trường số tầng theo mốc giá nhà như sau:

Giá > 139.000 USD thì Hướng ban công = Tây bắc

Giá <= 139.000 USD và Giá> 127.000 USD thì Hướng ban công = Tây nam

Giá <= 127.000 và Giá >124.400 USD thì Hướng ban công = Bắc

Giá <= 124.400 USD và Giá> 117.000 USD thì Hướng ban công = Đông bắc

Giá <= 117.000 USD và Giá> 111.000 USD thì Hướng ban công = Đông nam

Giá <= 111.000 và Giá >106.400 USD thì Hướng ban công = Nam

Giá <= 106.400 USD và Giá> 98.000 USD thì Hướng ban công = Tây

Giá <= 98.000 USD thì Hướng ban công = Đông

Sau khi sử lý xong số tầng kết quả thu được:



Hình 13: data trường hướng ban công sau khi được sử lý

**2.2.2.5: trường hướng của chính**

Trường hướng của chính: sẽ fill cùng trường usd để sử đưa ra kết quả tối ưu nhất:

Graphical user interface, application

Description automatically generated

Hình 14: sử lý data trường hướng của chính

Dựa vào thông tin hình 14, em sẽ sử lý trường số tầng theo mốc giá nhà như sau:

Giá > 150.000 USD thì Hướng của chính = Nam

Giá <= 150.000 USD và Giá> 135.000 USD thì Hướng của chính = Tây bắc

Giá <= 135.000 và Giá >125.000 USD thì Hướng của chính = Đông bắc

Giá <= 125.000 USD và Giá> 122.000 USD thì Hướng của chính = Đông nam

Giá <= 122.000 USD và Giá> 120.000 USD thì Hướng của chính = Bắc

Giá <= 120.000 và Giá >112.000 USD thì Hướng của chính = Tây nam

Giá <= 112.000 USD và Giá> 105.000 USD thì Hướng của chính = Tây

Giá <= 105.000 USD thì Hướng ban công = Đông

Sau khi sử lý xong số tầng kết quả thu được:

Graphical user interface, text, application, chat or text message

Description automatically generated

Hình 15: data trường hướng của chính sau khi được sử lý

. Bên trên là tất cả các trường em đã sử lý , và sau khi sử lý xong data đã được đầy đủ không bị missing trường nào nữa .

Text

Description automatically generated

**2.2.3: trức quan hóa dữ liệu và phân tích dữ liệu**

Để trực quan hóa dư liệu em dùng thư viện: matplotlib

Matplotlib là một trong những thư viện Python phổ biến nhất được sử dụng để trực quan hóa dữ liệu. Nó là một thư viện đa nền tảng để tạo các đồ thị 2D từ dữ liệu trong các mảng. Matplotlib được viết bằng Python và sử dụng NumPy, phần mở rộng toán học của Python. Nó cung cấp một API hướng đối tượng giúp nhúng các plot trong các ứng dụng và sử dụng bộ công cụ GUI Python như PyQt, WxPythonotTkinter. Ngoài ra có thể được sử dụng trong Python và IPython shell, Jupyter Notebook và các máy chủ web.

Matplotlib có giao diện được đặt tên là Pylab, được thiết kế giống với MATLAB - ngôn ngữ lập trình độc quyền được phát triển bởi MathWorks. Matplotlib cùng với NumPy có thể được coi là mã nguồn mở tương đương với MATLAB.

Matplotlib ban đầu được viết bởi John D. Hunter vào năm 2003. Phiên bản ổn định hiện tại là 2.2.0 được phát hành vào tháng 1 năm 2018.

Chart, histogram

Description automatically generatedHình 15: biểu đồ tần xuất căn nhà ở các quận tp.HCM

Biểu đồ trên thể hiện tần xuất các căn hộ được các chủ hoặc các nhà giáo bán hoặc mua trên website chợ tốt. nhìn vào biểu đồ ta có thể thấy ở quận 7 có rất nhiều căn hộ đang được giao bán hoặc mua, giúp những hộ gia đình có thể đến đây chọn những ngôi nhà mà mình yêu thích. tần xuất các căn hộ được phân tán xuống dần từ các quận như quận 9, quận 6, quận thủ đức… và quận củ chi và quận 11 có tần xuất các ngôi nhà được giao bán (mua) thấp nhất.

Chart, scatter chart

Description automatically generated

Hình 16: biểu đồ hiển thị dữ liệu quận với diện tích

Biểu đồ hiển thị dữ liệu giữa quận và giá nhà , chúng ta thấy các ngôi nhà tập chung ở bình thạnh , quận 7 , quận 1 và quận 2 với mọi mức giá từ thấp lên cao , các căn hộ ở đây đa dạng mọi người có nhiều lựa chọn phù hợp theo yêu cầu của mình ở các quận này đúng là vậy theo thống kê từ : <https://meeyland.com/thi-truong/nen-mua-nha-quan-nao-o-sai-gon/> Quận 2 của TPHCM sở dĩ được mệnh danh là "khu nhà giàu" bởi rất nhiều tòa chung cư cao cấp bậc nhất thành phố cùng vô vàn tiện nghi được xây dựng ở nơi này.

Tiêu biểu có thể kể đến "khu nhà giàu" Thảo Điền, nơi này tọa lạc tại vị trí đắc địa ngay trên mặt tiền trục đường huyết mạch của Sài Gòn - Xa lộ Hà Nội, được xem là cửa ngõ vào trung tâm thành phố.

**Quận 7: Nét đẹp đậm chất Singapore**

Quận 7 nổi tiếng với những khu đô thị mới văn minh, xanh sạch đẹp, đậm "chất Singapore". Các công trình kiến trúc tại quận 7 đều mang vẻ đẹp nghệ thuật hiện đại rất riêng như: Cầu Ánh Sao, Vivo City, bảo tàng tranh Artinus, cầu Phú Mỹ huyền thoại, ...

**Quận 1: Biểu tượng của Sài Thành**

Hầu như ai sống tại Sài Thành cũng ước mơ được sở hữu một căn nhà ở quận 1 - biểu tượng của Sài Gòn. Quận 1 là nơi tập trung nhiều cơ quan hành chính, những nhà hàng, khách sạn sang trọng cùng nhiều địa điểm du lịch nổi tiếng như: Nhà thờ Đức Bà, Dinh Độc lập...

, còn các quận khác thì các căn hộ thường ở mức trung bình trờ lại tạo điều kiện cho mọi gia đình muốn mua các căn hộ không quá cao nhưng vẫn rất tốt

Chart, histogram

Description automatically generatedHÌNH 17: Biểu đồ trên thể hiện tần xuất diện tích

Biểu đồ trên thể hiện tần xuất diện tích các căn hộ được các chủ hoặc các nhà giáo bán hoặc mua trên website chợ tốt. nhìn vào biểu đồ ta có thể thấy ở diện tích các căn hộ nhiều nhất ở khoảng 60-80 m2 khẳng 15000 dữ liệu, các căn hộ nhiều hơn 300m2 thì ít những vẫn có đồng nghĩa với việc giá các căn hộ này thường rất cao hoặc các căn hộ như này sẽ được cho thuê thay vì bán

Chart, scatter chart

Description automatically generated

Hình 18: biểu đồ hiển thị dữ liệu giá nhà với diện tích

Biểu đồ hiển thị dữ liệu giữa diện tích và giá nhà , chúng ta thấy các ngôi nhà tập chung ở diện tích từ 30-150 m2 thì giá cũng giao động từ 30.000- 400.000 USD 1 vài căn hộ tầm 80 – 200m giá cao là vì lý do nó thuộc loại Penthouse , Duplex hoặc Officetel , 1 số căn hộ diện tích 300-500 m2 mà giá thì rất thấp lý có thể là vì các căn hộ này thì đang được cho thuê chứ không phải là bán (điều đó cũng ảnh hưởng đến sai số khi em tiến hàng trainning model , nhưng vì do tệp dư liệu rất nhiều mà số căn hộ cho thuê như vậy lại không nhiều nên sai số bị ảnh hưởng không đáng kể .

Chart, histogram

Description automatically generated

Hình 19: Biểu đồ trên thể hiện tần xuất số tầng.

Biểu đồ trên thể hiện tần xuất căn hộ ở tầng bao các căn hộ được các chủ hoặc các nhà giáo bán hoặc mua trên website chợ tốt. nhìn vào biểu đồ ta có thể thấy ở các căn hộ thường ở các tầng 1 đến 25 các tâng cao hơn sẽ ít vì thường nó sẽ là các căn hộ Penthouse giá các hộ như vậy thường rất cao

Chart, scatter chart

Description automatically generated

Hình 20: biểu đồ hiển thị dữ liệu giá nhà với số tầng

Biểu đồ hiển thị dữ liệu giữa số tầng và giá nhà , chúng ta thấy các ngôi nhà tập chung ở tầng 0 -10 giá sẽ thấp hơn trên các ngôi nhà ở trên tầng cao , vì theo thống kê các căn hộ càng cao giá càng đắt theo nguồn : <https://cafebiz.vn/can-ho-tang-cang-cao-gia-cang-dat-20210303095738323.chn> Điều này không khó hiểu bởi khi sống trên cao, chủ nhân căn hộ được trải nghiệm những lợi ích đáng "đồng tiền bát gạo" mà căn hộ tầng thấp không có được.

Một nghiên cứu của các nhà khoa học thuộc Đại học Bern (Thụy Sĩ) cho biết, những người sống ở tầng cao thường sẽ có sức khỏe và tuổi thọ tốt hơn. Với vị trí cao vượt hẳn trong khu vực, gia chủ sẽ tránh được tối đa những ảnh hưởng từ tiếng ồn cũng như khói bụi từ mặt đường bên dưới. Cùng với đó, việc sở hữu độ cao lí tưởng, những căn hộ này cũng tránh được ảnh hưởng từ các loại côn trùng hay nấm mốc vốn có hại đến sức khỏe con người. Càng sống trên cao, chủ nhà lại càng có cơ hội đón nhận tối đa nguồn năng lượng tự nhiên mà "mẹ Trái Đất" ban tặng. Ánh mặt trời có thể tràn vào nhiều ngõ ngách trong nhà, tốc độ lưu thông không khí nhanh hơn, hiệu quả thông gió đồng thời cũng trở nên tốt hơn.

Chart

Description automatically generated

Hình 21: Biểu đồ trên thể hiện tần loại.

Biểu đồ trên thể hiện tần xuất loại căn hộ của các căn hộ được các chủ hoặc các nhà giáo bán hoặc mua trên website chợ tốt. nhìn vào biểu đồ ta có thể thấy các căn hộ đa phần là chung cư , không khó hiệu vì hiện nay các chung cư mang lại cho hộ gia định tối giá cả và đời sống xung quanh các tiện ích mà chung cư đem lại khá đầy đủ từ vinsmart , gym store , hay cả các bể bơi nhà thuốc đều rất đầy đủ , ở Việt Nam hiện nay có rất nhiều tòa nhà chung cư mọc lên giá cả rẻ cho các hộ gia đình lựa chọn , theo thống kê từ : <https://dantri.com.vn/bat-dong-san/ca-nuoc-co-khoang-3000-toa-chung-cu-tap-trung-chu-yeu-o-ha-noi-va-tphcm-20190310105551384.htm> Cả nước có khoảng 3.000 toà chung cư, tập trung chủ yếu ở Hà Nội và TPHCM Nguồn cung căn hộ chung cư cũng chiếm tỷ trọng chủ yếu trong tổng nguồn cung nhà ở tại 2 thành phố này.

Chart, line chart

Description automatically generated

Hình 22: biểu đồ hiển thị dữ liệu loại căn hộ với giá

Biểu đồ hiển thị dữ liệu giữa loại căn hộ và giá nhà, chúng ta thấy các ngôi nhà loại Chung cư rất phổ biến và phù hợp với mọi nhu cầu về giá các chung cư giá đắt đỏ thì thường là chung cư của các tòa có tiếng trong đất Sài Thành vd: Charmington Iris Quận 4, Đảo Kim Cương Quận 2 … các căn penthouse hay Duplex giá sẽ tập chung vào các giới thượng lưu hoặc sẽ cho thuê theo tháng hoặc năm

**2.2.4: chuẩn bị dữ liệu:**

Vì tệp dữ liệu để train model phải dạng nump.ndarray nên em sẽ chuyển các dư liệu dạng str hoặc 1 số dữ liệu quá to thu gọn lại ví dụ như giá nhà em sẽ rút gọn thành giá x 1000USD và xóa 1 số trường không cần thiết để training dữ liệu

Kết quả thu được:

Table

Description automatically generated

Hình 22: data sạch

Sau khi dữ liệu đã sách em sẽ tách dữ liệu đó thành 2 tệp: tệp data\_train và tệp data\_test và lưu lại để tránh trường hợp bị mất dữ liệu

Từ 2 tệp data\_train và data test em sẽ tiếp tục tách nó theo tỷ lệ 0.1 trong đó 90% là tập train và 10% tệp test:

-Data\_train: X\_train (21617 dòng và 10 cột), Y\_train (21617 dòng và 1 cột)

-Data\_Test: X\_test (2402 dòng và 10 cột), Y\_test (2402 dòng và 1 cột test)



Hình 23: data trainning

**2.2.5: Lựa chọn mô hình và training dữ liệu:**

**LinearRegression (Hồi quy tuyến tính) :**

**Linear Regression (Hồi quy tuyến tính)** là một trong những thuật toán cơ bản và phổ biến nhất của **Supervised Learning (Học có giám sát)**, trong đó **đầu ra dự đoán là liên tục**. Thuật toán này thích hợp để dự đoán các giá trị đầu ra là các đại lượng liên tục như **doanh số hay giá cả** thay vì cố gắng phân loại chúng thành các đại lượng rời rạc như màu sắc và chất liệu của quần áo, hay xác định đối tượng trong một bức ảnh là mèo hay chó, …

Thử lấy ví dụ sau: bạn đang có điểm thành phần về các môn như Nhập môn lập trình, OOP, Giải tích… và điều bạn đang cần là tính ra điểm trung bình cuối kỳ của mình. Rất đơn giản, bạn sẽ tính được chứ? Tất nhiên rồi! Bạn chỉ cần áp công thức tính điểm trung bình vào là ra. Tiếp tục, bạn lại muốn khảo sát, thống kê lại xem điểm thi giữa kỳ Nhập môn lập trình ảnh hưởng như thế nào đến điểm cuối kỳ của các bạn trong lớp, bạn muốn xác định xem quan hệ giữa điểm thành phần và điểm cuối kỳ thì phải làm sao? Đây có lẽ là một bài toán khó đối với những bạn chưa từng làm việc với Máy học hoặc Thống kê, tuy nhiên cũng đừng vội lo lắng, hãy cùng nhau khám phá và giải quyết các thắc mắc trong bài viết này nhé!

Trong Linear Regression chúng ta sẽ gặp hai loại bài toán đó là Hồi quy đơn biến và Hồi quy đa biến. Để đơn giản thuật toán, chúng ra sẽ tìm hiểu và phân tích kỹ toán học của bài hồi quy đơn biến. Vậy hồi quy tuyến tính đơn biến là gì? Univariate Linear Regression (hồi quy tuyến tính đơn biến) chính là mối quan hệ giữa hai biến số liên tục trên trục hoành \(x\) và trên trục tung \(y\). Phương trình hồi quy tuyến tính đơn biến có dạng như phương trình đường thẳng (y = ax + b \) với \(x\) là biến độc lập và \(y\) là biến phụ thuộc vào \(x\). Đối với Hồi quy tuyến tính đa biến, bạn có thể hiểu một cách đơn giản là sẽ có nhiều biến độc lập (x\_1, x\_2, \dots, x\_n\) và nhiều hệ số (a\_1, a\_2, \dots, a\_n\) thay vì chỉ một biến \(x\) duy nhất.

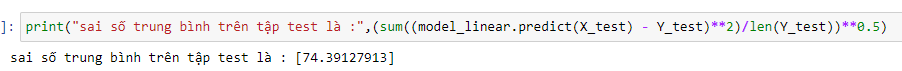
Dựa vào thuật toán này, chúng ta có thể sử dụng để giải các bài toán liên quan đến việc dự đoán mức lương trung bình sau khi ra trường dựa vào các đối số đầu vào là giới tính, điểm trung bình khóa học và số lượng các hoạt động ngoại khóa đã tham gia ,…

Hay trong các bài toán trả về giá trị nhiệt độ phòng với giá trị đầu vào là ngày, nhiệt độ ngoài trời và ánh sáng trong phòng, …

Cụ thể hơn và gần gũi với người dùng nhất, Facebook cũng sử dụng thuật toán này để dự đoán số lượng người share và bình luận dựa trên những tương tác trong bài viết trước đó của bạn hay số lượng bạn bè trên facebook, …

Sau khi training data bằng thuật toán ta thu được :

Sai số trung bình trên tập test sấp sỉ = 74.000 USD



Chart

Description automatically generated

Hình 24: Biểu đồ thể hiện sai số trên tệp test

**KneighborsRegressor**

K-nearest Neightbors (KNN) là một trong những thuật toán cơ bản nhất của supervised-learning trong Machine Learning. Thuật toán k - láng giềng (k-nearest Neightbor (kNN) ) được cho là thuật toán đơn giản nhất trong máy học. Mô hình được xây dựng chỉ bao gồm việc lưu trữ dữ liệu tập huấn (training dataset). Để dự đoán được một điểm dữ liệu mới, thuật toán sẽ tìm ra những láng giềng trong dữ liệu tập huấn (training dataset), đó là láng giềng (nearest neightbors)

Ở mức độ đơn giản, thuật toán chỉ quan tâm đến duy nhất một láng giềng, đó là điểm dữ liệu tập huấn gần nhất tới điểm mà chúng ta muốn đưa ra dự

đoán.

Các dự đoán tiếp theo chỉ đơn giản là đầu ra được biết cho điểm dữ liệu này.

Hình forge\_one\_neighbor minh họa điều này cho trường hợp phân loại các tập dữ liệu.

**Ưu và nhược điểm của KNN**

**Ưu điểm**

-Nó là một thuật toán rất đơn giản để hiểu và giải thích.

-Nó rất hữu ích cho dữ liệu phi tuyến vì không có giả định về dữ liệu trong thuật toán này.

-Nó là một thuật toán linh hoạt vì chúng ta có thể sử dụng nó để phân loại cũng như hồi quy.

-Nó có độ chính xác tương đối cao nhưng có nhiều mô hình học có giám sát tốt hơn KNN.

**Nhược điểm**

-Về mặt tính toán, nó là một thuật toán hơi tốn kém vì nó lưu trữ tất cả các dữ liệu huấn luyện.

-Yêu cầu bộ nhớ lưu trữ cao so với các thuật toán học có giám sát khác.

-Dự đoán chậm trong trường hợp N lớn.

-Nó rất nhạy cảm với quy mô dữ liệu cũng như các tính năng không liên quan.

-Các ứng dụng của KNN

**Sau đây là một số lĩnh vực mà KNN có thể được áp dụng thành công**

**Hệ thống ngân hàng**

KNN có thể được sử dụng trong hệ thống ngân hàng để dự đoán thời tiết một cá nhân có phù hợp để phê duyệt khoản vay không? Cá nhân đó có các đặc điểm giống với cá nhân mặc định không?

Tính toán xếp hạng tín dụng

Các thuật toán KNN có thể được sử dụng để tìm xếp hạng tín dụng của một cá nhân bằng cách so sánh với những người có đặc điểm tương tự.

**Chính trị**

Với sự trợ giúp của các thuật toán KNN, chúng tôi có thể phân loại cử tri tiềm năng thành nhiều lớp khác nhau như “Sẽ bỏ phiếu”, “Sẽ không bỏ phiếu”, “Sẽ bỏ phiếu cho Đảng 'Đại hội',“ Sẽ bỏ phiếu cho Đảng 'BJP'.

Các lĩnh vực khác mà thuật toán KNN có thể được sử dụng là Nhận dạng giọng nói, Phát hiện chữ viết tay, Nhận dạng hình ảnh và Nhận dạng video.

**CHƯƠNG 4 :KẾT QUẢ THỰC HIỆN ĐỀ TÀI**

**4.1. Kết quả đạt được**

Nghiên cứu được cách thức thu thập data , sử lý data , trực quan data và huấn luyện model cho bài toán

Xây dựng được chương trình có thể chạy thử nghiệm trong thực tế

**4,2. Kết quả chưa đạt được**

Mới chỉ sử dụng 2 thuật toán , chưa tối ưu được sai số như mong muốn

**4.3. Hướng phát triển**

Hoàn thiện và phát , tối ưu sai số nhất có thể

Phát triển đa nền tảng.

Xấy dựng hệ thống recommender system

4. Lời cam đoan của sinh viên:

Tôi –Trần Trọng Chính– cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của Giảng viên Hồ Thị Thảo Trang

Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

|  |  |
| --- | --- |
|  | Hà Nội, ngày tháng năm  Tác giả ĐATN  Ký, ghi rõ họ tên |

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:

|  |  |
| --- | --- |
|  | Hà Nội, ngày tháng năm  Cán bộ hướng dẫn  Ký, ghi rõ họ tên |

TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP

Phần tóm tắt nội dung đồ án tốt nghiệp được biên soạn sau khi bản đồ án được hoàn chỉnh, tóm tắt đồ án phản ánh trung thực nội dung chủ yếu của đồ án. Yêu cầu của tóm tắt là ngắn gọn, cô đọng, nêu được cấu trúc của đề tài, nêu bật được những nội dung chính của đồ án, nhấn mạnh được những nội dung cần thiết, những kết luận rút ra sau khi nghiên cứu đề tài cùng với các giải pháp, đề xuất, kiến nghị (nếu có). Phần tóm tắt nội dung đồ án được viết thành 01 đoạn văn, dài không quá 300 từ.