BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

ĐỒ ÁN
HỌC PHẦN MÁY HỌC ỨNG DỤNG

Đề Tài

DỰ ĐOÁN GIÁ TRỊ NHIỆT ĐỘ TRUNG BÌNH CỦA ANKARA TỪ 01/01/1994 ĐẾN 28/05/1998

Họ và Tên	MSSV	KHÓA
Nguyễn Huỳnh Hữu Nhâm	B2005809	46
Nguyễn Khả Siêu	B2005821	46
Trần Đình Sang	B2014876	46
Nguyễn Duy Khánh	B2014842	46

Cần Thơ, 04/2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

ĐỒ ÁN
HỌC PHẦN MÁY HỌC ỨNG DỤNG

Đề tài

DỰ ĐOÁN GIÁ TRỊ NHIỆT ĐỘ TRUNG BÌNH CỦA ANKARA TỪ 01/01/1994 ĐẾN 28/05/1998

Giảng viên hướng dẫn:
TS. Trần Nguyễn Minh Thư
TS. Mã Trường Thành

Sinh viên thực hiện:
Nguyễn Huỳnh Hữu Nhâm
Mã số: B2005809
Nguyễn Khả Siêu
Mã số: B2005821
Trần Đình Sang
Mã số: B2014876
Nguyễn Duy Khánh
Mã số: B2014842
Lớp/Khóa: 46

Cần Thơ, 04/2023

NHẬN XÉT CỦA GIẢNG VIÊN

Cần Thơ, ngày tháng năm
(Ký và ghi rõ họ tên)

Mục lục

PHÂN CÔNG CÔNG VIỆC

STT	MSSV	Tên Thành Viên	Nội Dung Thực Hiện	Vai Trò	Ghi Chú
1	B2005809	Nguyễn Huỳnh Hữu Nhâm	Phân công nhiệm vụ, tổng hợp nội dung, viết phần kết luận, đọc dữ liệu, phân tích dữ liệu và lựa chọn mô hình, hướng phát triển huấn luyện mô hình DecisionTree, vai trò máy tính 2 chạy đánh giá mô hình.	Nhóm trưởng	
2	B2014842	Nguyễn Duy Khánh	Lập bảng phân công công việc, tạo slide báo cáo, đọc dữ liệu, huấn luyện theo mô hình RandomForestRegres.	Thư ký	
3	B2005821	Nguyễn Khả Siêu	Phân tích dữ liệu chọn loại mô hình, nhận xét kết quả thực nghiệm, đọc dữ liệu, huấn luyện mô hình Lasso.	Thành viên	
4	B2014876	Trần Đình Sang	Tạo mục lục, viết mô tả dữ liệu, ý nghĩa dữ liệu, kết quả đạt được, đọc dữ liệu, huấn luyện mô hình LinearRegression, vai trò là máy tính 1 chạy đánh giá mô hình.	Thành viên	

Các thành viên trong nhóm xin cam kết thực hiện đúng công việc được giao, cũng như mục tiêu đề ra, đồng thời sẽ đồng hành, hỗ trợ lẫn nhau trong quá trình thực hiện dự án để đem lại hiệu quả tốt nhất.

Nhóm Trưởng	Thư Ký	Thành Viên	Thành Viên
Ký và ghi gõ họ tên	Ký và ghi gõ họ tên	Ký và ghi gõ họ tên	Ký và ghi gõ họ tên
Nguyễn Huỳnh Hữu Nhâm	Nguyễn Duy Khánh	Nguyễn Khả Siêu	Trần Đình Sang

PHẦN NỘI DUNG

1. Mô tả dữ liệu

- Dữ liệu được thực hiện có tên là : Bộ dữ liệu thời tiết Ankara
- Chứa thông tin về thời tiết của Ankara từ 01/01/1994 đến 28/05/1998. Từ các thuộc tính nhất định, mục tiêu là dự đoán giá trị nhiệt độ trung bình.
- Thuộc tính của các dữ liệu:

+ Nhiệt độ tối đa	[23.0, 100.0]
+ Nhiệt độ tối thiểu	[-7.1, 65.5]
+ Điểm sương	[-3.1, 57.6]
+ Lượng mưa	[0.0, 4.0]
+ Áp suất nước biển	[29.46, 30.6]
+ Áp suất tiêu chuẩn	[26.3, 27.18]
+ Hiển thị	[0.2, 11.5]
+ Tốc độ gió	[0.0, 18.0]
+ Tốc độ gió tối đa	[2.19, 57.4]
- Giá trị nhiệt độ trung bình:	[7.9, 81.8]
- Các thông tin khác	
+ Kiểu	Regression
+ Nguồn gốc	Thực tế
+ Trường hợp	1609
+ Đặc trưng	10
+ Giá trị bị mất	No

2. Ý nghĩa của dữ liệu

Đây là một bộ dữ liệu về nhiệt độ của Ankara vào những năm 1994 - 1998 ở đây dựa vào những yếu tố như : nhiệt độ tối đa, nhiệt độ tối thiểu, lượng sương, áp suất nước biển, áp suất tiêu chuẩn, tốc độ gió…. những yếu tố trên được thu thập ở nhiều thời điểm và đưa ra nhiều số liệu cụ thể và từ những số liệu trên sẽ được tổng hợp lại và đưa ra kết luận cuối cùng là nhiệt độ trung bình.

3. Phân tích dữ liệu và lựa chọn mô hình

- Bộ dữ liệu về nhiệt độ của Ankara gồm: 321 phần tử, 9 thuộc tính (Nhiệt độ tối đa, nhiệt độ tối thiểu, điểm sương, lượng mưa, áp suất nước biển, áp suất tiêu chuẩn, hiển thị, tốc độ gió, tốc độ gió tối đa) và 1 nhãn Mean_temperature (Nhiệt độ trung bình). được sử dụng để đánh giá mô hình theo Regression.
- Dữ liệu để dự đoán các giá trị mục tiêu là giá trị liên tục nên lựa chọn mô hình hồi quy (Regression). Các mô hình Regression được lựa chọn trong bài báo cáo là:

  + Decision Tree Regressor

+ Linear Regression

+ Lasso Regression

+ Random Forest Regressor

4. Cấu hình máy tính

- Cấu hình máy tính thứ nhất:

Device name	WINDOWS-11
Processor	AMD Ryzen 5 4600H with Radeon Graphics 3.00 GHz
Installed RAM	8,00 GB (7,40 GB usable)
Device ID	E8FC3A2B-F9FB-4B32-A6DF-C8E04CBEB752
Product ID	00331-10000-00001-AA177
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

- Cấu hình máy tính thứ hai:

Device name	WINDOWS-10
Processor	11th Gen Intel(R)Core(TM)i5-1135G7@ 2.40GHz 2.42 GHz
Installed RAM	8,00 GB (7,73 GB usable)
Device ID	071EA485-FD6D-497B-8A30-C4633252252A
Product ID	00327-35922-30475-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

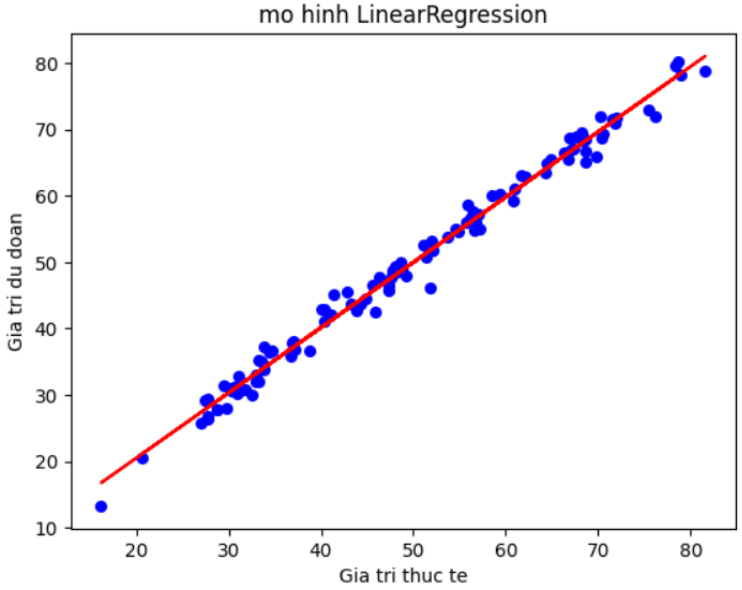
5. Huấn luyện và Kết quả thực nghiệm

Kết quả thu được là kết quả trải qua 50 lần lặp huấn luyện của cả 4 mô hình Decision Tree Regressor, Linear Regression, Lasso Regression, Random Forest Regressor sau khi thay đổi các tham số để đạt được kết quả tốt nhất, được thể hiện lần lược qua máy tính thứ nhất và máy thứ hai.

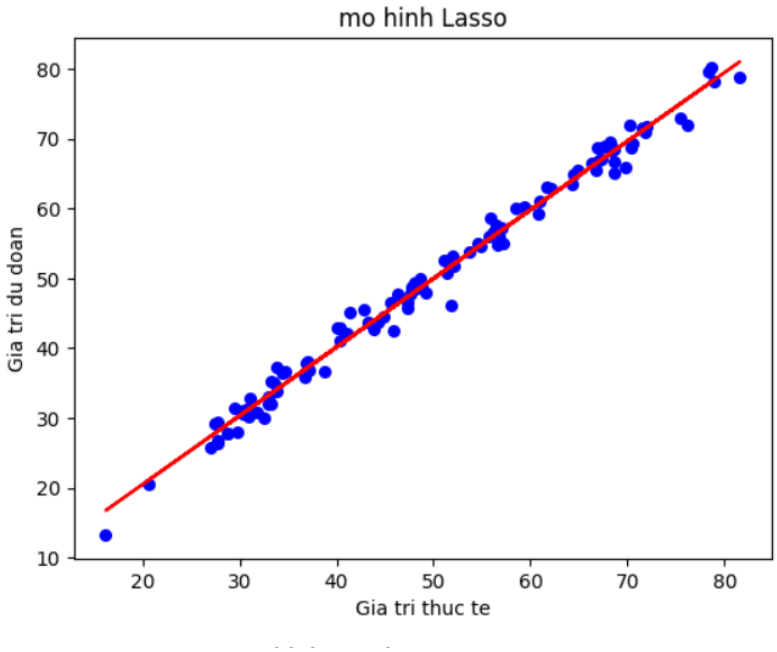
- Kết quả dự đoán máy tính thứ nhất:

```
#Load mô hình DecisionTreeRegressor[1]:
from sklearn.ensemble import BaggingRegressor
from sklearn.tree import DecisionTreeRegressor
tree = DecisionTreeRegressor(random_state= 0)
bagging_regtree = BaggingRegressor(estimator=tree,n_estimators=10,
random_state=42)
bagging_regtree.fit(X_Train,Y_Train)
```

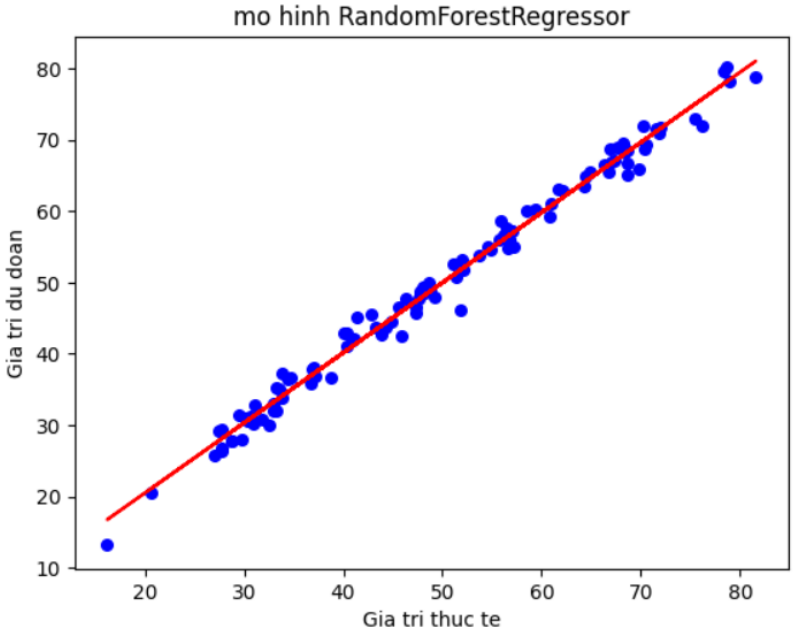
#Load mô hình LinearRegression[2]
lm = linear_model.LinearRegression()
lm.fit(dulieu_X, dulieu_Y)



#Load mô hình Lasso Regression
Lasso_reg = linear_model.Lasso(alpha=0.5)
Lasso_reg.fit(dulieu_X, dulieu_Y)

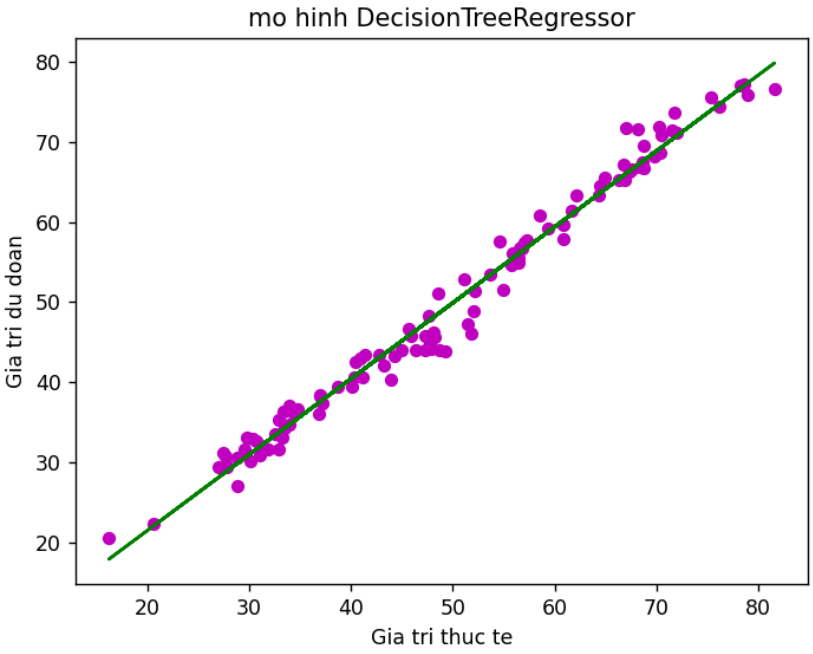



```
#Load mô hình RandomForestRegressor
RandomForest = RandomForestRegressor(n_estimators=100,
random_state=0)
RandomForest.fit(dulieu_X, dulieu_Y)
```

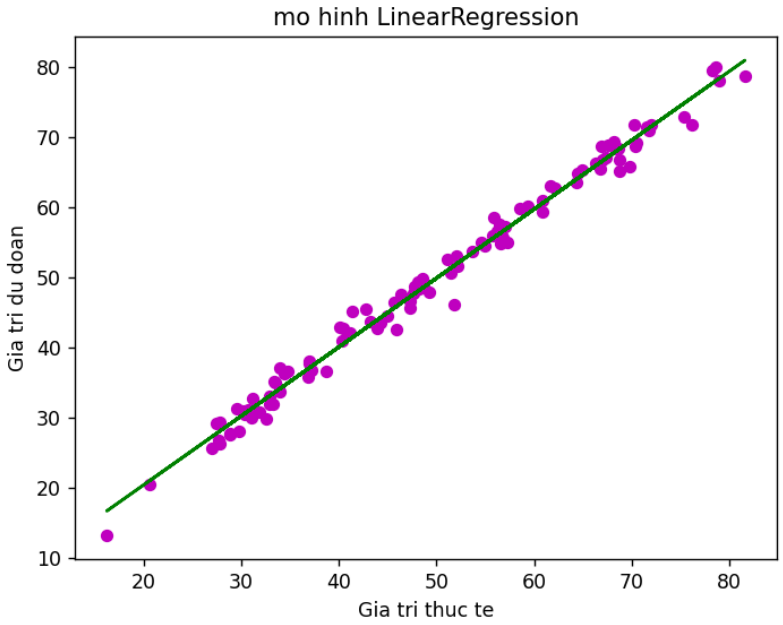


- Kết quả dự đoán máy tính thứ hai:

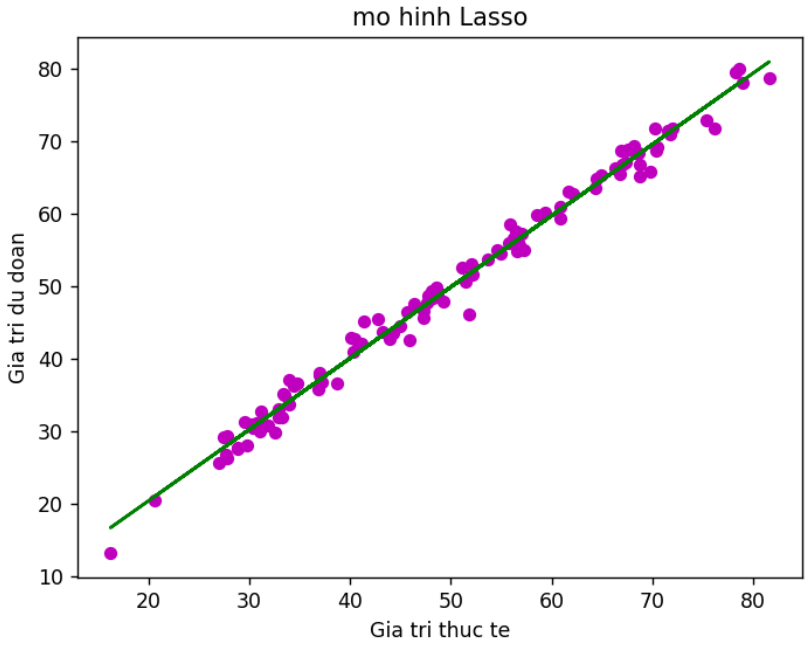
```
#Load mô hình DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.tree import DecisionTreeRegressor
tree = DecisionTreeRegressor(random_state= 0)

bagging_regtree=BaggingRegressor(estimator=tree,
n_estimators=10,random_state=42)
bagging_regtree.fit(X_Train,Y_Train)
```



#Load mô hình LinearRegression
lm = linear_model.LinearRegression()
lm.fit(dulieu_X, dulieu_Y)



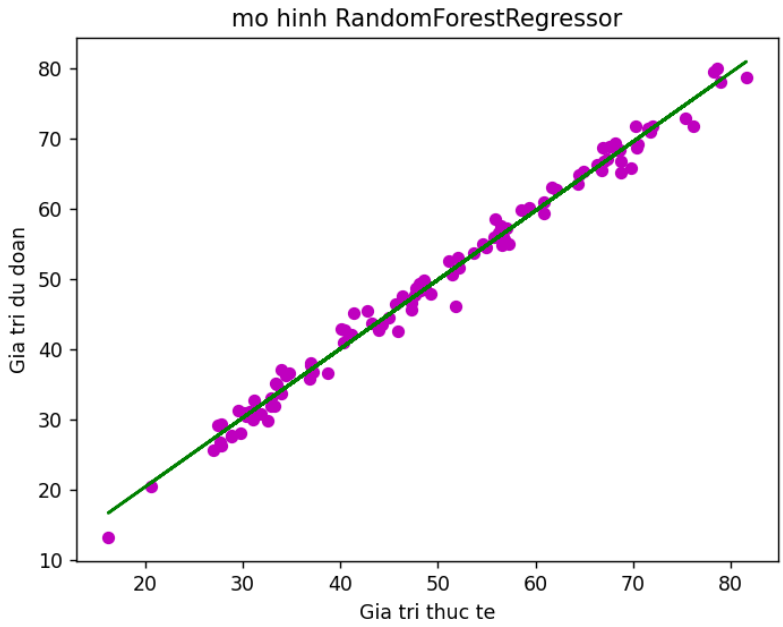
#Load mô hình LassoRegression[3]
Lasso_reg = linear_model.Lasso(alpha=0.5)
Lasso_reg.fit(dulieu_X, dulieu_Y)



#Load mo hinh RandomForestRegressor[4]

RandomForest = RandomForestRegressor(n_estimators=100, random_state=0)

RandomForest.fit(dulieu_X,dulieu_Y)

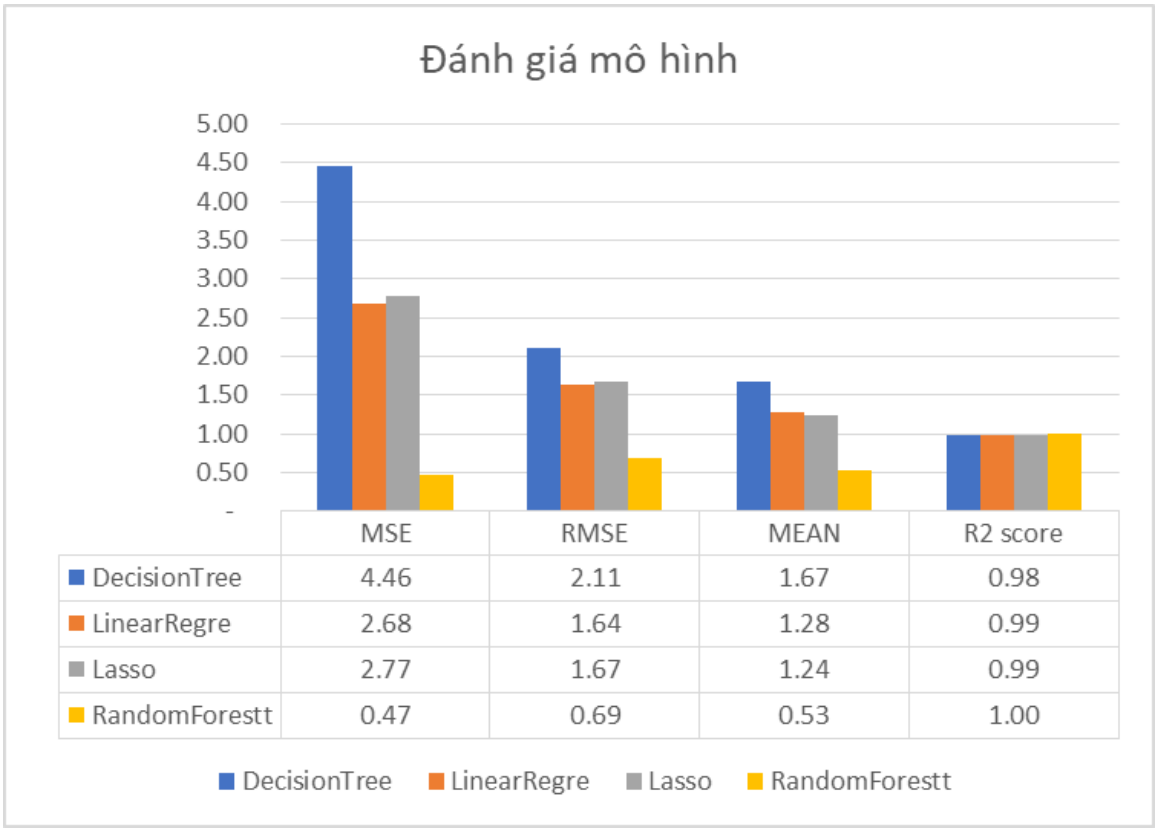


6. Đánh giá mô hình

6.1 Đánh giá mô hình Regression.

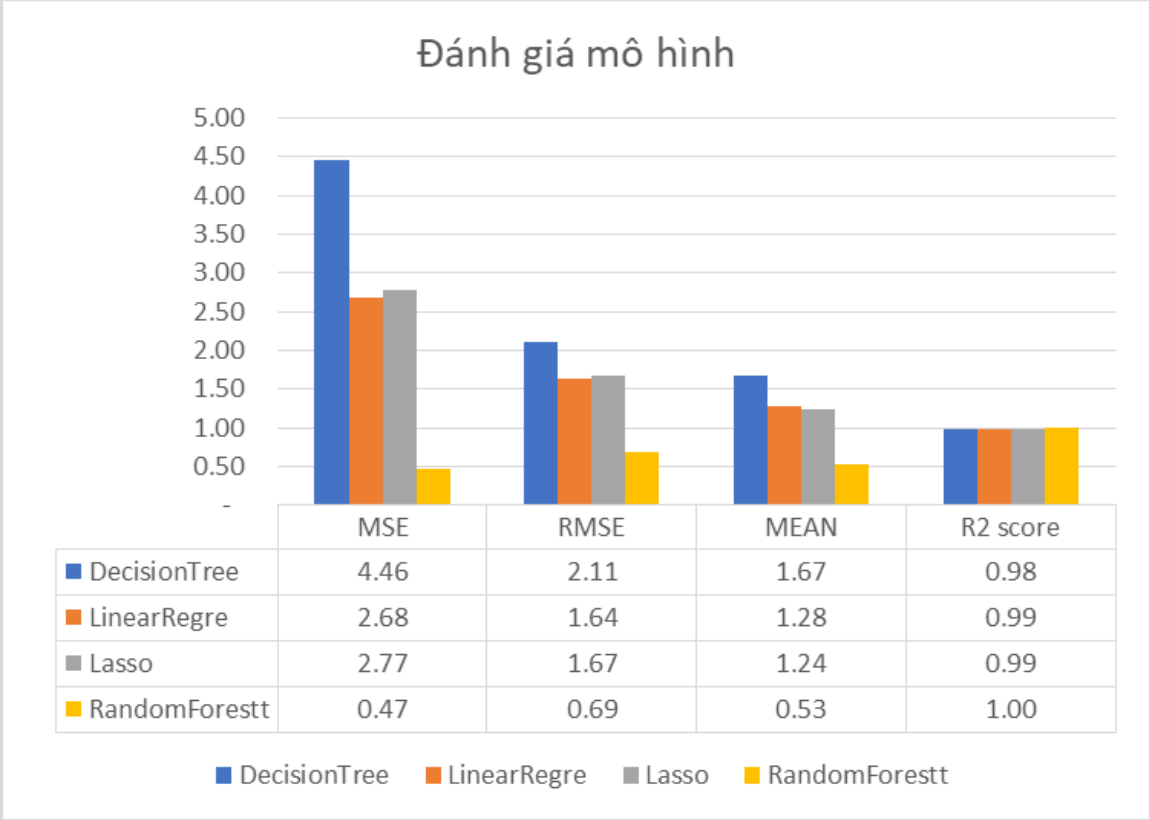
- Máy tính thứ nhất:

Mô Hình	MSE	RMSE	MEAN	R2 score
Decision Tree Regression	4,46	2,11	1,67	0,98
Linear Regression	2,68	1,64	1,28	0,99
Lasso Regression	2,77	1,67	1,24	0,99
Random Forestt Regression	0,47	0,69	0,53	1,00



- Máy tính thứ hai:

Mô Hình	MSE	RMSE	MEAN	R2 score
Decision Tree Regression	4,46	2,11	1,67	0,98
Linear Regression	2,68	1,64	1,28	0,99
Lasso Regression	2,77	1,67	1,24	0,99
Random Forestt Regression	0,47	0,69	0,53	1,00



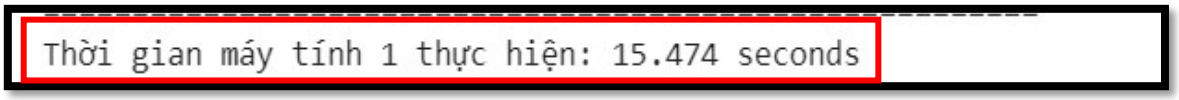
6.2. Nhận xét kết quả thực nghiệm:

- Dựa trên bảng số liệu trong phần đánh giá mô hình Regression, ta có thể nhận xét như sau:
  + Các mô hình đều có R2 score[5] cao, cho thấy khả năng dự đoán của chúng gần với dữ liệu thực tế.
  + Mô hình DecisionTree có kết quả kém nhất với các chỉ số MSE, RMSE, MEAN[6] cao nhất và R2 score thấp nhất trong bốn mô hình. Điều này cho thấy mô hình này có độ chính xác thấp và sai số lớn khi dự đoán.

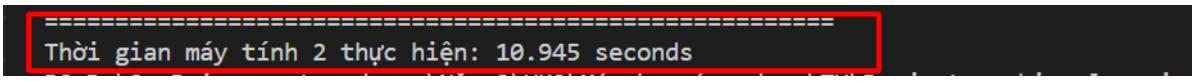
+ Mô hình LinearRegression và Lasso có kết quả tương đối tốt với các chỉ số MSE, RMSE, MEAN gần bằng nhau và R2 score cao trong bốn mô hình. Điều này cho thấy mô hình này có độ chính xác khá và sai số vừa phải khi dự đoán.
+ Mô hình RandomForestt có kết quả tốt nhất với các chỉ số MSE, RMSE, MEAN thấp nhất trong bốn mô hình và R2 score bằng 1. Điều này cho thấy mô hình này có độ chính xác cao và sai số nhỏ khi dự đoán.

Như vậy, ta có thể kết luận rằng trong bốn mô hình trên, mô hình RandomForest là mô hình phù hợp nhất để dự đoán dữ liệu cho bài toán này.

- Thời gian máy tính thứ 1 thực hiện mô hình.



- Thời gian máy tính thứ 2 thực hiện mô hình.



- Dựa vào cấu hình máy tính ta nhận thấy cấu hình máy tính là một yếu tố quan trọng ảnh hưởng đến kết quả mô hình trong máy học. Máy tính 2 có cấu hình cao sẽ có khả năng xử lý dữ liệu nhanh hơn, chính xác hơn và hiệu quả hơn so với một máy tính có cấu hình thấp hơn. Cấu hình máy tính bao gồm các thành phần như bộ vi xử lý, bộ nhớ, ổ cứng, card đồ họa và các thiết bị ngoại vi. Các thành phần này có vai trò khác nhau trong quá trình huấn luyện và kiểm tra mô hình máy học.

PHẦN KẾT LUẬN

1. Kết quả đạt được

- Đánh giá được nhiệt độ trung bình của wakana trong giai đoạn những năm 1990-1994
- Thông qua một số mô hình dự đoán cụ thể như RandomForest, DecisionTreeRegressor, LinearRegre, Lasso cho thấy rằng khả năng dự đoán về nhiệt độ trên bảng số liệu là tương đối chính xác so với thực tế. Điều này được minh chứng qua các thông số về đánh giá mô hình như MSE, MEAN, RMSE, R2 core. Các mô hình đều được đánh giá ở một mức giá trị khá tốt đặt biệt là mô hình RandomForest cho ta một số liệu thống kê tốt và gần với thực tế nhất , xếp sau là các mô hình Lasso, LinearRegr và cuối cùng là mô hình DecisionTree vì vậy sau khi thực hiện các đánh giá về mô hình cho ta thấy được rằng các số liệu thống kê trên bảng số liệu là đáng tin tưởng.

2. Hướng phát triển

Dựa trên kết quả đạt được, ta có thể đưa ra một số hướng phát triển cho các mô hình bên dưới:

- Mô hình DecisionTree có MSE và RMSE cao nhất trong số các mô hình, cho thấy nó có độ chính xác thấp nhất và có thể bị overfitting. Ta có thể cải thiện mô hình này bằng cách giảm độ sâu của cây quyết định, sử dụng phương pháp cross-validation để chọn các tham số tối ưu, hoặc kết hợp nhiều cây quyết định lại thành một mô hình ensemble như RandomForest.

- Mô hình LinearRegre có MSE và RMSE thấp hơn DecisionTree, cho thấy nó có độ chính xác cao hơn và ít bị overfitting hơn. Tuy nhiên, mô hình này có thể không phù hợp với các dữ liệu phi tuyến tính hoặc có nhiều biến độc lập. Ta có thể cải thiện mô hình này bằng cách sử dụng các phép biến đổi dữ liệu để tạo ra các đặc trưng mới, hoặc sử dụng các mô hình phi tuyến tính khác như Polynomial Regression hoặc Support Vector Regression.

- Mô hình Lasso có MSE và RMSE gần bằng với LinearRegre, cho thấy nó cũng có độ chính xác cao và ít bị overfitting. Lasso là một biến thể của LinearRegre với kỹ thuật regularization để giảm thiểu sự phụ thuộc của mô hình vào các đặc trưng không quan trọng. Ta có thể cải thiện mô hình này bằng cách điều chỉnh tham số alpha để tăng hoặc giảm mức độ regularization, hoặc sử dụng các phương pháp khác như Ridge Regression hoặc Elastic Net Regression.

- Mô hình RandomForestt có MSE và RMSE thấp nhất trong số các mô hình, cho thấy nó có độ chính xác cao nhất và không bị overfitting. RandomForest là một mô hình ensemble kết hợp nhiều cây quyết định để tạo ra kết quả trung bình. Ta có thể cải thiện mô hình này bằng cách tăng số lượng cây quyết định, điều chỉnh các tham số như max_depth, min_samples_split, min_samples_leaf, hoặc sử dụng các phương pháp khác như Gradient Boosting hoặc XGBoost.

TÀI LIỆU THAM KHẢO

[1]. DESAI, Nihal; PATEL, Vatsal. Linear Decision Tree Regressor: Decision Tree Regressor Combined with Linear Regressor.

[2]. ZOU, Kelly H.; TUNCALI, Kemal; SILVERMAN, Stuart G. Correlation and simple linear regression. *Radiology*, 2003, 227.3: 617-628.

[3]. RANSTAM, Jonas; COOK, J. A. LASSO regression. *Journal of British Surgery*, 2018, 105.10: 1348-1348.

[4]. EL MRABET, Zakaria, et al. Random forest regressor-based approach for detecting fault location and duration in power systems. *Sensors*, 2022, 22.2: 458.

[5]. PLONSKY, Luke; GHANBAR, Hessameddin. Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, 2018, 102.4: 713-731.

[6]. CHICCO, Davide; WARRENS, Matthijs J.; JURMAN, Giuseppe. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 2021, 7: e623.