Báo cáo đồ án cuối kỳ

1. Trần Nhật Huy – 1612272
2. Nguyễn Duy Hưng - 1512222

CONTENTS

- Phát biểu bài toán
- Thu thập dữ liệu
- Tiền xử lý dữ liệu
- Các thí nghiệm mà nhóm thực hiện
- Tổng kết

Phát biểu bài toán

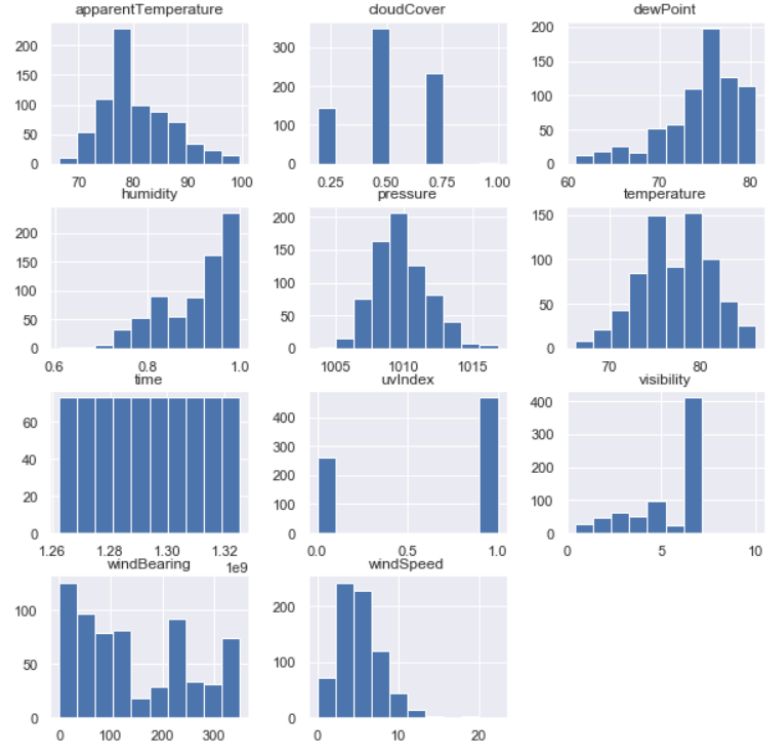
- Nhóm muốn dự đoán độ ẩm của ngày hôm sau dựa vào dữ liệu của ngày hôm trước
- Input: Dữ liệu thời tiết của ngày hôm trước (temperature, humidity, ...)
- Output: Humidity của ngày hôm sau
- Lợi ích: Chúng ta có thể dự đoán trước được humidity của một ngày nào đó. Phục vụ cho nông nghiệp hoặc dự báo thời tiết
- Nguồn gốc: Nhóm tự đặt ra câu hỏi này

Thu thập dữ liệu

- Dữ liệu của nhóm được lấy từ API Dark Sky
- Đây là dữ liệu thời tiết từ 01-01-2010 đến 31-12-2011 ở thành phố Hồ Chí Minh, Việt Nam

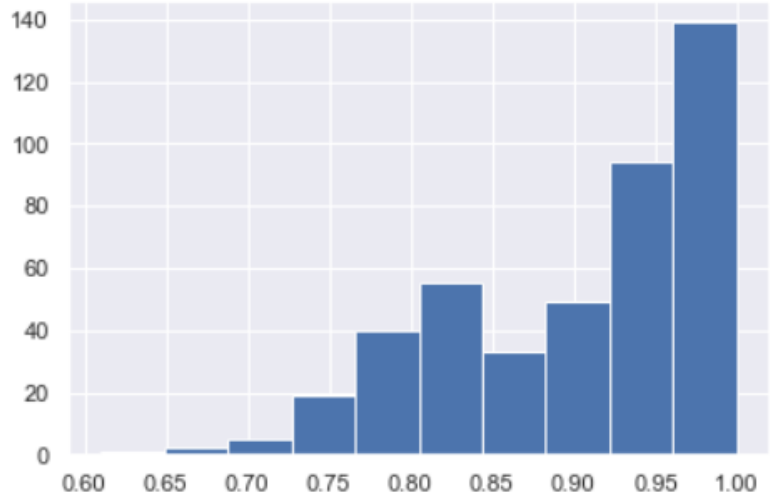
Tiền xử lý dữ liệu

- Các cột mà nhóm tiền xử lý: humidity, apparentTemperature, dewPoint, pressure, temperature, visibility, windBearing, windspeed.



Tiền xử lý dữ liệu

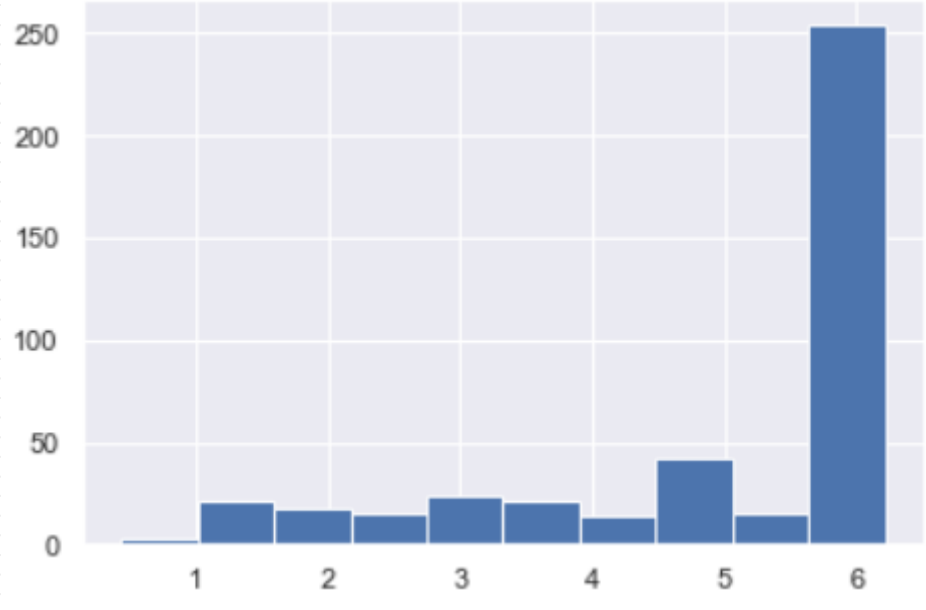
- Giá trị nào xuất hiện ít sẽ gom vào một nhóm và lấy trung bình nhóm đó (các cột trừ visibility)




```
1.00    137
0.98      1
0.97      1
0.94     94
0.92      1
0.89     48
0.88     31
0.87      1
0.85      1
0.84     28
0.83     26
0.82      1
0.79     20
0.78     18
0.77      2
0.74     16
0.73      3
0.70      2
0.69      3
0.65      2
0.61      1
Name: humidity, dtype: int64
```

Tiền xử lý dữ liệu

- Cột nào có giá trị xuất hiện quá trội so với các giá trị khác thì sẽ chia làm bin dựa vào ngưỡng như cột visibility

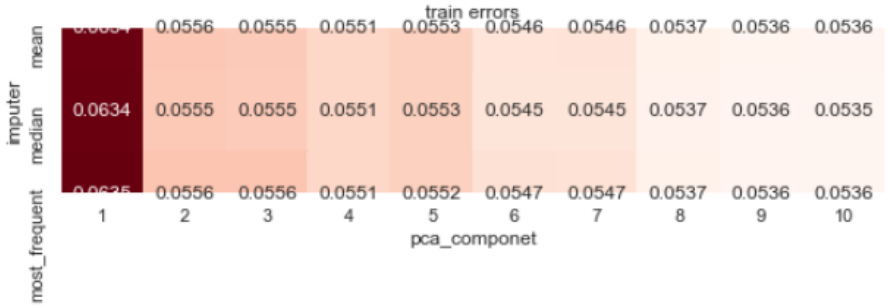


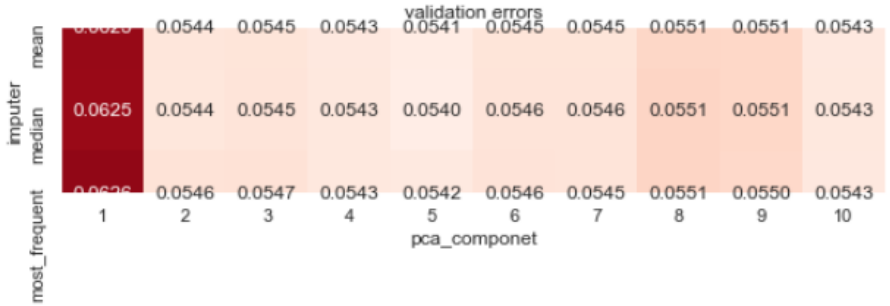
Các thí nghiệm mà nhóm thực hiện

- Nhóm sử dụng Linear Regression và MLPRegression
- Phương pháp đánh giá lỗi: MAE
- Các phương pháp điền giá trị thiếu: mean, median, mode
- Các phương pháp chọn đặc trưng: PCA, Pearson's Correlation, SelectKBest, Feature Importance và một số nhóm tự nghĩ

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 1: Nhóm tách thuộc tính time thành day và month. Bỏ thuộc tính visibility, apparentTemperature. Dùng PCA với n_components từ [1:số lượng cột]

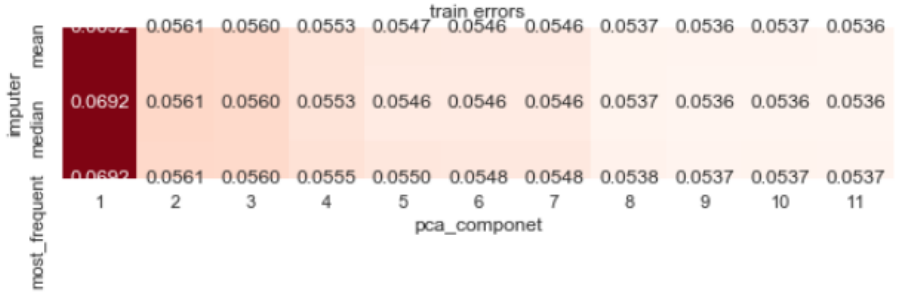


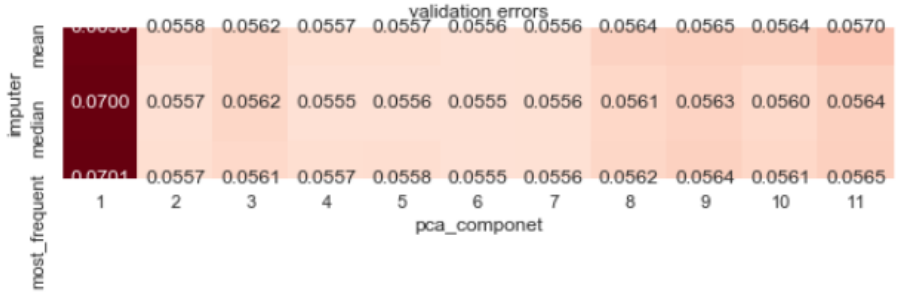


- NX: Độ lỗi MAE trên tập validation là 0.054 với median và PCA=5. Không tốt vì std của cột Humidity là 0.08

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 2: Nhóm giữ nguyên các thuộc tính ban đầu, không thêm hay xóa bất kỳ cột nào

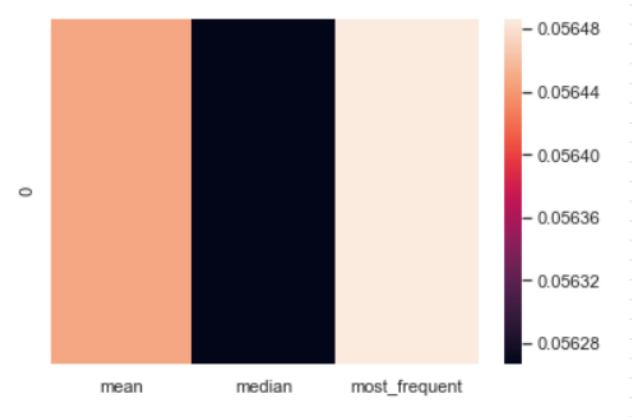


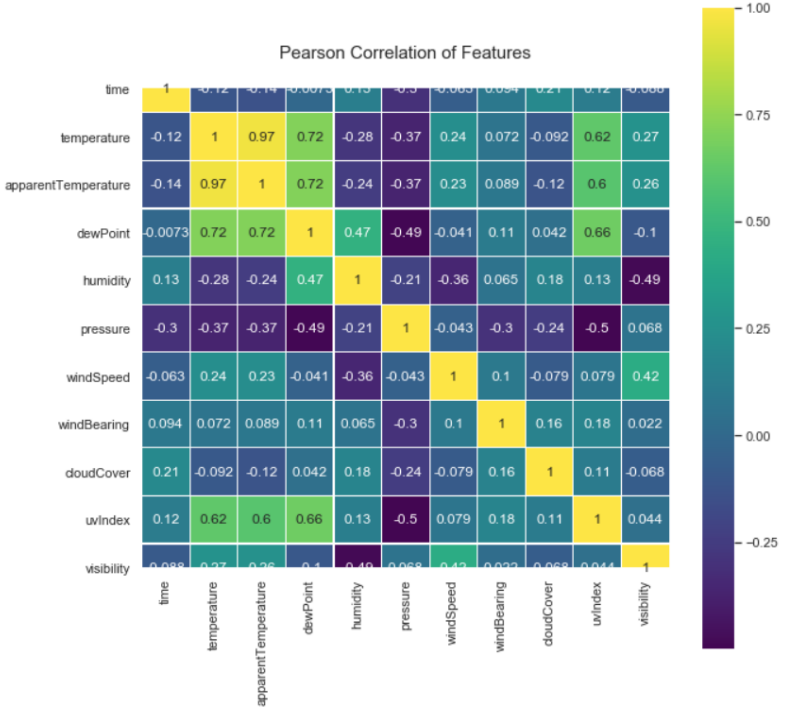


- NX: Độ lỗi là 0.055 với median và PCA=6

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 3: Nhóm sử dụng Pearson's Correlation để chọn đặc trưng (temperature, apparentTemperature, dewpoint, humidity, pressure, windSpeed, visibility

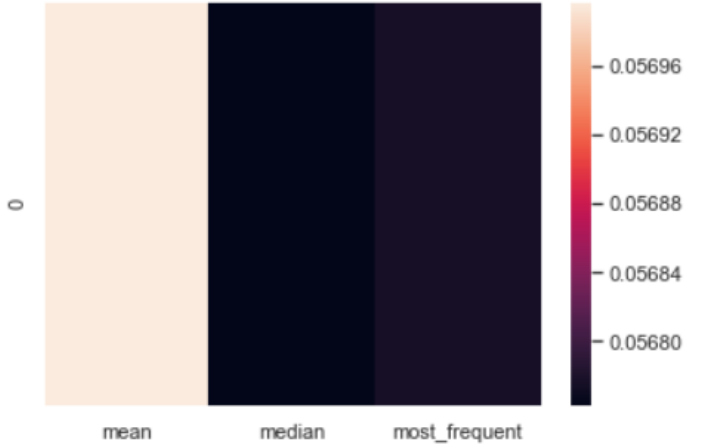


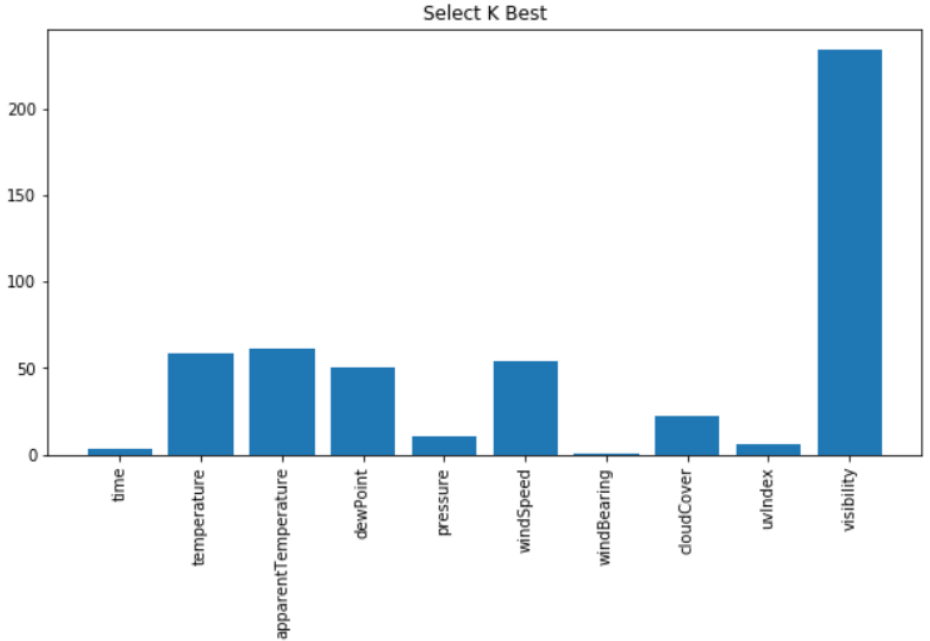


- NX: Độ lỗi là 0.056 với median

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 4: Nhóm lựa chọn đặc trưng bằng SelectKBest (temperature, apparentTemperature, dewPoint, windSpeed, visibility

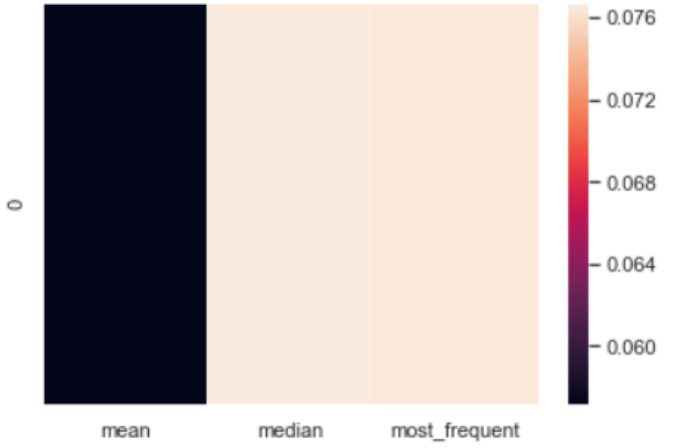


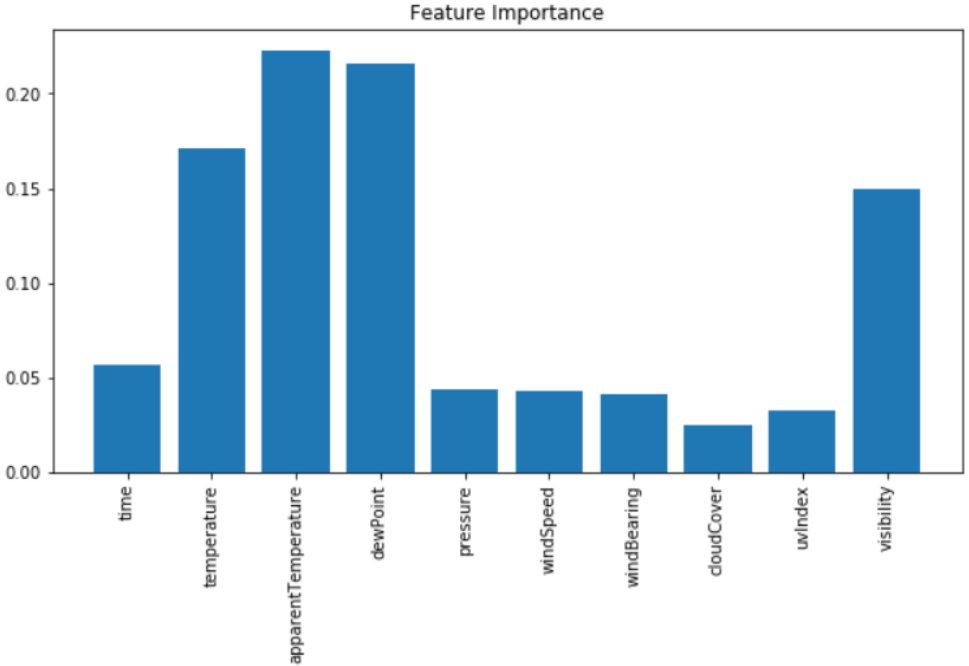


- NX: Độ lỗi là 0.056 với median

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 5: Nhóm lựa chọn đặc trưng bằng Feature Importance (time, temperature, apparentTemperature, dewPoint, humidity, visibility

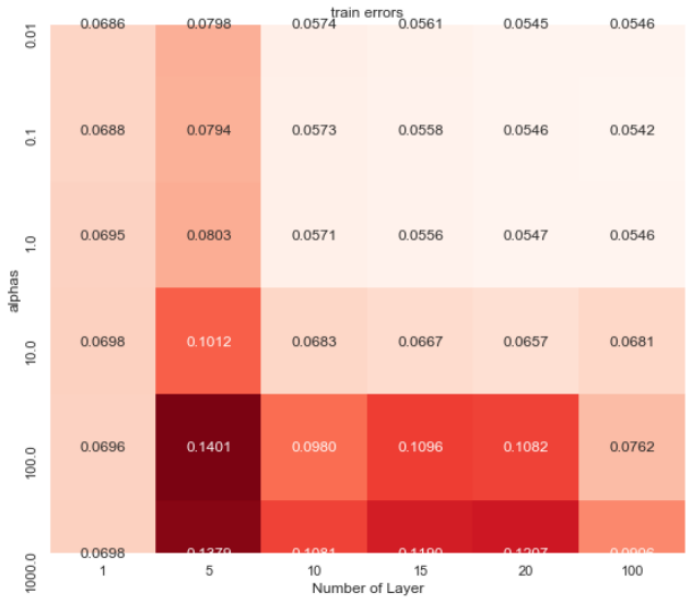


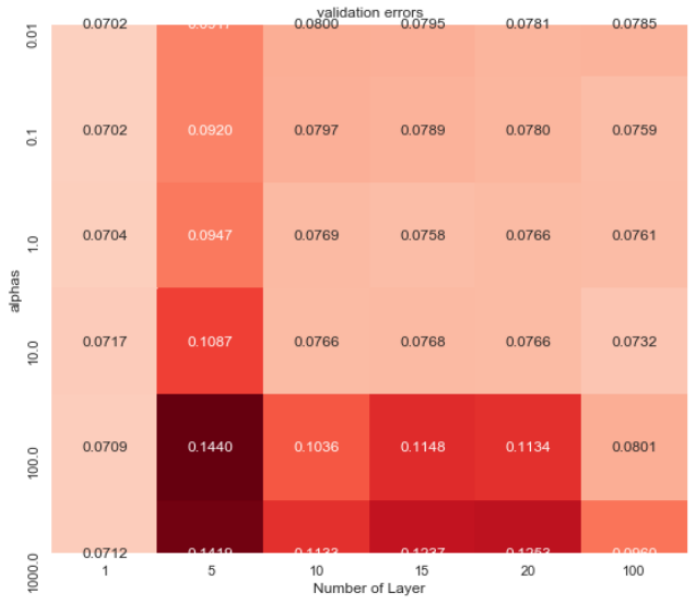


- NX: Độ lỗi là 0.057 với mean

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 6: Nhóm sử dụng mô hình Neural Network, với tất cả các đặc trưng

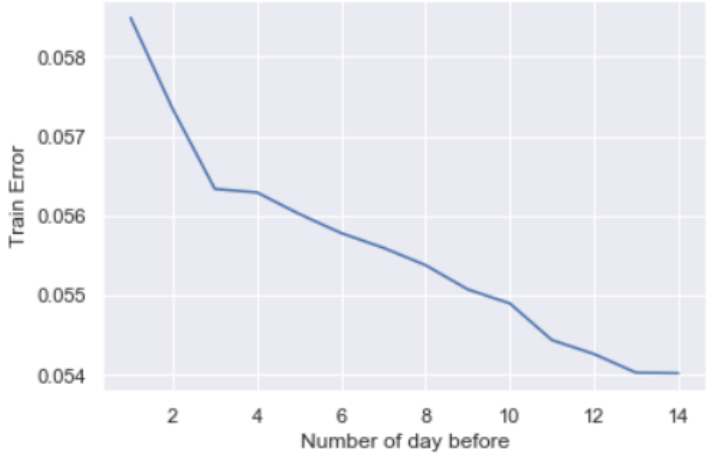


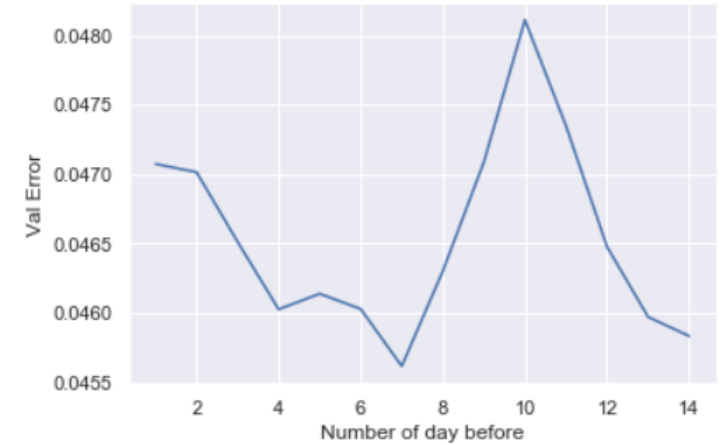


- NX: Độ lỗi là 0.07 với alpha=0.1 và hidden layer là 1

Các thí nghiệm mà nhóm thực hiện

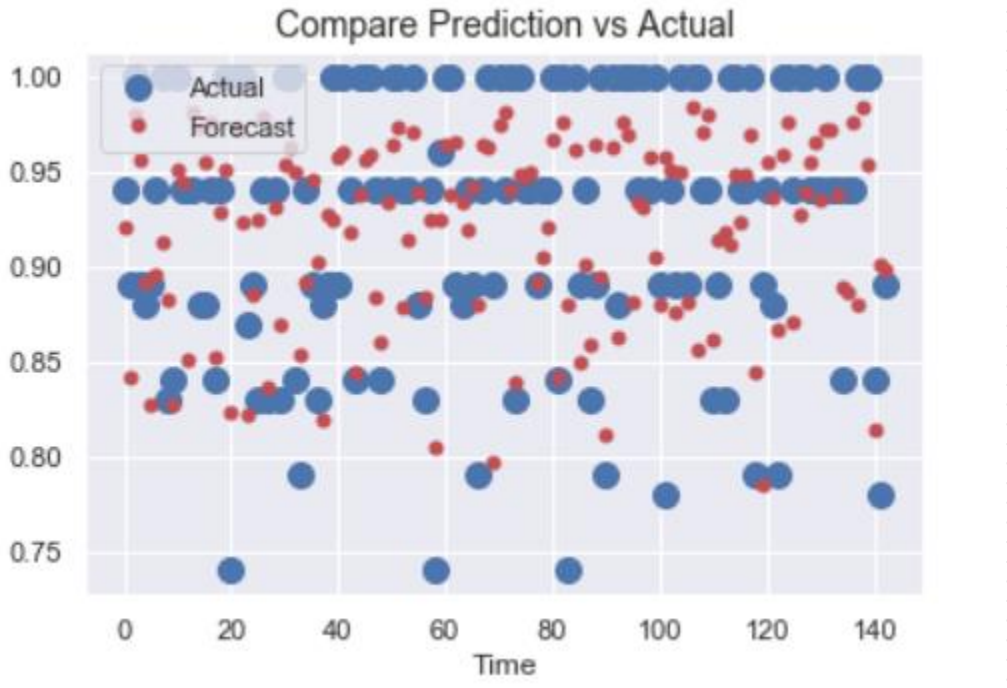
- Thí nghiệm 7: Nhóm sẽ lựa chọn humidity của k ngày trước đó với k chạy từ 1 đến 14





- NX: Độ lỗi là 0.045 với k=7 => **Đây là mô hình tốt nhất của nhóm**

Thí nghiệm trên tập test



- NX: Độ lỗi trên tập validation là 0.045

Tổng kết

- Các thí nghiệm mà nhóm thực hiện có kết quả không thay đổi nhiều khi sử dụng nhiều phương pháp khác nhau
- Nhóm nghĩ có thể nguyên nhân là:
  - Dữ liệu có nhiễu nhưng nhóm xử lý chưa tốt
  - Mô hình Linear Regression bị underfitting với dữ liệu này