|  |  |
| --- | --- |
| TRƯỜNG ĐH GIAO THÔNG VẬN TẢI PHÂN HIỆU TP.HCM  Bộ Môn Công Nghệ Thông Tin | CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM  Độc Lập – Tự Do -Hạnh Phúc |

**ĐỀ CƯƠNG CHI TIẾT ĐỒ ÁN TỐT NGHIỆP**

1. **Thông tin Sinh viên:**

Họ tên : Trà Minh Sơn Mã sinh viên : 6151071094

Lớp : CQ.61.CNTT Hệ : Chính quy

Ngành đào tạo : Công nghệ thông tin Khoá : 61

Email : 615171094@st.utc2.edu.vn Số điện thoại : 0396980021

1. **Thông tin Giảng viên hướng dẫn:**

Họ tên : Trần Phong Nhã Học vị : Thạc sĩ

Email : tpnha@utc2.edu.vn Số điện thoại : 0906761014

Đơn vị công tác: Đại học GTVT

phân hiệu TP HCM

**NỘI DUNG**

**I. Tên đề tài**

Nghiên cứu thuật toán Random Forest Regression, áp dụng vào việc dự đoán giá nhà cho một trang web mua bán nhà.

**II. Giới thiệu**

*Bối cảnh:*

Trong bối cảnh thị trường bất động sản ngày càng sôi động và phức tạp, việc áp dụng công nghệ vào việc dự đoán giá nhà đang trở nên cần thiết và nổi bật. Việc sử dụng các thuật toán thông minh và tiên tiến không chỉ hỗ trợ khách hàng và nhà đầu tư đưa ra quyết định chính xác hơn mà còn giúp các trang web mua bán nhà tăng cường tính cạnh tranh và hiệu quả hoạt động.

*Cần thiết:*

Phát triển một thuật toán Random Forest Regression trên nền tảng phân tích dữ liệu giá nhà sẽ giúp trang web không chỉ nâng cao khả năng dự đoán giá bán, mà còn cung cấp dịch vụ chuẩn xác và đáng tin cậy cho người dùng. Thuật toán này, với khả năng xử lý đồng thời nhiều biến và định hướng dự đoán chính xác cao thông qua việc phân tích ngẫu nhiên nhiều cây quyết định, sẽ là công cụ đắc lực giúp định hình cách mà giá nhà được đánh giá.

*Mục tiêu:*

Mục tiêu nghiên cứu này là để thiết kế và triển khai thuật toán Random Forest Regression, nhằm mục đích dự đoán giá nhà một cách chính xác hơn. Cụ thể, điều này không chỉ giúp trang web mua bán nhà cung cấp dự đoán chất lượng cao cho khách hàng, mà còn thúc đẩy việc đưa ra quyết định khoa học dựa trên dữ liệu thực tế.

*Phạm vi và đối tượng nghiên cứu:*

Phạm vi nghiên cứu này bao gồm việc thiết kế, triển khai và kiểm định thuật toán Random Forest Regression trong môi trường dữ liệu thực của một trang web mua bán nhà. Đối tượng nghiên cứu chính là các nhà phát triển, nhà phân tích dữ liệu và người dùng của trang web, những người sẽ trực tiếp tương tác và được hưởng lợi từ hệ thống dự đoán này.

**III. Cơ sở lý thuyết**

*Thuật toán cây quyết định(Decision Tree):*

Thuật toán cây quyết định là một phương pháp máy học rất mạnh mẽ và linh hoạt được sử dụng cho cả các bài toán phân loại và hồi quy. Đây là một trong những thuật toán dễ hiểu và dễ giải thích nhất trong lĩnh vực học máy.

Cây quyết định là một cấu trúc cây có các nút trong đó mỗi nút đại diện cho một thuộc tính và mỗi cạnh từ nút đại diện cho một quyết định hoặc một bài toán con. Ý tưởng chính của thuật toán này là phân tách tập dữ liệu thành các phần nhỏ hơn và nhỏ hơn dựa trên các thuộc tính của dữ liệu đó cho đến khi mỗi phần chỉ chứa một loại hoặc giá trị mong muốn.

Ưu điểm:

Dễ giải thích: Kết quả của cây quyết định có thể được diễn giải một cách dễ dàng và trực quan, giúp người dùng hiểu rõ về quá trình ra quyết định của thuật toán.

Không cần tiền xử lý dữ liệu phức tạp: Cây quyết định có thể xử lý dữ liệu có mất mát hoặc bị nhiễu mà không cần phải tiền xử lý dữ liệu một cách phức tạp như một số thuật toán khác.

Phù hợp cho cả dữ liệu dạng số và dạng rời rạc: Cây quyết định có thể xử lý cả dữ liệu dạng số và dạng rời rạc một cách hiệu quả.

Hạn chế:

Dễ bị quá mức phức tạp: Cây quyết định có thể trở nên quá mức phức tạp nếu không được kiểm soát cẩn thận, dẫn đến hiện tượng quá khớp.

Không ổn định: Một số thay đổi nhỏ trong dữ liệu đầu vào có thể dẫn đến sự thay đổi lớn trong cấu trúc của cây quyết định.

*Phương Pháp Bagging:*

Phương pháp bagging (Bootstrap Aggregating) là một kỹ thuật kết hợp (ensemble learning) được sử dụng trong máy học để cải thiện hiệu suất của các mô hình dự đoán. Ý tưởng chính của bagging là tạo ra nhiều mô hình dự đoán độc lập từ các tập dữ liệu con được lấy mẫu ngẫu nhiên từ tập dữ liệu huấn luyện ban đầu, sau đó kết hợp kết quả từ các mô hình này để tạo ra một dự đoán cuối cùng.

Quá trình bagging diễn ra như sau:

1. Lấy mẫu dữ liệu: Từ tập dữ liệu huấn luyện ban đầu, tạo ra các tập dữ liệu con bằng cách lấy mẫu ngẫu nhiên với hoàn lại (bootstrap sampling). Mỗi tập dữ liệu con thường có kích thước bằng với kích thước của tập dữ liệu huấn luyện ban đầu.
2. Xây dựng mô hình: Sử dụng mỗi tập dữ liệu con để huấn luyện một mô hình dự đoán độc lập. Các mô hình này có thể là bất kỳ loại mô hình nào phù hợp với bài toán cụ thể, chẳng hạn như cây quyết định, hồi quy tuyến tính, hoặc mạng nơ-ron.
3. Kết hợp dự đoán: Khi tất cả các mô hình đã được huấn luyện, dự đoán cuối cùng được tạo ra bằng cách kết hợp kết quả từ các mô hình con. Trong trường hợp của phân loại, dự đoán cuối cùng thường được tính bằng cách lấy trung bình hoặc lấy phần đa số của các dự đoán từ các mô hình con.

Ưu điểm của phương pháp bagging bao gồm:

Giảm overfitting: Bằng cách sử dụng nhiều mô hình độc lập và kết hợp kết quả, bagging giảm nguy cơ overfitting so với việc sử dụng một mô hình đơn lẻ.

Ổn định: Kỹ thuật bagging làm giảm sự phụ thuộc vào dữ liệu đầu vào, giúp cải thiện tính ổn định và độ tin cậy của mô hình.

Hiệu suất: Thường thì, bagging tạo ra các mô hình có hiệu suất dự đoán tốt hơn so với mô hình đơn lẻ, đặc biệt là đối với các bài toán có nhiều biến phức tạp hoặc dữ liệu nhiễu.

*Thuật toán Random Forest Regression:*

Thuật toán Random Forest Regression là một phương pháp trong máy học được sử dụng cho các bài toán dự đoán giá trị liên tục, dựa trên cấu trúc của cây quyết định (decision trees) và kỹ thuật ensemble learning.

Một số lý thuyết của thuật toán Random Forest Regression:

Cây quyết định (Decision Trees)

Tập dữ liệu con ngẫu nhiên (Random Subspace): Trong thuật toán Random Forest, từ tập dữ liệu huấn luyện, một tập dữ liệu con được tạo ra bằng cách lấy mẫu ngẫu nhiên mà không cần thiết phải lấy mẫu có hoàn lại. Tập dữ liệu con này thường có kích thước nhỏ hơn tập dữ liệu huấn luyện gốc.

Xây dựng nhiều cây quyết định (Decision Trees): Random Forest sử dụng nhiều cây quyết định độc lập nhau, mỗi cây được xây dựng trên một tập dữ liệu con khác nhau. Điều này giúp giảm thiểu overfitting và tăng tính đa dạng của các cây.

Phương pháp tập hợp (Ensemble Method): Kết quả từ các cây quyết định được kết hợp lại để tạo ra một dự đoán cuối cùng. Trong trường hợp của Regression, kết quả thường được tính bằng cách lấy trung bình của các dự đoán từ các cây con.

Đánh giá độ quan trọng của đặc trưng (Feature Importance): Random Forest cung cấp thông tin về độ quan trọng của từng đặc trưng trong việc đưa ra dự đoán. Điều này có thể giúp hiểu rõ hơn về cách mà các đặc trưng ảnh hưởng đến kết quả dự đoán.

Ưu điểm của thuật toán Random Forest Regression bao gồm khả năng xử lý dữ liệu lớn, khả năng làm việc tốt với dữ liệu không đồng nhất và khả năng tự động chọn đặc trưng quan trọng. Tuy nhiên, việc hiểu rõ về siêu tham số và sự cần thiết của việc điều chỉnh chúng là một điểm cần lưu ý khi triển khai thuật toán này.

*Các thông số đánh giá hiệu suất mô hình:*

**Mean Absolute Error (MAE):**

MAE là một phép đo đánh giá mức độ trung bình của sai số tuyệt đối giữa các dự đoán của mô hình và giá trị thực tế.

Công thức tính MAE là trung bình của các giá trị tuyệt đối của sự khác biệt giữa dự đoán và giá trị thực tế cho mỗi điểm dữ liệu.

**Root Mean Squared Error (RMSE):**

RMSE là một phép đo đánh giá mức độ trung bình của sai số giữa các dự đoán của mô hình và giá trị thực tế, với trọng số cao hơn đối với các sai số lớn hơn.

Công thức tính RMSE là căn bậc hai của trung bình của bình phương của sự khác biệt giữa dự đoán và giá trị thực tế cho mỗi điểm dữ liệu.

**R-squared (R2):**

R2 là một phép đo đánh giá mức độ mà mô hình giải thích phương sai của dữ liệu so với một mô hình dự đoán trung bình đơn giản.

R2 có giá trị từ 0 đến 1, với 0 có nghĩa là mô hình không giải thích được phương sai nào cả và 1 có nghĩa là mô hình giải thích được toàn bộ phương sai của dữ liệu.

Công thức tính R2 liên quan đến tỉ lệ giữa phương sai của dự đoán của mô hình và phương sai của giá trị thực tế.

**IV. Phương pháp nghiên cứu** *Phương pháp nghiên cứu sẽ tập trung vào các bước chính như sau:*

Thu thập dữ liệu về giá nhà từ các nguồn đáng tin cậy và cần thiết.

Xây dựng mô hình Random Forest Regression dựa trên dữ liệu thu thập được.

Tinh chỉnh và đánh giá hiệu suất của mô hình Random Forest Regression.

Triển khai mô hình vào trang web mua bán nhà để cung cấp dự đoán giá nhà cho người dùng.

**V. Kết quả dự kiến**

Xây dựng thành công mô hình thuật toán Random Forest Regression.

Áp dụng thuật toán Random Forest Regression để dự đoán giá nhà cho 1 trang web mua bán nhà ở một cách chính xác.

**VI. Đóng góp của đề tài**

Đề tài sẽ đóng góp vào việc tối ưu hóa quá trình dự đoán giá nhà trên trang web mua bán nhà và cung cấp một công cụ hỗ trợ quyết định dựa trên bộ phân tích kết quả dự đoán. Bằng cách này, trang web sẽ cung cấp thông tin đáng tin cậy và giúp người dùng đưa ra quyết định mua bán nhà một cách hiệu quả hơn.

**VII. Cấu trúc đồ án**

Mô tả chi tiết về cấu trúc của đồ án, bao gồm các chương, phần, mục.

**CHƯƠNG 1: MỞ ĐẦU**

**1.1 Giới thiệu đề tài**

**1.2 Mục tiêu nghiên cứu**

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

**2.1 Thuật toán Random Forest**

**2.2 Cây quyết định và Bagging**

**2.3 Đánh giá hiệu quả của mô hình**

**CHƯƠNG 3: QUÁ TRÌNH NGHIÊN CỨU**

**3.1 Thu thập dữ liệu**

**3.2 Xây dựng mô hình Random Forest Regression**

**3.3 Kiểm định mô hình**

**CHƯƠNG 4: TRIỂN KHAI VÀ KẾT QUẢ THỰ NGHIỆM**

**4.1 Triển khai mô hình trên trang web**

**4.2 Đánh giá kết quả thực nghiệm**

**KẾT LUẬN VÀ KIẾN NGHỊ**

**5.1 Kết luận**

**5.2 Hạn chế và thách thức**

**5.3 Kiến nghị**

**VIII. Tài liệu tham khảo**

[1]"*Introduction to Machine Learning with Python: A Guide for Data Scientists*" by Andreas C. Müller and Sarah Guido: Cuốn sách này cung cấp một cái nhìn tổng quan về machine learning và các thuật toán phổ biến, bao gồm cả Random Forest Regression, cùng với các ví dụ minh họa và hướng dẫn thực hành.

[2]"*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*" by Aurélien Géron: Tài liệu này cung cấp một cái nhìn chi tiết về machine learning với Scikit-Learn, Keras và TensorFlow. Nó bao gồm các phương pháp và kỹ thuật thực tiễn để triển khai thuật toán Random Forest Regression và đánh giá hiệu suất của mô hình.

[3]"*Random Forests*" by Leo Breiman and Adele Cutler: Đây là bài báo gốc về thuật toán Random Forest được viết bởi các tác giả sáng lập thuật toán này. Bài báo này cung cấp cái nhìn sâu sắc về nguyên lý hoạt động của Random Forest và cách áp dụng nó trong thực tế.

**IX. Kế Hoạch thực hiện và tiến độ nghiên cứu**

Thời gian và nội dung công công việc theo tuần.

|  |  |  |
| --- | --- | --- |
| **Thời gian** | **Nội dung công việc** | **Ghi chú** |
| Tuần 1  (18/03 - 24/03) | Chọn đề tài, xây dựng đề cương |  |
| Tuần 2 - 3  (25/03 - 07/04) | Nghiên cứu về thuật toán Random Forest và các nguyên lý cơ bản của mô hình học có giám sát.  Thu thập dữ liệu và tài liệu phục vụ cho việc xây dựng và phân tích mô hình.. |  |
| Tuần 4 - 8  (08/04 - 12/05) | Xử lý dữ liệu  Xây dựng mô hình  Kiểm thử hiệu quả mô hình  Phát triển cơ sở dữ liệu: xây dựng cấu trúc và thiết kế cơ sở dữ liệu giá nhà.  Thiết kế và phát triển giao diện người dùng cho hệ thống dự đoán giá nhà trực tuyến. |  |
| Tuần 9 - 10  (06/05 - 19/05) | Triển khai và tích hợp thuật toán Random Forest Regression vào hệ thống.  Thực hiện kiểm thử ban đầu về chức năng và hiệu năng của hệ thống. |  |
| Tuần 11 - 12  (20/05 - 02/06) | Tiếp tục kiểm thử chức năng và hiệu năng của hệ thống.  Thực hiện các chỉnh sửa nhỏ dựa trên kết quả kiểm thử.  Bắt đầu viết báo cáo. |  |
| Tuần 13  (03/06 - 15/06) | Sửa chữa và hoàn thiện đồ án.  Hoàn thiện báo cáo.  Chuẩn bị cho buổi bảo vệ và trình bày đồ án. |  |

|  |  |  |
| --- | --- | --- |
|  |  | ……ngày….tháng….năm…. |
| **Trưởng Bộ Môn** | **Ý kiến của GVHD** | **Sinh viên thực hiện** |
| **Th.S Trần Phong Nhã** | **Th.S Trần Phong Nhã** | **Trà Minh Sơn** |