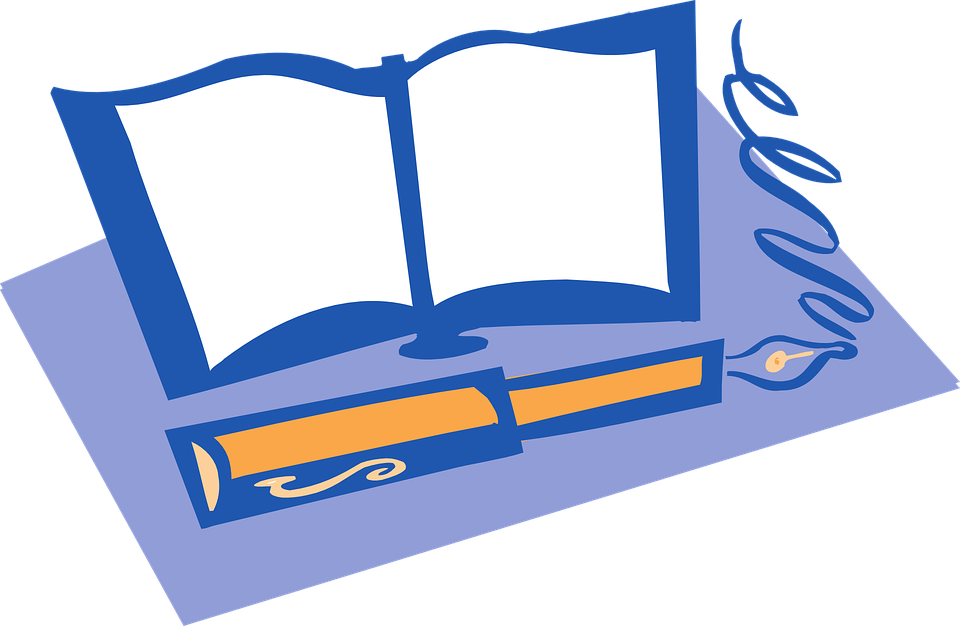
**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP HỒ CHÍ MINH**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

**BÁO CÁO ĐỒ ÁN CUỐI KỲ**



**Nhóm 21**

1612272 - Trần Nhật Huy

1512222 - Nguyễn Duy Hưng

1. **Họp nhóm và phân công công việc**

* Lần 1

Ngày: 24/11/2019

Địa điểm: Facebook

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Ngày thực hiện** | **Người làm** |
| Crawl dữ liệu | 26/11/2019 | Nguyễn Duy Hưng |
| Viết báo cáo | 30/11/2019 | Trần Nhật Huy |

* Lần 2

Ngày: 02/01/2020

Địa điểm: Facebook

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Ngày thực hiện** | **Người làm** |
| Thống nhất kế hoạch làm | 02/01/2020 | Nguyễn Duy Hưng  Trần Nhật Huy |
| Viết kế hoạch và chỉnh sửa | 03/01/2020 | Nguyễn Duy Hưng  Trần Nhật Huy |
| Thí nghiệm 1 | 05/01/2020 | Trần Nhật Huy |
| Thí nghiệm 2 | 05/01/2020 | Nguyễn Duy Hưng |
| Thí nghiệm 3 | 05/01/2020 | Trần Nhật Huy |
| Thí nghiệm 4, 5, 6 | 06/01/2020 | Nguyễn Duy Hưng |
| Thí nghiệm 07 | 06/01/2020 | Trần Nhật Huy |
| Viết báo cáo | 06/01/2020 | Nguyễn Duy Hưng |
| Slide | 06/01/2020 | Trần Nhật Huy |

1. **Bài toán đặt ra và hướng giải quyết**
2. **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

1. **Thu thập dữ liệu**

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

1. **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, Feature Importance + với tự tạo thêm đặc trưng ở một số thí nghiệm
* Chuẩn hóa dữ liệu nhóm sử dụng StandardScalar

1. **Các thí nghiệm đã 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].

Độ lỗi MAE trên tập validation là 0.054. Không tốt vì std của cột Humidity là 0.08.

* Thí nghiệm 2

Nhóm giữ nguyên các thuộc tính ban đầu, không thêm hay xóa bất kỳ cột nào. Độ lỗi MAE trên tập validation là 0.055

* Thí nghiệm 3

Nhóm lựa chọn đặc trưng bằng Pearson’s Correlation

Độ lỗi MAE trên tập validation là 0.056

* Thí nghiệm 4

Nhóm lựa chọn đặc trưng bằng SelectKBest

Độ lỗi MAE trên tập validation là 0.056

* Thí nghiệm 5

Nhóm lựa chọn đặc trưng bằng Feature Importance

Độ lỗi MAE trên tập validation là 0.057

* Thí nghiệm 6

Nhóm sử dụng mô hình Neural Network

Độ lỗi MAE trên tập validation là 0.07 với alpha=0.1 và hidden layer là 1

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

* Thí nghiệm 7

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

Độ lỗi MAE trên tập validation là 0.045

1. Tổng kết và đánh giá đóng góp của các thành viên trong nhóm



* Độ lỗi là 0.045
* 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 xử lý chưa tốt
* Mô hình không phù hợp, bị underfitting