TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

KHOA ĐIỆN TỬ

A circular logo with a star and a blue star with a red and yellow circle

Description automatically generated with medium confidence

**TIỂU LUẬN**

**KHOA HỌC DỮ LIỆU**

**BỘ MÔN: CÔNG NGHỆ THÔNG TIN**

**HỆ : ĐẠI HỌC CHÍNH QUY**

ĐỀ TÀI: XÂY DỰNG CHƯƠNG TRÌNH NHẬN DẠNG KHUÔN MẶT ỨNG DỤNG VÀO BÀI TOÁN ĐIỂM DANH

THÁI NGUYÊN - 2024

TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

KHOA ĐIỆN TỬ

A circular logo with a blue star and a yellow and red circle

AI-generated content may be incorrect.

**BÀI TẬP KHOA HỌC DỮ LIỆU**

**BỘ MÔN : CÔNG NGHỆ THÔNG TIN**

ĐỀ TÀI: XÂY DỰNG CHƯƠNG TRÌNH NHẬN DẠNG KHUÔN MẶT ỨNG DỤNG VÀO BÀI TOÁN ĐIỂM DANH

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn** | **:** | **TS. Nguyễn Văn Huy** |
| **Họ và tên sinh viên** | **:** | **La Đức Thắng** |
| **Ngành học** | **:** | **Kỹ thuật Máy tính** |
| **MSSV** | **:** | **K215480106120** |
| **Lớp** | **:** | **K57KMT.01** |

THÁI NGUYÊN - 2024

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

TIỂU LUẬN

KHOA HỌC DỮ LIỆU

**Sinh viên:** La Đức Thắng

**Lớp:** K57KMT **Khoá:** 2021-2025

**Bộ môn:** Công Nghệ Thông Tin

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

**Tên đề tài** : Xây dựng mô hình nhận diện khuôn mặt ứng dụng vào bài toán điểm danh

Yêu cầu :

* Phân tích yêu cầu, dữ liệu
* Trình bày phương pháp học máy được sử dụng
* Huấn luyện và kiểm thử mô hình
* Giao diện chương trình ứng dụng: Web, android, Windows app

|  |  |  |
| --- | --- | --- |
| BCN KHOA | TRƯỞNG BỘ MÔN | GIÁO VIÊN HƯỚNG DẪN |
| (Ký và ghi rõ họ tên) | (Ký và ghi rõ họ tên) | (Ký và ghi rõ họ tên) |

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

Xếp loại:……………… Điểm :……

Thái Nguyên, ngày….tháng…..năm 20....

GIÁO VIÊN HƯỚNG DẪN

(Ký và ghi rõ họ tên)

MỤC LỤC

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 4](#_Toc199336673)

[DANH MỤC TỪ VIẾT TẮT 6](#_Toc199336674)

[LỜI NÓI ĐẦU 7](#_Toc199336675)

[LỜI CẢM ƠN 8](#_Toc199336676)

[CHƯƠNG I: GIỚI THIỆU ĐẦU BÀI 9](#_Toc199336677)

[CHƯƠNG II: CƠ SỞ LÝ THUYẾT 10](#_Toc199336678)

[2.1. Pandas 10](#_Toc199336679)

[2.2. Matplotlib 10](#_Toc199336680)

[2.3. Scikit-Learn 11](#_Toc199336681)

[2.4. Flask 11](#_Toc199336682)

[2.5. Hồi quy tuyến tính 12](#_Toc199336683)

[2.6. Joblib 12](#_Toc199336684)

[CHƯƠNG III: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 13](#_Toc199336685)

[3.1. Sơ đồ khối hệ thống 13](#_Toc199336686)

[3.2. Sơ đồ khối các thuật toán chính 14](#_Toc199336687)

[3.3. Cấu trúc dữ liệu 14](#_Toc199336688)

[3.4. Chương trình 15](#_Toc199336689)

[**3.4.1. Chương trình huấn luyện 15**](#_Toc199336690)

[**3.4.2. Chương trình kiểm thử 16**](#_Toc199336691)

[CHƯƠNG IV: THỰC NGHIỆM VÀ KẾT LUẬN 19](#_Toc199336692)

[4.1. Thực nghiệm 19](#_Toc199336693)

[4.2. Kết luận 20](#_Toc199336694)

[KẾT LUẬN 22](#_Toc199336695)

[TÀI LIỆU THAM KHẢO 23](#_Toc199336696)

# DANH MỤC TỪ VIẾT TẮT

1. CSDL: Cơ sở dữ liệu.
2. CNTT: Công nghệ thông tin

LỜI NÓI ĐẦU

Trong thời đại công nghệ số phát triển mạnh mẽ, trí tuệ nhân tạo và học máy (machine learning) đang đóng vai trò ngày càng quan trọng trong việc hỗ trợ ra quyết định trong nhiều lĩnh vực của đời sống – đặc biệt là trong bất động sản. Việc định giá nhà ở không chỉ dựa trên cảm tính mà ngày nay có thể thực hiện một cách khách quan, chính xác hơn nhờ vào các mô hình học máy.

Đề tài *“Xây dựng ứng dụng dự đoán giá nhà”* được thực hiện nhằm khai thác sức mạnh của các thuật toán như Linear Regression và Random Forest để phân tích, xử lý dữ liệu, và dự đoán giá nhà dựa trên các đặc điểm đầu vào như diện tích, số phòng, năm xây dựng, chất lượng, tiện ích,…

Trong quá trình thực hiện đề tài, em đã cố gắng tìm hiểu, nghiên cứu và áp dụng những kiến thức, kỹ năng đã học để xây dựng một chương trình hoàn chỉnh, có khả năng hoạt động ổn định và đáp ứng được yêu cầu thực tế. Tuy nhiên, do kiến thức và kinh nghiệm còn hạn chế, chắc chắn sẽ không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý, đánh giá từ quý thầy cô và các bạn để đề tài được hoàn thiện hơn.

LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện bài tập, em đã nhận được sự giúp đỡ tận tình của thầy giáo TS. Nguyễn Văn Huy trong bộ môn Tin học công nghiệp – 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 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 thầy giáo, cô giáo và các bạn.

Em xin chân thành cảm ơn!

CHƯƠNG I: GIỚI THIỆU ĐẦU BÀI

**Đầu bài:**

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 và các tiện ích khác.

**Đầu vào:**

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Đầu ra: Giá nhà dự đoán, đồ thị phân phối giá nhà theo các yếu tố.

**Các tính năng:**

* 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

**Hướng dẫn:**

* Sử dụng Pandas đọc và xử lý dữ liệu khuyết thiếu.
* Feature engineering: sử dụng các đặc trưng phù hợp để dự báo.
* Dùng scikit-learn huấn luyện mô hình Linear Regression hoặc Random Forest.
* Hiển thị kết quả bằng đồ thị histogram, scatter plots để phân tích giá.

CHƯƠNG II: CƠ SỞ LÝ THUYẾT

## 2.1. Pandas

Pandas là một thư viện mã nguồn mở mạnh mẽ của Python, được thiết kế đặc biệt để xử lý và phân tích dữ liệu dễ dàng và hiệu quả. Tên "Pandas" bắt nguồn từ "Panel Data" (dữ liệu bảng) hoặc "Python Data Analysis".

Cấu trúc dữ liệu chính:

* Series: Giống như một mảng 1 chiều (có nhãn index)
* DataFrame: Cấu trúc 2 chiều giống bảng tính Excel hoặc bảng SQL

Tính năng chính:

* Đọc/ghi dữ liệu từ nhiều định dạng (CSV, Excel, SQL, JSON,...)
* Lọc, sắp xếp, nhóm dữ liệu
* Xử lý dữ liệu thiếu (NaN)
* Thống kê cơ bản và nâng cao
* Hợp nhất và biến đổi dữ liệu

Pandas có thể tích hợp tốt với Matplotlib/Seaborn để trực quan hóa. Phù hợp với các công việc liên quan tới khoa học dữ liệu hoặc phân tích dữ liệu.

## 2.2. Matplotlib

Matplotlib là thư viện đồ họa mạnh mẽ và phổ biến nhất của Python, dùng để trực quan hóa dữ liệu dưới dạng các biểu đồ 2D và 3D chuyên nghiệp. Nó được thiết kế để mô phỏng cách vẽ đồ thị trong MATLAB nhưng với cú pháp Python.

`pyplot` là module thường dùng nhất, cung cấp cách tiếp cận giống MATLAB – thuận tiện khi viết code nhanh.

## 2.3. Scikit-Learn

Scikit-Learn là một thư viện mã nguồn mở nổi bật trong Python dùng để triển khai thuật toán ML, cung cấp:

- Các mô hình học máy chuẩn hóa: classification, regression, clustering,...

- Giao diện API thống nhất, dễ học, dễ dùng

- Tích hợp với NumPy, Pandas, Matplotlib

## 2.4. Flask

Flask là một framework web nhẹ (micro web framework. Nó được thiết kế để đơn giản, dễ học, và linh hoạt, phù hợp cho cả những ứng dụng nhỏ lẫn những dự án lớn khi được mở rộng với các thư viện khác.

Các đặc điểm chính của Flask:

* Micro framework: Flask không tích hợp sẵn ORM (Object Relational Mapping), hệ thống xác thực người dùng, hay các tính năng khác như một số framework "full-stack" (ví dụ: Django). Tuy nhiên, bạn có thể thêm vào khi cần.
* Dễ học và sử dụng: Rất phù hợp cho người mới học lập trình web bằng Python.
* Modular & mở rộng dễ dàng: Bạn có thể sử dụng các tiện ích mở rộng (extension) như Flask-SQLAlchemy, Flask-Login, Flask-WTF, v.v.
* Routing mạnh mẽ: Cho phép xử lý URL một cách linh hoạt.
* Tích hợp sẵn server phát triển: Thuận tiện cho quá trình xây dựng và kiểm thử ứng dụng.

## 2.5. Hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression) là một phương pháp thống kê dùng để mô hình hóa mối quan hệ tuyến tính giữa một biến phụ thuộc (biến cần dự đoán) và một hoặc nhiều biến độc lập (biến đầu vào).

Hồi quy tuyến tính là một kỹ thuật cơ bản và mạnh mẽ trong học máy dùng để mô hình hóa mối quan hệ giữa một biến đầu ra liên tục và một hoặc nhiều biến đầu vào. Mô hình có dạng:

## 2.6. Joblib

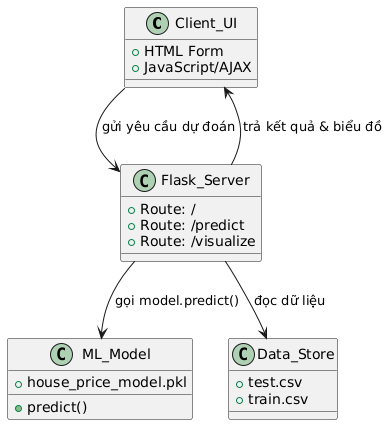
Joblib là một thư viện trong Python được dùng để lưu trữ (serialize) và tải lại (deserialize) các đối tượng Python, đặc biệt là những đối tượng lớn như:

* Mô hình machine learning đã huấn luyện (trained models)
* Mảng dữ liệu lớn (numpy arrays, pandas DataFrames)
* Hàm hoặc đối tượng phức tạp

Trong bài này, joblib sẽ được sử dụng để lưu trữ mô hình đã được huấn luyện nhằm sử dụng cho các chương trình khác.

CHƯƠNG III: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Sơ đồ khối hệ thống



Hình 3.1: Sơ đồ khối hệ thống

## 3.2. Sơ đồ khối các thuật toán chính



Hình 3.2: Sơ đồ các thuật toán hoàn chỉnh

## 3.3. Cấu trúc dữ liệu

Vì file .csv có quá nhiều cột giá trị nên em sẽ chỉ lọc ra một số các giá trị quan trọng.

Dữ liệu đầu vào:

|  |  |
| --- | --- |
| **Tên biến** | **Ý nghĩa** |
| GrLivArea | Diện tích sử dụng |
| LotArea | Diện tích đất |
| OverallQual | Chất lượng tổng thể |
| YearBuilt | Năm xây dựng |
| YearRemodAdd | Năm gần nhất được sửa chữa |
| GarageCars | Sức chứa xe trong gara |
| FullBath | Số phòng tắm |
| TotalBsmtSF | Diện tích tầng hầm |

Dữ liệu đầu ra:

|  |  |
| --- | --- |
| **Tên biến** | **Ý nghĩa** |
| SalePrice | Giá nhà dự đoán |
| PriceRange | Khoảng giá |

## 3.4. Chương trình

### 3.4.1. Chương trình huấn luyện

Đầu tiên, chúng ta cần huấn luyện mô hình để sử dụng. Để huấn luyện mô hình cần nạp dữ liệu, sau đó tách các đặc trưng cần huấn luyện ra.

df = pd.read\_csv('data/test.csv')

features = [

'GrLivArea', #Diện tích sử dụng

'LotArea', #Diện tích đất

'OverallQual', #Chất lượng tổng thể

'YearBuilt', #Năm xây dựng

'YearRemodAdd', #Năm gần nhất được sửa chữa

'GarageCars', #Sức chứa xe trong gara

'FullBath', #Số phòng tắm

'TotalBsmtSF' #Diện tích tầng hầm

]

Sử dụng Pandas, loại bỏ các cột có giá trị thiếu và tạo thêm cột 'SalePrice' để lưu kết quả dự đoán.

target = 'SalePrice'

df = df[features + [target]].dropna()

Tiếp đến, sử dụng tính năng chia dữ liệu và huấn luyện mô hình của thư viện Scikit-Learn:

#Gán nhãn

X = df[features]

y = df[target]

# Chia dữ liệu

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=42)

# Huấn luyện

model = LinearRegression()

model.fit(X\_train, y\_train)

Cuối cùng, lưu mô hình huấn luyện bằng joblib để sử dụng cho các chương trình khác.

joblib.dump(model, 'house\_price\_model.pkl')

### 3.4.2. Chương trình kiểm thử

Bây giờ ta sẽ sử dụng mô hình đã được huấn luyện để kiểm thử với tập dữ liệu test.csv.

Trước tiên, trích xuất đặc trưng và loại bỏ các cột thiếu như file train ở trên. Sau đó sử dụng mô hình 'house\_price\_model.pkl' đã lưu để dự đoán giá nhà dựa trên các tiêu chí đó. Cuối cùng lưu kết quả vào cột SalePrice.

predicted\_prices = model.predict(df)

df['SalePrice'] = predicted\_prices

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.3: Kết quả dự đoán

Vì kết quả không thể hiển thị hết nên ta sẽ cần trực quan hóa kết quả dự đoán. Ở đây em sẽ dùng matplotlib để vẽ một biểu đồ tròn vì biểu đồ tròn là dạng biểu đồ dễ nhìn, dễ hiểu và dễ hình dung.

Đầu tiên, chia giá nhà ra từng khoảng

bins = [0, 100000, 200000, 300000, 400000, 500000, float('inf')]

labels = ['<100k', '100k–200k', '200k–300k', '300k–400k', '400k–500k', '>500k']

df['PriceRange'] = pd.cut(df['SalePrice'], bins=bins, labels=labels)

Tiếp đến, đếm số lượng các giá trị dự đoán như từng khoảng ở trên.

range\_counts = df['PriceRange'].value\_counts().sort\_index()

total = range\_counts.sum()

Cuối cùng, tiến hành vẽ biểu đồ bằng matplotlib.

Thêm bảng chú giải:

legend\_labels = [

f"{label}: {count} ({count/total:.1%})"

for label, count in zip(range\_counts.index, range\_counts)

]

Vẽ biểu đồ:

colors = plt.cm.Set3.colors[:len(range\_counts)]

plt.figure(figsize=(10, 6))

patches, texts, autotexts = plt.pie(range\_counts, autopct='', startangle=90, colors=colors)

plt.legend(patches, legend\_labels, title="Khoảng giá", loc="center left", bbox\_to\_anchor=(1, 0.5))

plt.title('Tỷ lệ số lượng nhà theo từng khoảng giá dự đoán')

plt.tight\_layout()

plt.show()

A pie chart with different colored circles

AI-generated content may be incorrect.

Hình 3.4: Tỷ lệ số lượng nhà theo từng khoảng giá dự đoán

CHƯƠNG IV: THỰC NGHIỆM VÀ KẾT LUẬN

## 4.1. Thực nghiệm

Sử dụng mô hình trên, em đã xây dựng một ứng dụng web bằng Flask.

Chương trình Flask sẽ có các hàm sau:

* Hàm home(): Dùng để tải mô hình, dữ liệu vào giao diện.
* Hàm predict(): Dùng để dự đoán kết quả.

Sau đó, ta cần thiết kế một giao diện. Em sẽ sử dụng html kết hợp với css và JavaScript.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 4.1: Cửa sổ giao diện

A screenshot of a computer

AI-generated content may be incorrect.

Hình 4.2: Giá nhà dự đoán, bao gồm cả giá USD và VND

## 4.2. Kết luận

Kết quả đã đạt được:

* Ứng dụng web hoàn chỉnh bằng Flask
* Cho phép nhập thông tin nhà qua giao diện người dùng
* Trả về kết quả giá dự đoán bằng USD và VND

Kiến thức đã học được:

* Đọc và xử lý dữ liệu với Pandas.
* Dùng Linear Regression để huấn luyện mô hình.
* Trực quan hóa dữ liệu bằng matplotlib.
* Tạo ứng dụng bằng Flask.

Hướng phát triển của đề tài:

* Phát triển lên một ứng dụng kinh doanh bất động sản hoàn chỉnh.
* Thêm chức năng vẽ biểu đồ dựa trên nhiều yếu tố khác nhau.

Link GitHub của chương trình:

A qr code with a dinosaur

AI-generated content may be incorrect.

KẾT LUẬN

Kết quả dự đoán của mô hình cho thấy độ chính xác tương đối cao khi áp dụng các kỹ thuật xử lý dữ liệu hợp lý và lựa chọn đặc trưng (feature engineering) phù hợp. Đồng thời, việc kết hợp giao diện trực quan giúp người dùng không chuyên cũng có thể tiếp cận và khai thác ứng dụng một cách hiệu quả. Tuy nhiên do trình độ của em còn hạn chế nên đề tài của em vẫn chưa thể đáp ứng được nhu cầu thực tế, còn nhiều những thiếu sót hạn chế nhất định. Nhưng qua đó em đã rút ra rất nhiều kinh nghiệm cho bản thân.

Em xin chân thành cảm ơn thầy Nguyễn Văn Huy đã giúp em hoàn thành đề tài này.

TÀI LIỆU THAM KHẢO

<https://chatgpt.com>