**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**ĐỒ ÁN**

**HỌC PHẦN MÁY HỌC ỨNG DỤNG**

**Đề tài**

**XÂY DỰNG MÔ HÌNH­­ MÁY HỌC ĐỂ DỰ ĐOÁN SỐ LƯỢNG XE ĐẠP ĐƯỢC THUÊ**

**Nhóm sinh viên thực hiện:**

**1. Nguyễn Quốc Việt B2111908**

**2. Trường Huỳnh Tú Như B2111893**

**3. Nguyễn Hồng Tuấn Phát B2111894**

**4. Lê Thị Tiến B2105647**

#### Cần Thơ, 4/2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**ĐỒ ÁN**

**HỌC PHẦN MÁY HỌC ỨNG DỤNG**

**Đề tài**

**XÂY DỰNG MÔ HÌNH­­ MÁY HỌC ĐỂ DỰ ĐOÁN SỐ LƯỢNG XE ĐẠP ĐƯỢC THUÊ**

**Giảng viên hướng dẫn: Nhóm sinh viên thực hiện:**

**Huỳnh Gia Khương 1. Nguyễn Q Việt B2111908**

**Nhóm: 6 2. Trương H.T Như B2111893**

**3. Nguyễn H.T Phát B2111894**

**4. Lê Thị Tiến B2105647**

#### Cần Thơ, 4/2024

**NHẬN XÉT CỦA GIẢNG VIÊN**

---------------------------------------------------------------------------------

---------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------

*Cần Thơ, ngày tháng năm*

(Ký và ghi rõ họ tên)

**MỤC LỤC**

**PHÂN CÔNG CÔNG VIỆC**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ tên SV** | **Công việc** | **Thời gian** | **Ký tên nhận việc** | **Ghi chú** |
| 1 | B2111908 | Nguyễn Quốc Việt | Trưởng nhóm, phân tích, xây dựng giải thuật, Làm Slide, Thuyết trình | 31/3 – 10/4 | A close up of a piece of paper  Description automatically generated |  |
| 2 | B2111893 | Trương Huỳnh Tú Như | Thư ký, phân tích, xây dựng giải thuật, Làm Slide, Thuyết trình | 31/3 – 10/4 | A close up of a signature  Description automatically generated |  |
| 3 | B2111893 | Nguyễn Hồng Tuấn Phát | phân tích, xây dựng giải thuật, Làm Slide, Thuyết trình | 31/3 – 10/4 | A piece of paper with blue writing  Description automatically generated |  |
| 4 | B21105647 | Lê Thị Tiến | phân tích, xây dựng giải thuật, Làm Slide, Thuyết trình | 31/3 – 10/4 | A close up of a paper  Description automatically generated |  |

**PHẦN NỘI DUNG**

## Mô tả dữ liệu

Tập dữ liệu Bike Sharing Demand được thu thập từ Kaggle, tập dữ liệu Bike Sharing Demand là một trong những cuộc thi của Kaggle diễn ra vào năm 2015 với câu hỏi được đặt ra là: Dựa vào data của một hãng cung cấp cho thuê xe đạp ở Washington D.C, người tham gia cần phải dự báo được số lượng xe đạp sẽ được thuê.

## 2. Ý nghĩa của dữ liệu

a

## 3. Phân tích dữ liệu và lựa chọn mô hình

### 3.1. Phân tích dữ liệu

#### 3.1.1. Ý nghĩa của các đặc trưng

datetime: cho biết thời gian theo dịnh dạng %m/%d/%Y %H:%M:%hS

season: cho biết mùa

1: mùa xuân

2: mùa hè

3: mùa thu

4: mùa đông

workingday: cho biết có phải là ngày làm việc không bao gồm cả cuối tuần và ngày lễ

1: ngày làm việc

0: ngày không làm việc

weather: Cho biết thời tiết

1: Trời trong, có mây, có mây rải rác

2: Sương mù + Đầy mây, sương mù + ít mây, sương mù

3: Tuyết nhẹ, mưa nhẹ + sấm sét + mây rải rác, mưa nhẹ + mây rải rác

4: Mưa lớn + mảnh băng + sấm sét + sương mù, tuyết + sương mù

temp: Nhiệt độ thực tế tính theo độ Celsius (độ °C)

atemp: Nhiệt độ mà con người cảm thấy tính theo độ Celsius (độ °C)

humidity: Độ ẩm tương đối

windspeed: Tốc độ gió

casual: Số lượng xe thuê mà không có đăng ký dịch vụ

registered: Số lượng xe thuê mà có đăng ký dịch vụ

count: Tổng số lượng xe được thuê

month: tháng (1-12)

year: năm (2011, 2012)

hour: giờ trong ngày (0-23 giờ)

weekday: ngày trong tuần

0: thứ 2

1: thứ 3

2: thứ 4

3: thứ 5

4: thứ 6

5: thứ 7

6: chủ nhật

#### 3.1.2 Tiền xử lý và phân tích dũ liệu

A screenshot of a computer screen

Description automatically generated

Hình 1: Bản đồ nhiệt của các đặc trưng trước khi xử lý

A screenshot of a computer screen

Description automatically generated

Hình 2: Kiểu dữ liệu của các đặc trưng trước khi xử lý

Do biến datetime có kiểu dữ liệu là object nên sẽ chuyển biến datetime sang các biến year, month, weekday, hour để thực hiện tính toán.

Do đề tài là dự đoán số lượng xe đạp được thuê mà biến casual + registered = count nên sẽ xoá biến casual + registered để xây dựng mô hình.

A screenshot of a computer

Description automatically generated

Hình 3: Bản đồ nhiệt của các đặc trưng sau khi xử lý

A screenshot of a computer

Description automatically generated

Hình 4:Kiểu dữ liệu của các đặc trưng sau khi xử lý

Vậy tập thuộc tính để xây dựng mô hình là: season, holiday, workingday, holiday, weather, temp, atemp, humidity, windspeed, hour, month, weekday, year, count. Trong đó count là thuộc tính mục tiêu cần dự đoán. Sau đây là biểu đồ phân phối của các thuộc tính.

[A screenshot of a computer

Description automatically generated](#hinh5)

Hình 5: Kiểm tra xem có thuộc tính nào có giá trị bị thiếu không

Do dữ liệu từ Bike Sharing Demand đã được xử lý làm sạch nên không có giá trị nào bị thiếu. Vì vậy không cần xử lý giá trị thiếu.

A graph showing the different seasons

Description automatically generated with medium confidence

Hình 6: Biểu đồ phân phối giá trị của thuộc tính season

Dựa vào biểu đồ trên ta thấy dữ liệu season được phân phối đều nhau.

A white rectangular object with black text

Description automatically generated

Hình 7: Biểu đồ phân phối giá trị của thuộc tính holiday

Dựa vào biểu đồ trên ta thấy dữ liệu holiday phân phối ở cột not holiday tương đối cao trong khi cột holiday chiếm rất ít.

A white rectangular object with black text

Description automatically generated

Hình 8: Biểu đồ phân phối giá trị của thuộc tính workingday

Biểu đồ trên cho thấy dữ liệu workingday được phân phối chủ yếu ở cột working day.

A graph showing the different weather conditions

Description automatically generated with medium confidence

Hình 9: Biểu đồ phân phối giá trị của thuộc tính weather

Biểu đồ trên cho thấy dữ liệu weather được phân phối tương đối cao ở cột Clear (trời trong), rải rác ở các cột Mist (sương mù) và Light Snow (tuyết nhẹ), hầu như không có ở cột Heavy Rain (mưa lớn).

A graph showing the distribution of temperature

Description automatically generated

Hình 10: Biểu đồ phân phối giá trị của thuộc tính temp

Biểu đồ trên cho thấy dữ liệu temp được phân phối rải rác ở các cột, trong đó tương đối cao ở khoảng 10°C đến 30°C.

A graph of a bar graph

Description automatically generated with medium confidence

Hình 11: Biểu đồ phân phối giá trị của thuộc tính atemp

Biểu đồ trên cho thấy dữ liệu atemp được phân phối gần tương tự như temp.

A graph showing the amount of humidity

Description automatically generated

Hình 12: Biểu đồ phân phối giá trị của thuộc tính humidity

Trong biểu đồ, ta thấy hầu hết các ngày trong dữ liệu humidity đều có độ ẩm từ 50 đến 90.

A graph of a distribution of windspeed

Description automatically generated

Hình 13: Biểu đồ phân phối giá trị của thuộc tính windspeed

Biểu đồ trên cho thấy dữ liệu windspeed dường như được phân phối chủ yếu ở khoảng 5 đến 20.

A diagram of a distribution of hours

Description automatically generated

Hình 14: Biểu đồ phân phối giá trị của thuộc tính hour

Dựa vào biểu đồ trên ta thấy dữ liệu hour được phân phối khá đều ở các cột tuy nhiên có vài cột lại khá cao.

A graph of blue bars with white text

Description automatically generated

Hình 15: Biểu đồ phân phối giá trị của thuộc tính month

Dựa vào biểu đồ ta thấy dữ liệu month được phân phối đều nhau.

A graph showing the number of days and months

Description automatically generated with medium confidence

Hình 16: Biểu đồ phân phối giá trị của thuộc tính weekday

Biểu đồ trên cho thấy dữ liệu weekday được phân phối tương đối đều nhau.

A white rectangular object with text

Description automatically generated

Hình 17: Biểu đồ phân phối giá trị của thuộc tính year

Dựa vào biểu đồ trên ta thấy dữ liệu year được phân phối đều nhau ở cả 2 năm 2011 và 2012.

A graph of a number of bars

Description automatically generated

Hình 18: Biểu đồ phân phối giá trị của thuộc tính count

Biểu đồ trên cho thấy dữ liệu count được phân phối chủ yếu ở khoảng 0 đến 200 và giảm dần qua các cột.

A graph showing different seasons

Description automatically generated

Hình 19: Biểu diễn mối quan hệ giữa Season và Count

Nhu cầu xe được thuê cao nhất vào mùa thu (fall) và thấp nhất vào mùa xuân (spring).

A graph of a graph with a purple and blue rectangle

Description automatically generated with medium confidence

Hình 20: Biểu diễn mối quan hệ giữa Holiday và Count

Nhu cầu thuê xe vào ngày bình thường dường như cao hơn ngày lễ.

A blue and purple squares

Description automatically generated

Hình 21: Biểu diễn mối quan hệ giữa Workingday và Count

Nhu cầu xe được thuê của working day cao hơn not working day.

A graph showing different weather conditions

Description automatically generated

Hình 22: Biểu diễn mối quan hệ giữa Weather và Count

Nhu cầu thuê xe khi trời trong (clear) cao nhất, trong khi có tuyết nhẹ (light snow) sẽ ít được thuê.

A graph showing a graph of a temperature

Description automatically generated with medium confidence

Hình 23: Biểu diễn mối quan hệ giữa Temp và Count

Nhu cầu cao khi thời tiết tốt và nhiệt độ dễ chịu (20°C đến 30°C).

A graph showing a number of blue dots

Description automatically generated

Hình 24: Biểu diễn mối quan hệ giữa Atemp và Count

Nhu cầu thuê xe cao khi nhiệt độ trong khoảng 20°C đến 40°C.

A graph showing the amount of humidity

Description automatically generated

Hình 25: Biểu diễn mối quan hệ giữa Humidily và Count

Nhu cầu thuê xe cao khi độ ẩm nằm trong khoảng 30 đến 80.

A graph showing a number of wind speed

Description automatically generated

Hình 26: Biểu diễn mối quan hệ giữa Windspeed và Count

Nhu cầu xe được thuê cao khi tốc độ gió trong khoảng 10 đến 30.

### 3.2. Lựa chọn mô hình

Sử dụng nghi thức Hold-out và K-fold để phân chia tập dữ liệu và 5 mô hình: Phương pháp tổng hợp mô hình, KNN, Linear ReGressor, Ramdomforest, Decision Tree để huấn luyện tập dữ liệu.

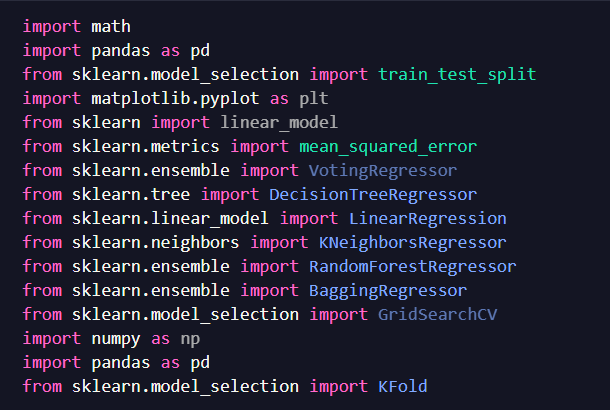
## 4. Cấu hình máy tính

|  |  |
| --- | --- |
| **CPU** | **Intel Core i3-8145U 2.10Ghz** |
| **RAM** | **8GB** |
| **ROM** | **SSD 512GB** |
| **GPU** | **Intel UHD Graphics 620** |

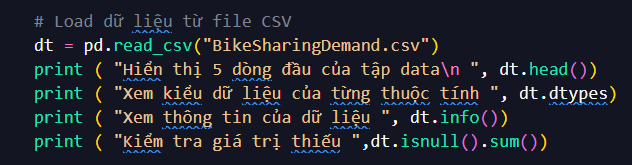
### 5. Huấn luyện và Kết quả thực nghiệm

## 5.1. Các bước huấn luyện đầu tiên

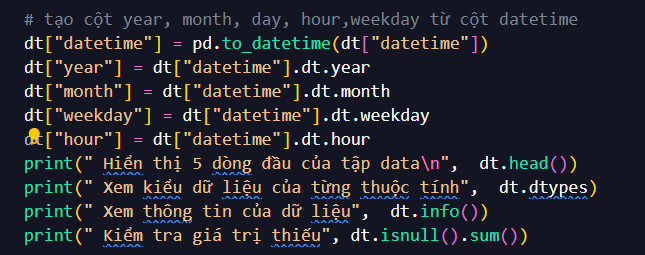
**Bước 1:** Nhập tất cả thư viện cần thiết.



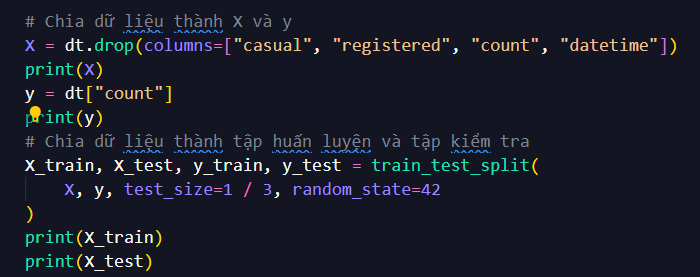
**Bước 2:** Đọc file BikeSharingDemand.csv và in ra các thông tin cần thiết.

****

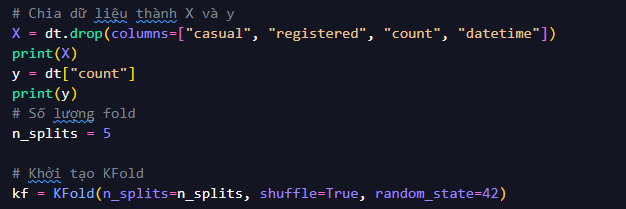
**Bước 3:** Xử lý data



**Bước 4:** Chia tập dữ liệu theo Hold-out train 2/3 và test 1/3.

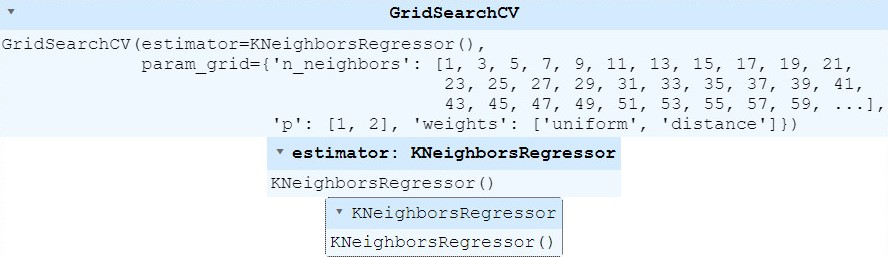


**Bước 5:** Chia tập dữ liệu theo K-fold.



### 5.2. KneighborsRegressor

Sử dụng GridSearchCV để tìm ra tham số tốt nhất cho mô hình, ta được:

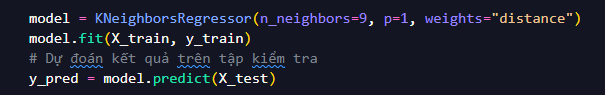


Hình 27: Kết quả của GridSearchCV đối với KneighborsRegressor

Kết quả sau khi tìm tham số tốt nhất:



Huấn luyện mô hình:



Hình 28: Huấn luyện mô hình KneighborsRegressor

Hold-out:



Hình 29: Kết quả của KneighborsRegressor với nghi thức Hold-out

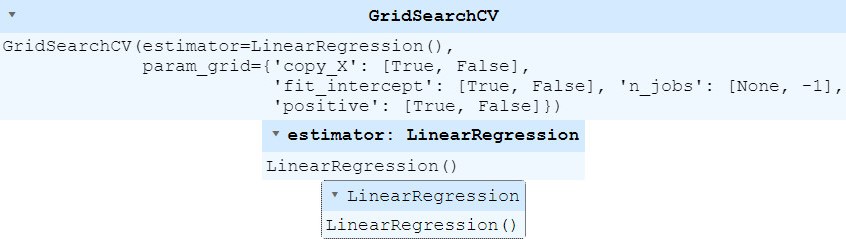
K-fold:



Hình 30: Kết quả của KneighborsRegressor với nghi thức K-fold

### 5.3. Linear Regression

Sử dụng GridSearchCV để tìm ra tham số tốt nhất cho mô hình, ta được:

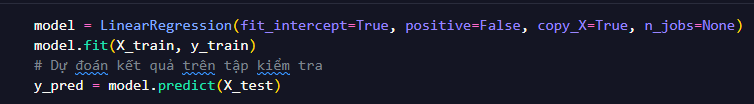


Hình 31: Kết quả của GridSearchCV đối với Linear Regression

Kết quả sau khi tìm tham số tốt nhất:



Huấn luyện mô hình:



Hình 32: Huấn luyện mô hình Linear Regression

Hold-out:



Hình 33: Kết quả của Linear Regression với nghi thức Hold-out

K-fold:



Hình 34: Kết quả của Linear Regression với nghi thức K-fold

### 5.4. DecisionTreeRegressor

Sử dụng GridSearchCV để tìm ra tham số tốt nhất cho mô hình, ta được:

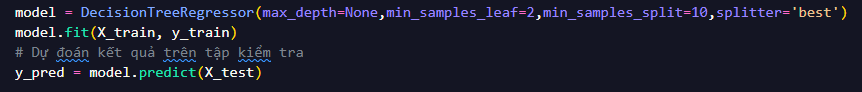
A screenshot of a computer

Description automatically generated

Kết quả sau khi tìm tham số tốt nhất:



Huấn luyện mô hình:



Hình 35: Huấn luyện mô hình DecisionTreeRegressor

Hold-out:



Hình 36: Kết quả của DecisionTreeRegressor với nghi thức Hold-out

K-fold:



Hình 37: Kết quả của DecisionTreeRegressor với nghi thức K-fold

### 5.5. RandomForestRegressor

Sử dụng GridSearchCV để tìm ra tham số tốt nhất cho mô hình, ta được:

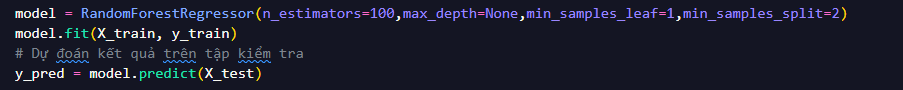
A screenshot of a computer code

Description automatically generated

Kết quả sau khi tìm tham số tốt nhất:



Huấn luyện mô hình:



Hình 38: Huấn luyện mô hình RandomForestRegressor

Hold-out:



Hình 39: Kết quả của RandomForestRegressor với nghi thức Hold-out

K-fold:



Hình 40: Kết quả của RandomForestRegressor với nghi thức K-fold

### 5.6. Phương pháp tổng hợp mô hình (BaggingRegressorRegressor)

Huấn luyện mô hình bằng phương pháp BaggingRegressorRegressor với mô hình cơ sở là KneighborsRegressor:



A computer screen shot of text

Description automatically generated

Hold-out:



K-fold:



Huấn luyện mô hình bằng phương pháp BaggingRegressor với mô hình cơ sở là Linear Regression:



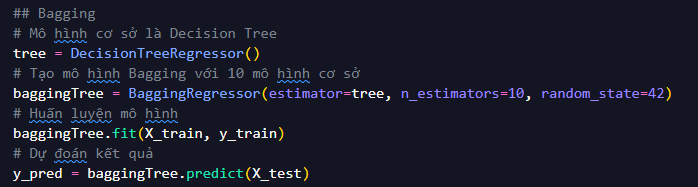
Hold-out:



K-fold:



Huấn luyện mô hình bằng phương pháp BaggingRegressor với mô hình cơ sở là DecisionTreeRegressor:



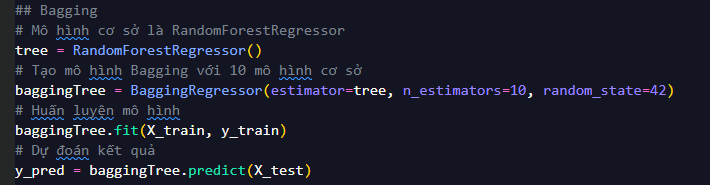
Hold-out:



K-fold:



Huấn luyện mô hình bằng phương pháp BaggingRegressor với mô hình cơ sở là RandomForestRegressor:



Hold-out:



K-fold:



## 6. Đánh giá mô hình

### 6.1 Đánh giá mô hình Regressor

### 6.2. Nhận xét kết quả thực nghiệm

## PHẦN KẾT LUẬN

### 1. Kết quả đạt được

### 2. Hướng phát triển

# TÀI LIỆU THAM KHẢO

[1].

[2].