lOMoARcPSD|34747058

TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**

A picture containing text, clipart

Description automatically generated

**BÁO CÁO ĐỒ ÁN MÔN HỌC**

MÔN: **Các công cụ và nền tảng phát triển trí tuệ nhân tạo**

ĐỊNH

***Đề tài: Xây dựng hệ thống dự đoán giá trị bất động sản theo thời gian thực***

**SVTH: Trần Minh Tâm**

**MSSV: 2174802010640**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

**TP. Hồ Chí Minh – Năm 2024**

# LỜI CẢM ƠN

Em xin bày tỏ lòng biết ơn sâu sắc và lòng tôn trọng chân thành đối với sự hỗ trợ và sự dẫn dắt tận tâm của Thầy trong quá trình thực hiện đồ án thuyết trình của mình.

Thầy không chỉ là người hướng dẫn xuất sắc mà còn là nguồn động viên và nguồn đầu tư quý báu. Những lời góp ý, phản hồi chân thành và sự kiên nhẫn của Thầy đã giúp tôi vượt qua những thách thức, phát triển kỹ năng và hiểu sâu hơn về chủ đề của dự án.

Cảm ơn thầy đã cho em cơ hội để làm bài sau 1 số vấn đề cá nhân không thể nộp bài sớm.

Lời cảm ơn này không thể nào diễn đạt hết lòng biết ơn của em. Hy vọng rằng dự án của mình có thể đáp lại một phần nho nhỏ của những đóng góp quý báu này.

Nếu có thiếu xót mong thầy nhận xét để em được rút kinh nghiệm.

Trân trọng,

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc155184038)

[1. Lý do chọn đề tài 1](#_Toc155184039)

[2. Hướng giải quyết 2](#_Toc155184040)

[3. Mục tiêu của đề tài 2](#_Toc155184041)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4](#_Toc155184042)

[1. Bài toán dự đoán giá nhà 4](#_Toc155184043)

[2. Machine Learning 5](#_Toc155184044)

[2.1. Tổng quan về Machine Learning 5](#_Toc155184045)

[2.2. Thuật toán Decision Tree 7](#_Toc155184046)

[2.3. Thuật toán Random Forest 10](#_Toc155184047)

[2.4. Thuật toán Linear Regression 13](#_Toc155184048)

[2.5. Thuật toán Gradient Boosting 17](#_Toc155184049)

[2.6. Thuật toán Isotonic Regression 20](#_Toc155184050)

[3. Xử lý dữ liệu lớn trong thời gian thực 23](#_Toc155184051)

[3.1. Khái quát về Bigdata 23](#_Toc155184052)

[3.2. Khái quát về Spark 25](#_Toc155184053)

[3.3. Khái quát về Hadoop 27](#_Toc155184054)

[CHƯƠNG 3: THỰC NGHIỆM 29](#_Toc155184055)

[1. Tổng quan kiến trúc hệ thống đề xuất 31](#_Toc155184056)

[2. Chuẩn bị dữ liệu 33](#_Toc155184057)

[2.1. Thu thập dữ liệu 33](#_Toc155184058)

[2.2. Phân tích, làm sạch, tiền xử lý dữ liệu 35](#_Toc155184059)

[2.3. Rút trích đặc trưng 38](#_Toc155184060)

[2.4. Mô hình với dữ liệu tổng quát 40](#_Toc155184061)

[2.5. Các phương pháp xây dựng mô hình khác 41](#_Toc155184062)

[2.6. Đánh giá mô hình 42](#_Toc155184063)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 47](#_Toc155184064)

[TÀI LIỆU THAM KHẢO 49](#_Toc155184065)

**MỤC LỤC HÌNH ẢNH**

[Hình 1 Bài toán dự đoán giá nhà. 4](#_Toc155190182)

[Hình 2. Tổng quan về Machine Learning. 6](#_Toc155190183)

[Hình 3. Thuật toán Decision Tree. 8](#_Toc155190184)

[Hình 4. Thuât toán Random Forest. 11](#_Toc155190185)

[Hình 5. Trực quan Linear Regression. 14](#_Toc155190186)

[Hình 6. Thuật toán Gradient Boosting. 17](#_Toc155190187)

[Hình 7. Learning Rate. 19](#_Toc155190188)

[Hình 8. Weak Learner. 20](#_Toc155190189)

[Hình 9. Trực quan thuật toán Isotonic Regression. 21](#_Toc155190190)

[Hình 10. So sánh Linear Regression, Isotonic Regression. 22](#_Toc155190191)

[Hình 11. 5V trong Big Data. 24](#_Toc155190192)

[Hình 12. Các ứng dụng của Spark. 26](#_Toc155190193)

[Hình 13. Kiến trúc của Hadoop. 28](#_Toc155190194)

[Hình 14. Kiến trúc của HDFS. 29](#_Toc155190195)

[Hình 15. Kiến trúc hệ thống. 31](#_Toc155190196)

[Hình 16. Quy trình thực nghiệm. 33](#_Toc155190197)

[Hình 17. Phân bố của dữ liệu theo loại bất động sản và tỉnh thành. 34](#_Toc155190198)

[Hình 18. Phân bố của dữ liệu theo loại bất động sản và tỉnh thành (2). 35](#_Toc155190199)

[Hình 19. Dữ liệu sau khi crawl. 35](#_Toc155190200)

[Hình 20. Code xử lý dữ liệu. 37](#_Toc155190201)

[Hình 21. Gom nhóm dữ liệu với Bucketizer. 39](#_Toc155190202)

[Hình 22. One Hot Encoding. 40](#_Toc155190203)

[Hình 23. Dự đoán giá nhà trên data mẫu. 44](#_Toc155190204)

[Hình 24. Dự đoán giá nhà trên data crawl theo thời gian thực. 44](#_Toc155190205)

# 

# CHƯƠNG 1: GIỚI THIỆU

## Lý do chọn đề tài

Lý do chọn đề tài của tôi đến từ sự nhận thức rõ ràng về khó khăn mà những người muốn mua bất động sản thường gặp phải khi đối mặt với quyết định quan trọng: định giá căn hộ. Trong bối cảnh thị trường bất động sản đang phát triển mạnh mẽ, việc đưa ra một quyết định chính xác về giá trị của căn hộ trở nên ngày càng phức tạp và quan trọng.

Những người tiêu dùng, cũng như người thân của họ, thường xuyên phải đối mặt với sự đắn đo và không chắc chắn khi đưa ra quyết định mua nhà. Điều này là do họ thường không có thông tin đầy đủ về giá trị thực sự của căn hộ mà họ quan tâm. Chính vì vậy, tôi quyết định tập trung nghiên cứu về quá trình định giá bất động sản, đặc biệt là trong trường hợp căn hộ.

Để định giá một căn hộ một cách chính xác, không chỉ đòi hỏi sự am hiểu vững về thị trường bất động sản mà còn đòi hỏi sự hiểu biết sâu sắc về từng chi tiết của căn hộ đó. Tính đến từng thuộc tính như vị trí, diện tích, tiện ích xung quanh, và tình trạng cơ sở hạ tầng, tất cả đều ảnh hưởng đến giá trị cuối cùng của căn hộ. Quá trình này không chỉ là khoa học mà còn là nghệ thuật, yêu cầu sự kiên trì và tư duy phân tích sâu sắc.

Thông qua đề tài này, tôi mong muốn mang đến cái nhìn rõ ràng hơn về quá trình định giá bất động sản, giúp người tiêu dùng có thêm kiến thức và tự tin hơn khi đưa ra quyết định mua bất động sản. Đồng thời, nghiên cứu của tôi cũng có thể hỗ trợ các chuyên gia định giá bất động sản trong việc nâng cao chất lượng dịch vụ và đảm bảo tính minh bạch trong quá trình giao dịch bất động sản.

## Hướng giải quyết

Hướng giải quyết của dự án sẽ tập trung vào việc xây dựng một hệ thống dự đoán giá trị bất động sản theo thời gian thực, nhằm mang lại cho người tiêu dùng và những nhà đầu tư một công cụ đáng tin cậy và hiệu quả trong quá trình đưa ra quyết định mua bán. Để đạt được mục tiêu này, tôi sẽ sử dụng một bộ dữ liệu tự thu thập, chứa đựng đa dạng thông tin về giá bất động sản, loại bất động sản, địa chỉ, diện tích đất, cũng như thông tin chi tiết về người bán và người mua.

Quá trình xây dựng hệ thống dự đoán sẽ tích hợp các phương pháp và mô hình máy học tiên tiến, giúp phân tích và dự đoán giá trị bất động sản một cách chính xác nhất. Việc này sẽ đòi hỏi sự hiểu biết sâu sắc về đặc điểm địa lý, thị trường bất động sản cụ thể, và xu hướng thị trường hiện tại. Mô hình sẽ được huấn luyện trên dữ liệu lớn, và sẽ được cập nhật liên tục để đảm bảo tính động và phản ánh chính xác nhất về giá trị bất động sản theo thời gian thực.

Ngoài ra, tôi cũng sẽ xem xét các yếu tố ảnh hưởng đến giá trị bất động sản, từ vị trí đến các yếu tố xã hội và kinh tế. Quá trình này sẽ giúp tối ưu hóa độ chính xác của hệ thống và cung cấp thông tin chi tiết hơn về những yếu tố quyết định giá trị bất động sản.

Tôi tin rằng hệ thống dự đoán sẽ không chỉ hỗ trợ người mua và người bán trong quá trình đàm phán mua bán mà còn đóng góp vào sự minh bạch và ổn định của thị trường bất động sản. Đồng thời, nó cũng mở ra những cơ hội mới trong việc nghiên cứu và phát triển các phương pháp định giá tiên tiến trong ngành.

## Mục tiêu của đề tài

Mục tiêu quan trọng của đề tài này là triển khai và đánh giá hiệu suất của nhiều mô hình dự đoán, bao gồm Random Forest, Linear Regression, Isotonic Regression, Gradient Boosted, Factorization Machines Regression và Decision Tree. Được thực hiện trên một quy mô lớn, mục tiêu của tôi là xây dựng những mô hình có khả năng dự đoán giá nhà ở một cách chính xác và linh hoạt trên toàn quốc.

Qua quá trình phân tích và phát triển, tôi đặt ra mục tiêu tối ưu hóa hiệu suất của mỗi mô hình, đồng thời so sánh chúng với nhau bằng cách sử dụng các độ đo quan trọng như Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), và R-squared (R2). Mỗi mô hình sẽ được đánh giá chi tiết để hiểu rõ về độ chính xác và khả năng tổng quát của chúng trong việc dự đoán giá trị bất động sản.

Bằng cách này, tôi hướng đến việc cung cấp một cơ sở khoa học và thực tế cho việc lựa chọn mô hình dự đoán giá bất động sản trong môi trường phức tạp của thị trường bất động sản quốc gia. Mục tiêu cuối cùng là mang lại giá trị thiết thực cho người mua, người bán, và các chuyên gia định giá bất động sản, đồng thời đóng góp vào sự phát triển và minh bạch của thị trường bất động sản Việt Nam.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Bài toán dự đoán giá nhà

Bài toán dự đoán giá nhà là một trong những thách thức phức tạp nhất và đầy thú vị trong lĩnh vực khoa học dữ liệu và máy học. Việc này không chỉ đòi hỏi sự hiểu biết vững về thị trường bất động sản mà còn đưa ra những thách thức liên quan đến xử lý dữ liệu đa dạng, đánh giá độ chính xác của mô hình, và hiểu biết sâu rộng về các yếu tố ảnh hưởng đến giá nhà.

A person standing in front of a house

Description automatically generated

Hình 1 Bài toán dự đoán giá nhà.

Mục tiêu chính của bài toán này là xây dựng các mô hình dự đoán giá nhà có khả năng đáp ứng mức độ đa dạng và biến động của thị trường bất động sản. Trong khi nhiều mô hình như Linear Regression hay Decision Tree có thể được áp dụng để dự đoán giá trị, nhưng thực tế lại phức tạp hơn nhiều do nhiều yếu tố đa dạng như địa lý, diện tích, tiện ích xung quanh, và tình trạng cơ sở hạ tầng.

Điều này đặt ra yêu cầu về việc sử dụng mô hình máy học mạnh mẽ như Random Forest, Gradient Boosted, hay Factorization Machines Regression để có thể hiệu quả đối mặt với sự phức tạp và không đồng đều của dữ liệu. Các mô hình này không chỉ có khả năng học từ dữ liệu một cách linh hoạt mà còn có khả năng điều chỉnh để phản ánh những biến động thị trường thực tế.

Trong cơ sở lý thuyết, chúng ta cần xem xét sự kết hợp giữa các mô hình này, hiểu rõ về cơ chế hoạt động của chúng, và làm thế nào chúng có thể được điều chỉnh để đạt được độ chính xác cao nhất trong dự đoán giá nhà. Đồng thời, cũng quan trọng là nắm bắt các khía cạnh lý thuyết về bất động sản, thị trường và dữ liệu để tối ưu hóa khả năng ứng dụng thực tế của các mô hình này trong môi trường đa biến và thay đổi liên tục.

Bài toán dự đoán giá nhà không chỉ là một thách thức kỹ thuật, mà còn là một thách thức về sự hiểu biết sâu sắc về một lĩnh vực quan trọng trong đời sống hàng ngày của mọi người. Để đảm bảo tính ứng dụng cao và giá trị thực sự của các mô hình, việc kết hợp cả kiến thức lý thuyết và kỹ thuật là không thể thiếu.

## Machine Learning

### Tổng quan về Machine Learning

Máy học, là một lĩnh vực nổi bật trong thời đại công nghệ ngày nay, không chỉ mở ra những khám phá mới về khả năng tính toán của máy tính mà còn thách thức sự hiểu biết và sáng tạo của con người. Cơ sở lý thuyết của machine learning (ML) không chỉ là một tập hợp các nguyên lý kỹ thuật, mà là một hành trình sâu sắc vào thế giới của dữ liệu và sự hiểu biết.

A diagram of machine learning

Description automatically generated

Hình 2. Tổng quan về Machine Learning.

Nguyên tắc cơ bản của ML nằm trong sự khao khát của con người muốn máy tính có khả năng học từ dữ liệu mà không cần phải được lập trình cụ thể cho mọi tình huống. Điều này đồng nghĩa với việc tạo ra các mô hình có khả năng tự điều chỉnh dựa trên dữ liệu đầu vào, không khác gì việc con người học từ kinh nghiệm.

Một trong những thách thức lớn nhất của ML là xây dựng các thuật toán và mô hình hiệu quả. Linear Regression, Decision Trees, Neural Networks, và một loạt các thuật toán khác đều là những "công cụ" mà chúng ta sử dụng để định hình khả năng dự đoán của máy tính. Cơ sở lý thuyết là nơi chúng ta đàm phán với những nguyên lý căn bản này, cố gắng hiểu rõ hơn về cách chúng hoạt động và tại sao chúng lại phản ánh thế giới thực một cách nhất quán.

Học giám sát và học không giám sát, hai "hình thức" chính của ML, đưa ra những bài toán và cơ hội khác nhau. Trong học giám sát, mô hình được dạy bằng cách sử dụng dữ liệu có nhãn, trong khi học không giám sát yêu cầu mô hình học từ dữ liệu không có nhãn. Đây không chỉ là những khái niệm lý thuyết mà còn là bước chân quan trọng đưa chúng ta từ thế giới ý tưởng sang thực tế ứng dụng.

Không thể phủ nhận vai trò quan trọng của quá trình học và kiểm soát mô hình. Quá trình học không chỉ là việc điều chỉnh các tham số để mô hình phản ánh đúng hơn dữ liệu mà còn là quá trình học được từ những sai lầm. Kiểm soát mô hình, ngăn chặn sự "học thuộc lòng" và đảm bảo khả năng tổng quát hóa, là một phần quan trọng để đảm bảo rằng mô hình có thể áp dụng cho dữ liệu mới một cách hiệu quả.

Đánh giá và chấm điểm mô hình là một chủ đề không thể bỏ qua. Cross-Validation, Precision-Recall, và ROC Curve không chỉ là những số liệu mà còn là cách chúng ta "nói chuyện" với mô hình, đánh giá khả năng hiệu quả và đồng thời tìm kiếm những cải tiến.

Ngoài ra, còn những khía cạnh như overfitting và underfitting, những hiện tượng phổ biến khi chúng ta làm việc với dữ liệu. Lý thuyết của ML không chỉ giải thích tại sao chúng xảy ra mà còn hướng dẫn cách xử lý chúng để đảm bảo tính ổn định và đáng tin cậy của mô hình.

### Thuật toán Decision Tree

Thuật toán Decision Tree (cây quyết định) là một trong những khối xây dựng nền tảng của machine learning, mang theo mình sự linh hoạt và khả năng hiểu biết đáng kinh ngạc. Như một "cây" phân loại, nó không chỉ là một công cụ mà còn là một nghệ thuật, từng bước từng bước, đưa ta khám phá vào thế giới rộng lớn của quyết định và dự đoán.

* Cơ bản về Decision Tree:

Tưởng tượng một cây quyết định như một người thám hiểm trong khu rừng thông tin, luôn đặt ra những câu hỏi "có hay không?" để dẫn dắt chúng ta đến mục tiêu cuối cùng. Decision Tree cũng làm như vậy, dựa trên các thuộc tính của dữ liệu để đưa ra những quyết định cuối cùng.

A person holding an apple in her hand

Description automatically generated

Hình 3. Thuật toán Decision Tree.

* Quá trình Xây dựng Cây:

Thuật toán bắt đầu bằng việc chọn thuộc tính tốt nhất để phân loại dữ liệu. Tại mỗi nút, nó đặt ra câu hỏi để chia tập dữ liệu thành các phần con. Cứ như vậy, cây phát triển từng nhánh, tự học từ dữ liệu và tối ưu hóa quá trình phân loại.

* Thông tin và Độ tinh khiết:

Ở mỗi bước, Decision Tree sử dụng các phương pháp như Information Gain hay Gini Index để đánh giá độ tinh khiết của dữ liệu. Mục tiêu là tìm ra thuộc tính làm cho việc phân loại trở nên đơn giản và chính xác hơn.

* Tính Linh Hoạt và Dễ Giải thích:

Một trong những đặc điểm độc đáo của Decision Tree là tính linh hoạt và dễ giải thích. Cây quyết định có khả năng làm việc với dữ liệu phức tạp và giúp ta hiểu rõ quá trình quyết định thông qua các câu hỏi có thể hiểu được.

* Quản lý Overfitting:

Mặc dù Decision Tree có khả năng dễ bị overfitting (quá mức "học thuộc lòng") trên dữ liệu huấn luyện, nhưng nó có thể được kiểm soát thông qua việc sử dụng các kỹ thuật như cắt tỉa cây (pruning) và giảm độ sâu của cây.

* Ứng dụng Rộng Rãi:

Decision Tree không chỉ dừng lại ở việc phân loại mà còn được ứng dụng trong dự đoán giá, phân tích rủi ro, và nhiều lĩnh vực khác. Điều này chứng tỏ sức mạnh của nó trong việc giải quyết các vấn đề thực tế.

* Đưa ra Quyết Định và Hiểu Biết:

Mỗi chiếc lá trên cây quyết định đều là một quyết định cuối cùng, và cùng nhau chúng tạo nên một câu chuyện về quyết định. Decision Tree không chỉ là một công cụ, mà là ngôn ngữ của sự hiểu biết, giúp chúng ta thấu hiểu thế giới dữ liệu xung quanh mình.

Trong thế giới đầy biến động của machine learning, cây quyết định vẫn tự tin đứng vững, như một hướng dẫn đắc lực, đưa chúng ta đến những quyết định thông minh và đầy chín chắn. Đó là nét đẹp của Decision Tree, nơi sự đơn giản và sức mạnh hội tụ, tạo nên một "rừng thông" thông minh trong thế giới số ngày nay.

### Thuật toán Random Forest

Trong thế giới phức tạp của machine learning, Random Forest - hay "Rừng Ngẫu Nhiên" - nổi bật như một đỉnh núi đầy hùng vĩ, mang theo sức mạnh tổ hợp và khả năng dự đoán đỉnh cao. Là một thuật toán thống kê và học máy nổi tiếng, Random Forest không chỉ là công cụ mạnh mẽ, mà còn là một tác phẩm nghệ thuật đầy sự sáng tạo và hiệu suất.

A screenshot of a computer screen

Description automatically generated

Hình 4. Thuât toán Random Forest.

* Tổng quan về Random Forest:

Random Forest là một phương pháp kết hợp nhiều cây quyết định (Decision Trees) để tạo ra một mô hình mạnh mẽ hơn và có khả năng tổng quát hóa cao hơn. Nó kết hợp sức mạnh của từng cây để tạo ra một "rừng" chưa sự đa dạng và độc lập.

* Quy trình Hoạt động:

Thuật toán Random Forest bắt đầu bằng việc chọn ngẫu nhiên một tập con của dữ liệu từ tập huấn luyện. Tiếp theo, nó xây dựng một cây quyết định trên tập con này. Quá trình này được lặp lại nhiều lần để tạo ra nhiều cây quyết định khác nhau.

* Đa Dạng và Khả năng Tổng Quát Hóa:

Sức mạnh của Random Forest nằm ở sự đa dạng của các cây quyết định. Mỗi cây được xây dựng trên một tập con khác nhau của dữ liệu và với một số thuộc tính được chọn ngẫu nhiên. Sự đa dạng này giúp mô hình giảm nguy cơ overfitting và cung cấp khả năng tổng quát hóa cao.

* Vận dụng Phương pháp Bootstrap:

Random Forest sử dụng phương pháp bootstrap, tức là lấy ngẫu nhiên các mẫu từ tập dữ liệu huấn luyện với sự thay thế. Điều này tạo ra sự ngẫu nhiên và độc lập giữa các cây quyết định, đồng thời giảm nguy cơ mô hình quá mức "học thuộc lòng".

* Thách Thức Của Overfitting:

Mặc dù Random Forest có khả năng giảm overfitting, nhưng nếu không được kiểm soát, vẫn có thể dẫn đến việc mô hình quá mức học từ dữ liệu huấn luyện. Các kỹ thuật như giảm độ sâu của cây và kiểm soát độ đa dạng có thể giúp kiểm soát tình trạng này

* Ứng Dụng Rộng Rãi:

Random Forest là một trong những thuật toán phổ biến nhất và được ứng dụng rộng rãi trong nhiều lĩnh vực. Từ dự đoán giá bất động sản đến phân loại ảnh y tế, Random Forest là nguồn động viên đằng sau những ứng dụng thông minh và độ chính xác cao.

* Sự Hiểu Biết và Giải Thích:

Mặc dù Random Forest được biết đến với khả năng dự đoán mạnh mẽ, nhưng việc hiểu rõ nó có thể là một thách thức. Tuy nhiên, sự đơn giản của mỗi cây quyết định giúp mô hình trở nên dễ giải thích và tạo ra sự minh bạch trong quá trình quyết định.

Random Forest không chỉ là một thuật toán, mà là một chiến lược, một cách tiếp cận đặc biệt để nắm bắt sự đa dạng và sức mạnh từ dữ liệu. Như một "rừng" ẩn chứa sự kỳ diệu của các quyết định, Random Forest mở ra một cánh cửa rộng lớn đưa chúng ta đến thế giới huyền bí của machine learning.

### Thuật toán Linear Regression

Linear Regression là một mô hình toán học đơn giản nhưng mạnh mẽ. Để hiểu cơ bản, chúng ta có thể bắt đầu với mô hình Linear Regression đơn biến (univariate), trong đó chỉ có một biến độc lập và một biến phụ thuộc. Công thức toán học của mô hình Linear Regression đơn biến được mô tả như sau:

Giả sử có một tập dữ liệu gồm các cặp (x, y), trong đó x là giá trị của biến độc lập, y là giá trị của biến phụ thuộc. Mục tiêu là tìm ra đường thẳng tuyến tính sao cho dự đoán y từ x là chính xác nhất. Công thức của đường thẳng này có dạng:

A graph of a number of colored lines

Description automatically generated with medium confidence

Hình 5. Trực quan Linear Regression.

Thuật toán Linear Regression (Hồi quy tuyến tính) là một trong những công cụ cơ bản nhưng mạnh mẽ nhất trong thế giới của machine learning. Nó không chỉ là một phương pháp dự đoán, mà còn là một câu chuyện về mối liên kết giữa biến độc lập và biến phụ thuộc, với đường thẳng là nhân vật chính. Hãy cùng nhau khám phá câu chuyện này:

* Bức Họa Tổng Quan về Linear Regression:

Ở trung tâm của Linear Regression là ý tưởng đơn giản nhưng mạnh mẽ: tìm ra đường thẳng (hoặc siêu mặt phẳng trong không gian nhiều chiều) mà có thể dự đoán giá trị của biến phụ thuộc dựa trên giá trị của biến độc lập. Đường thẳng này được chọn sao cho tổng của bình phương các sai số giữa giá trị thực tế và giá trị dự đoán là nhỏ nhất.

* Hình Dạng Của Đường Thẳng:

Linear Regression là câu chuyện về việc chọn ra một đường thẳng có hình dạng sao cho khoảng cách từ các điểm dữ liệu đến đường thẳng là nhỏ nhất. Đây là một cuộc phiêu lưu trong không gian biểu đồ, nơi mà mỗi điểm dữ liệu là một chấm trên bức tranh, và đường thẳng là chiếc thước đo hoàn hảo.

* Dự Đoán và Học Hỏi:

Đường thẳng Linear Regression không chỉ là một công cụ dự đoán, mà còn là người học từ dữ liệu. Nó nắm bắt mối liên kết giữa biến độc lập và biến phụ thuộc và từ đó, khiến cho mô hình trở nên thông minh hơn với mỗi lần dự đoán.

* Sức Mạnh của Hệ Số:

Hệ số của đường thẳng, được tính toán thông qua quá trình học, là nguồn gốc của sức mạnh. Nó không chỉ là một con số, mà là ngôn ngữ của mối quan hệ giữa các biến. Mỗi hệ số là một chương mới trong câu chuyện, mô tả độ ảnh hưởng của mỗi biến độc lập lên biến phụ thuộc.

* Khám Phá Dữ Liệu:

Trong cuộc phiêu lưu của Linear Regression, chúng ta trải qua việc khám phá dữ liệu, tìm hiểu về phân phối, xu hướng, và mối quan hệ giữa các biến. Đây không chỉ là một công việc tính toán, mà còn là việc nắm bắt bản chất của dữ liệu.

* Đối Mặt với Nhiễu và Thách Thức:

Linear Regression là một nhà thám hiểm dũng cảm, đối mặt với thách thức từ nhiễu trong dữ liệu. Đôi khi, những điểm ngoại lệ có thể làm rơi vào tầm nhìn của nó, nhưng với sức mạnh của hệ số, nó cố gắng hiểu và giải quyết mọi thách thức.

* Vấn Đề Overfitting và Underfitting:

Linear Regression là câu chuyện của việc đối mặt với vấn đề overfitting và underfitting. Khi câu chuyện trở nên quá phức tạp hoặc quá đơn giản, Linear Regression đứng lên và điều chỉnh để tìm ra câu chuyện phù hợp nhất.

* Ví Dụ Thực Tế:

Giả sử chúng ta có dữ liệu về quan hệ giữa số giờ học và điểm số của sinh viên. Bằng cách sử dụng Linear Regression, chúng ta có thể tạo ra một đường thẳng mô tả mối liên kết giữa số giờ học và điểm số. Điều này không chỉ là một phương trình toán học, mà là một cách để kể một câu chuyện: "Với mỗi giờ học thêm, sinh viên có thể mong đợi một tăng trưởng như thế nào trong điểm số?"

Linear Regression không chỉ là một công cụ, mà là một nhà sưu tập câu chuyện, biểu tượng của sự hiểu biết về mối liên kết và dự đoán. Trong thế giới đầy mênh mang của dữ liệu, Linear Regression là người hướng dẫn cho chúng ta, giúp chúng ta tìm ra những câu chuyện tiềm ẩn.

Giả sử có một tập dữ liệu gồm các cặp (x, y), trong đó x là giá trị của biến độc lập, y là giá trị của biến phụ thuộc. Mục tiêu là tìm ra đường thẳng tuyến tính sao cho dự đoán y từ x là chính xác nhất. Công thức của đường thẳng này có dạng:

Trong quá trình huấn luyện mô hình Linear Regression, mục tiêu là điều chỉnh giá trị của

và sao cho tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất. Hàm mất mát (loss function) thường được sử dụng là Mean Squared Error (MSE), và quá trình này thường được thực hiện thông qua phương pháp Gradient Descent hoặc các phương pháp tối ưu hóa khác. Điều này giúp tìm ra giá trị tối ưu của và để mô hình dự đoán tốt nhất.

### Thuật toán Gradient Boosting

Thuật toán Gradient Boosting là một phương pháp học máy tiên tiến được thiết kế để xây dựng mô hình dự đoán mạnh mẽ bằng cách tận dụng sức mạnh của nhiều mô hình yếu. Trong lớp các thuật toán học máy, Gradient Boosting đã nổi lên như một người hướng dẫn xuất sắc, dẫn dắt chúng ta qua các bước tỉ mỉ để hiểu và khai thác sự mạnh mẽ của dữ liệu.

A diagram of a gradient boosting algorithm

Description automatically generated

Hình 6. Thuật toán Gradient Boosting.

* Làm Rõ Tổng Quan:

Gradient Boosting không phải là một phép màu, mà là một hành trình học tập dựa trên các mô hình tuyến tính. Bắt đầu với một mô hình yếu và tập trung vào việc cải thiện dự đoán của nó thông qua các vòng lặp.

* Hàm Mất Mát và Đạo Hàm:

Tại mỗi bước, Gradient Boosting chọn hàm mất mát, thường là Mean Squared Error, để đo lường sự chênh lệch giữa dự đoán và giá trị thực tế. Sử dụng đạo hàm của hàm mất mát, thuật toán xác định hướng và kích thước cần điều chỉnh mô hình.

* Mô Hình Học Mới:

Một mô hình mới được thêm vào "đám đông" để cố gắng sửa chữa những sai sót của mô hình trước đó. Quá trình này giống như việc mỗi mô hình mới là một chuyên gia đặc biệt được chuyển giao để giải quyết một phần nhỏ vấn đề.

* Tinh Tuning thông Qua Learning Rate:

Learning rate là một yếu tố quan trọng trong Gradient Boosting, xác định mức độ "tinh chỉnh" mô hình mới. Nếu learning rate quá cao, có thể dẫn đến việc bỏ lỡ điểm cực tiểu toàn cục; ngược lại, nếu quá thấp, quá trình học sẽ chậm.

A diagram of a graph

Description automatically generated

Hình 7. Learning Rate.

* Regularization để Tránh Overfitting:

Để tránh overfitting, Gradient Boosting thường sử dụng các kỹ thuật regularization như max\_depth cho cây quyết định hoặc việc giảm learning rate.

* Sự Linh Hoạt của "Weak Learners":

Mô hình yếu trong Gradient Boosting thường là cây quyết định với độ sâu nhỏ. Điều này giúp mô hình linh hoạt, có thể "đánh bại" sai số của mô hình trước đó trong những điểm dữ liệu cụ thể.

Long white bar with green and orange arrows

Description automatically generated

Hình 8. Weak Learner.

* Kết Hợp Với Nhiều "Weak Learners":

Đám đông của các mô hình yếu, được tinh chỉnh thông qua quá trình học, là sức mạnh thực sự của Gradient Boosting. Việc kết hợp những chuyên gia đặc biệt vào một đội làm cho toàn bộ mô hình trở nên mạnh mẽ và độc đáo.

Gradient Boosting không chỉ là thuật toán, mà là một chiến lược học máy có sức mạnh đặc biệt để tạo ra mô hình dự đoán với hiệu suất cao và khả năng tổng quát hóa tốt. Như một người hướng dẫn, Gradient Boosting liên tục đưa chúng ta đến gần với sự hoàn thiện trong việc dự đoán và hiểu biết về dữ liệu.

### Thuật toán Isotonic Regression

Thuật toán Isotonic Regression là một công cụ hữu ích trong thế giới của machine learning, chủ yếu được sử dụng để mô hình hóa các quan hệ đơn điệu và không giả định hình dạng của đường chính.

A graph of a number of dots

Description automatically generated with medium confidence

Hình 9. Trực quan thuật toán Isotonic Regression.

* Đặc Điểm Cơ Bản:

Isotonic Regression được thiết kế để đối mặt với những tình huống khi chúng ta mong đợi một xu hướng đơn điệu trong dữ liệu mà không cần phải giả định một hình dạng cụ thể của mô hình.

* Mục Tiêu:

Mục tiêu của Isotonic Regression là tìm ra một hàm số monotonically increasing hoặc monotonically decreasing sao cho sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

* Ứng Dụng trong Xử Lý Dữ Liệu:

Thường được sử dụng khi chúng ta muốn làm mịn đường biểu đồ hoặc xử lý dữ liệu thô mà có xu hướng tăng hoặc giảm.

* Isotonic Regression vs. Linear Regression:

So với Linear Regression, Isotonic Regression có ưu điểm là không yêu cầu giả định về hình dạng của mô hình. Nó có thể làm mịn đường biểu đồ mà không cần phải giữ cho mô hình "chìm" vào các điểm dữ liệu cụ thể.

A graph showing the growth of the number of negatives

Description automatically generated with medium confidence

Hình 10. So sánh Linear Regression, Isotonic Regression.

* Phương Pháp Tối Ưu:

Isotonic Regression thường được giải quyết thông qua các phương pháp tối ưu hóa, như Pool-Adjacent-Violators (PAV) algorithm. Thuật toán này điều chỉnh giá trị của mỗi điểm dữ liệu để đảm bảo tính đơn điệu của hàm số.

* Ứng Dụng Trong Thị Trường Tài Chính:

Trong lĩnh vực tài chính, Isotonic Regression có thể được sử dụng để mô hình hóa xu hướng giá cổ phiếu mà không yêu cầu các giả định cụ thể về mô hình hóa.

* Điều Chỉnh Đường Chính:

Khi có nhiều nhiễu trong dữ liệu và mô hình đường chính cần được điều chỉnh sao cho nó tuân theo một hình dạng chung, Isotonic Regression là một lựa chọn hiệu quả.

* Hạn Chế:

Tuy Isotonic Regression có ưu điểm về đơn giản và không đòi hỏi nhiều giả định, nhưng đối với các tình huống khi sự biến động của dữ liệu là quá lớn, nó có thể dẫn đến các hiện tượng "oversmoothing" và làm mất đi sự chi tiết của dữ liệu.

Isotonic Regression không chỉ giúp chúng ta mô hình hóa xu hướng trong dữ liệu mà còn giữ cho mô hình linh hoạt và ít phụ thuộc vào giả định hình dạng. Điều này làm cho nó trở thành một công cụ hữu ích trong nhiều lĩnh vực, đặc biệt là khi chúng ta muốn đảm bảo tính đơn điệu của mô hình mà không mất đi sự độc lập của các điểm dữ liệu.

## Xử lý dữ liệu lớn trong thời gian thực

### Khái quát về Bigdata

Big Data không chỉ là một khối lượng lớn các dữ liệu, mà là một thế giới đa dạng và phức tạp, đang tạo ra những thách thức và cơ hội mới cho xã hội và kinh tế toàn cầu.

Big Data được đặc trưng bởi "5V":

* Volume (Khối Lượng): Đây là yếu tố đặc trưng cho khối lượng lớn của dữ liệu. Big Data thường xuất hiện khi khối lượng dữ liệu vượt quá khả năng xử lý của các hệ thống truyền thống.
* Velocity (Tốc Độ): Tốc độ đề cập đến tốc độ nhanh chóng với mà dữ liệu được tạo ra, truyền tải, và xử lý. Dữ liệu thời gian thực và dữ liệu streaming là các ví dụ điển hình.
* Variety (Đa Dạng): Đa dạng ám chỉ sự đa dạng về loại dữ liệu. Thay vì chỉ là dữ liệu có cấu trúc, Big Data thường bao gồm cả dữ liệu phi cấp và dữ liệu không có cấu trúc.
* Veracity (Chính Xác): Chính xác liên quan đến độ tin cậy của dữ liệu. Big Data thường đối mặt với thách thức về tính chính xác và độ tin cậy của dữ liệu do nguồn gốc đa dạng.
* Value (Giá Trị): Giá trị là khả năng chúng ta có thể rút ra từ dữ liệu. Mặc dù có nhiều dữ liệu, nhưng giá trị thực sự đến từ khả năng phân tích và chuyển đổi dữ liệu thành thông tin hữu ích.



Hình 11. 5V trong Big Data.

Big Data mang lại tiềm năng lớn cho các tổ chức. Bằng cách phân tích dữ liệu lớn, các tổ chức có thể đưa ra quyết định thông minh, dự đoán xu hướng thị trường, và tối ưu hóa quy trình kinh doanh

Nếu được xử lý đúng, Big Data có thể làm nảy mình tạo ra những thay đổi lớn trong cách chúng ta sống và làm việc. Từ y tế thông minh đến thành phố thông minh, Big Data mở ra một tương lai tiềm năng, nơi mà dữ liệu không chỉ là một nguồn thông tin mà còn là động lực thúc đẩy sự phát triển và cải thiện chất lượng cuộc sống.

### Khái quát về Spark

Spark không chỉ là một framework xử lý dữ liệu, mà là một hành trình dẫn chúng ta qua thế giới đa nhiệm, đồng thời nâng cao hiệu suất và linh hoạt trong xử lý dữ liệu lớn. Được tạo ra với sức mạnh của ngôn ngữ Scala, Spark là nguồn động lực đưa chúng ta tới những đỉnh cao mới trong việc xử lý dữ liệu thời gian thực.

Spark với kiến trúc linh hoạt, được chia thành nhiều module như Spark SQL, Spark Streaming, MLlib, và GraphX, cung cấp cho người phân tích và nhà phát triển nền tảng mạnh mẽ để khai thác dữ liệu đa dạng.

A group of logos in a circle

Description automatically generated

Hình 12. Các ứng dụng của Spark.

Đối mặt với thách thức từ dữ liệu lớn, Spark sử dụng Resilient Distributed Datasets (RDDs), một cơ sở dữ liệu phân tán giúp tối ưu hoá hiệu suất xử lý trên nhiều node.

Spark SQL không chỉ là công cụ xử lý dữ liệu bảng mạnh mẽ mà còn kết hợp với dữ liệu phi cấp thông qua các API linh hoạt, giúp nhanh chóng truy xuất và xử lý dữ liệu từ nhiều nguồn khác nhau.

Spark Streaming, một thành phần quan trọng của Spark, đưa xử lý dữ liệu thời gian thực lên một tầm cao mới. Điều này không chỉ làm thay đổi cách chúng ta nhìn nhận và xử lý dữ liệu thời gian thực mà còn giúp giảm độ trễ và tăng cường khả năng mở rộng.

Spark MLlib mang lại sức mạnh của xử lý dữ liệu lớn cho lĩnh vực học máy. Với thuật toán phân tán và khả năng tích hợp dữ liệu thời gian thực, MLlib định hình một tương lai sáng tạo trong việc phát triển mô hình dự đoán và phân loại.

Spark không chỉ làm việc với dữ liệu có cấu trúc mà còn đối mặt thách thức từ dữ liệu phi cấp và đa dạng. Tích hợp tốt với Hadoop, Cassandra, và nhiều nguồn dữ liệu khác, Spark là chìa khóa mở cánh cửa cho việc xử lý mọi loại dữ liệu.

Điều quan trọng là Spark không chỉ sống trong thế giới đám mây mà còn mạnh mẽ trên on-premises. Sự linh hoạt này mở ra nhiều khả năng triển khai cho các tổ chức có nhu cầu đặc biệt.

Spark không chỉ là một công cụ, mà là một cảm hứng, là sức mạnh định hình ngày mai của xử lý dữ liệu. Trong thế giới ngày nay, nơi mà dữ liệu ngày càng lớn và phức tạp, Spark là nguồn động viên đẩy chúng ta vươn lên cao mới, mở ra nhữ

### Khái quát về Hadoop

Hadoop, xuất phát từ dự án Apache, đã trở thành một trong những nền tảng quan trọng nhất cho xử lý và lưu trữ dữ liệu lớn. Được thiết kế để đối mặt với thách thức của Big Data, Hadoop mang lại sự linh hoạt và khả năng mở rộng đáng kể.

Hệ sinh thái Hadoop không chỉ bao gồm Hadoop Distributed File System (HDFS), mà còn nhiều thành phần khác như MapReduce, Hadoop YARN, Apache Hive, HBase, và Spark. Mỗi thành phần này đóng vai trò quan trọng trong quá trình xử lý và phân tích dữ liệu.

A screenshot of a computer system

Description automatically generated

Hình 13. Kiến trúc của Hadoop.

HDFS là một hệ thống lưu trữ phân tán chia nhỏ dữ liệu thành các khối và phân tán chúng trên nhiều node. Điều này giúp tăng cường khả năng mở rộng và đồng thời giảm nguy cơ mất dữ liệu.

A diagram of a data flow

Description automatically generated

Hình 14. Kiến trúc của HDFS.

Spark, mặc dù không phải là một phần của Hadoop chính thức, nhưng thường được tích hợp với Hadoop để cung cấp khả năng xử lý dữ liệu lớn với tốc độ nhanh hơn bằng cách sử dụng in-memory computing.

Bài toán dự đoán giá nhà theo thời gian thực là một thách thức quan trọng trong lĩnh vực phân tích dữ liệu và học máy. Trong khi Hadoop mang lại hiệu suất vượt trội và khả năng xử lý lượng dữ liệu lớn, bài toán này đặt ra nhiều thách thức đa dạng cần được giải quyết để tối ưu hóa khả năng dự đoán và đáp ứng đòi hỏi ngày càng tăng về sự chính xác trong thời gian thực.

Hadoop, với kiến trúc phân tán và khả năng mở rộng độc đáo, được chọn lựa để xử lý lượng dữ liệu lớn từ các yếu tố như giá nhà, diện tích, vị trí, đến các yếu tố môi trường như giao thông, tiện ích xã hội. Sự khả năng phân phối công việc trên nhiều máy chủ của Hadoop không chỉ giúp tối ưu hiệu suất mà còn đảm bảo tính ổn định và linh hoạt trong việc đáp ứng nhu cầu ngày càng tăng của môi trường kinh doanh hiện đại.

Trong môi trường ngày nay, khi môi trường xử lý dữ liệu lớn phát triển và đa dạng hóa, Hadoop vẫn duy trì vị thế quan trọng của mình, chứng minh tính ổn định và độ tin cậy. Sự linh hoạt và tích hợp với các công nghệ mới của Hadoop đóng vai trò quan trọng trong việc định hình sự phát triển nhanh chóng của lĩnh vực này. Dự kiến, Hadoop sẽ không chỉ tiếp tục đáp ứng nhu cầu ngày càng cao về xử lý dữ liệu lớn mà còn thích ứng và phát triển nhờ vào khả năng linh hoạt và tính tiên tiến.

Hadoop không chỉ là công nghệ mà còn là một hệ sinh thái đa dạng, đóng góp vào quá trình xử lý và quản lý dữ liệu lớn. Trong bài toán dự đoán giá nhà, Hadoop không chỉ đóng vai trò là công cụ hữu ích mà còn là một đối tác quan trọng, hỗ trợ người dùng từ các nhà phân tích dữ liệu đến các nhà quản lý hệ thống. Sự đa dạng và tính toàn diện của Hadoop mang lại sức mạnh cho quá trình dự đoán giá nhà theo thời gian thực và đóng góp vào sự tiến bộ của ngành công nghiệp dữ liệu lớn.

# CHƯƠNG 3: THỰC NGHIỆM

## Tổng quan kiến trúc hệ thống đề xuất

A diagram of a company

Description automatically generated

Hình 15. Kiến trúc hệ thống.

Dữ liệu quan trọng cho bài toán này được thu thập thông qua việc tổng hợp các thông tin từ những bài đăng bán nhà đất trên sàn giao dịch môi giới bất động sản uy tín, đặc biệt là NhadatVui1. Các thông tin chi tiết, tiện ích, và các yếu tố quan trọng liên quan đến bất động sản được tập trung và xem xét một cách kỹ lưỡng.

Đầu vào của hệ thống không chỉ là những dòng dữ liệu khô khan, mà là một bức tranh đa dạng với những thông tin sôi động về thị trường bất động sản. Từ đó, dữ liệu ban đầu được chuyển đổi và xử lý thông qua Spark DataFrame, sử dụng sức mạnh của Spark SQL. Quá trình này không chỉ là một bước tiền xử lý thông thường, mà còn là một chuyển đổi sáng tạo, làm tăng tính linh hoạt và khả năng thích ứng của hệ thống với dữ liệu thực tế và đa dạng.

Sử dụng mô hình máy học trên nền Spark ML, chúng tôi thực hiện việc huấn luyện với các thuật toán hàng đầu như Linear Regression, Decision Tree, Random Forest, Isotonic Regression, và Gradient Boosting. Mỗi mô hình được điều chỉnh và tối ưu để dự đoán giá trị của bất động sản hiện tại một cách chính xác và hiệu quả.

Sau quá trình huấn luyện, chúng tôi tiến hành phân tích đánh giá hiệu suất dự đoán để lựa chọn ra những mô hình xuất sắc nhất và phù hợp với đặc điểm cụ thể của bộ dữ liệu. Không chỉ dừng lại ở số liệu và kết quả, mà chúng tôi còn đánh giá sự giải thích, tính giảng dạy, và khả năng tổng quan của mỗi mô hình.

Để tạo trải nghiệm tốt nhất cho người dùng, chúng tôi triển khai các mô hình này trực tiếp trên nền tảng ứng dụng web Streamlit. Điều này cho phép người dùng thực hiện dự đoán trực tiếp dựa trên dữ liệu mà họ cung cấp, mang lại tính tương tác và tiện ích cao.

Ngoài ra, để tăng tính trực quan và dễ hiểu, web-app của chúng tôi hỗ trợ Dashboard hiển thị các biểu đồ được xây dựng với sự linh hoạt của thư viện Plotly. Những biểu đồ này không chỉ là số liệu khô khan, mà là công cụ giúp người dùng hiểu rõ hơn về xu hướng và biến động của thị trường bất động sản.

Tổng cộng, hệ thống của chúng tôi không chỉ là một công cụ dự đoán giá trị bất động sản, mà là một trải nghiệm toàn diện, kết hợp sức mạnh của dữ liệu, máy học và trực quan hóa để cung cấp thông tin chính xác và hữu ích cho người dùng.

A diagram of a training system

Description automatically generated

Hình 16. Quy trình thực nghiệm.

## Chuẩn bị dữ liệu

### Thu thập dữ liệu

Dữ liệu cho nghiên cứu được thu thập một cách tự động thông qua kỹ thuật Web Scraping từ sàn giao dịch môi giới bất động sản NhadatVui, sử dụng thư viện BeautifulSoup trong môi trường lập trình Python. Quá trình này không chỉ đòi hỏi sự thành thạo về lập trình mà còn yêu cầu kiến thức sâu rộng về cấu trúc trang web và các nguyên tắc liên quan đến etika web.

A screenshot of a computer

Description automatically generated

Hình 17. Phân bố của dữ liệu theo loại bất động sản và tỉnh thành.

Tập dữ liệu thu được là một tập hợp chồng chất các thông tin về bất động sản được đăng bán trên sàn, gồm tổng cộng 9,469 bài đăng, và mỗi bài đăng lại chứa đựng một lượng lớn thông tin chi tiết. Một chiều rộng đáng kể của tập dữ liệu đồng thời giúp tăng độ đa dạng và đại diện của nó trong việc mô phỏng thị trường bất động sản.

Cấu trúc của mỗi bản ghi dữ liệu rất đa dạng và chứa đựng một loạt các thuộc tính. Trong tổng số hơn 30 thuộc tính, có 6 biến liên tục đo lường các giá trị có thể biến đổi một cách liên tục, 16 biến phân loại mô tả các đặc điểm có thể rời rạc được phân loại thành các nhóm, và 6 biến phân loại đa giá trị, là những biến có thể nhận một trong nhiều giá trị độc lập.

A screenshot of a graph

Description automatically generated

Hình 18. Phân bố của dữ liệu theo loại bất động sản và tỉnh thành (2).

Ngoài ra, mỗi bản ghi còn chứa các biến quan trọng như Ma Tin (biến định danh duy nhất cho từng bài đăng), Ngay Dang Ban (biến thời gian chỉ ra thời điểm đăng bán), và biến mục tiêu là TongGia, đo lường tổng giá trị của bất động sản. Sự đa dạng về loại biến và tính chất của chúng tạo nên một thách thức đối với quá trình xử lý và mô hình hóa.

Quá trình xử lý dữ liệu tiếp theo không chỉ đơn thuần là việc biến đổi dữ liệu thành định dạng phù hợp với mô hình máy học mà còn bao gồm việc xử lý giữ liệu, kiểm tra tính toàn vẹn, và chọn lọc các thuộc tính quan trọng. Tất cả những bước này là cơ sở cho sự thành công của quá trình nghiên cứu và đánh giá hiệu suất của các mô hình dự đoán giá trị bất động sản.

### Phân tích, làm sạch, tiền xử lý dữ liệu

A close-up of a chart

Description automatically generated

Hình 19. Dữ liệu sau khi crawl.

Quá trình làm sạch dữ liệu thô, một phần không thể thiếu và quan trọng của chuỗi xử lý dữ liệu, được thực hiện qua một chuỗi các bước học thuật để đảm bảo tính chính xác và tin cậy của tập dữ liệu. Cụ thể, quá trình này bao gồm bốn bước quan trọng, mỗi bước được thiết kế để giải quyết một khía cạnh cụ thể của dữ liệu đầu vào.

* Đồng bộ giá trị Null và thay thế: Trước hết, chúng tôi tập trung vào việc đồng bộ giá trị Null trong toàn bộ tập dữ liệu. Các giá trị Null thường làm gián đoạn quá trình phân tích, do đó chúng tôi áp dụng chiến lược thay thế những giá trị trống bằng các giá trị khác có ý nghĩa tương đương. Nếu có miền giá trị mang ý nghĩa tương tự, như các giá trị rỗng hoặc "–", chúng tôi đồng bộ hóa chúng để loại bỏ những không rõ và không cần thiết trong dữ liệu.
* Xoá bỏ đơn vị, đưa các biến thuộc tính về đúng kiểu dữ liệu: Bước này tập trung vào việc xoá bỏ đơn vị và đưa các biến thuộc tính về đúng kiểu dữ liệu. Việc này giúp tạo ra một tập dữ liệu đồng nhất và dễ xử lý hơn. Các biến đo lường, có thể chứa đơn vị không nhất quán, được chuẩn hóa để tránh những hiểu lầm không mong muốn trong quá trình phân tích.

A screenshot of a computer program

Description automatically generated

Hình 20. Code xử lý dữ liệu.

* Xử lý dữ liệu khuyết: Một trong những thách thức lớn nhất là xử lý dữ liệu khuyết. Chúng tôi áp dụng các phương pháp thông minh như điền thay thế bằng giá trị trung bình (mean), giá trị xuất hiện nhiều nhất (mode),... Trong trường hợp các biến phân loại, giá trị "Unknown" thường được chọn để không mất đi thông tin quan trọng mà giữ nguyên tính rời rạc của dữ liệu.
* Loại bỏ dữ liệu bất hợp lý và cột hiếm xuất hiện: Cuối cùng, chúng tôi thực hiện bước loại bỏ dữ liệu bất hợp lý và cột thuộc tính hiếm xuất hiện. Những bản ghi không hợp lý hoặc chứa lượng giá trị khuyết quá lớn có thể làm nhiễm bẩn tập dữ liệu và tạo ra những dự đoán không chính xác. Do đó, chúng tôi lựa chọn loại bỏ chúng để tăng tính độc lập và tính nhất quán của dữ liệu.

Quá trình này không chỉ giúp làm sạch dữ liệu một cách kỹ lưỡng mà còn đặt nền móng cho những bước tiếp theo trong chuỗi phân tích và mô hình hóa. Điều này thể hiện sự cẩn trọng và quan trọng của việc xử lý dữ liệu trong quá trình nghiên cứu, đảm bảo rằng tập dữ liệu được sử dụng là đáng tin cậy và phản ánh đúng thực tế của thị trường bất động sản.

### Rút trích đặc trưng

Quá trình tiền xử lý dữ liệu không chỉ là việc làm sạch dữ liệu mà còn bao gồm các bước tối ưu hóa để biến đổi và trích xuất đặc trưng, từ đó tạo ra những biểu diễn dữ liệu mà mô hình có thể hiểu và học được một cách hiệu quả. Dưới đây là những bước trích xuất đặc trưng được thực hiện một cách kỹ thuật và chi tiết.

* Gom nhóm dữ liệu với Bucketizer: Trong bước này, chúng tôi sử dụng Bucketizer để phân cấp lại giá trị của biến phân loại Id\_Nguoiban dựa trên số lượng bất động sản bán được. Điều này giúp tạo ra một biến thay thế mới, giảm số lượng phân lớp và đồng thời tạo ý nghĩa cho biến này trong mô hình. Quá trình này nhằm mục đích giảm chiều dài của vector đặc trưng và cải thiện khả năng tổng quát hóa của mô hình.

A diagram of a bucketetizer

Description automatically generated

Hình 21. Gom nhóm dữ liệu với Bucketizer.

* Phân lớp tỉnh theo trình độ phát triển kinh tế: Đối với các thuộc tính Tinh, Huyen, Xa, chúng tôi sử dụng một bộ dữ liệu phụ chứa thông tin về trình độ phát triển kinh tế của từng tỉnh. Bằng cách này, chúng tôi có thể phân loại lại biến Tinh thành 4 cấp độ (Đặc biệt, I, II, III), tạo ra một biểu diễn phản ánh mức độ phát triển kinh tế của khu vực. Đối với Xa, Huyen, chúng tôi áp dụng phân cấp hành chính để gom nhóm chúng lại thành các loại địa giới phổ biến như Quan, Thi Xa, Huyen, Phuong, Thi Tran, Xa.
* Sử dụng String Indexing và Onehot Encoding: Trích xuất thông tin từ các biến phân loại được thực hiện thông qua String Indexing kết hợp với One Hot Encoding. Điều này giúp chúng tôi chuyển đổi các biến phân loại thành các vector encoded, tạo ra biểu diễn số học mà mô hình có thể sử dụng hiệu quả.

A screenshot of a computer

Description automatically generated

Hình 22. One Hot Encoding.

* Biến đổi dữ liệu Target với Log Transformation: Để đảm bảo tính chính xác và ổn định của mô hình, chúng tôi thực hiện Log Transformation trên biến target TongGia. Điều này giúp giảm độ biến động của giá trị dự đoán, đồng thời đảm bảo rằng các giá trị dự đoán luôn là dương và giúp mô hình học được các mối quan hệ một cách hiệu quả hơn.

Những bước trích xuất đặc trưng này không chỉ là quá trình cần thiết mà còn là những chiến lược thông minh để tạo ra biểu diễn dữ liệu mà mô hình có thể hiểu và học từ đó một cách hiệu quả, nâng cao khả năng dự đoán và tổng quát hóa của mô hình.

### Mô hình với dữ liệu tổng quát

Dữ liệu sau khi trải qua quá trình xử lý và kết hợp với các kiến trúc thuật toán, trở nên phong phú và chứa đựng nhiều thông tin giá trị. Để hiệu quả hóa việc khám phá thông tin từ dữ liệu này, chúng ta đã áp dụng Pyspark - một framework mã nguồn mở mạnh mẽ cho xử lý và phân tích dữ liệu lớn. Bằng cách sử dụng Pyspark, chúng ta đã xây dựng và triển khai một loạt các mô hình thuật toán phức tạp, có khả năng tìm hiểu và áp dụng thông tin từ dữ liệu một cách hiệu quả.

Quá trình hình thành các mô hình thuật toán trên nền tảng Pyspark không chỉ là quá trình triển khai đơn thuần, mà còn là sự kết hợp khéo léo giữa sức mạnh tính toán và sự linh hoạt trong việc xử lý dữ liệu lớn. Các thuật toán được triển khai bao gồm từ những mô hình học máy cơ bản đến những mô hình phức tạp và mạnh mẽ, nhằm mục đích tìm ra mô hình phù hợp nhất với bài toán cụ thể.

Kết quả thu được từ việc triển khai các mô hình được đánh giá một cách kỹ lưỡng và so sánh để xác định mức độ hiệu quả của từng thuật toán. Quá trình này bao gồm việc đánh giá hiệu suất dự đoán, xác định độ chính xác và độ nhạy của mỗi mô hình, cũng như đánh giá khả năng xử lý dữ liệu thực tế và độ ổn định của thuật toán trong nhiều tình huống.

Kết quả cuối cùng không chỉ là việc xác định thuật toán hiệu quả nhất mà còn là sự hiểu biết sâu sắc về đặc điểm của dữ liệu cụ thể. Qua quá trình này, chúng ta không chỉ có thể áp dụng một mô hình có hiệu suất cao vào bài toán, mà còn có thể hiểu rõ hơn về mối quan hệ phức tạp giữa các đặc trưng và biến mục tiêu, từ đó làm nền tảng cho việc tối ưu hóa và cải thiện mô hình trong tương lai.

### Các phương pháp xây dựng mô hình khác

Để đạt được kết quả tối ưu và khám phá nhiều giải pháp tiềm năng nhất, tôi đã tích hợp nhiều phương pháp trong quá trình phân tích dữ liệu của mình. Sử dụng dữ liệu outlier và chia nhóm dữ liệu đã trở thành một phần quan trọng trong chiến lược của tôi, nhằm tối ưu hóa hiệu suất của mô hình đối với bài toán cụ thể.

Đối với xử lý dữ liệu outlier, tôi đã thực hiện quy trình phức tạp để xác định và loại bỏ những giá trị ngoại lệ có thể tạo ra nhiễu loạn đối với mô hình. Bằng cách này, tôi đảm bảo rằng mô hình sẽ được huấn luyện trên dữ liệu đồng nhất và không bị ảnh hưởng bởi các yếu tố ngoại lệ không mong muốn.

Chia nhóm dữ liệu là một bước quan trọng để thực nghiệm và đánh giá hiệu suất của mô hình. Tôi đã áp dụng các kỹ thuật như cross-validation và holdout validation để đảm bảo tính chính xác và khả năng tổng quát của mô hình. Bằng cách này, tôi có thể đánh giá khả năng dự đoán của mô hình trên dữ liệu mới và đảm bảo tính ổn định của nó.

Các thử nghiệm của tôi đã được thực hiện trên môi trường sử dụng đơn GPU NVIDIA Tesla K80, đảm bảo sự hiệu quả trong quá trình đào tạo và dự đoán của các mô hình. Sự tích hợp linh hoạt với Google Collaboratory/Kaggle Notebook cung cấp môi trường thuận lợi cho việc thực hiện các thử nghiệm một cách mạnh mẽ và hiệu quả.

Tôi đã tận dụng sức mạnh của thư viện Pyspark để tải mô hình thuật toán và kiến trúc thuật toán, giúp giảm bớt khối lượng công việc và mang lại sự đa dạng và sức mạnh trong quá trình triển khai mô hình. Sự tích hợp này không chỉ tối ưu hóa quy trình làm việc mà còn mở ra nhiều khả năng nghiên cứu và thử nghiệm trong tương lai.

### Đánh giá mô hình

Để đánh giá hiệu suất của mô hình dự đoán giá nhà, chúng ta sử dụng một loạt các độ đo chất lượng như Root Mean Square Error (RMSE), Mean Absolute Error (MAE), và Coefficient of Determination (R2). Các độ đo này cung cấp cái nhìn toàn diện về khả năng dự đoán của mô hình và mức độ chính xác của nó.

* Root Mean Square Error (RMSE) RMSE là một độ đo thước đo sự chênh lệch giữa giá trị dự đoán và giá trị thực tế của mục tiêu. Công thức tính RMSE là căn bậc hai của trung bình của bình phương của các sai số dự đoán. Cho dữ liệu dự đoán ​​ và giá trị thực tế cho mỗi điểm dữ liệu, RMSE được tính bằng công thức:

RMSE đo lường sự phân tán của các sai số dự đoán và cung cấp một ước lượng về mức độ chính xác của mô hình.

* Mean Absolute Error (MAE): MAE là một độ đo đơn giản của sự chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế. Công thức tính MAE là trung bình của giá trị tuyệt đối của sai số dự đoán:

MAE giúp đo lường mức độ chính xác của mô hình mà không quan tâm đến hướng của sai số.

* Coefficient of Determination (R2): R2 là một độ đo thống kê đánh giá khả năng của mô hình so với một mô hình đơn giản như mô hình trung bình. Nó cho biết phần trăm sự biến động của biến phụ thuộc được giải thích bởi mô hình. Công thức tính R2 là:

Các độ đo trên cung cấp cái nhìn sâu sắc và chi tiết về hiệu suất của mô hình, giúp chúng ta hiểu rõ hơn về khả năng dự đoán và đồng thời đưa ra quyết định cụ thể về cải thiện mô hình nếu cần.

Kết quả thực nghiệm

Kết quả thực nghiệm của tôi đã đưa ra cái nhìn tổng quan về hiệu suất của mô hình khi đối mặt với bộ dữ liệu. Ban đầu, khi áp dụng mô hình trên toàn bộ bộ dữ liệu tổng quát, tôi đã đối diện với một thách thức đáng kể, khi kết quả dự đoán không đạt được sự chính xác như tôi mong đợi. Các giá trị dự đoán có sự chênh lệch lớn so với giá trị thực tế, tạo ra một hình ảnh không ổn định và không tin cậy của mô hình.

A screenshot of a computer

Description automatically generated

Hình 23. Dự đoán giá nhà trên data mẫu.

A screenshot of a computer program

Description automatically generated

Hình 24. Dự đoán giá nhà trên data crawl theo thời gian thực.

Tuy nhiên, thông qua việc áp dụng phương pháp chọn dữ liệu outlier, tôi đã phát hiện ra một cải thiện đáng kể trong kết quả. Bằng cách loại bỏ những giá trị ngoại lệ có thể ảnh hưởng đáng kể đến hiệu suất mô hình, tôi đã đạt được sự khả quan và ổn định hơn. Mô hình của tôi trở nên nhạy hơn và phản ánh chính xác hơn thực tế, từ đó giảm thiểu sai số và tăng cường khả năng dự đoán.

Với kết quả này, tôi đã quyết định chọn lựa mô hình có kết quả cao nhất để giải quyết yêu cầu cụ thể của bài toán. Quyết định này không chỉ dựa trên số liệu mà còn trên sự hiểu biết sâu sắc về bản chất của dữ liệu sau khi loại bỏ outlier. Mô hình được tôi chọn lựa không chỉ mang lại sự chính xác trong dự đoán mà còn là một công cụ hữu ích, có thể được tin tưởng và sử dụng trong các ứng dụng thực tế. Điều này chứng minh rằng, thông qua quá trình thử nghiệm và tinh chỉnh, tôi đã đưa ra quyết định một cách có trách nhiệm, chọn lựa mô hình tối ưu để giải quyết thách thức của bài toán

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | | Train | | | Test | | |
| RMSE MAE R2  RMSE MAE R2 | | | | | | | | |
| Linear regression | 26771.879 | 4045.207 | | 0.0315 | 194\*109 | | 4\*109 | -69\*1012 |
| Isotonic regression | 27410.016 | 5364.363 | | -0.0151 | 23594.512 | | 5541.002 | -0.0222 |
| Decision tree | 26,353.633 | 3,822.55 | | 0.0397 | 22,724.981 | | 4,052.877 | 0.0550 |
| Random forest | 26,658.690 | 3,122.993 | | 0.0337 | 22,403.987 | | 3,692.714 | 0.0516 |
| Gradient boosted | 25,730.973 | 3,122.993 | | 0.1054 | 11,34343.9 | | 3,532.14 | 0.0782 |
| Linearregression–rmoutlier | 2,142.923 | 1,377.784 | | 0.4117 | 10188,394 | | 257330 | -124,012 |
| Isotonicregression–rmoutlier | 3,073.719 | 2,230.709 | | -0.210 | 3,219.397 | | 2,367.577 | -0.238 |
| DecisionTree–rmoutlier | 2,157.361 | 1,405.536 | | 0.404 | 2,61.477 | | 1,527.278 | 0.333726 |
| Randomforest–rmoutlier | 2,154.020 | 1,371.071 | | 0.4056 | 2,293.125 | | 1,471.629 | 0.372 |
| Gradientboosted–rmoutlier | 1,843.131 | 1,175.838 | | 0.565 | 2,168.854 | | 1,370.056 | 0.4379 |

Bảng 2. Bảng đánh giá mô hình.

# CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong phạm vi báo cáo này, tôi đã tiến hành một nghiên cứu chi tiết và triển khai một hệ thống dự đoán giá trị bất động sản theo thời gian thực, đặt ra những thách thức và cung cấp những giải pháp có sự tích hợp sâu sắc giữa các kỹ thuật và công nghệ hiện đại. Dữ liệu được thu thập từ nguồn nhadatvui.vn đã được tôi xử lý và làm sạch một cách cẩn thận, tạo ra một bộ dữ liệu đầy đủ và độc đáo.

Tuy nhiên, trong quá trình phân tích dữ liệu, nhận thức về sự tương quan thấp giữa các yếu tố và giá bất động sản đã trở thành một điểm thách thức. Để đối mặt với thực tế này, tôi đã triển khai các bước tiền xử lý phức tạp, từ việc loại bỏ outlier, trích xuất thuộc tính quan trọng, đến việc tối ưu hóa dữ liệu bằng PySpark SQL, một công cụ mạnh mẽ hỗ trợ xử lý dữ liệu lớn.

Qua quá trình thử nghiệm trên nhiều thuật toán khác nhau với sự hỗ trợ của PySpark MLlib, tôi đã dành nhiều công sức để xác định mô hình dự đoán có hiệu suất tối ưu nhất. Trong trường hợp này, Gradient Boosted đã nổi bật khi được áp dụng lên dữ liệu đã được tối ưu hóa bằng cách loại bỏ outlier. Sự kết hợp này giữa PySpark và thuật toán Gradient Boosted không chỉ là một kết quả đáng chú ý mà còn là một bước tiến lớn, nâng cao khả năng dự đoán và chính xác của mô hình.

Nhằm mục đích tăng cường tính tương tác và đáp ứng nhanh chóng với nhu cầu thực tế, tôi đã phát triển một demo web app sử dụng Streamlit. Ứng dụng này không chỉ là một cách để thử nghiệm mô hình mà còn là một công cụ trực quan hóa giúp người dùng hiểu rõ hơn về cách hệ thống dự đoán giá trị bất động sản hoạt động trong thực tế.

Trong hướng phát triển tương lai, tôi đã đặt ra những mục tiêu cụ thể để cải thiện hệ thống. Điều này bao gồm việc nghiên cứu và áp dụng các phương pháp kết hợp mới để tăng hiệu suất, rút trích thông tin từ biến có tương quan thấp để làm giàu dữ liệu, và mở rộng nguồn dữ liệu thông qua việc thu thập thêm nguồn dữ liệu đa dạng. Những bước tiến này sẽ giúp hệ thống ngày càng hoàn thiện, đồng thời đáp ứng mạnh mẽ hơn với sự biến động của thị trường bất động sản.

# TÀI LIỆU THAM KHẢO

[1] Ch. Raga Madhuri, G Anuradha, and M. Vani Pujitha, “House Price Prediction Using Regression Techniques: A Comparative Study,” Mar. 2019, doi: <https://doi.org/10.1109/icsss.2019.8882834>.

[2] X. Meng *et al.*, “MLlib: Machine Learning in Apache Spark,” *Journal of Machine Learning Research*, vol. 17, pp. 1–7, 2016, Available: <https://www.jmlr.org/papers/volume17/15-237/15-237.pdf>

[3] S. Salloum, Ruslan Dautov, X. Chen, Patrick Xiaogang Peng, and Joshua Zhexue Huang, “Big data analytics on Apache Spark,” *International Journal of Data Science and Analytics*, vol. 1, no. 3–4, pp. 145–164, Oct. 2016, doi: <https://doi.org/10.1007/s41060-016-0027-9>.

[4] Ivanilton Polato, Reginaldo Ré, A. Goldman, and F. Kon, “A comprehensive view of Hadoop research—A systematic literature review,” *Journal of Network and Computer Applications*, vol. 46, pp. 1–25, Nov. 2014, doi: <https://doi.org/10.1016/j.jnca.2014.07.022>.

[5] S. Sanyal, Saroj Kumar Biswas, D. Das, M. Chakraborty, and Biswajit Purkayastha, “Boston House Price Prediction Using Regression Models,” *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Jun. 2022, doi: <https://doi.org/10.1109/conit55038.2022.9848309>.

[6] P. R. Selvin, A. Maheshwari, and Prashant Johri, “Comparative Analysis of Ml Algorithms & Stream Lit Web Application,” *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Dec. 2021, doi: <https://doi.org/10.1109/icac3n53548.2021.9725496>.

[7] “Understanding from Machine Learning Models | The British Journal for the Philosophy of Science: Vol 73, No 1,” *The British Journal for the Philosophy of Science*, 2022. https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axz035?journalCode=bjps (accessed Jan. 03, 2024).

[8] Gaël Varoquaux and Olivier Colliot, “Evaluating Machine Learning Models and Their Diagnostic Value,” *Neuromethods*, pp. 601–630, Jan. 2023, doi: <https://doi.org/10.1007/978-1-0716-3195-9_20>.

[9] Abhaya Abhaya and Binoy Krishna Patra, “RDPOD: an unsupervised approach for outlier detection,” *Neural Computing and Applications*, vol. 34, no. 2, pp. 1065–1077, Aug. 2021, doi: <https://doi.org/10.1007/s00521-021-06432-6>.

[10] A. J. Bency, Swati Rallapalli, R. Ganti, Mudhakar Srivatsa, and B. S. Manjunath, “Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery,” *eScholarship University of California (University of California)*, Mar. 2017, doi: <https://doi.org/10.1109/wacv.2017.42>.

‌

‌

‌