**BÁO CÁO LẦN 1**

**ĐỒ ÁN CUỐI KỲ MÔN KHAI THÁC DỮ LIỆU**

1. **Câu hỏi:**

* Câu hỏi là: Thời tiết của nửa giờ sau có liên quan như thế nào với thời tiết hiện tại?
* Nếu trả lời được câu hỏi này thì chúng ta sẽ có cơ hội chuẩn bị kỹ càng hơn cho những sự cố về mặt thời tiết trong tương lai gần
* Ý tưởng do nhóm tự nghĩ ra.

1. **Thu thập dữ liệu:**

* Dữ liệu được thu thập từ trang: timeanddate.com
* Cụ thể từ link: “ https://www.timeanddate.com/weather/vietnam/ho-chi-minh/historic?month={m}&year=2018 ”
* Với m = chạy từ 1 đến 12.
* Đã kiểm tra và dữ liệu lấy được được cho phép
* Dữ liệu train (raw) trong file wth\_data.csv, dữ liệu test (raw) trong file testdata.csv
* Dữ liệu gồm 16607 dòng và 8 cột bao gồm:
  + Thời điểm trong ngày: dạng float (ví dụ 10:30 = 10.5)
  + Tháng: dạng int
  + Nhiệt độ (độ C): dạng int
  + Tốc độ gió (km/h): dạng int
  + Độ ấm (%): dạng int
  + Áp suất : dạng int
  + Thời tiết hiện tại: dạng int trong khoảng [0,4] gồm 5 loại thời tiết sấm, mưa, nhiều mây, trời trong (có nắng), và sương mù. 5 loại thời tiết này được gom nhóm từ các loại thời tiết khác nhau, chi tiết trong file Class.docx
  + Label: thời tiết của thởi điểm tiếp theo (lấy giá trị thời tiết của dòng dữ liệu ở phía sau (nửa tiếng sau))
* Dữ liệu áp suất bị thiếu đáng kể, dữ liệu gió thiếu 2 dòng
* Giải pháp cho đến hiện tại: để nguyên dữ liệu sức gió, dữ liệu trống của áp suất thay bằng trung vị