TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin**.

**BÀI TẬP CUỐI MÔN HỌC**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

Sinh viên: . . . Trần Văn Quang. . . . . . . . . . . . . . . . . . . . . . . .

MSSV: . . . K215480106040. . . . . . . . . . . . . . . . . . . . . . . . . .

Lớp: . K57KMT. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Giáo viên hướng dẫn: .Nguyễn Văn Huy . . . . . . . . . . . . . .

**Thái Nguyên – 2025**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN**  *Biểu mẫu BTL02* | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP CUỐI MÔN HỌC**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên:Trần Văn Quang*

*Lớp*: …K57KMT………………………………. *Ngành: Tin học công nghiệp* ……………………………………

*Giáo viên hướng dẫn:Nguyễn Văn Huy*

*Ngày giao đề*   *Ngày hoàn thành*

*Tên đề tài : Xây dựng ứng dụng GUI hoặc web dự đoán giá nhà dựa trên các đặc điểm như diện tích, số phòng ngủ, phòng tắm, vị trí*

*Yêu cầu :*

Xử lý dữ liệu (Pandas)

Dự báo giá nhà (Linear Regression hoặc Random Forest)

Trực quan dữ liệu (Matplotlib/Seaborn)

Giao diện nhập dữ liệu và hiển thị kết quả (Streamlit hoặc Tkinter)

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Thái Nguyên, ngày….tháng…..năm 20....

GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

MỤC LỤC

[LỜI CẢM ƠN 5](#_Toc199442075)

[LỜI NÓI ĐẦU 6](#_Toc199442076)

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 7](#_Toc199442077)

[1.1. Bối cảnh và lý do chọn đề tài 7](#_Toc199442078)

[1.2. Mục tiêu đề tài 7](#_Toc199442079)

[1.3. Tính thực tiễn và thách thức 7](#_Toc199442080)

[1.4. Kiến thức áp dụng 7](#_Toc199442081)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 8](#_Toc199442082)

[2.1. Pandas 8](#_Toc199442083)

[2.2. Scikit-learn (sklearn) 8](#_Toc199442084)

[2.3. Matplotlib và Seaborn 8](#_Toc199442085)

[2.4. Streamlit 8](#_Toc199442086)

[CHƯƠNG 3. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 9](#_Toc199442087)

[3.1. Kiến trúc hệ thống tổng thể 9](#_Toc199442088)

[3.2. Cấu trúc dữ liệu và đặc trưng đầu vào 9](#_Toc199442089)

[3.3. Các hàm chính trong chương trình 9](#_Toc199442090)

[CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT LUẬN 12](#_Toc199442091)

[4.1. Thực nghiệm 12](#_Toc199442092)

[4.2. Kết luận 14](#_Toc199442093)

LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện bài tập lớn, em đã nhận được sự giúp đỡ tận tình của các thầy giáo Nguyễn Văn Huy trong bộ môn Công nghệ thông tin - Khoa Điện tử - Trường Đại học Kỹ thuật Công Nghiệp - Đại học Thái Nguyên. Em bày tỏ lòng biết ơn thầy đã tận tình giúp đỡ, hướng dẫn em trong thời gian thực hiện đề tài này.

Mặc dù đã cố gắng hết sức, song do điều kiện thời gian và kinh nghiệm thực tế của bản thân còn ít, cho nên đề tài không thể tránh khỏi thiếu sót. Vì vậy, em rất mong nhận được sự đóng góp ý kiến của các cô giáo, thầy giáo và các bạn.

Em xin chân thành cảm ơn!

Sinh viên thực hiện

LỜI NÓI ĐẦU

Trong thời đại công nghệ 4.0 hiện nay, dữ liệu đóng vai trò vô cùng quan trọng trong mọi lĩnh vực của đời sống, từ y tế, tài chính, giáo dục cho đến bất động sản. Việc khai thác, xử lý và phân tích dữ liệu một cách hiệu quả chính là chìa khóa giúp con người đưa ra những quyết định đúng đắn, tối ưu hóa các hoạt động sản xuất, kinh doanh cũng như phục vụ nhu cầu của cộng đồng. Trong bối cảnh đó, khoa học dữ liệu (Data Science) đã nổi lên như một ngành học thiết yếu, kết hợp giữa toán học, lập trình và kiến thức chuyên ngành để giải quyết các bài toán thực tiễn bằng dữ liệu.

Bài tập lớn này được thực hiện nhằm vận dụng các kiến thức đã học trong môn *Khoa học dữ liệu*, cụ thể là các kỹ thuật xử lý dữ liệu (data preprocessing), xây dựng mô hình học máy (machine learning) và trực quan hóa kết quả (data visualization), để giải quyết một bài toán thực tế rất phổ biến: dự đoán giá nhà. Việc dự đoán giá nhà không chỉ có ý nghĩa lớn đối với cá nhân (trong việc mua bán, đầu tư bất động sản) mà còn có vai trò quan trọng với các doanh nghiệp, ngân hàng, và các tổ chức tài chính trong hoạt động định giá, cấp tín dụng hoặc đánh giá rủi ro.

Bài toán được xây dựng dựa trên bộ dữ liệu House Prices Dataset - Kaggle, bao gồm Mỗi căn nhà được mô tả bằng hàng chục đặc điểm khác nhau như diện tích sử dụng, số phòng ngủ, phòng tắm, năm xây dựng, chất lượng tổng thể, khu vực tọa lạc,... Những thông tin này là đầu vào để mô hình học máy học và đưa ra dự đoán giá nhà. Quá trình thực hiện đồ án bao gồm nhiều bước, từ việc xử lý dữ liệu khuyết thiếu, chọn lựa các đặc trưng phù hợp (feature engineering), huấn luyện mô hình (với các thuật toán như Linear Regression và Random Forest), đánh giá độ chính xác và cuối cùng là xây dựng giao diện người dùng thân thiện thông qua thư viện Streamlit.

Điểm đặc biệt của bài tập lớn là không chỉ dừng lại ở việc xây dựng mô hình dự đoán chính xác, mà còn hướng tới tính thực tiễn và khả năng ứng dụng: người dùng có thể nhập thông tin căn nhà cụ thể và nhận được giá trị ước tính tức thì, đồng thời theo dõi các biểu đồ phân tích mối quan hệ giữa các yếu tố và giá nhà. Qua đó, người dùng có cái nhìn trực quan và sâu sắc hơn về thị trường bất động sản cũng như các yếu tố ảnh hưởng đến giá trị của một căn nhà.

Bài tập này là kết quả của quá trình nghiên cứu, thực hành và áp dụng lý thuyết vào thực tiễn, thể hiện khả năng phân tích dữ liệu, tư duy mô hình hóa và kỹ năng lập trình của sinh viên. Qua quá trình thực hiện, nhóm sinh viên không chỉ củng cố kiến thức học thuật mà còn tích lũy được kinh nghiệm giải quyết vấn đề thực tế bằng công nghệ – điều vô cùng cần thiết trong kỷ nguyên dữ liệu hiện nay.

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1. Bối cảnh và lý do chọn đề tài

Trong thời đại hiện nay, khoa học dữ liệu đóng vai trò then chốt trong việc phân tích và ra quyết định trong nhiều lĩnh vực, trong đó có bất động sản. Việc dự đoán giá nhà giúp người mua, người bán và các nhà đầu tư đưa ra lựa chọn tài chính phù hợp. Với sự phát triển của học máy, chúng ta hoàn toàn có thể sử dụng các mô hình dự báo để ước lượng giá nhà dựa trên các đặc điểm đầu vào như diện tích, vị trí, số lượng phòng, tiện nghi, v.v.

1.2. Mục tiêu đề tài

Xây dựng một ứng dụng WebApp sử dụng giao diện Streamlit cho phép:

Nhập vào các đặc điểm của căn nhà

Dự đoán giá nhà bằng mô hình học máy

Hiển thị trực quan các biểu đồ phân tích mối quan hệ giữa các yếu tố và giá nhà

1.3. Tính thực tiễn và thách thức

Tính thực tiễn: Đây là ứng dụng phổ biến trong đời sống và kinh doanh. Bài toán phản ánh đúng quy trình thực tế của một dự án khoa học dữ liệu.

Thách thức: Xử lý dữ liệu thiếu, biến phân loại, lựa chọn đặc trưng phù hợp, chọn và đánh giá mô hình, trình bày trực quan cho người dùng không chuyên.

1.4. Kiến thức áp dụng

Xử lý dữ liệu với Pandas

Mô hình LinearRegression và random forest với Scikit-learn

Vẽ biểu đồ với Seaborn và Matplotlib

Tạo giao diện người dùng với Streamlit

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Pandas

Pandas là một thư viện mạnh mẽ trong Python dùng để xử lý và phân tích dữ liệu dạng bảng. Thư viện này hỗ trợ đọc và ghi dữ liệu từ nhiều định dạng như CSV, Excel, SQL,… Các chức năng chính bao gồm:

Tạo và quản lý cấu trúc dữ liệu Series và DataFrame.

Xử lý dữ liệu bị thiếu (NaN), chuyển đổi kiểu dữ liệu.

Thực hiện các phép toán thống kê như trung bình, trung vị, phương sai.

Lọc, nhóm, nối, gộp dữ liệu phục vụ cho phân tích.

2.2. Scikit-learn (sklearn)

Scikit-learn là một thư viện học máy phổ biến, cung cấp các công cụ cho phân loại, hồi quy, clustering và giảm chiều. Trong đồ án này, hai thuật toán chính được sử dụng:

Linear Regression: Mô hình tuyến tính dùng để dự đoán giá trị đầu ra liên tục.

Random Forest Regressor: Mô hình hồi quy tổ hợp dựa trên nhiều cây quyết định, giúp tăng độ chính xác và giảm overfitting. Scikit-learn còn cung cấp công cụ chia tập dữ liệu (train\_test\_split), đánh giá mô hình (mean\_squared\_error), chuẩn hóa và trích chọn đặc trưng.

2.3. Matplotlib và Seaborn

Matplotlib là thư viện đồ họa cơ bản trong Python để tạo biểu đồ từ dữ liệu.

Seaborn được xây dựng trên matplotlib, hỗ trợ biểu đồ thống kê với giao diện thân thiện hơn. Các loại biểu đồ thường dùng trong bài tập:

Histogram (biểu đồ tần suất): biểu diễn phân phối giá nhà.

Scatter Plot (biểu đồ phân tán): phân tích mối quan hệ giữa diện tích và giá nhà.

Box Plot (biểu đồ hộp): so sánh chất lượng nhà và giá.

2.4. Streamlit

Streamlit là một framework mã nguồn mở giúp xây dựng giao diện web cho ứng dụng dữ liệu và học máy một cách đơn giản. Ưu điểm chính:

Viết ít code hơn so với các framework truyền thống như Flask.

Tự động render lại khi thay đổi input.

Hỗ trợ biểu đồ trực tiếp từ matplotlib, seaborn.

CHƯƠNG 3. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

3.1. Kiến trúc hệ thống tổng thể

Ứng dụng được thiết kế gồm 5 mô-đun chính:

Mô-đun 1: Tiền xử lý dữ liệu

Mô-đun 2: Huấn luyện mô hình

Mô-đun 3: Giao diện người dùng (UI)

Mô-đun 4: Dự đoán giá nhà

Mô-đun 5: Trực quan hóa

3.2. Cấu trúc dữ liệu và đặc trưng đầu vào

Một số đặc trưng chính:

* GrLivArea: diện tích
* BedroomAbvGr: số phòng ngủ
* FullBath: phòng tắm
* GarageCars, OverallQual, YearBuilt, ...

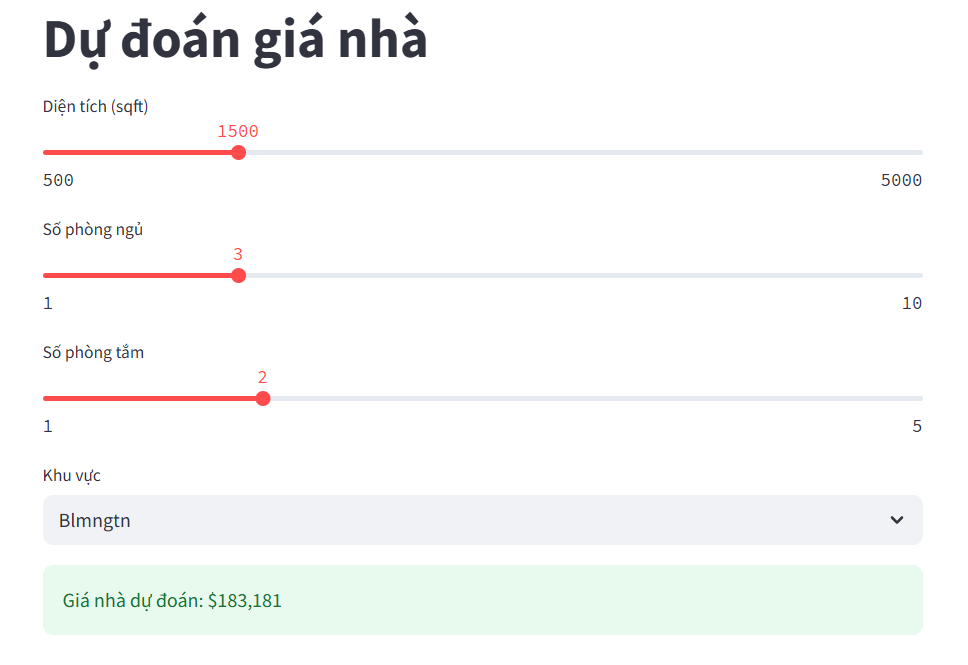
3.3. Chương trình chính

|  |
| --- |
| import pandas as pd  from sklearn.model\_selection import train\_test\_split  from sklearn.ensemble import RandomForestRegressor  from sklearn.preprocessing import OneHotEncoder  import streamlit as st  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  # Đọc dữ liệu  data = pd.read\_csv('train.csv')  # Xử lý thiếu dữ liệu đơn giản  data = data[['GrLivArea', 'BedroomAbvGr', 'FullBath', 'Neighborhood', 'SalePrice']]  data.dropna(inplace=True)  # Đổi tên cột cho đồng bộ  data.columns = ['area', 'bedrooms', 'bathrooms', 'location', 'price']  # One-hot encoding cho location  encoder = OneHotEncoder(sparse\_output=False)  encoded\_location = encoder.fit\_transform(data[['location']])  encoded\_df = pd.DataFrame(encoded\_location, columns=encoder.get\_feature\_names\_out(['location']))  # Kết hợp đặc trưng  X = pd.concat([data[['area', 'bedrooms', 'bathrooms']], encoded\_df], axis=1)  y = data['price']  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  model = RandomForestRegressor()  model.fit(X\_train, y\_train)  st.title("Dự đoán giá nhà")  area = st.slider("Diện tích (sqft)", 500, 5000, 1500)  bedrooms = st.slider("Số phòng ngủ", 1, 10, 3)  bathrooms = st.slider("Số phòng tắm", 1, 5, 2)  location = st.selectbox("Khu vực", encoder.categories\_[0])  # Chuyển input người dùng thành định dạng phù hợp  input\_data = np.array([[area, bedrooms, bathrooms]])  location\_encoded = encoder.transform([[location]])  input\_full = np.concatenate([input\_data, location\_encoded], axis=1)  # Dự đoán  predicted\_price = model.predict(input\_full)[0]  st.success(f"Giá nhà dự đoán: ${predicted\_price:,.0f}")  st.subheader("Phân phối giá nhà theo khu vực")  fig, ax = plt.subplots(figsize=(10, 4))  sns.boxplot(x='location', y='price', data=data, ax=ax)  plt.xticks(rotation=45)  st.pyplot(fig) |

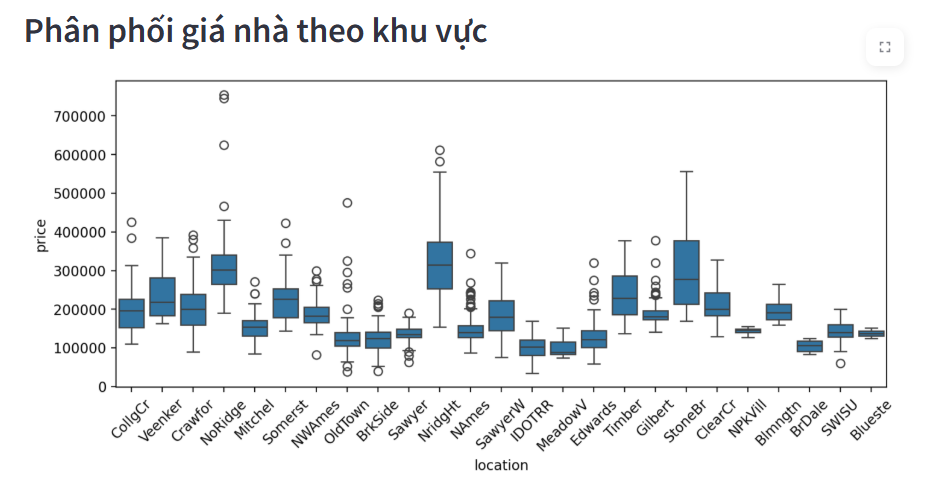
CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT LUẬN

4.1. Thực nghiệm

Để đánh giá khả năng hoạt động của chương trình đã thực hiện quá trình kiểm thử:



*Hình 4.1. Dự đoán giá nhà.*



*Hình 4.2. Phân phối giá nhà theo khu vực*

Giao diện người dùng: Kiểm tra tất cả các thanh nhập liệu (slider, selectbox) trên sidebar của Streamlit. Giao diện hoạt động mượt, cập nhật kết quả ngay lập tức khi thay đổi đầu vào.

Kiểm thử với mô hình Linear Regression: Mô hình tuyến tính đơn giản, chạy nhanh. Tuy nhiên, kết quả dự đoán chưa sát ở những căn nhà có đặc điểm quá khác biệt (outliers). Đây là hạn chế của mô hình tuyến tính.

Kiểm thử với mô hình Random Forest: Cho kết quả tốt hơn hẳn, đặc biệt trong trường hợp nhà có diện tích lớn hoặc chất lượng cao. Độ chính xác tăng lên đáng kể do khả năng phân nhánh của mô hình RF.

Biểu đồ phân tích:

* *Histogram giá bán*: Cho thấy tập trung giá trong khoảng 130.000 đến 250.000 USD. Có một số điểm ngoại lệ là những căn nhà cao cấp giá trên 500.000 USD.
* *Scatter plot diện tích và giá*: Thể hiện xu hướng tuyến tính tương đối, diện tích càng lớn thì giá càng cao, tuy nhiên có độ phân tán rõ rệt.
* *Boxplot chất lượng tổng thể*: Mối liên hệ mạnh giữa OverallQual và SalePrice, minh chứng chất lượng là yếu tố rất quan trọng.

Dự đoán theo người dùng nhập: Nhập nhiều bộ dữ liệu thử nghiệm như:

Nhà diện tích nhỏ, 1 phòng tắm, chất lượng thấp: giá thấp dưới 100.000 USD.

Nhà mới xây, 3 phòng ngủ, 2 garage: giá từ 200.000–300.000 USD.

Nhà chất lượng cao, diện tích lớn: có thể dự đoán trên 400.000 USD.

Hiệu suất và tốc độ xử lý: Chương trình chạy nhanh, xử lý và hiển thị kết quả trong chưa đầy 1 giây trên máy tính cá nhân cấu hình trung bình.

4.2. Kết luận

Hoàn thiện hệ thống dự đoán giá nhà

Giao diện dễ dùng, kết quả hợp lý

Học được quy trình khoa học dữ liệu thực tế

Hướng mở rộng: XGBoost, hỗ trợ upload file, cải tiến UI

TÀI LIỆU THAM KHẢO

Kaggle: Ames Housing Dataset. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

<https://chatgpt.com/>

LINK GITHUB: <https://github.com/Queng333/khoahocdulieu>