Báo cáo đồ án cuối kỳ

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

CONTENTS

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

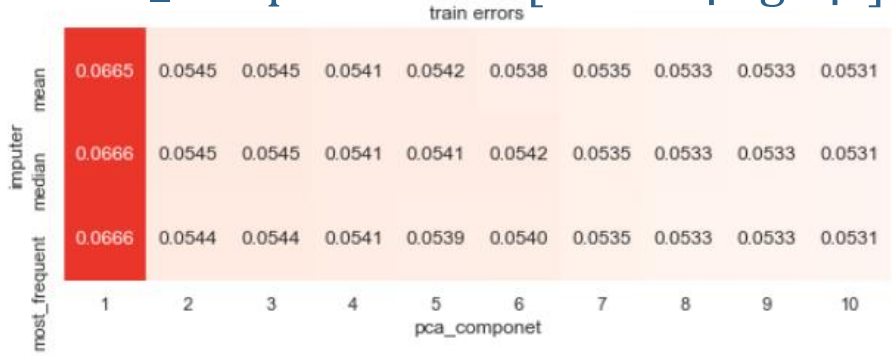
- Điền giá trị thiếu, nhóm sử dụng SimpleImputer với các strategy lần lượt là mean, median, most_frequent
- Lựa chọn đặc trưng nhóm sử dụng PCA, SelectKBest, Pearson + với tự tạo thêm đặc trưng ở một số thí nghiệm
- Chuẩn hóa dữ liệu nhóm sử dụng StandardScalar

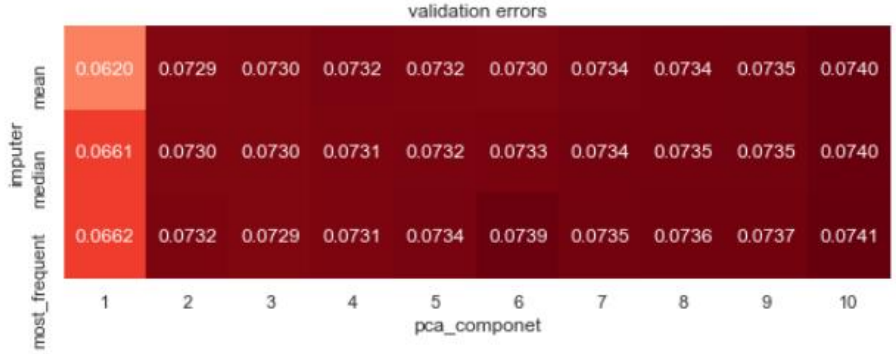
Các thí nghiệm mà nhóm thực hiện

- Nhóm sử dụng Linear Regression và MLPRegression
- Phương pháp đánh giá lỗi: MAE

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]





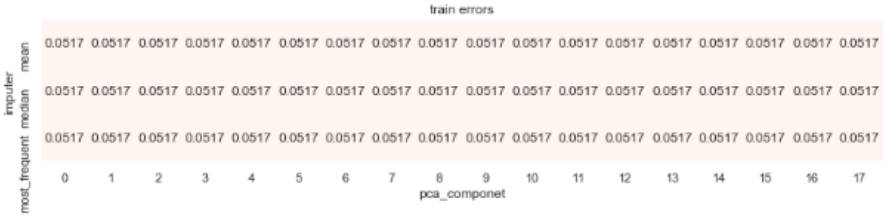
- NX: Độ lỗi MAE trên tập validation là 0.06. Không tốt vì std của cột Humidity là 0.08

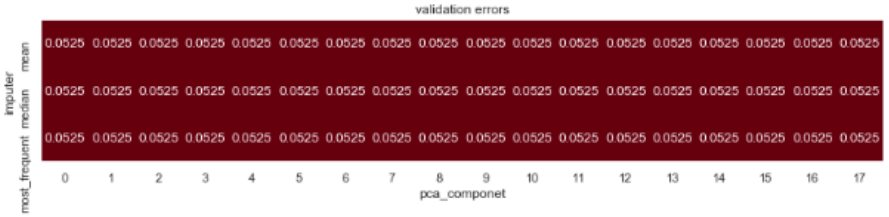
Các thí nghiệm mà nhóm thực hiện

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

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 3: Nhóm dự đoán có thể sử dụng dữ liệu của một ngày chưa đủ nên thử thêm độ ẩm của k ngày trước đó. Nhóm thử nghiệm với $k = 7$

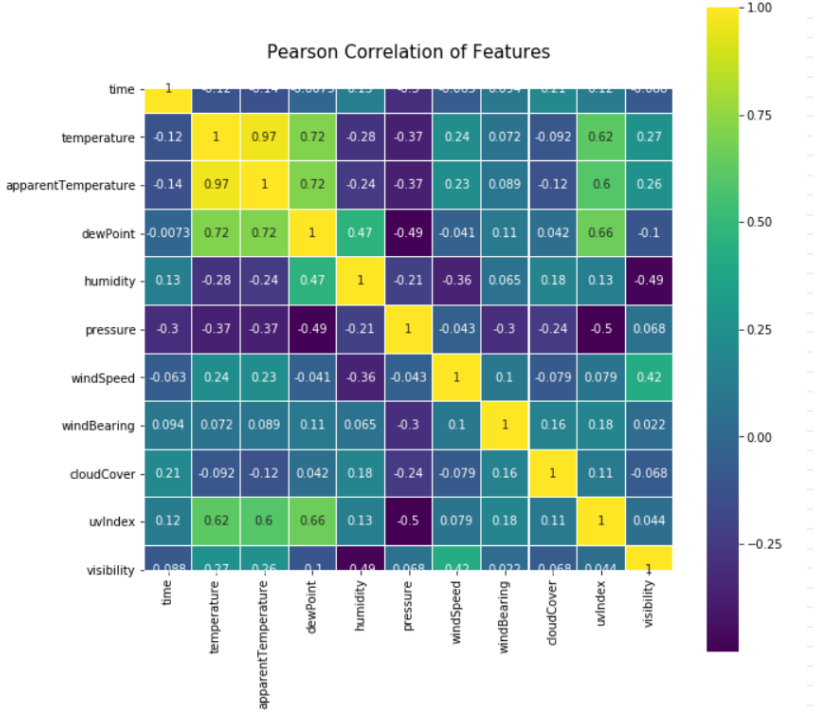




- NX: Có tốt hơn khi độ lỗi MAE trên tập validation là khoảng 0.05

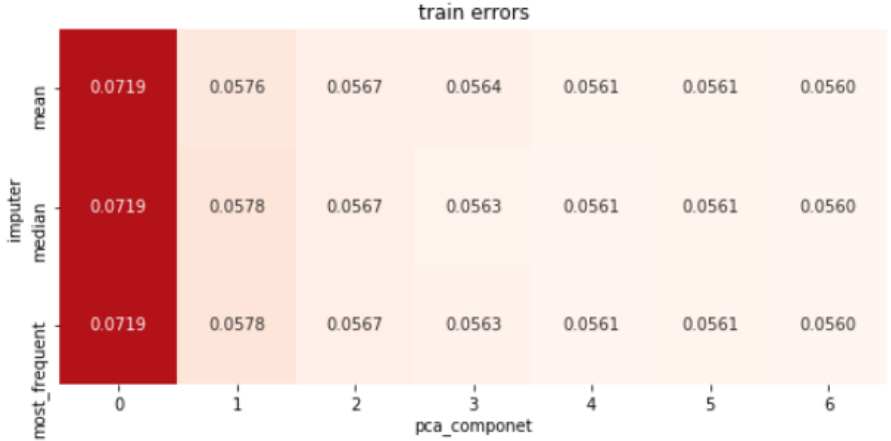
Các thí nghiệm mà nhóm thực hiện

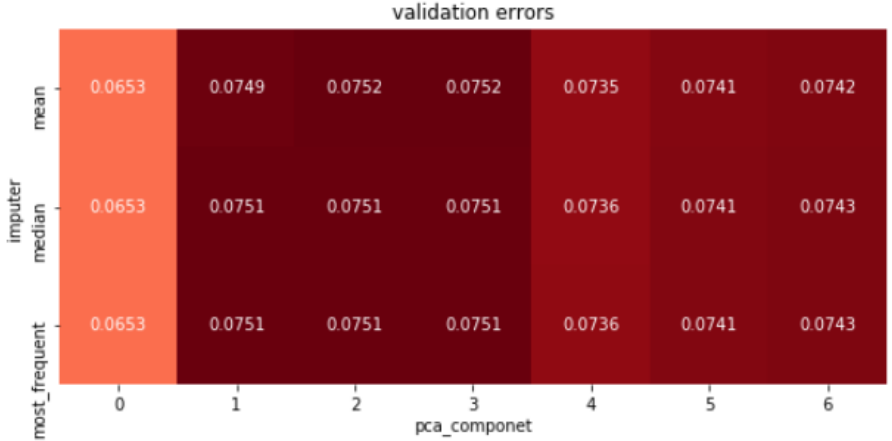
- Thí nghiệm 4: Nhóm lựa chọn đặc trưng bằng Pearson's Correlation



Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 4: Nhóm lựa chọn đặc trưng bằng Pearson's Correlation

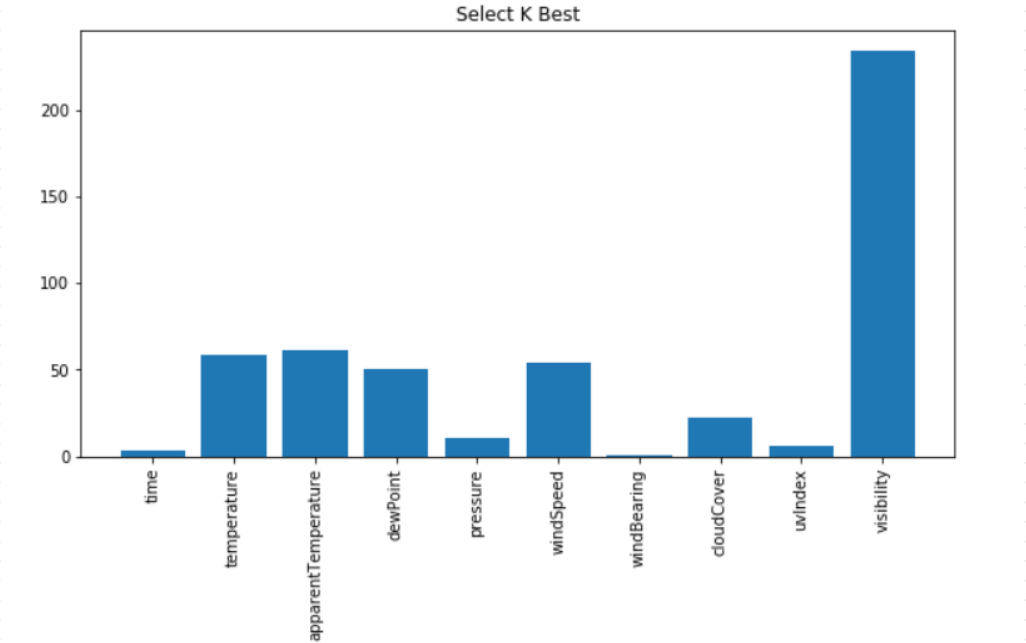




- NX: Độ lỗi trên tập train vẫn không cải thiện vẫn vào khoảng 0.0653

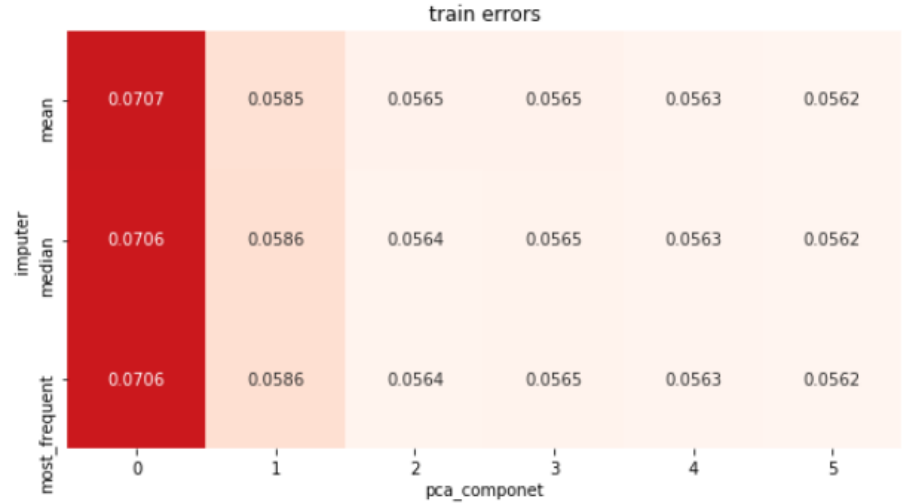
Các thí nghiệm mà nhóm thực hiện

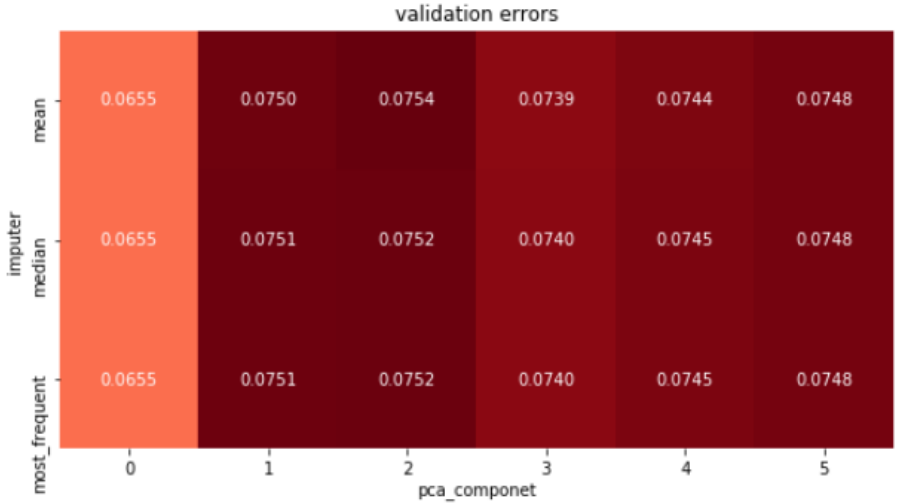
- Thí nghiệm 5: Nhóm lựa chọn đặc trưng bằng SelectKBest



Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 5: Nhóm lựa chọn đặc trưng bằng SelectKBest

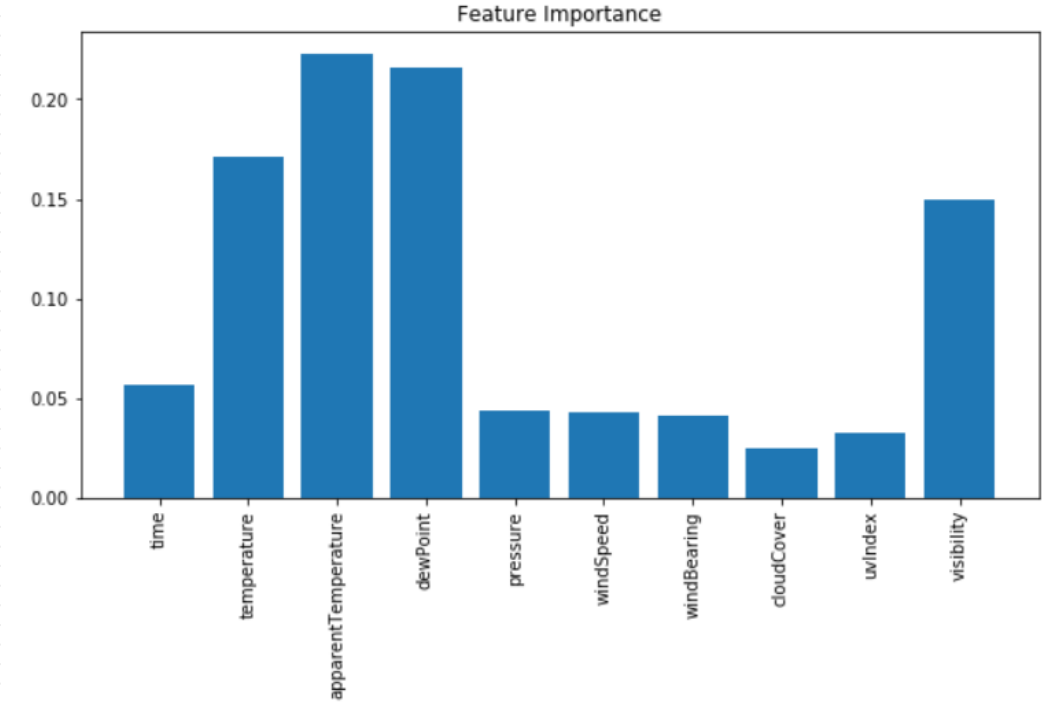




- NX: Độ lỗi trên tập train vẫn không có thay đổi, vẫn vào khoảng 0.06

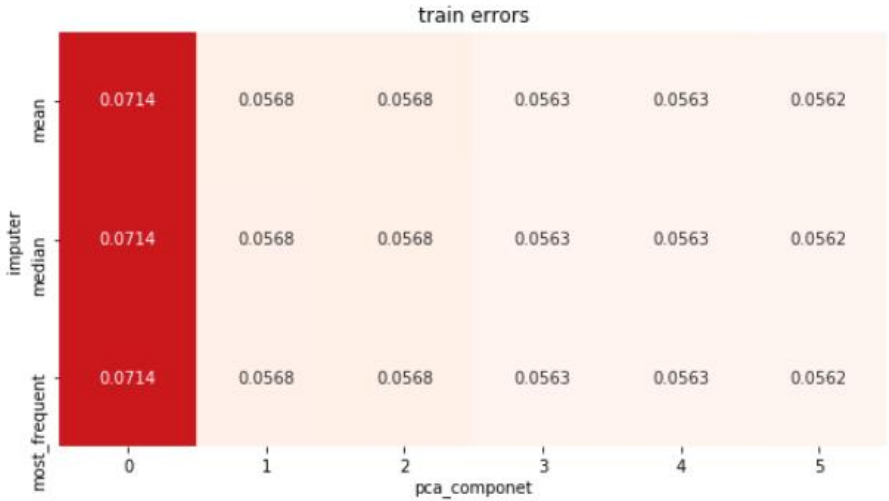
Các thí nghiệm mà nhóm thực hiện

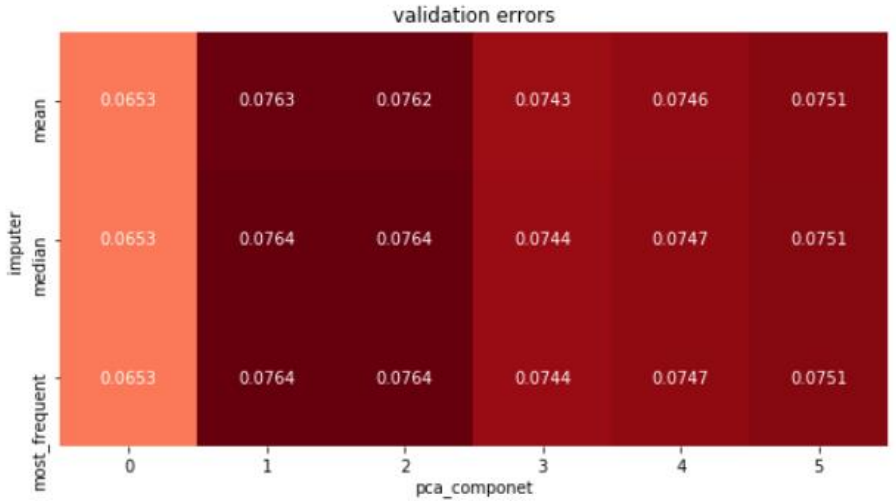
- Thí nghiệm 6: Nhóm lựa chọn đặc trưng bằng Feature Importance



Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 6: Nhóm lựa chọn đặc trưng bằng Feature Importance





- NX: Kết quả khá giống với SelectKBest vào khoảng 0.0653

Các thí nghiệm mà nhóm thực hiện

- Thí nghiệm 7: Nhóm sử dụng mô hình Neural Network

```
print(best_val_err)
print(best_alpha)
```

```
0.5945205479452055
0.1
```

- NX: Độ lỗi của mô hình này còn lớn hơn trước. Có thể bị overfitting

Tổng kết

- Các thí nghiệm mà nhóm thực hiện có kết quả không được tốt
- Nhóm nghĩ có thể nguyên nhân là:
  - Dữ liệu có nhiễu nhưng nhóm không biết xử lý như thế nào
  - Dữ liệu humidity bị lệch, không cân bằng
  - Mô hình không phù hợp, bị underfitting